Scientia Psychiatrica
Vol. 7, No. 2, 2026, pp. 111–121 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Nonlinear Actigraphy Complexity Does Not Improve Participant-Aware Prediction of Depression Severity Change

Desiree Montesinos[1], Muhammad Yoshandi[2*], Muhammad Azzam Awais[3]
[1]*Department of Women and Child Welfare, Lira State Hospital, Lira, Uganda*
[2]*Department of Health Sciences, Tembilahan Community Health Center, Tembilahan, Indonesia*
[3]*Department of Community Medicine, Shifa College of Medicine, Islamabad, Pakistan*

ARTICLE INFO

Article history
Received: February 27, 2026
Revised: March 19, 2026
Accepted: April 12, 2026
Available online: April 26, 2026
Published: April 30, 2026

Keywords
Actigraphy
Depression
Machine learning
Nonlinear dynamics
Prediction

***Corresponding author**
Muhammad Yoshandi
E-mail: muh.yoshandi@gmail.com
ORCID: 0009-0005-1272-3331

DOI
10.37275/scipsy.v7i2.216

ABSTRACT

Background. Wrist actigraphy records motor behavior and circadian organization continuously, but it is unclear whether nonlinear complexity summaries can predict within-person change in depressive symptoms.

Objective. To test whether nonlinear actigraphy features improve participant-aware prediction of change in Montgomery–Åsberg Depression Rating Scale (MADRS) scores beyond a training-fold mean benchmark and clinical covariates.

Methods. This secondary analysis used the open OBF-Psychiatric depression subset: 23 adults with unipolar or bipolar depression, 551,716 minute-level activity observations, 359 complete days, and paired MADRS assessments. The outcome was endpoint minus baseline MADRS. Five daily nonlinear complexity features were averaged per participant. Ridge models with clinical, nonlinear, or combined predictors were evaluated by nested leave-one-subject-out validation; mean absolute error (MAE) was primary. Sensitivity analyses allowed an intercept-only option and used metadata-defined recording intervals.

Results. MADRS fell by a mean of 2.74 points (95% confidence interval [CI], −4.35 to −1.22; Wilcoxon p = 0.004). Benchmark MAE was 3.19 (95% CI, 2.26–4.28); the clinical, nonlinear, and combined models had MAEs of 3.31, 3.21, and 3.22 and negative cross-validated R^2. The nonlinear model's MAE exceeded the benchmark's by 0.021 (95% CI, 0.008–0.035). No nonlinear feature was associated with MADRS change after multiplicity correction. With an intercept-only option, nonlinear and combined models selected it in all 23 folds and matched the benchmark, as did metadata-defined intervals.

Conclusion. In this small cohort, nonlinear actigraphy complexity added no predictive value for short-term change in depression severity. Such features need participant-separated validation against a simple benchmark before being considered clinically informative.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Depressive disorders involve more than affective and cognitive symptoms. Sleep, psychomotor activity, energy, and the regularity of daily behavior change as well. These changes fluctuate over time, and episodic clinical assessments may capture them only partly. Wrist actigraphy can follow motor behavior continuously in everyday life at little burden to the wearer; the open OBF-Psychiatric dataset combines actigraphy with clinical information for several psychiatric conditions and offers a sound basis for reproducible analysis;[1] its depression component holds minute-resolution activity records and clinician-rated symptom severity. The Montgomery–Åsberg Depression Rating Scale (MADRS) is still widely used to quantify depressive severity and treatment-related change.

Most actigraphy research on depression has examined the level, timing, fragmentation, and circadian organization of activity. In clinical and population samples, lower daytime activity, irregular rhythms, altered sleep–wake patterns, and less stable behavior have all been linked to symptom burden.[2–8] Other studies have looked at seasonal depression, treatment-resistant depression, multimodal assessment, physical-

activity patterns, relapse, and accelerometer-defined sleep irregularity.[9–14] Such work supports the clinical relevance of behavioral time series. It does not show that a compact summary of the signal can predict prospective symptom change in a new participant.

The field has recently shifted from cross-sectional association toward individualized prediction. Investigators have tested mood-disorder phenotyping, bipolar–unipolar classification, chronotype and sleep, relapse prediction, and combined smartphone–wearable models.[15–19] Related studies have applied passive movement, sleep, physiology, mood-symptom, affect, circadian, student, and population-level wearable data to depression phenotypes.[20–28] Yet performance depends heavily on the unit of analysis and on how models are validated. If observations from one person fall in both the development and evaluation sets, the model may pick up person-specific signatures. It then fails to learn a relationship that holds for an unseen participant.

Nonlinear time-series methods can describe properties that conventional averages miss. Entropy measures irregularity or unpredictability. Fractal methods assess scale-dependent organization. Spectral entropy reflects how signal power is spread across frequencies. Recent studies have examined longitudinal actigraphy classes, daily steps, mobile severity assessment, adolescent sensing, wearable screening, fractal motor activity, impending recurrence, pediatric tracking, accelerometer clustering, adolescent circadian disruption, and day-to-day activity regularity.[29–39] Chronobiological predictors and symptom-linked activity patterns make complexity features plausible candidates for tracking clinically relevant change.[40,41] Complexity, however, is not the same as clinical information. Estimates vary with recording length, missing data, preprocessing, parameter choices, daily context, and the way minute-level signals are aggregated into participant-level outcomes.

Predicting MADRS change adds a further difficulty. A change score mixes baseline heterogeneity, treatment exposure, measurement variability, and regression toward the mean. Recent smartphone and wearable studies illustrate personalized prediction, day-level physical-activity states, objectively measured depression subtypes, longitudinal smartphone monitoring, and momentary affect estimation.[42–46] With few participants, though, a flexible model may fit idiosyncratic variation and then do worse than the training-fold mean on a held-out participant. Participant-aware evaluation, explicit safeguards against leakage, and full reporting of the prediction model are therefore needed.[47–49] Regularization and feature standardization must be learned only within each outer training fold. A lone correlation or an in-sample fit statistic cannot show incremental predictive value.

This study contributes a direct test of five complementary nonlinear complexity domains as predictors of within-person MADRS change; the test uses fully participant-separated nested validation, a transparent training-fold benchmark, uncertainty intervals, multiplicity-controlled feature analyses, and sensitivity checks. The checks let the model collapse to an intercept-only solution and align activity extraction with metadata-defined recording intervals. The study therefore aimed to establish whether participant-level nonlinear actigraphy features improve prediction of MADRS change beyond a training-fold mean benchmark and compact clinical covariates, and to quantify the uncertainty and practical size of any incremental value.

