

Original Paper

Individual-Level Modeling of Depressive Symptom Severity Using Smartphone and Wearable Data: Time-Aware 1-Year Study With Feature-Group Contributions

Simon Schreynemackers^{1,2*}, MSc; Johannes Leimhofer^{1,2,3*}, MSc; Milica Petrovic^{1,2}, PhD; Hanna Reich^{1,2}, Dr rer nat; Sascha Ludwig⁴, MSc; Andreas Dominik⁵, Prof Dr; Dominik Heider⁶, Prof Dr rer nat; Ulrich Hegerl^{1,2,7}, Prof Dr med; the MONDY Consortium⁸

¹German Foundation for Depression and Suicide Prevention, Leipzig, Germany

²Research Centre of the German Foundation for Depression and Suicide Prevention, Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Frankfurt, Hesse, Germany

³Institute of Computer Science, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, North Rhine-Westphalia, Germany

⁴Institute for Applied Informatics, Leipzig University, Leipzig, Saxony, Germany

⁵Department of Mathematics Natural Sciences and Informatics, Life Science Informatics Group, University of Applied Sciences, Giessen, Germany

⁶Institute of Medical Informatics, University of Münster, Münster, North Rhine-Westphalia, Germany

⁷Goethe Research Professorship, Department for Psychiatry, Psychosomatics and Psychotherapy, University Hospital Frankfurt, Frankfurt, Hesse, Germany

⁸See Acknowledgments

*these authors contributed equally

Corresponding Author:

Johannes Leimhofer, MSc

Institute of Computer Science, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf

Universitätsstraße 1

Düsseldorf, North Rhine-Westphalia 40225

Germany

Phone: +49 69 6301 86342

Email: johannes.leimhofer@hhu.de

Abstract

Background: Smartphones and wearables can continuously capture behavioral and physiological data in everyday life. Such mobile-sensing data may help track depressive symptoms more closely than occasional retrospective questionnaires, but prior findings have been mixed. Inconsistent findings do not preclude the presence of predictive relationships in specific individuals or time periods. In addition, it remains unclear which broader sensor domains, rather than single features, contribute most to prediction at the individual level.

Objective: The objective of this study is to evaluate whether smartphone and wearable data improve prospective prediction of daily depressive symptom severity beyond a person-specific baseline and, where improvement is observed, to identify contributing sensor feature groups.

Methods: Data from the MONDY (secure and open platform for AI-based health care apps) n-of-1 study were analyzed, including adults with recurrent major depressive disorder who provided up to 12 months of daily self-reported depressive symptom ratings, together with continuous smartphone and smartwatch data. To capture individual-specific patterns, separate participant-specific (idiographic) linear and nonlinear models were developed for each individual. Performance was evaluated using repeated temporally separated training and test periods against a person-specific baseline model predicting the mean symptom score from the training period. Explainability analyses were restricted to models showing improvement over baseline, and summarized at the level of sensor feature groups.

Results: Among 11 participants with sufficient data, linear models improved predictive performance over the person-specific baseline in 8 (73%) participants, and nonlinear models did so in a different set of 8 (73%) participants. Improvements over the person-specific baseline, expressed as reductions in mean absolute error (change in mean absolute error relative to baseline), ranged from 0.05 to 1.23 points for linear models and from 0.12 to 1.07 points for nonlinear models on the adapted Patient

Health Questionnaire-2 scale. Predictive performance varied substantially between individuals, and although nonlinear models exhibited greater aggregate feature-group contributions, this did not translate into a consistent overall predictive advantage over linear models. Feature-group importance profiles were heterogeneous, and no single sensor domain dominated across all participants.

Conclusions: Multimodal smartphone and wearable data can provide individual-specific predictive signals for daily depressive symptom severity in a subset of patients when evaluated under prospective, time-aware conditions. Both predictive performance and contributing sensor domains vary markedly across individuals, and no single modeling approach is uniformly superior. These findings suggest that previously reported associations in mobile-sensing research, often arising from heterogeneous study designs and feature representations, may not consistently translate into prospective prediction, and highlight the importance of evaluation frameworks that assess predictive signal at the temporal individual-participant level.

Trial Registration: German Clinical Trials Register DRKS00032618; <https://drks.de/search/en/trial/DRKS00032618/details>

JMIR Form Res 2026;10:e99204; doi: [10.2196/99204](https://doi.org/10.2196/99204)

Keywords: major depressive disorder; digital phenotyping; smartphone data; wearable devices; idiographic modeling; feature importance; mobile phone

Introduction

Background

Major depressive disorder is a highly heterogeneous and recurrent condition in which symptom burden can fluctuate substantially within the same person over time [1]. In clinical practice, severity is commonly tracked with self-report questionnaires administered at comparatively sparse intervals (eg, weekly to monthly) to guide measurement-based care [2]. However, retrospective symptom reports are susceptible to recall biases and cognitive heuristics, and they often capture different information than high-frequency, momentary assessments [3]. These limitations have motivated interest in objective, ecologically valid markers that can be collected continuously in daily life.

Mobile sensing (MS) aims to infer mental states from passively captured behavioral and physiological signals such as activity [4-8], sleep-wake patterns [4,7-10], social interaction proxies [4], heart rate variability (HRV) [7,9], and speech characteristics [8,11] recorded by smartphones and wearables. Recent reviews conclude that multimodal passive sensing can, in principle, relate to depressive symptoms and trajectories, but results vary widely across studies due to differences in data quality, sensor coverage, preprocessing choices, outcome definitions, and evaluation procedures [12-14]. Long-term MS data are often missing for behavioral rather than purely technical reasons: missingness not at random may reflect withdrawal or reduced engagement, and disengagement from digital monitoring has been associated with symptom severity or relapse risk [15,16]. Such adherence-related signal is often not treated as an explicit domain in model evaluation and interpretation. Importantly, evidence increasingly suggests that personalized (idiographic) modeling and multimodal integration can improve predictive performance compared with one-size-fits-all approaches, consistent with the person-specific nature of depression dynamics [17]. However, this evidence spans a wide range of outcome definitions (eg, daily symptom severity [4], symptom profiles [6], binary or categorical states [5,9,11,18], or longer-horizon symptom change [10]) and model classes

(from linear and mixed models [4,8] to nonlinear machine learning and deep learning [5-7,9-11,18]), which complicates direct comparisons and may contribute to inconsistent findings [12,17].

