Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Participant Overlap and Apparent Accuracy in Behavioral Depression Classification

Sofia Moretti

sofiamoretti079@gmail.com

University of Bologna

Short Report

Keywords: depression classification, behavioral machine learning, repeated measurements, participant overlap, cross-validation, simulation

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11592916/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Behavioral machine-learning studies can include several observations from each participant. When observations from the same person appear in both training and evaluation data, a classifier may recognize participant-specific characteristics instead of learning relationships that generalize to new people.

Objective

To quantify how participant overlap affects apparent classification accuracy under controlled conditions with and without a class-related signal.

Methods

Six simulation scenarios combined three levels of participant-specific variation with two levels of class-related signal. Each scenario contained 100 Monte Carlo replicates, each comprising 60 simulated participants, 10 observations per participant, and 10 features. A one-nearest-neighbor classifier was evaluated using five-fold record-wise and participant-wise cross-validation. Both approaches used 480 training observations and 120 evaluation observations per fold. Outcomes were pooled out-of-fold accuracy, the paired accuracy difference between evaluation approaches, and the frequency with which the record-wise classifier selected a neighbor from the same participant.

Results

Without class-related signal, increasing participant-specific variation raised mean record-wise accuracy from 49.88% to 94.81%. Corresponding participant-wise accuracy ranged from 49.90% to 51.27%. At the strongest participant-specific variation, the mean accuracy difference was 43.54 percentage points, with a Monte Carlo standard error of 0.41 percentage points. In that scenario, 89.81% of selected neighbors belonged to the same participant as the evaluated observation.

Conclusions

Under the simulated conditions, high record-wise accuracy occurred without a generalizable class-related signal. Evaluation intended to estimate performance in previously unseen people should preserve participant separation. These findings demonstrate an evaluation mechanism and do not establish depression-detection performance.

1 Introduction

Research on behavioral depression classification includes the work of Gajbhiye and Singh [1] and the activity-data study of Ahmed and colleagues [2]. Evaluating such systems requires attention to the relationship between individual participants and the multiple observations they may contribute.

Two evaluation strategies address different prediction settings. Record-wise splitting permits observations from a participant to appear in both training and evaluation sets. Participant-wise splitting assigns all observations from a participant to one fold, allowing evaluation on people absent from training. Saeb and colleagues emphasized that evaluation should reflect the intended use case [3]. Chaibub Neto and colleagues demonstrated how stable participant characteristics can create identity confounding in diagnostic machine learning [4].

The present study provides a controlled replication of this established concern. Its contribution is a transparent factorial experiment that separately varies participant-specific variation and class-related signal while keeping the number of training observations constant. It does not assess whether references [1] or [2] contain this evaluation problem.

2 Methods

Simulation design

Each simulated dataset contained 60 participants, equally divided between two artificial classes. Every participant contributed 10 observations with 10 continuous features, producing 600 observations per dataset. Labels remained constant within participants.

For participant i, observation j, and feature p, data were generated as:

$$x_{ijp} = \sigma u_{ip} + \epsilon_{ijp} + \delta (y_i - 0.5) 1_{p \leq 3}$$

Here, the participant effects and observation errors were independent standard-normal variables. Each participant effect remained constant across that participant's observations. The observation error varied independently between observations. The indicator equals one for the first three features and zero otherwise.

Participant-specific variation, σ, took values 0, 1, and 2. Class-related signal, δ, took values 0 and 1. These settings produced six scenarios. At δ = 0, the feature-generating distribution contained no class-dependent term. At δ = 1, the first three features had class means differing by one unit.

The parameters were chosen as illustrative stress-test conditions. They were not estimated from depression measurements, and the artificial labels were not clinical diagnoses.

Classification and evaluation

Classification used one-nearest-neighbor prediction with squared Euclidean distance. The evaluated observation received the label of its nearest eligible training observation. Participant identifiers were used for fold construction and the identity diagnostic, but were not included as predictor features.

No feature selection, normalization, hyperparameter search, or oversampling was performed. All simulated feature dimensions used the same noise scale.

For record-wise cross-validation, each participant's observations were randomly distributed across five folds, with two observations per fold. Consequently, every evaluated participant also contributed eight observations to the corresponding training set.

For participant-wise cross-validation, participants were randomly allocated to five folds within each class. Each fold contained six participants from each class. All observations from a participant remained together.

Both strategies therefore evaluated 120 observations against 480 training observations per fold. The record-wise training set represented 60 participants; the participant-wise training set represented 48.

Replication and outcomes

Each scenario used 100 Monte Carlo replicates. Replicate seeds were 20261101 through 20261200. The same random draws were reused across parameter settings for each replicate to facilitate comparison.

Accuracy was calculated after pooling out-of-fold predictions across all 600 observations. Because both classes contained exactly 300 observations, this accuracy also equaled balanced accuracy.