2. Methods

2.1. Study design and setting

This was a secondary analysis of openly available observational data from the OBF-Psychiatric dataset.[1] Only the depression subset was used. The source files contain minute-level actigraphy, demographic and clinical descriptors, diagnostic grouping, inpatient or outpatient status, and two MADRS observations for each eligible participant. The participant was the analytical unit, because the target outcome was a participant-level change score. Minute-level measurements and daily feature estimates served only to build participant summaries. They were not treated as independent clinical outcomes.

2.2. Participants

The source subset comprised 23 adults: 15 with unipolar depression, seven with bipolar II disorder, and one with bipolar I disorder. Five were inpatients and 18 were outpatients; ten were female and 13 were male. Age was given in five-year bands, not exact years. Every participant with paired MADRS measurements and usable activity files was kept. No case was dropped because of observed symptom response, predictor value, or model performance.

2.3. Variables and instruments

Outcome definition. The primary outcome was change in depression severity, defined as $\Delta MADRS = MADRS_{endpoint} - MADRS_{baseline}$. Negative values mean improvement, and positive values mean worsening. This direction was fixed before model evaluation and used in every descriptive, association, and prediction analysis. Baseline and endpoint MADRS were also summarized separately so that the clinical meaning stayed clear. A paired change score can depend on baseline severity. For that reason baseline MADRS entered the clinical predictor set; the outcome itself was not redefined.

2.4. Data collection procedure

Actigraphy processing and day-level eligibility. Each complete day was analyzed on its own. That kept the sequence length the same for all participants, so a long recording did not carry a different estimator scale from a short one. Day-level estimates were then averaged within each participant to match the participant-level outcome. The number of complete days stayed in the model as a clinical/recording covariate, which made differences in observation opportunity explicit.

Processing was deterministic, and all participants shared the same feature definitions.

Each complete day was analyzed on its own; that kept the sequence length the same for all participants, so a long recording did not carry a different estimator scale from a short one. Day-level estimates were then averaged within each participant to match the participant-level outcome; the number of complete days stayed in the model as a clinical/recording covariate, which made differences in observation opportunity explicit. Processing was deterministic, and all participants shared the same feature definitions.

Nonlinear feature extraction. Five features were chosen, each representing a distinct and interpretable aspect of temporal organization. Sample entropy gives the likelihood that similar short activity patterns stay similar when extended by one observation, and lower values mean greater regularity under fixed parameters.[50] Permutation entropy measures the diversity of ordinal patterns and is fairly insensitive to monotonic transformations;[51] the detrended fluctuation analysis exponent summarizes long-range scale relationships once local trends are removed.[52] Higuchi fractal dimension estimates geometric complexity from the series itself across several interval lengths.[53] Spectral entropy normalizes the power spectrum and shows whether activity power is concentrated or diffuse across frequencies.

The daily feature vector thus held sample entropy, permutation entropy, the detrended fluctuation analysis exponent, Higuchi fractal dimension, and spectral entropy. Constant or numerically undefined daily sequences were handled by the deterministic extraction routine, and none led to an outcome-dependent exclusion. Each participant-level feature was the arithmetic mean of the valid daily values for that domain. This aggregation asks whether a person's typical daily complexity over the observed interval predicts clinical change. It cannot speak to hour-specific patterns, day-to-day dynamics, or treatment-phase trajectories.

2.5. Statistical analysis

Predictor sets and models. Four predefined prediction strategies were compared. The benchmark, called the training-fold mean below, predicted for the held-out participant the mean MADRS change of the training folds. The clinical model used baseline MADRS, age-band coding, sex, diagnostic grouping, inpatient/outpatient status, and the number of complete days; the nonlinear model used the five aggregated complexity features, and the combined model used clinical and nonlinear predictors together. Ridge regression was chosen because the number of candidate predictors was not small relative to the number of participants, and because shrinkage gives stable estimates when predictors are correlated.

Every learned model was built as a pipeline. Any missing predictor values within a training fold were handled by rules derived from the training data. Continuous predictors were standardized with training-fold means and standard deviations, and the fitted transformations were then applied to the held-out participant. Categorical variables were encoded without the outcome. No minute, day, or participant record from the outer test fold touched preprocessing, feature scaling, model fitting, or tuning.

Participant-aware validation and performance measures. Generalization was assessed by leave-one-subject-out (LOSO) cross-validation. In each of 23 outer folds one participant was held out, and the remaining participants formed the development set. Within that set an inner LOSO procedure chose the ridge regularization; The primary grid ran from weak to strong regularization, and the value with the lowest inner-fold error was selected. The outer prediction came only after tuning and refitting on the full outer development set. This nested design keeps model selection apart from performance estimation.

Mean absolute error (MAE), expressed in MADRS points, was the primary performance measure. Root mean squared error (RMSE) was added to give more weight to large errors. Cross-validated R^2 compared the sum of squared outer-fold errors with the variation around the overall observed outcome mean; a negative R^2 means the predictions were less accurate than that reference. Incremental value was defined as benchmark MAE minus model MAE, so positive values favor the model; the abstract reports the opposite difference (model minus benchmark) as an excess error. Participant-level bootstrap resampling with 5,000 replicates gave 95% confidence intervals for MAE and for paired differences in absolute error.

Participant-aware validation is essential when repeated wearable measurements are collapsed or sampled from the same individual.[47] Without it, leakage between a participant's development and evaluation observations can inflate apparent performance;[48] the analysis and its reporting followed the principles of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, extended for artificial intelligence (TRIPOD+AI) statement: a defined target population and outcome, explicit handling of predictors, separated validation, a benchmark comparison, and uncertainty reporting.[49]

Statistical analyses. Continuous variables were summarized as mean and standard deviation where that was useful for scale, and as median and interquartile range where distributional shape or small-sample robustness mattered. Categorical variables were summarized as counts and percentages. Paired MADRS change was tested with the Wilcoxon signed-rank test, and a bootstrap confidence interval was computed for the mean change. This test was supportive and did not drive predictor selection.