Even when idiographic models outperform one-size-fits-all baselines on average, prospective performance can remain unstable because the sensor-symptom relationship is not fixed over time [12,19]. One key methodological reason is that the association between sensors and symptoms is plausibly time-varying: routines, treatment effects, seasonality, and life events can change the behavioral meaning of the same sensor feature over time. This phenomenon is closely related to concept drift and model degradation in nonstationary environments [19]. Complementing this, intensive longitudinal clinical research shows that within-person symptom dynamics and relationships among clinically relevant variables can shift over time rather than remaining stable [20]. Together, these considerations imply that a single static model can appear worse than the mean under nonstationarity in prospective evaluation even when informative structure exists intermittently or only in specific phases, making time-aware validation and rolling assessment critical for determining whether there is reliable predictive signal.

Beyond predictive performance, interpretability represents a second major challenge. Many mobile-sensing studies report importance at the level of individual predictors (eg, coefficient lists [4,8], impurity-based rankings [10,21], and SHAP [Shapley Additive Explanations] summaries [9,18]), but high-dimensional sensing pipelines routinely produce large sets of correlated features. In such settings, per-feature attribution can be unstable and difficult to communicate, especially if explanations rely on assumptions that are strained by correlated predictors. For example, widely used SHAP variants can yield misleading attributions when feature dependence is not handled appropriately [22]. Relatedly, some work derives patient-specific physiological biosignatures from wearables, but these approaches typically summarize selected predictors rather than quantifying domain contributions via prospective, performance-based perturbations over time [7].

A pragmatic and clinically interpretable alternative is to shift from single-feature explanations toward sensor-domain (feature-group) explanations, where predictors are aggregated into interpretable modality groups (eg, activity regularity, HRV, sleep, communication, app interaction, and audio), because domains align with clinically meaningful behaviors (sleep, activity, and social withdrawal). Grouped importance supports cross-model comparability because it can be implemented in a model-agnostic manner and summarized consistently across linear and nonlinear approaches. Methodologically, permutation-based grouped importance has been formalized and reviewed as a general framework for model-agnostic explanation at the feature-group level [23]. This perspective is particularly relevant for the multimodal estimation of depressive symptom severity using smartphones and wearable data, where the scientific question often concerns which domains are informative rather than which single derived feature dominates [8].

Against this background, it remains necessary to examine whether multimodal passive sensing can predict day-to-day symptom variation under conditions that more closely reflect real-world use. In particular, even when idiographic models outperform one-size-fits-all approaches on average, prospective performance may remain unstable because sensor-symptom relationships are likely to vary over time. At the same time, a key limitation of the existing literature is that predictive performance is often evaluated using nonprospective or insufficiently time-aware validation strategies and only rarely benchmarked against simple person-specific baseline models. As a result, it remains unclear whether reported associations translate into meaningful predictive signal beyond trivial reference models under real-world longitudinal conditions.

To address these limitations, this study combines (1) strictly prospective, time-aware evaluation, (2) explicit comparison against a person-specific baseline, and (3) model-agnostic, domain-level explainability restricted to cases with demonstrated predictive value using rolling, block-wise permutation-based attribution that respects temporal structure. This design enables a more rigorous assessment of when and for whom MS provides incremental predictive information, while linking interpretability analyses explicitly to demonstrated predictive signal.

Objective

This study aims to evaluate whether multimodal passive sensing features can predict day-to-day variation in depressive symptom severity at the individual level using up to 12 months of daily self-reported symptoms and continuous smartphone and smartwatch data from adults with recurrent major depressive disorder participating in the MONDY (secure and open platform for AI-based health care apps) trial [24]. To address this aim, elastic net (EN) regression as the primary linear modeling approach and random forest (RF) regression as a complementary nonlinear approach were applied to examine whether passive sensing features provide predictive information beyond a simple person-spe-

cific baseline and whether nonlinear modeling improves detection of such participant-specific signal.

The following research questions are addressed:

Primary research question 1 (RQ1): In how many patients do passive sensing models show evidence of improved prospective prediction of daily depressive symptom severity beyond a simple person-specific baseline?

Additionally, the following research questions were investigated:

Research question 2 (RQ2): For patients in whom models show evidence of improvement over the person-specific baseline, which sensor-derived feature groups contribute most to the prediction of individual depressive symptom severity?

Research question 3 (RQ3): Do nonlinear models change (1) the number of patients showing evidence of improved prediction beyond baseline and (2) the relative contribution of sensor feature groups compared with linear models?

Methods

Study Design and Cohort

The data for the present investigation were prospectively collected as part of the MONDY exploratory n-of-1 trial (German Clinical Trials Register DRKS00032618). This study's design involved repeated measurements of self-reported, behavioral, and physiological outcomes over 1 year (365 d) within individuals without manipulation of variables by the investigators. The work was carried out in accordance with the World Medical Association Declaration of Helsinki [25].

In total, 15 adults with recurrent major depression participated in this study and self-monitored their symptoms for up to 12 months using self-report questionnaires and smartphone and smartwatch data. Participants completed the Patient Health Questionnaire-2 (PHQ-2 [26]) every evening and a full Patient Health Questionnaire-9 [27] once per week, providing repeated self-reported measures of depressive symptom severity. MS was carried out with a Samsung Galaxy Watch 5 and the companion "iTrackDepression" smartphone app, which continuously recorded the following data modalities: telephone activity, daily prompted voice recordings (30-60 s) in response to randomly selected questions from a predefined set of 16 prompts for acoustic analysis, step counts, wrist accelerometry, app-usage logs, message traffic, and optical heart-rate-derived HRV, as detailed in the MS section of the MONDY study protocol [24].

Feature Extraction and Grouping

Daily Segmentation and Grouping of Features

To examine within-person relationships between passive sensing data and depressive symptoms, all sensor-derived features were aligned to 2 fixed daily windows relative to the evening PHQ-2 rating: the preceding night (6 PM-6 AM) and the same-day daytime period (6 AM-6 PM). This segmentation captures established diurnal patterns in behavior

and physiology while ensuring temporal alignment between predictors and outcomes [28,29].