Results are summarized as the mean and standard deviation across replicate datasets. The paired accuracy difference was calculated within each replicate. Its Monte Carlo standard error was the standard deviation of the paired differences divided by the square root of 100. This quantifies simulation precision, not uncertainty about clinical performance.

As a mechanism diagnostic, the proportion of record-wise predictions whose nearest training neighbor came from the same participant was recorded. Programmatic checks verified fold separation. Selected predictions were also checked against a separate brute-force neighbor calculation.

3 Results

Table 1
Classification accuracy across simulation scenarios

Class signal δ	Participant variation σ	Record-wise accuracy (%)	Participant-wise accuracy (%)	Mean difference (pp)
Absent (0)	0	49.88 ± 2.21	49.90 ± 2.27	−0.02
Absent (0)	1	65.58 ± 3.12	50.63 ± 3.01	14.95
Absent (0)	2	94.81 ± 1.65	51.27 ± 4.29	43.54
Present (1)	0	69.43 ± 2.68	69.15 ± 2.73	0.28
Present (1)	1	73.00 ± 3.45	60.48 ± 3.85	12.52
Present (1)	2	95.40 ± 1.55	55.23 ± 4.65	40.18

Values following ± are standard deviations across 100 replicate datasets. Differences are record-wise minus participant-wise accuracy; pp denotes percentage points.

Without participant-specific variation, the evaluation strategies produced similar mean accuracy. Both were approximately at chance without class signal and approximately 69% with class signal.

When participant-specific variation was introduced, record-wise accuracy increased substantially. Without class signal, it reached 94.81% at σ = 2, while participant-wise accuracy was 51.27%. The paired difference was 43.54 percentage points, with a Monte Carlo standard error of 0.41 percentage points.

The identity diagnostic supported the proposed mechanism. In the no-signal scenarios, the record-wise classifier selected a neighbor from the same participant for 1.70%, 31.35%, and 89.81% of observations at σ = 0, 1, and 2, respectively.

With class signal present, participant-wise accuracy decreased as participant-specific variation increased. At σ = 2, record-wise accuracy was 95.40% and participant-wise accuracy was 55.23%. The paired difference was 40.18 percentage points, with a Monte Carlo standard error of 0.42 percentage points.

4 Discussion

The strongest result is the high record-wise accuracy obtained when the data-generating process contained no class-related signal. Stable participant characteristics allowed the nearest-neighbor classifier to retrieve another observation from the same person. Because labels remained constant within participants, recognizing that person frequently yielded the correct label.

This finding is consistent with previously described identity confounding [4]. It also illustrates why performance estimates must be interpreted in relation to the intended prediction setting [3]. A record-

wise evaluation with overlapping participants does not directly estimate performance in previously unseen people.

The comparison held training-observation counts constant, but the two strategies intentionally differed in the number and identity of training participants. Their performance difference should therefore be interpreted as a contrast between evaluation settings, rather than a universal correction factor for reported accuracy.

The practical implication for behavioral depression research is to report the number of independent participants alongside observation counts and describe exactly how participants are assigned to folds. Any data-dependent preprocessing or model selection should also respect the evaluation boundary. When the goal is prediction in new people, participant separation is necessary, although it does not by itself establish transportability to a new clinic or population.

Limitations

This experiment used artificial Gaussian features and a single nearest-neighbor classifier. It did not analyze speech, facial expressions, wearable recordings, or clinically assessed depression. The simulated labels were balanced and fixed within participants. Temporal drift, changing symptoms, missing observations, measurement artifacts, and demographic confounding were not modeled.

Consequently, the numerical results cannot estimate the amount of inflation in any particular published study. They also cannot establish that an alternative classifier would exhibit the same magnitude of effect.

The underlying mechanism is already established. This report is a computational replication and stress test, not a claim of discovering identity confounding. An empirical extension using an appropriately licensed behavioral dataset would improve its relevance to depression research.

5 Conclusion

In this controlled experiment, participant overlap produced apparent accuracy above 94% despite the absence of class-related signal. Keeping participants separate reduced performance to approximately chance under the same no-signal conditions. Behavioral depression-classification studies should align their evaluation design with the intended prediction population and avoid interpreting participant recognition as evidence of diagnostic generalization.

References

1. Gajbhiye N, Singh KK. An Optimized Depression Detection Technique Using Behavioral Analysis and Machine Learning. In: 2025 6th International Conference on Recent Advances in Information Technology (RAIT). 2025:1–5. https://doi.org/10.1109/RAIT65068.2025.11089439

2. Ahmed I, Brahmacharimayum A, Ali RH, Khan TA, Ahmad MO. Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework. JMIR Mental Health. 2025;12:e72038. https://doi.org/10.2196/72038
3. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6(5):1–9. https://doi.org/10.1093/gigascience/gix019
4. Chaibub Neto E, Pratap A, Perumal TM, et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digital Medicine. 2019;2:99. https://doi.org/10.1038/s41746-019-0178-x