Spearman rank correlations measured the monotonic association between each participant-level nonlinear feature and MADRS change, with two-sided confidence intervals from a participant-level bootstrap. The Benjamini–Hochberg procedure controlled the false discovery rate across the five feature tests. Change in

the bipolar group (bipolar I and II combined) was compared with the unipolar group by an exact Mann–Whitney test, with a bootstrap interval for the difference in group means. Because of the small sample, subgroup results were read as descriptive and were not used to justify more model complexity.

Sensitivity analyses. Two sensitivity analyses addressed the decisions most likely to affect interpretation. In the first, the ridge grid was extended to a penalty of 1,000,000 and an explicit intercept-only candidate was added. A model could then record that the available predictors did not improve inner-fold performance, instead of being forced to assign nonzero feature effects. The intercept-only candidate was selected inside each outer development set. In the second, nonlinear features were re-extracted after restricting each participant to the chronological recording interval defined by the source metadata, starting at the first activity time stamp. Complete-day criteria, feature definitions, participant aggregation, nested LOSO validation, and performance metrics were otherwise unchanged. The interval-aligned analysis held 268 complete days, with 4–17 days per participant.

Reproducibility and reporting. The analytical dataset, feature tables, outer-fold predictions, performance summaries, and validation checks were kept in the study record. The reproduced primary analysis passed all 18 predefined computational validation checks, covering participant counts, outcome direction, feature dimensions, fold separation, metric reconstruction, and the expected output structure. Random procedures used fixed seeds. Software details and machine-readable outputs accompany the analytical archive.

2.6. Ethical approval

The source dataset reports review by the Norwegian Regional Medical Research Ethics Committee West under protocols 150.01 and 251.08, with written informed consent for the originating studies.[1] This secondary analysis of de-identified open data was reviewed and approved by the CMHC Research Ethics Committee (021/CMHC/Ethics/08/2026). No attempt was made to re-identify participants. Reporting followed the TRIPOD+AI principles that apply to prediction-model studies.[49]

3. Results

3.1. Participants, recordings, and symptom change

All 23 participants with paired MADRS values were analyzed. The sample included 15 participants with unipolar depression and eight with bipolar depression, of whom seven had bipolar II disorder and one had bipolar I disorder. The female proportion was 43.5%; 78.3% were treated as outpatients. The modal age band was 45–49 years. Activity coverage comprised 551,716 minute records and 359 complete 24-hour days. The median participant contributed 14 complete days (interquartile range, 14–15; range, 12–28). That gave substantial within-participant signal length, yet only 23 independent clinical outcomes. Table 1 summarizes participant and recording details.

Mean baseline MADRS was 22.74 (standard deviation [SD], 4.80), and mean endpoint MADRS was 20.00 (SD, 4.73). Mean change was −2.74 points (95% CI, −4.35 to −1.22); the median was −3 points (interquartile range, −5 to 0). The paired distribution differed from zero in the Wilcoxon analysis ($p = 0.004$). Figure 1 shows participant-level baseline and endpoint values. At cohort level, the interval indicates a mean reduction in symptom score. Individual changes, however, remained heterogeneous. Some participants showed little change or worsening.

Table 1. Participant and recording characteristics.

Characteristic	Value
Participants	23
Female / male	10 (43.5%) / 13 (56.5%)
Unipolar / bipolar II / bipolar I	15 / 7 / 1
Outpatient / inpatient	18 (78.3%) / 5 (21.7%)
Age bands: 20–39 / 40–59 / 60–69 years	9 / 12 / 2
Minute-level activity observations	551,716
Complete 24-hour days	359
Complete days per participant	Median 14; IQR 14–15; range 12–28
Baseline MADRS	Mean 22.74; SD 4.80
Endpoint MADRS	Mean 20.00; SD 4.73
MADRS change	Mean −2.74; 95% CI −4.35 to −1.22

Notes: CI, confidence interval; IQR, interquartile range; MADRS, Montgomery–Åsberg Depression Rating Scale; SD, standard deviation. Change is endpoint minus baseline; negative values indicate improvement.

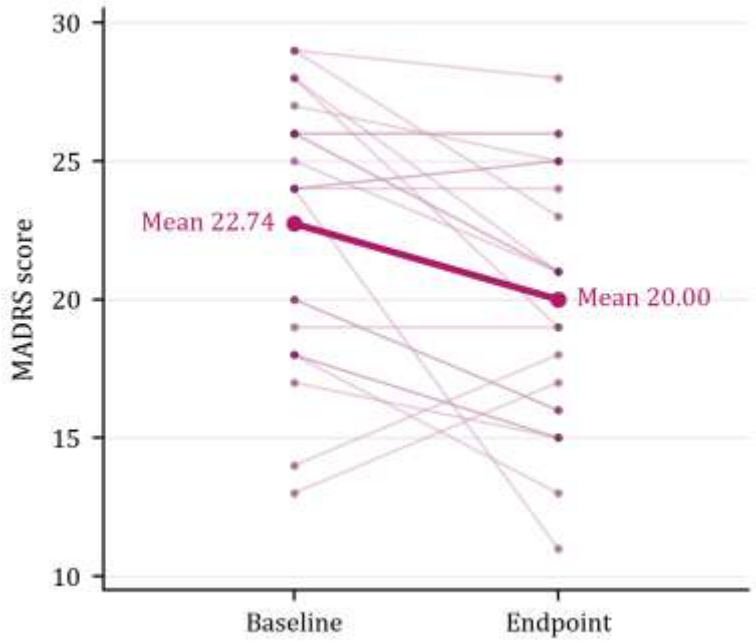


Figure 1. Participant-level change in MADRS scores. Thin lines connect baseline and endpoint values for each participant (n = 23; identical values overlap); the thick line shows the cohort means (22.74 and 20.00). Lower scores indicate lower depressive symptom severity. MADRS, Montgomery–Åsberg Depression Rating Scale.

3.2. Nonlinear features and diagnostic subgroup

None of the five participant-level nonlinear features showed a material monotonic association with MADRS change. Spearman coefficients ranged from −0.033 for Higuchi fractal dimension to 0.104 for the detrended fluctuation analysis exponent. Every confidence interval included zero, and all false-discovery-rate-adjusted p values were 0.966. The directions were inconsistent across complexity domains: higher sample entropy and Higuchi dimension were weakly associated with more negative change, whereas the other three estimates were weakly associated with more positive change. Every coefficient was close to zero, so multiplicity correction alone did not explain the result.