To facilitate interpretable modeling and explainability analyses, derived variables were organized into modality-specific feature groups representing literature-based clusters of behavioral or physiological information. Feature groups served as units for later permutation-based domain contribution analyses.

To reduce redundancy in the high-dimensional feature space, we applied a 2-step collinearity control procedure. First, within homogeneous feature families, we removed 1 variable from pairs with very high Spearman correlation ($|r| \geq 0.90$), retaining the more interpretable or statistically stable representative [30]. Second, the remaining variables underwent iterative variance inflation factor pruning using a threshold of 10 until multicollinearity was reduced or only domain-anchored features remained [31]. Domain-specific rules further avoided algebraic dependencies (eg, simultaneous inclusion of raw values, normalized variants, and ratios) and prioritized features consistent with established measurement conventions in HRV and accelerometry research [32,33]. After feature extraction, grouping, and collinearity reduction, the final configured feature set comprised 130 predictors across 12 feature groups. All retained predictors, their feature-group assignments, and the preprocessing scaler applied to each predictor are reported in [Multimedia Appendix 1](#).

Outcome Variable

For the outcome variable, daily depressive symptom severity was assessed using an adapted version of the PHQ-2 [26], capturing the 2 core symptoms of depression (loss of interest or joylessness and feeling down, depressed, or hopeless). Following previously established procedures [20,34], answers were given on a visual analog scale ranging from 0-10 with labeled end points (0=never, 10=all the time). Therefore, the outcome variable was a daily adapted PHQ-2 symptom sum score based on the 2 PHQ-2 core symptoms assessed each evening on 0-10 visual analog scales.

Time Features

To account for secular trends and calendar-related effects, a baseline time-related features (eg, day of week, month of year, and elapsed time; TIME) block was included containing elapsed study time, day of week, and month of year.

Speech and Acoustic Features

Participants provided prompted free-speech recordings in response to up to 4 randomly selected questions each day from a predefined set of 16 prompts. No standardized text was read, and only acoustic characteristics, not speech content, were analyzed. Acoustic descriptors were extracted from the 30-60 s evening Daily-Dialog recordings using the openSMILE toolkit [35]. The resulting speech and acoustic feature group captured key dimensions of vocal expression, including loudness and its variability (vocal energy and emphasis), perturbation measures such as jitter and shimmer (voice stability), and the first 4 mel-frequency cepstral

coefficients representing the spectral envelope and timbral properties of speech. These parameters summarize prosodic, spectral, and voice-quality characteristics that have been associated with mood and suicide risk in prior work [36-41].

Phone Communication Features

Phone behavior was summarized in the phone communication features feature group using 4 complementary metrics computed separately for day and night windows: total call duration (interaction intensity), call frequency (activity volume), number of unique contacts (network breadth), and missed calls (responsiveness or opportunity for interaction).

Activity Features

Wrist accelerometer data were processed using the SciKit-Digital-Health framework [32] following standard preprocessing steps including resampling, gravity correction, and reduction to the Euclidean Norm Minus One activity signal. The resulting activity features (accelerometry-derived features) feature set was grouped into 3 domains capturing core motor correlates of depression: fragmentation (irregularity and burstiness of activity [activity fragmentation]), intensity (overall motor output and time spent in activity states [activity intensity]), and rhythmicity (circadian regularity and temporal organization of activity patterns [activity rhythmicity]).

App Usage and Social Interaction Features (App Traffic Features and App Usage Features)

Smartphone interaction features were derived from passive Android sensors capturing app traffic and foreground usage. Daily transmitted and received data volumes and total foreground durations were aggregated for communication and social media apps identified via Google Play meta-data. These variables formed the feature groups app traffic features, app usage of communication apps, and app usage of social media apps, representing digital interaction density and active engagement with socially oriented apps. This grouping reflects evidence that app traffic volume provides an indirect proxy for digital interaction density, whereas app foreground time quantifies active engagement, and that such digital behavioral markers are associated with stress, anxiety, and depressive symptoms [42-44].

HRV Features

Interbeat interval series were preprocessed using NeuroKit2 for beat detection, artifact correction, and quality control [33]. From valid segments, time-domain, frequency-domain, and nonlinear HRV metrics capturing overall variability, autonomic balance, and signal complexity were derived. These variables were combined into the HRV feature group.

Sleep Features

Sleep duration was estimated using HRV-derived sleep staging with the deep-learning classifier wrn-gru-mesa [45]. Cleaned interbeat interval sequences were segmented into epochs and classified into sleep stages, and total daily sleep time (hours) was computed. This variable defined the sleep

duration feature group representing overall nightly sleep duration.

Adherence Indicator Features

To capture behaviorally informative absence patterns, we derived minimal missing not at random (MNAR) indicators across sensing modalities. These daily flags captured sensor dropout, diurnal imbalance, low coverage, or consecutive absence streaks in HRV, accelerometry, calls, audio recordings, sleep estimation, and app usage streams. All adherence indicators were combined into a single adherence indicator features feature group reflecting sensor availability and digital engagement patterns.

Data Analysis

Participant Exclusion Due to Insufficient Data

Participants were anonymized using single-letter identifiers (eg, A, B, and C) and excluded if their time series did not meet the minimal length needed to construct at least one temporally valid train-test split. A time-based validation scheme with a minimum 60-day training window and a subsequent 30-day test window was chosen to reflect the intended prospective prediction scenario, prevent temporal leakage, and account for the nonstationary nature of depressive symptom dynamics. Consequently, 4 of the 15 participants with fewer than 90 valid daily assessments were excluded and not used for model training. This requirement was determined by the sum of the minimum training window (60 d) and the minimum test-window constraints (30-d holdout window).

Data Preprocessing

All preprocessing steps were performed separately for each participant and implemented as a single preprocessing pipeline within each cross-validation fold. To preserve the temporal order of observations and prevent information leakage, every preprocessing component (eg, imputation and scaling) was fitted exclusively on the training data and subsequently applied unchanged to the corresponding test window.

Daily sensor features were first aligned to the corresponding evening PHQ-2 score, and days without a valid outcome were removed. As mobile and wearable sensor data naturally contain extreme or physiologically implausible values, often due to recording artifacts, intermittent device use, or measurement noise, outliers were identified using the Tukey rule [46]. Values lying more than 1.5 times the IQR beyond the first or third quartile were set to missing. This reduced the influence of extreme observations on subsequent modeling.

No explicit feature exclusion threshold based on missingness was applied. Missing values were imputed using multivariate imputation by chained equations [47], implemented with scikit-learn's `IterativeImputer` (maximum 50 iterations, convergence tolerance of 1×10^{-2}) [48]. Consequently, all participant-specific models were fitted using the same set of 130 features. A participant-level summary of feature missingness before imputation is provided in

[Multimedia Appendix 2](#). Among the 11 participants included in modeling, the median number of features exceeding 50% missingness was 0 (IQR 0.0-1.5), and the median number exceeding 70% missingness was 0 (IQR 0.0-0.5). The 50% and 70% thresholds were used descriptively and were not applied as feature-exclusion criteria. Features were retained despite high missingness because the proportion of missing data alone was not considered a sufficient criterion for exclusion. Previous work [49] demonstrated that multivariate imputation by chained equations can yield unbiased estimates even with up to 90% missingness when auxiliary information is available. To partially account for informative missingness, adherence-related indicators were deterministically derived from the original missingness patterns before imputation and subsequently included as auxiliary predictor variables; these indicators themselves were not imputed. In MS and mental health contexts, missingness is often MNAR and reflects user adherence or disengagement, making omission of such variables potentially biasing.

Continuous predictors were transformed using a 2-stage scaling process, designed both to harmonize feature distributions and to create optimal conditions for EN regularization. In the first stage, each continuous feature underwent 1 of 3 possible transformations depending on its distributional properties: right-skewed variables were stabilized using a natural logarithm transformation ($\log_{1p}$), variables with bounded ranges were normalized to the [0, 1] interval using minimum-maximum scaling, and approximately symmetric variables were standardized to have mean 0 and SD 1. These transformations ensured that features spanning different physical units (eg, app usage seconds, HRV indices, and accelerometer-based counts) were brought to a common numerical scale (the scaler used for each feature is described in [Multimedia Appendix 1](#)). In the second stage, all features, regardless of the transformation applied in the first stage, were again z-standardized. The second standardization step ensured centered and variance-normalized inputs for EN regularization. For consistency, the same preprocessed feature matrix was supplied to RF models.

Model Training and Validation

Prediction Models

Daily adapted PHQ-2 sum scores served as the dependent variable. Two complementary regression model types were considered: EN regression [50] and RF regression [51], following prior work by Balliu et al [4] and Price et al [52].

Model development followed a nested time-series cross-validation procedure designed to preserve temporal ordering while preventing optimistic performance estimates.

Outer Cross-Validation (Model Evaluation)

The outer loop implemented a repeated time-series cross-validation scheme with 10 splits and 5 repetitions. Each split trained on at least 60 consecutive days and evaluated predictions on a subsequent window of up to 30 days (minimum 14 days of valid observations). To increase robustness, the starting point of the cross-validation procedure

was randomly offset within the first 20% of the time series in each repetition. This produced multiple forecast origins while maintaining chronological ordering.

Within each outer fold, the model was trained only on past observations and evaluated on future observations, approximating a prospective forecasting scenario.

Inner Cross-Validation (Hyperparameter Tuning)

Hyperparameters were tuned exclusively within the training portion of each outer fold using an inner time-series cross-validation procedure based on TimeSeriesSplit. For EN, the grid search explored combinations of the regularization strength α and the L1-L2 mixing parameter (see [Multimedia Appendix 3](#)). For RF, grid search was performed over tree depth, number of estimators, and minimum leaf size. Model selection in the inner loop used the coefficient of determination (R^2) as the optimization criterion.

After hyperparameter selection, the model was refitted on the full outer training window using the selected hyperparameters and evaluated on the corresponding outer test window.

Eligibility Analysis

To address research question 1 (RQ1) and research question 3 (RQ3), the first analytical step determined whether EN and RF models captured predictive structure beyond a trivial baseline for each participant.

Within each outer fold, predictions from the tuned EN and RF models were compared with a constant baseline model implemented as a DummyRegressor [48] that always predicted the participant's mean PHQ-2 score observed in the training window. This simple person-specific baseline was intentionally chosen to represent prediction based on symptom history alone. Accordingly, improvement beyond this baseline indicates incremental predictive information from passive sensing, but not necessarily clinically useful prediction accuracy.

Two performance metrics were computed for each fold: (1) R^2 reflecting predictive performance relative to the test-set mean and (2) mean absolute error (MAE), representing the average absolute prediction error. Performance was expressed relative to the baseline model: $\Delta R^2 = R^2_{\text{model}} - R^2_{\text{baseline}}$; $\Delta \text{MAE} = \text{MAE}_{\text{baseline}} - \text{MAE}_{\text{model}}$. As the baseline model was fitted on the training window and evaluated on temporally subsequent test observations, its R^2 was not constrained to zero and could vary across folds, including negative values. Positive values therefore indicated improved predictive performance relative to the constant predictor.

To quantify the reliability of these improvements at the participant level, a weighted bootstrap procedure with 20,000 resamples was applied. For each participant and model type, folds were resampled with replacement and the weighted mean improvement in change in coefficient of determination relative to baseline (ΔR^2) and change in mean absolute error relative to baseline (ΔMAE) was computed, using the number of test observations in each fold as weights. This yielded

an average improvement estimate and a 1-sided P value. To account for multiple testing across participants and model types, P values were adjusted using the Benjamini-Hochberg procedure controlling the false discovery rate (FDR) at $\alpha = .05$. Adjustment was performed separately for each performance metric (ΔR^2 and ΔMAE) across both model types. This approach was selected to limit the expected proportion of false-positive findings across the multiple participant-level comparisons while maintaining sensitivity for this exploratory analysis.