Participants in the bipolar group had a mean MADRS change of −4.38 points, compared with −1.87 points in the unipolar group. The between-group mean difference was −2.51 points (95% CI, −6.16 to 0.82), and the exact Mann–Whitney comparison was not statistically significant (p = 0.357). Table 2 presents the nonlinear-feature associations and the exploratory diagnostic contrast. The wide interval includes both a clinically relevant difference and a small one, which reflects the limited subgroup size. The comparison is exploratory. It gives context for heterogeneity but does not establish diagnosis-specific response.

Table 2. Associations with MADRS change.

Predictor / contrast	Estimate	95% CI	p value
Sample entropy	ρ = −0.009	−0.447 to 0.432	0.966
Permutation entropy	ρ = 0.097	−0.355 to 0.500	0.966
DFA exponent	ρ = 0.104	−0.409 to 0.545	0.966
Higuchi fractal dimension	ρ = −0.033	−0.505 to 0.481	0.966
Spectral entropy	ρ = 0.041	−0.457 to 0.534	0.966
Bipolar minus unipolar mean change	−2.508	−6.158 to 0.817	0.357†

Notes: CI, confidence interval; DFA, detrended fluctuation analysis; ρ, Spearman rank correlation. Feature p values use Benjamini–Hochberg false-discovery-rate adjustment. †Exact Mann–Whitney p value, unadjusted; exploratory comparison.

3.3. Participant-aware prediction

The training-fold mean benchmark achieved an MAE of 3.19 MADRS points (95% CI, 2.26–4.28), RMSE of 4.06, and cross-validated R^2 of −0.093. The clinical ridge model produced an MAE of 3.31, RMSE of 4.28, and R^2 of −0.216. The nonlinear ridge model produced an MAE of 3.21, RMSE of 4.08, and R^2 of −0.106, and the combined model an MAE of 3.22, RMSE of 4.09, and R^2 of −0.111. Table 3 details model performance, and Figure 2 visualizes it with participant-bootstrap uncertainty. No learned predictor set improved any of the three metrics relative to the benchmark.

For the predefined primary comparison, benchmark MAE minus nonlinear-model MAE was −0.021 points (95% CI, −0.035 to −0.008). Equivalently, the nonlinear model's absolute error was approximately 0.02 MADRS points greater than the benchmark. The interval excludes zero in the direction of the benchmark, but the magnitude is negligible in clinical terms, and the intercept-only analysis in Section 3.4 carries the interpretive weight. The primary ridge grid selected its

maximum penalty of 1,000 in the median outer fold for every learned model. Inner validation therefore consistently preferred extensive coefficient shrinkage.

Table 3. Participant-aware prediction performance.

Strategy	MAE (95% CI)	RMSE	R^2
Training-fold mean benchmark	3.194 (2.259–4.285)	4.057	−0.093
Clinical ridge	3.309 (2.296–4.485)	4.279	−0.216
Nonlinear ridge	3.214 (2.273–4.313)	4.082	−0.106
Combined ridge	3.224 (2.277–4.323)	4.090	−0.111

Notes: CI, confidence interval; MAE, mean absolute error; RMSE, root mean squared error. Metrics use outer leave-one-subject-out predictions. Confidence intervals are based on 5,000 participant-bootstrap replicates. For the primary nonlinear comparison, benchmark MAE minus model MAE was −0.021 (95% CI, −0.035 to −0.008); positive values would favor the nonlinear model. Differences were computed from unrounded values.

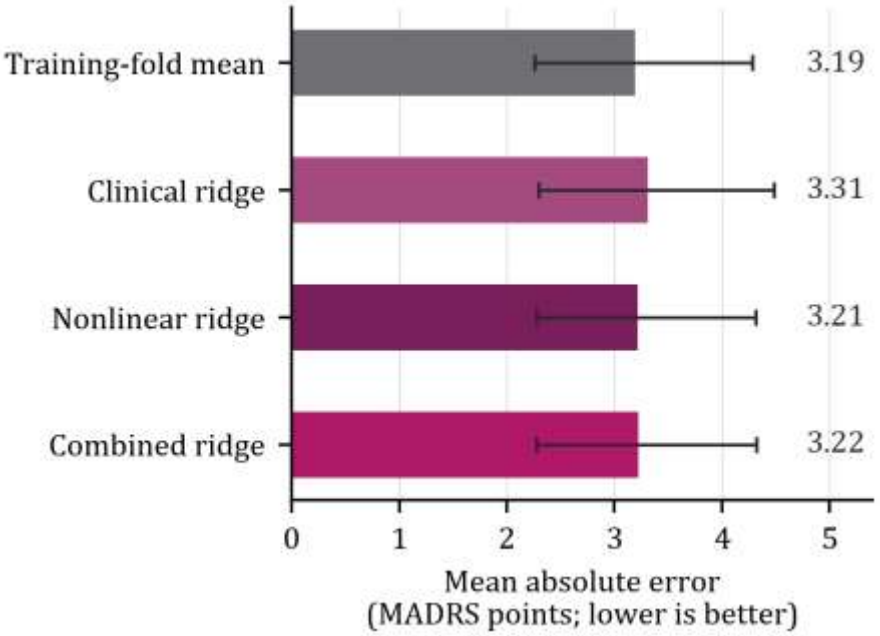


Figure 2. Participant-aware mean absolute error by prediction strategy. Bars show outer leave-one-subject-out MAE; whiskers show 95% participant-bootstrap confidence intervals (5,000 replicates). Lower values indicate better prediction. MAE, mean absolute error; MADRS, Montgomery–Åsberg Depression Rating Scale.

3.4. Sensitivity analyses

When the regularization range was expanded and an explicit intercept-only candidate was allowed, the nonlinear and combined strategies selected the intercept-only solution in all 23 outer folds. Both then exactly reproduced benchmark performance: MAE 3.19, RMSE 4.06, and cross-validated R^2 −0.093. The clinical strategy selected the intercept-only solution in 22 of 23 folds and had an MAE of 3.31, closely matching the primary clinical result. These findings clarify that the boundary selection in the primary ridge grid represented a preference for essentially no predictor contribution.

Restricting actigraphy to metadata-defined chronological intervals reduced the number of complete days from 359 to 268, with 4–17 complete days per participant. Under this extraction window, the nonlinear and combined models again selected the intercept-only option in all 23 folds and exactly matched the benchmark. The clinical model selected it in 21 folds and had an MAE of 3.20. Table 4 summarizes the primary and metadata-window results sequentially. Consequently, the substantive conclusion did not depend on including complete calendar days beyond the metadata-defined interval.