A model type was considered to significantly outperform the baseline if either ΔR^2 or ΔMAE exhibited an FDR-adjusted P value $< .05$. This inclusive criterion was chosen because the 2 metrics capture complementary aspects of predictive performance (relative model fit [ΔR^2] and absolute prediction error), and the objective of this screening step was to identify participant-model combinations showing any evidence of predictive information beyond the person-specific baseline.

Participant-model combinations that significantly outperformed the baseline were retained for the corresponding model-specific explainability analysis. Participants for whom neither model type satisfied the eligibility criterion were classified as noneligible. Comparisons between EN and RF feature-group contributions were restricted to participants whose models were eligible under both approaches.

The nested cross-validation procedure served as the primary participant-level eligibility and model-comparison stage, providing leakage-safe and temporally prospective estimates of predictive performance relative to baseline. The subsequent 50% train-holdout split was not used as an additional confirmatory test of eligibility, but rather as a standardized framework for retuning eligible models and conducting rolling explainability analyses under a fixed prospective reference period.

Explainability Analysis

To address RQ2 and the feature-importance component of RQ3, a rolling, time-aware permutation-importance analysis was performed for all eligible participant-model combinations.

For these analyses, an initial training window covering the first 50% of the participant's observation period (in calendar time) was defined, with the remaining 50% reserved as a prospective holdout period. The preprocessing pipeline was fitted solely on the initial training window and applied to both training and holdout data. Hyperparameters were retuned within this initial window using time-series cross-validation and subsequently fixed for the explainability stage.

Using these fixed models, expanding rolling-origin evaluation windows were constructed across the observation period. Starting from the initial training window, the training set was extended in 30-day increments, each paired with a subsequent 30-day test window. In each fold, a new model was trained on the current training window and evaluated on the corresponding test window.

Feature relevance was assessed using block-wise permutation importance, a model-agnostic method that quantifies the deterioration in predictive performance when feature information is disrupted. To preserve temporal dependencies and short-term autocorrelation, permutations were applied within contiguous 7-day blocks rather than individual observations. For each feature, the column was permuted in 7-day blocks across 50 repetitions, and the resulting change in R^2 and MAE relative to the original model was recorded.

Relative importance scores were then computed by dividing the change in MAE by the baseline MAE of the fold (relative ΔMAE) and the change in R^2 by the absolute baseline R^2 (relative ΔR^2). These relative metrics express the proportional degradation in predictive accuracy caused by permuting a feature and allow comparison across participants, folds, and model types. As relative ΔR^2 can become unstable when baseline R^2 is close to 0, interpretation focused primarily on relative ΔMAE .

To mitigate distortions caused by multicollinearity among individual variables, permutation importance was also computed at the feature-group level by jointly permuting all variables belonging to the same sensor or behavioral category (eg, speech and acoustic features, app usage features [APP_USAGE], and HRV). Group-level importance was calculated analogously to feature-level importance and aggregated across folds. These importance scores quantify the predictive reliance of the fitted model on a feature group under temporal perturbation and should not be interpreted as causal effects of that domain on depressive symptoms. Negative contributions, rare cases where permutation accidentally improved predictions, were truncated at zero before aggregation.

For each participant and model type, importance scores were finally aggregated across rolling windows to obtain participant-level summary estimates of feature and feature-group contributions. These participant-level importance profiles were then summarized across eligible participants to compare the distribution and rank order of behavioral and physiological information sources between EN and RF models.

This framework allowed evaluation of whether a common pattern of dominant feature groups emerged across individuals and whether nonlinear RF models exploited the available behavioral signals differently from linear EN models.

Protocol Deviations

Although this study closely followed the preregistered methods protocol [53], several methodological refinements were introduced to enhance statistical validity and interpretability. The first adjustment concerned the definition of eligibility for downstream explainability analyses. In the preregistered plan, participants were to be considered eligible for further analysis only if their models achieved an R^2 of at least 0.30 and an MAE below 1. However, preliminary analyses revealed that no participant met both criteria, despite clear evidence of systematic within-person associations in

several cases [54]. To avoid excluding all participants while maintaining rigorous evaluation, the fixed thresholds were replaced by a bootstrap-based statistical test that determined whether a model's predictive accuracy exceeded that of a constant baseline predictor. Participants whose models failed to show a statistically significant improvement over this baseline were excluded from further modeling, whereas those with significant improvements were retained. This inferential approach allowed for individualized assessment of model performance and preserved the intent of the original research question.

The second modification involved the explainability procedure. The preregistered protocol proposed using the model-agnostic Kernel SHAP framework. During implementation, this was replaced by rolling block-wise permutation importance for both model types because Kernel SHAP relies on feature-independence assumptions that are difficult to reconcile with strongly autocorrelated behavioral data and grouped sensor features, whereas block-wise permutation importance remains model-agnostic, respects temporal structure, and scales tractably across all participants and folds.

All analyses were conducted in Python using scikit-learn [48]. The complete analysis pipeline has been made publicly available [55], which also provides the software environment and dependency specifications required to reproduce the analyses.

Ethical Considerations

Ethical approval has been granted by the Ethics Committee of the Department of Medicine at Goethe University Frankfurt, Germany (date: August 24, 2023, reference number: 2023-1309).

Participants were recruited at the Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital of Goethe University Frankfurt, Germany. Treating physicians and psychologists were informed about this study and invited eligible patients with affective disorders to participate. Interested patients provided their contact details and agreed to be contacted by this study's center. All participants provided written informed consent before inclusion in this study.

The legal basis for data processing was voluntary informed consent in accordance with Article 6 (1)(a) of the General Data Protection Regulation (GDPR). Data were collected and processed only after participants had provided written informed consent following detailed study and data protection information. Additional permissions for access to individual smartphone sensors were obtained within the iTrackDepression app.

Results

RQ1: How Many Patients Show Predictable Depression Patterns?

Table 1 summarizes participant-level improvements in predictive accuracy for the EN models, expressed as the mean

change in explained variance (ΔR^2) and the mean reduction in prediction error (ΔMAE) compared to a constant baseline predictor.

Among the evaluable participants, 7 participants showed significant improvement in ΔR^2 and 8 participants in ΔMAE ; 7 participants satisfied both criteria and 1 participant satisfied only the ΔMAE criterion. Participants A, C, F, G, I, J, K, and O showed significant gains in either R^2 or MAE (all FDR-adjusted P values $<.05$), classifying them as eligible participants under the revised inferential definition. Conversely, participants H, M, and N did not show significant improvement in either performance measure and were classified as ineligible participants. Participant I exhibited substantially larger ΔR^2 values than the remaining participants. These values resulted from comparison against a participant-specific baseline model with strongly negative prospective R^2 values rather than from explained variance exceeding its theoretical upper bound. Accordingly, the reported ΔR^2 values should be interpreted

as large improvements relative to the baseline predictor rather than as conventional measures of explained variance. As shown in [Multimedia Appendix 4](#), these improvements were observed consistently across the rolling evaluation folds rather than being driven by a small number of extreme observations. The fold-wise MAE plots likewise demonstrate consistent reductions in prediction error relative to the baseline, indicating that the large ΔR^2 values reflect sustained improvement over a poorly performing baseline rather than isolated outliers or numerical artifacts. The corresponding fold-wise eligibility analysis is provided in [Multimedia Appendix 4](#).

Overall, 8 of the 11 (73%) participants with sufficient data (95% Wilson CI 43%-90%) showed evidence of improved predictive performance of the EN models over the baseline model. This indicates that, for several individuals, smartphone and wearable features added incremental predictive information beyond the person-specific baseline.

Table 1. Participant-level elastic net performance relative to a constant baseline predictor.

Participant ID	Mean ΔR^2 (SD)	ΔR^2 P value (FDR ^a -adjusted)	Mean ΔMAE ^b (SD)	ΔMAE P value (FDR-adjusted)
A	0.549 (0.986)	.001	0.147 (0.408)	.03
C	1.079 (3.577)	.002	0.096 (0.214)	.001
F	2.645 (4.309)	<.001	0.763 (0.848)	<.001
G	0.039 (0.119)	.02	0.047 (0.110)	.002
H	-0.049 (0.158)	>.99	-0.046 (0.115)	>.99
I	15.416 (17.114)	<.001	1.232 (0.653)	<.001
J	0.058 (0.094)	<.001	0.084 (0.102)	<.001
K	1.165 (1.510)	<.001	0.217 (0.293)	<.001
M	0.053 (0.221)	.34	0.043 (0.134)	.23
N	-0.071 (0.221)	>.99	-0.085 (0.276)	>.99
O	0.045 (0.307)	.24	0.072 (0.212)	.01

^aFDR: false discovery rate.

^bMAE: mean absolute error.

RQ2: Which Sensor Domains Contribute Most to Individual-Level Predictions?

Permutation-based explainability analyses revealed substantial variation in the relative importance of different feature groups across participants. [Figure 1](#) provides a participant-level stacked bar visualization of relative feature-group contributions, where each bar represents 1 participant and each color segment corresponds to a feature group.

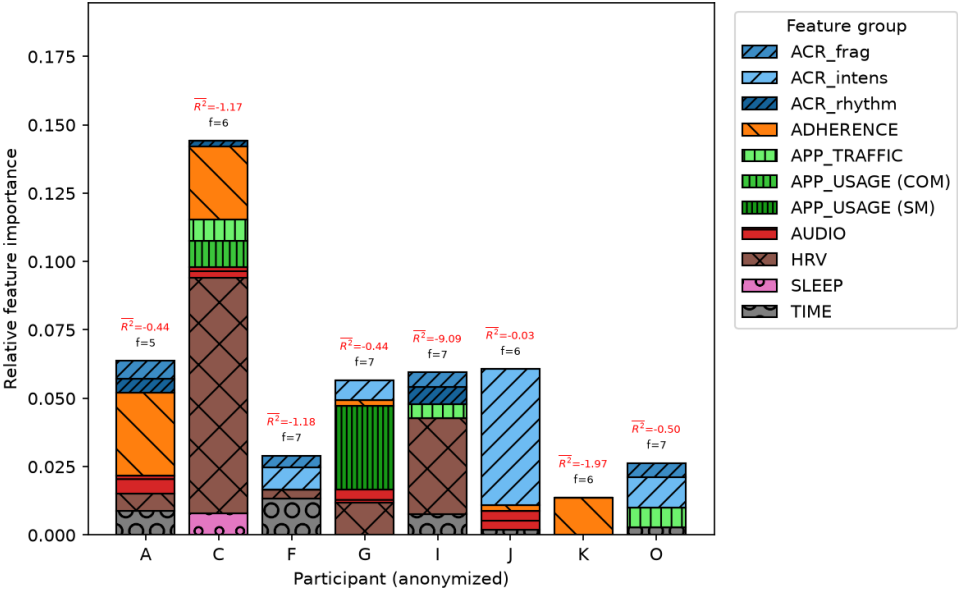
Relative contributions reflect the proportion of prediction degradation attributable to permuting a feature group, expressed using the relative ΔMAE metric, which normalizes contributions within folds and across participants.

Across participants, the permutation-based feature-importance profiles demonstrated substantial heterogeneity in the relative contributions of different feature groups, underscoring the individualized nature of the learned associations. As shown in [Figure 1](#), no single feature group was consistently

dominant across all participants. Instead, the relative influence of sensor modalities varied markedly from person to person. For example, participant G showed comparatively strong contributions from APP_USAGE features, whereas participants C and I exhibited greater influence from HRV-related features. In participants A, C, G, and K, adherence-related features contributed noticeably to model performance. In addition, accelerometry-derived activity regularity measures were among the more influential features for all participants except K. A subset of participants (A, C, G, and J) also showed meaningful contributions from audio features, although these patterns appeared idiosyncratic rather than systematic.

Fold-wise permutation importance profiles for individual participants, illustrating temporal variability in feature-group contributions across expanding windows, are shown for EN models in [Multimedia Appendix 4](#).