Table 4. Sensitivity analyses.

Strategy	MAE	Intercept-only folds
A. All complete days		
Training-fold mean	3.194	Not applicable
Clinical	3.305	22/23
Nonlinear	3.194	23/23
Combined	3.194	23/23
B. Metadata-defined interval		
Training-fold mean	3.194	Not applicable
Clinical	3.200	21/23
Nonlinear	3.194	23/23
Combined	3.194	23/23

Notes: MAE, mean absolute error. The all-complete-day analysis included 359 complete days; the metadata-defined interval included 268. Intercept-only candidates and penalties up to 1,000,000 were selected exclusively within each outer development set.

4. Discussion

4.1. Principal findings

This participant-separated secondary analysis found that nonlinear complexity summaries, built from hundreds of thousands of minute-level actigraphy observations, did not improve prediction of MADRS change in 23 adults with unipolar or bipolar depression; the training-fold mean benchmark had an MAE of about 3.19 points. The nonlinear and combined ridge models erred slightly more and had negative cross-validated R^2. Clinical predictors did not improve generalization either. Once uncertainty and multiple testing were taken into account, the univariate associations between single complexity features and change were close to zero; the sensitivity analyses resolved the preference for the largest primary ridge penalty. With an explicit intercept-only option, the nonlinear and combined models selected it in every outer fold, and this held under metadata-aligned activity windows.

The central point is not that motor behavior is irrelevant to depression; the cohort did show a measurable mean fall in clinician-rated severity, and earlier work documents links between activity patterns and depressive states. What these participant-level summaries lacked was stable incremental information about short-term symptom change, beyond what a training-fold central tendency already offered in this sample. Association, discrimination of prevalent diagnosis, and prediction of future within-person change are different tasks. A feature can describe current illness or separate groups and still fail to forecast how an unseen participant's score will change.

4.2. Comparison with the existing literature

These results add a deliberately conservative held-out-participant test to a literature that has often reported promising associations for activity, sleep, circadian, physiological, and multimodal digital measures.[2,4,11,28,41] Such associations are plausible in biology and behavior, because psychomotor slowing, disturbed sleep, fatigue, and loss of routine alter the timing and amplitude of movement. They justify further work on wearable measures. They do not imply that any single summary will predict a continuous change score.

More recent personalized studies have drawn on wider information: longitudinal relapse signatures, repeated symptom labels, smartphone behavior, and multimodal sensing.[18,21,42,45,46] Here there was one change outcome per participant and five time-averaged complexity predictors; the question was therefore narrower: did average nonlinear organization across the observed days carry a reproducible participant-level signal about MADRS change? In these data it did not.

Recent activity and wearable studies report diagnostic discrimination, subtype characterization, severity assessment, and clustering of change trajectories.[27,31,33,37,44] Their outcome, validation target, and estimand differ from ours. Case-control classification can use stable differences between people with depression and healthy controls. Predicting symptom change has to explain heterogeneity inside a clinical group. A dataset can thus support diagnostic discrimination and still say little about individual change; the present study complements earlier classification results and does not contradict them.

4.3. Interpretation and mechanism

Sample entropy, permutation entropy, detrended fluctuation analysis, Higuchi fractal dimension, and spectral entropy summarize different mathematical properties.[50–53] Their variety is a strength. It also shows why "complexity" should not be read as one biological construct. Entropy may rise when behavior becomes less regular. Even so, flexible adaptation and disorganized routine can both produce irregularity. A fractal estimate may reflect scale-free organization, but it can also change with inactivity, finite sequence length, device processing, or the distribution of counts. Spectral entropy responds to the balance among circadian, ultradian, and irregular components and does not reveal their clinical origin.

Averaging over participants changes the question again. It steadies daily estimates and limits the influence of an atypical day. However, it discards temporal order and within-person evolution. If behavioral organization shifts gradually with treatment, a slope or a phase-specific contrast could be more informative than the mean. If response begins only after a medication change, an aggregate spanning days before and after the change

may dilute the signal. Day-level modeling without repeated clinical labels, on the other hand, would create pseudo-replication. Future studies need denser clinical assessments and time-aligned treatment information to examine dynamic coupling properly.

The strong shrinkage chosen during nested validation is informative in itself. With few independent outcomes and several correlated predictors, the development folds could not support stable coefficients. An intercept-only candidate made that finding explicit, so the result no longer rested on a boundary penalty. This is preferable methodologically: the inner loop picks the simplest adequate strategy, and arbitrarily small nonzero coefficients are not over-read. The negative cross-validated R^2 values make the same point. The models did not explain held-out variation relative to an outcome-mean reference. MAE alone could have hidden this comparison, and R^2 alone could have hidden the absolute error scale.

Several alternative explanations deserve mention. The link between behavioral complexity and symptom change may be nonlinear or conditional, not monotonic. Highly stereotyped inactivity and highly fragmented activity might both accompany marked symptoms, for example, giving a U-shaped relation that a small ridge model cannot estimate reliably. Complexity may also interact with baseline psychomotor phenotype, since a change in entropy could mean one thing in a participant with marked retardation and another in a participant with agitation or preserved activity. Sedating medication, hospital routines, work schedules, physical comorbidity, and season may alter activity independently of mood response. The data did not allow these interactions to be estimated reliably. Adding them post hoc would have widened the predictor space far beyond what 23 outcomes can support.

Timing offers another explanation. The predictors summarized the observed activity interval, whereas the outcome contrasted two clinical assessments. If part of the actigraphy fell after meaningful improvement, a feature may partly reflect the outcome and not forecast it. If sensing took place before treatment worked, it may miss how response shows up in behavior; the estimand is best read as participant-level prediction from the recorded interval. It is not a mechanistic test showing that complexity precedes improvement. A future prospective protocol should define an index window before the prediction time, set a prediction horizon, record treatment changes, and collect outcomes at fixed intervals; such a design would separate prognostic prediction from concurrent digital measurement.

Statistical and clinical importance. The paired MAE difference between the nonlinear model and the benchmark had a bootstrap interval that excluded zero. It amounted to about two-hundredths of a MADRS point, and it favored the benchmark. A difference of that size has no plausible clinical importance. The finding shows why statistical evidence, direction, and practical size need to be read together. It does not support a claim that nonlinear features cause worse outcomes or are harmful in themselves; it shows only that the fitted prediction procedure added a very small amount of error in this validation sample.

The cohort-level mean MADRS reduction of 2.74 points was larger, and can be read as a modest average change, although the study was not designed to judge treatment efficacy. Treatment regimens, timing, adherence, and other clinical events were not modeled. The bipolar–unipolar contrast also had a nontrivial point estimate, but its uncertainty was wide. Neither result should be turned into a causal or subgroup-treatment conclusion. Both serve to describe the outcome distribution that the prediction models tried to learn.

An MAE near three points may look modest against the MADRS scale, yet usefulness depends on the intended decision. A prediction meant to trigger treatment escalation needs evidence that it improves decisions, calibration, and outcomes. A small average error is not enough. The learned models here did not beat a central-tendency benchmark. Clinical use would add processing and interpretive burden with no demonstrated gain. Within this dataset and target, the benchmark remains the preferred strategy.

4.4. Implications for practice and policy

For research, the results point to three priorities. First, wearable prediction studies should define the independent unit before splitting data. When the intended use concerns new individuals, participant-aware validation should be the default, and analyses with repeated sensor observations should include leakage audits.[47,48] Second, simple benchmarks must be reported. Seen alone, the nonlinear MAE of 3.21 points might look acceptable. The benchmark shows that this error carried no incremental predictive value.

Third, future datasets should align sensing, symptoms, and treatment events over time. Repeated MADRS or brief validated symptom measures would allow dynamic models to ask whether changes in movement precede, accompany, or follow changes in symptoms. Sample sizes should be planned around the number of independent participants and events, not the number of minute records. External validation across devices, care settings, climates, and cultural activity patterns would then be needed before transportability can be claimed.

Feature development could also become more clinically targeted. In place of averaging complexity over all complete days, future work could examine circadian phase, sleep–wake transitions, weekday–weekend contrasts, response trajectories, and interactions between activity amplitude and irregularity. Multimodal measures such as light exposure, heart rate, sleep staging, and ecological momentary assessment could help show whether irregular movement reflects adaptive engagement, disrupted routine, medication effects, or residual symptoms. Any added flexibility should come with nested tuning, shrinkage, and a prespecified analysis plan.

Once sample size permits, model assessment should go beyond discrimination-like summaries. Calibration-in-the-large would reveal systematic over- or

underprediction, a calibration slope would show whether estimated changes are too extreme, and decision-curve analysis could test whether predictions improve an explicit clinical action. Repeated external validation could separate device-specific error from population shift. These evaluations were not estimated here. Calibration models built on 23 outer predictions would be unstable, and no treatment threshold or decision rule had been prespecified. Leaving them out avoids a false appearance of precision and marks the next evidentiary step.

For practice, the findings do not support using these five aggregated nonlinear features to estimate an individual patient's MADRS change. They do support actigraphy as a feasible measurement platform. Participants contributed many complete days, and the data were adequate for reproducible signal processing. Actigraphy may still help with monitoring sleep–wake timing, activity goals, or treatment engagement when read alongside clinical assessment. Predicting symptom change, however, needs incremental evidence at the level of the intended patient and decision.

4.5. Strengths and limitations

The study has several strengths. It used an open dataset, a clinically interpretable continuous outcome, deterministic feature extraction, and a clearly defined direction of change; the five nonlinear measures covered complementary domains and were not variants of one statistic. Evaluation was separated by participant, tuning took place inside each outer development fold, and performance was compared with a training-derived benchmark. Uncertainty was quantified for absolute performance, paired incremental value, feature associations, and subgroup differences. The computational reproduction included 18 validation checks. The sensitivity analyses covered both boundary regularization and the definition of the activity window, and their agreement sharpened interpretation.

The main limitation is the number of independent participants. The sensor archive held more than half a million minute observations, yet a participant-level outcome still rests on only 23 independent units. This limited precision, model complexity, subgroup analysis, and calibration assessment. LOSO validation makes efficient use of the sample. Even so, predictions across folds come from heavily overlapping development sets, and they do not replace external validation. Bootstrap intervals describe uncertainty within this sample and should not be read as evidence of transportability.

Second, paired MADRS observations say nothing about the shape or timing of change between assessments. Treatment details and time-aligned clinical events were unavailable for modeling. The source cohort mixed unipolar and bipolar depression and inpatient and outpatient settings. This created clinically meaningful heterogeneity that could not be modeled reliably at this sample size. Age came in bands, and device or protocol characteristics may limit direct generalization to current consumer wearables.

Third, daily complexity was summarized by the participant mean. That suited the available participant-level outcome, but it left out variability, trend, circadian phase, and episode-specific changes. The complete-day requirement avoided variable-length estimates and also excluded partial-day information. Feature parameters were fixed, and other parameterizations could give different numerical values. Trying many alternatives in the same small sample would, however, add researcher degrees of freedom and would need independent confirmation.

Finally, the conclusion applies to these features and this target only; it does not rule out predictive information in raw time series, conventional activity summaries, multimodal sensors, repeated clinical measurements, or larger samples. Nor does it estimate treatment effects. The appropriate inference is that average nonlinear actigraphy complexity, as operationalized here, added no stable predictive value for the defined MADRS change outcome under participant-aware validation.

5. Conclusion

Five participant-level nonlinear actigraphy complexity features did not improve prediction of MADRS change over a training-fold mean benchmark in 23 adults with unipolar or bipolar depression. The clinical, nonlinear, and combined ridge models all had negative cross-validated R^2. Individual complexity features correlated with change at near-zero levels. With stronger shrinkage and an explicit intercept-only option available, the nonlinear and combined models chose the benchmark-equivalent solution in every fold, and metadata-defined activity intervals gave the same conclusion. Minute-level observations are plentiful. They cannot stand in for an adequate number of independent participants. Claims about wearable prediction call for participant-separated validation, uncertainty estimates, and meaningful benchmarks.

Declarations

Use of Generative Artificial Intelligence. The authors declare that no generative artificial intelligence tool was used in the conception, research, analysis, drafting, revision, or preparation of this manuscript.

Author Contributions. D.M.: conceptualization, methodology, investigation, writing – original draft. M.Y.: software, formal analysis, data curation, visualization, project administration, writing – review & editing. M.A.A.: validation, supervision, writing – review & editing.