Figure 1. Relative contribution of feature groups in elastic net models across participants. Feature groups are color-coded according to the legend. Stacked bars show positive relative feature importance based on Δ mean absolute error (MAE) for each participant after thresholding. Participants are anonymized by letters along the x-axis. The height of each stacked segment represents the mean relative MAE contribution of the respective feature group across rolling evaluation folds. Annotations above the bars indicate the mean R^2 of the participant-specific model and the number of evaluation folds (f); negative mean R^2 values are shown in red. ACR_frag: activity fragmentation; ACR_intens: activity intensity; ACR_rhythm: activity rhythmicity; ADHERENCE: adherence indicator features; APP_TRAFFIC: app traffic features; APP_USAGE (COM): app usage (communication apps); APP_USAGE (SM): app usage (social media apps); AUDIO: speech and acoustic features; HRV: heart rate variability; SLEEP: sleep duration feature group; TIME: time-related features (eg, day of week, month of year, and elapsed time).



RQ3: Do Nonlinear Models Improve Prediction and Change Domain Importance?

Model Performance of RF Models Relative to Baseline

RF models exhibited a somewhat different performance profile compared to EN models, consistent with the possibility that nonlinear relationships were relevant for some participants. Table 2 reports participant-level improvements over the baseline mean predictor. As with EN models, participants were considered eligible when performance gains were statistically significant.

A total of 6 participants showed significant improvement in ΔR^2 and 8 participants in Δ MAE; 6 participants satisfied

both criteria and 2 participants satisfied only the Δ MAE criterion. Participants A, F, G, I, J, K, M, and O showed statistically significant improvements in at least 1 of the 2 performance metrics, classifying them as eligible under the revised inferential criteria. For most of these participants, improvements were observed in both explained variance (ΔR^2) and prediction error (Δ MAE), with FDR-adjusted P values below .05 indicating consistently positive out-of-sample gains across cross-validation folds. Participants G and M were considered eligible based on a significant reduction in MAE despite a nonsignificant ΔR^2 . In contrast, participants C, H, and N did not exhibit improvements on either metric, showing no evidence of predictive structure detectable by the RF model.

Table 2. Participant-level random forest performance relative to constant baseline predictor.

Participant ID	Mean ΔR^2 (SD)	ΔR^2 P value (FDR ^a -adjusted)	Mean Δ MAE ^b (SD)	Δ MAE P value (FDR-adjusted)
A	0.624 (0.838)	<.001	0.187 (0.345)	.001
C	0.529 (2.741)	.16	0.092 (0.600)	.20
F	2.615 (4.158)	<.001	0.746 (0.776)	<.001
G	0.033 (0.260)	.25	0.120 (0.257)	.002
H	-0.121 (0.505)	>.99	-0.092 (0.469)	>.99
I	14.012 (16.403)	<.001	1.069 (0.551)	<.001
J	0.099 (0.160)	<.001	0.206 (0.196)	<.001
K	1.913 (2.213)	<.001	0.392 (0.413)	<.001
M	0.257 (0.505)	.12	0.241 (0.208)	.001
N	-0.270 (0.556)	>.99	-0.325 (0.792)	>.99
O	0.278 (0.443)	<.001	0.250 (0.324)	<.001

^aFDR: false discovery rate.

^bMAE: mean absolute error.

Overall, 8 of 11 (73%) participants with sufficient data (95% CI 43%-90%) demonstrated statistically reliable improvements over baseline when modeled with RF models, consistent with the overall rate of eligible participants observed for the EN models. However, the sets of eligible participants were not identical across modeling approaches, indicating that the 2 model classes captured predictive signal in partly different individuals. This finding suggests that both linear and nonlinear approaches were able to detect incremental within-person predictive signal in some individuals, although the magnitude and composition of these gains differed across participants.

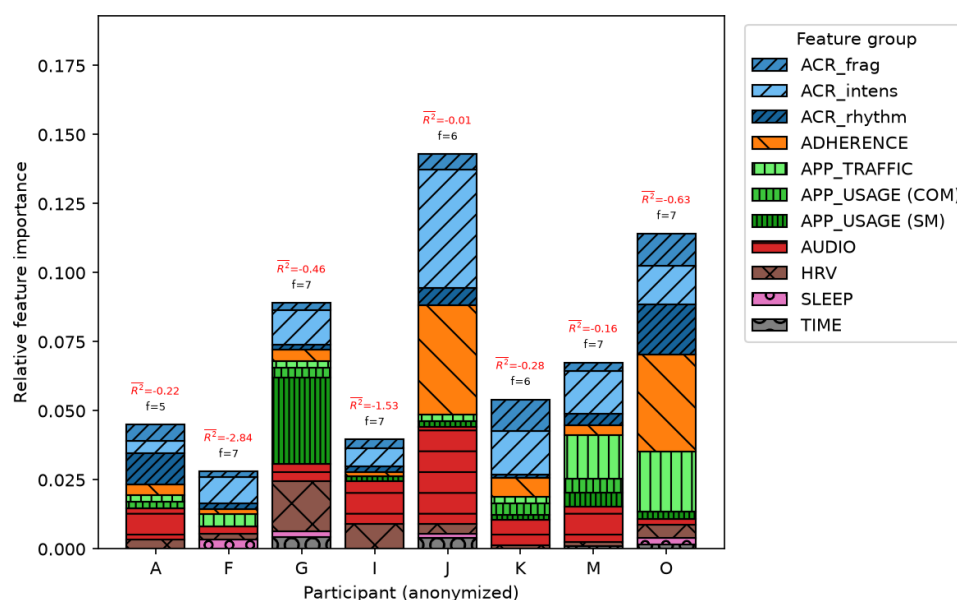
Detailed per-participant results underlying the eligibility analysis, including fold-wise performance of EN and RF models relative to the baseline and the corresponding bootstrap-based improvements (ΔR^2 and ΔMAE), are provided in [Multimedia Appendix 4](#).

Relative Contribution of Feature Groups in RF Models

Permutation-based feature-importance patterns for the RF did not indicate a uniform increase in feature-group contributions

across participants, but rather revealed pronounced inter-individual differences. As shown in [Figure 2](#), for specific participants (notably K and O), the RF models were able to extract incremental predictive information from multiple feature domains. In contrast, the corresponding EN models for these participants showed very limited feature-group contributions, suggesting that the linear models were unable to leverage the available signals effectively in these cases. For other participants, however, feature-group contributions were comparable between model types. Despite this heterogeneous pattern, a paired Wilcoxon signed-rank test showed a statistically detectable difference in aggregated feature-group importance between RF and EN ($P=.03$). To assess the influence of participant I, the aggregate comparison was repeated after excluding this participant. The resulting feature-group importance profiles remained qualitatively similar, and the paired Wilcoxon signed-rank test likewise remained significant ($P=.005$), indicating that the observed aggregate difference was not driven by this participant alone ([Multimedia Appendix 5](#)).