Funding. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare that they have no competing interests, financial or non-financial, in relation to this work.

Ethical Approval. The originating studies were reviewed by the Norwegian Regional Medical Research Ethics Committee West under protocols 150.01 and 251.08. This secondary analysis of de-identified open data was approved by the CMHC Research Ethics Committee

(2026/021). No attempt was made to re-identify participants.

Informed Consent. The source dataset reports written informed consent for participation in the originating studies. This secondary analysis used de-identified open data.

Data Availability. The source OBF-Psychiatric dataset is openly available at Zenodo (record 13754984) under a Creative Commons Attribution 4.0 license. Reproduced analysis code and derived non-identifying outputs accompany the study record.

Acknowledgements. The authors acknowledge the investigators and participants who contributed to the OBF-Psychiatric and Depresjon datasets.

References

1. Garcia-Ceja E, Stautland A, Riegler MA, et al. OBF-Psychiatric, a motor activity dataset of patients diagnosed with major depression, schizophrenia, and ADHD. Sci Data. 2025;12(1):32. doi:10.1038/s41597-025-04384-3
2. Lekkas D, Gyorda JA, Price GD, et al. Depression deconstructed: Wearables and passive digital phenotyping for analyzing individual symptoms. Behav Res Ther. 2023;168:104382. doi:10.1016/j.brat.2023.104382
3. Difrancesco S, Penninx BWJH, Riese H, et al. The role of depressive symptoms and symptom dimensions in actigraphy-assessed sleep, circadian rhythm, and physical activity. Psychol Med. 2022;52(13):2760-2766. doi:10.1017/S0033291720004870
4. Akbarova S, Im M, Kim S, et al. Improving Depression Severity Prediction from Passive Sensing: Symptom-Profiling Approach. Sensors (Basel). 2023;23(21):8866. doi:10.3390/s23218866
5. Qiu S, Xing Z. Dose-response relationship between accelerometer-measured physical activity and depression: evidence from the UK Biobank. Transl Psychiatry. 2025;15(1):297. doi:10.1038/s41398-025-03543-9
6. Zou B, Zhang X, Xiao L, et al. Sequence Modeling of Passive Sensing Data for Treatment Response Prediction in Major Depressive Disorder. IEEE Trans Neural Syst Rehabil Eng. 2023;31:1786-1795. doi:10.1109/TNSRE.2023.3260301
7. Zhuparris A, Maleki G, van Londen L, et al. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. Sci Rep. 2023;13(1):18844. doi:10.1038/s41598-023-46075-2
8. Culbreth AJ, Barch DM, Nepal S, et al. Passive sensing of anhedonia and amotivation in a transdiagnostic sample. J Psychopathol Clin Sci. 2025;134(8):893-901. doi:10.1037/abn0001000
9. Wescott DL, Wallace ML, Hasler BP, et al. Sleep and circadian rhythm profiles in seasonal depression. J Psychiatr Res. 2022;156:114-121. doi:10.1016/j.jpsychires.2022.10.019
10. Punturieri C, Duncan WC, Greenstein D, et al. An exploration of actigraphy in the context of ketamine and treatment-resistant depression. Int J Methods Psychiatr Res. 2023;33(1):e1984. doi:10.1002/mpr.1984
11. Chen J, Chan NY, Li CT, et al. Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. Transl Psychiatry. 2024;14(1):150. doi:10.1038/s41398-024-02873-4
12. Xu S, Zhao C, Hu L. Association of accelerometer-derived physical activity pattern with depression: Weekend warrior or regular activity? J Affect Disord. 2025;381:92-99. doi:10.1016/j.jad.2025.04.016
13. Tonon AC, Nexha A, Cunningham JEA, et al. One-Year Actigraphy Study of Sleep and Rest-Activity Rhythms as Markers of Relapse in Depression. JAMA Psychiatry. 2026;83(4):379-388. doi:10.1001/jamapsychiatry.2025.4453
14. Zhao B, Huang X, Xing Y, et al. Sleep irregularity and depression among U.S. adults: An accelerometer-based cross-sectional study in NHANES. Sleep Health. 2026;12(4):612-618. doi:10.1016/j.sleh.2026.04.007
15. Breitinger S, Gardea-Resendez M, Langholm C, et al. Digital Phenotyping for Mood Disorders: Methodology-Oriented Pilot Feasibility Study. J Med Internet Res. 2023;25:e47006. doi:10.2196/47006
16. Faurholt-Jepsen M, Rohani DA, Busk J, et al. Using digital phenotyping to classify bipolar disorder and unipolar disorder - exploratory findings using machine learning models. Eur Neuropsychopharmacol. 2024;81:12-19. doi:10.1016/j.euroneuro.2024.01.003
17. Lamprou E, Kivelä LMM, Rohling JHT, et al. Chronotype, sleep quality, depression and pre-sleep rumination: A diary and actigraphy study. Eur J Neurosci. 2024;60(11):6593-6604. doi:10.1111/ejn.16551
18. Vairavan S, Rashidisabet H, Li QS, et al. Personalized relapse prediction in patients with major depressive disorder using digital biomarkers. Sci Rep. 2023;13(1):18596. doi:10.1038/s41598-023-44592-8
19. Moshe I, Terhorst Y, Opoku Asare K, et al. Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. Front Psychiatry. 2021;12:625247. doi:10.3389/fpsyt.2021.625247
20. Price GD, Heinz MV, Zhao D, et al. An unsupervised machine learning approach using passive movement data to understand depression and schizophrenia. J Affect Disord. 2022;316:132-139. doi:10.1016/j.jad.2022.08.013
21. Matcham F, Carr E, Meyer N, et al. The relationship between wearable-derived sleep features and relapse in Major Depressive Disorder. J Affect Disord. 2024;363:90-98. doi:10.1016/j.jad.2024.07.136
22. Shui X, Xu H, Tan S, et al. Depression Recognition Using Daily Wearable-Derived Physiological Data. Sensors (Basel). 2025;25(2):567. doi:10.3390/s25020567
23. Wu CT, Hsieh MH, Chen IM, et al. Using Wearable Device and Machine Learning to Predict Mood Symptoms in Bipolar Disorder: Development and Usability Study. JMIR Med Inform. 2025;13:e66277. doi:10.2196/66277
24. Phanord CS, Ruzic LL, Kalyanasundaram S, et al. Digital phenotyping of affect and stress in emerging adults. Front Digit Health. 2026;8:1799541. doi:10.3389/fdgth.2026.1799541
25. Salvatore P, Indic P, Khalsa HK, et al. Circadian Activity Rhythms and Psychopathology in Major Depressive Episodes. Psychopathology. 2024;57(1):1-9. doi:10.1159/000530768
26. Melcher J, Lavoie J, Hays R, et al. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. J Am Coll Health. 2023;71(3):736-748. doi:10.1080/07448481.2021.1905650
27. Price GD, Heinz MV, Collins AC, et al. Detecting major depressive disorder presence using passively-collected wearable movement data in a nationally-representative sample. Psychiatry Res. 2024;332:115693. doi:10.1016/j.psychres.2023.115693
28. Carpenter JS, Crouse JJ, Park SH, et al. Actigraphy-derived circadian rhythms, sleep-wake patterns, and physical activity across clinical stages and pathophysiological subgroups in young people presenting for mental health care. J Psychiatr Res. 2025;186:396-406. doi:10.1016/j.jpsychires.2025.03.003

29. Sequeira L, Fadaiefard P, Seat J, et al. Latent class analysis of actigraphy within the depression early warning (DEW) longitudinal clinical youth cohort. Child Adolesc Psychiatry Ment Health. 2024;18(1):149. doi:10.1186/s13034-024-00843-8
30. Ramsey CM, Lynch KG, Gehrman PR, et al. Daily steps and depressive symptoms: A longitudinal evaluation of patients with major depressive disorder in the precision medicine in mental health care study. J Affect Disord. 2022;300:334-340. doi:10.1016/j.jad.2021.12.116
31. Lamichhane B, Moukaddam N, Sabharwal A. Mobile sensing-based depression severity assessment in participants with heterogeneous mental health conditions. Sci Rep. 2024;14(1):18808. doi:10.1038/s41598-024-69739-z
32. Mullick T, Radovic A, Shaaban S, et al. Predicting Depression in Adolescents Using Mobile and Wearable Sensors: Multimodal Machine Learning-Based Exploratory Study. JMIR Form Res. 2022;6(6):e35807. doi:10.2196/35807
33. Rykov Y, Thach TQ, Bojic I, et al. Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling. JMIR Mhealth Uhealth. 2021;9(10):e24872. doi:10.2196/24872
34. Minaeva O, Riese H, Booij SH, et al. Fractal motor activity during wakefulness and sleep: a window into depression recency and symptom recurrence. Psychol Med. 2024;54(15):4429-4437. doi:10.1017/S0033291724002769
35. Lee HJ, Cho CH, Lee T, et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. Psychol Med. 2023;53(12):5636-5644. doi:10.1017/S0033291722002847
36. Unzueta Saavedra J, Deaso EA, Austin M, et al. Smartphone Ecological Momentary Assessment and Wearable Activity Tracking in Pediatric Depression: Cohort Study. JMIR Form Res. 2025;9:e66187. doi:10.2196/66187
37. Ross MK, Tulabandhula T, Bennett CC, et al. A Novel Approach to Clustering Accelerometer Data for Application in Passive Predictions of Changes in Depression Severity. Sensors (Basel). 2023;23(3):1585. doi:10.3390/s23031585
38. Tonon AC, Constantino DB, Amando GR, et al. Sleep disturbances, circadian activity, and nocturnal light exposure characterize high risk for and current depression in adolescence. Sleep. 2022;45(7):zsac104. doi:10.1093/sleep/zsac104
39. Yokoyama S, Kagawa F, Takamura M, et al. Day-to-day regularity and diurnal switching of physical activity reduce depression-related behaviors: a time-series analysis of wearable device data. BMC Public Health. 2023;23(1):34. doi:10.1186/s12889-023-14984-6
40. Xavier NB, Abreu ACVO, Amando GR, et al. Chronobiological parameters as predictors of early treatment response in major depression. J Affect Disord. 2023;323:679-688. doi:10.1016/j.jad.2022.12.002
41. Spulber S, Elberling F, Svensson J, et al. Patterns of activity correlate with symptom severity in major depressive disorder patients. Transl Psychiatry. 2022;12(1):226. doi:10.1038/s41398-022-01989-9
42. Balliu B, Douglas C, Seok D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. NPJ Digit Med. 2024;7(1):49. doi:10.1038/s41746-024-01035-6
43. Walsh RFL, Smith LT, Titone MK, et al. The relationship between physical activity states and depressive symptoms: Using ambulatory assessment to characterize day-to-day associations among individuals with and without bipolar spectrum disorder. Depress Anxiety. 2022;39(12):835-844. doi:10.1002/da.23290
44. Rovero M, Preisig M, Marques-Vidal P, et al. Subtypes of major depressive disorders and objectively measured physical activity and sedentary behaviors in the community. Compr Psychiatry. 2024;129:152442. doi:10.1016/j.comppsych.2023.152442
45. Ikäheimonen A, Luong N, Baryshnikov I, et al. Predicting and Monitoring Symptoms in Patients Diagnosed With Depression Using Smartphone Data: Observational Study. J Med Internet Res. 2024;26:e56874. doi:10.2196/56874
46. Akre-Bhide S, Cohen ZD, Welborn A, et al. Detecting momentary reward and affect with real-time passive digital sensor data. JAMIA Open. 2026;9(1):ooag005. doi:10.1093/jamiaopen/ooag005
47. Karbalaie A, Abtahi F, Häger CK. Participant-Aware Model Validation for Repeated-Measures Data: Comparative Cross-Validation Study. JMIR AI. 2026;5:e87728. doi:10.2196/87728
48. Lee HT, Cheon HR, Lee SH, et al. Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders. Sci Rep. 2023;13(1):16633. doi:10.1038/s41598-023-43542-8
49. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
50. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol. 2000;278(6):H2039-H2049. doi:10.1152/ajpheart.2000.278.6.H2039
51. Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett. 2002;88(17):174102. doi:10.1103/PhysRevLett.88.174102
52. Peng CK, Buldyrev SV, Havlin S, et al. Mosaic organization of DNA nucleotides. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1994;49(2):1685-1689. doi:10.1103/PhysRevE.49.1685
53. Higuchi T. Approach to an irregular time series on the basis of the fractal theory. Physica D. 1988;31(2):277-283. doi:10.1016/0167-2789(88)90081-4