Figure 2. Relative contribution of feature groups in random forest models across participants. Feature groups are color-coded according to the legend. Stacked bars show positive relative feature importance based on Δ mean absolute error (MAE) for each participant after thresholding. Participants are anonymized by letters along the x-axis. The height of each stacked segment represents the mean relative MAE contribution of the respective feature group across rolling evaluation folds. Annotations above the bars indicate the mean R^2 of the participant-specific model and the number of evaluation folds (f); negative mean R^2 values are shown in red. ACR_frag: activity fragmentation; ACR_intens: activity intensity; ACR_rhythm: activity rhythmicity; ADHERENCE: adherence indicator features; APP_TRAFFIC: app traffic features; APP_USAGE (COM): app usage (communication apps); APP_USAGE (SM): app usage (social media apps); AUDIO: speech and acoustic features; HRV: heart rate variability; SLEEP: sleep duration feature group; TIME: time-related features (eg, day of week, month of year, and elapsed time).



Across participants, feature groups related to app usage, activity regularity (activity features [accelerometry-derived features]), and audio characteristics frequently accounted for a substantial proportion of the predictive signal ([Figure 2](#)). At the same time, nearly all feature groups, including HRV, adherence, temporal, and sleep features, showed

notable permutation importance in at least some individuals, highlighting the broad range of behavioral information captured by the models. For example, participants A, F, G, I, J, K, M, and O showed marked contributions from app-usage and activity-related features, whereas sleep features were particularly influential for participants F, G, J, and

O. Overall, no single feature group consistently dominated across participants, reflecting substantial interindividual heterogeneity in the learned feature-importance profiles.

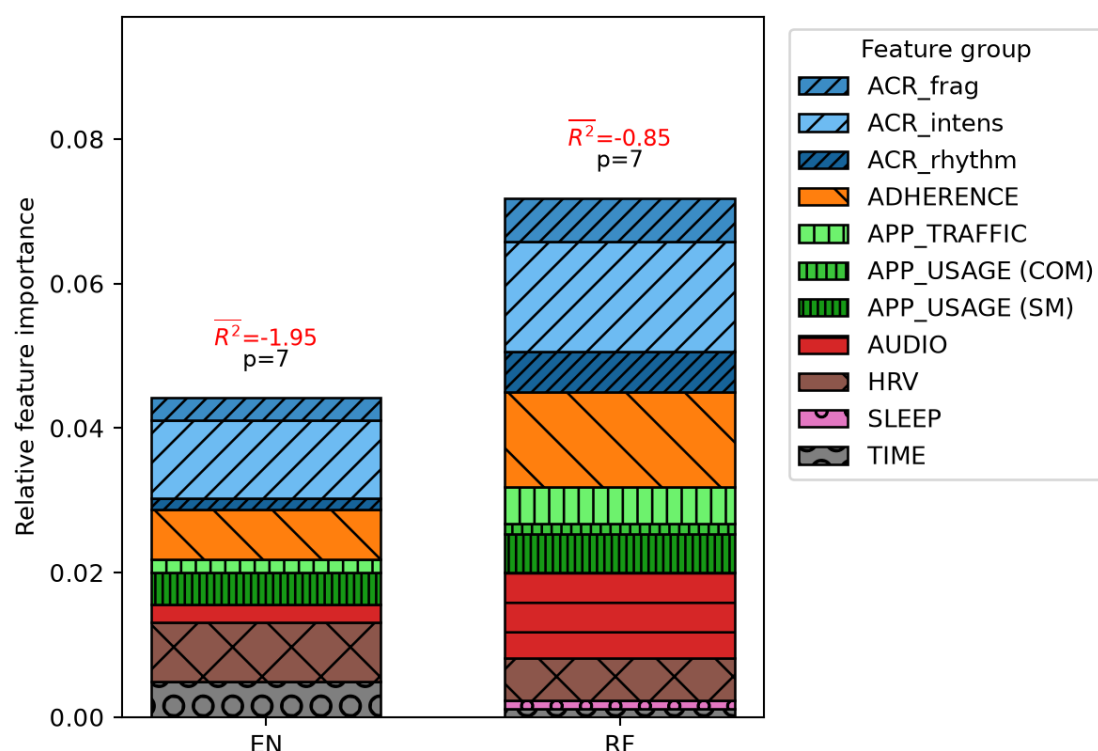
Corresponding fold-wise permutation importance plots for RF models at the participant level are provided in [Multimedia Appendix 4](#).

Comparison of Feature-Group Contributions Between Model Types

Figure 3 presents an aggregated comparison of relative feature-group contributions for the EN and RF models. Each stacked bar summarizes the mean importance values across participants who showed significant predictive improvements in both model types. As this comparison was restricted to participants showing significant predictive improvement with both model types, the aggregated comparison was based on only 7 individuals and should therefore be interpreted as exploratory.

Feature-group permutation importance showed broadly similar contribution patterns across EN and RF models (**Figure 3**). Accelerometry-derived features (activity fragmentation, activity intensity, and activity rhythmicity) contributed most to prediction in both modeling approaches. Smartphone interaction features (APP_USAGE and app traffic features) showed moderate contributions, whereas physiological features such as sleep duration feature group were generally less influential. A notable difference between models was the greater relative importance of calendar-based covariates (TIME) in the EN models, whereas RF models generally showed larger normalized permutation-importance values across the remaining feature groups, suggesting that the nonlinear models were able to leverage predictive information more effectively across multiple sensing domains.

Figure 3. Comparison of aggregated feature-group contributions between EN and RF models. The feature groups are color-encoded according to the legend on the right. Stacked bars represent the positive relative feature importance (Δ mean absolute error [MAE]) aggregated over participants, with each bar showing the combined contribution of all feature groups retained after thresholding. The height of each stacked segment reflects the mean relative MAE contribution of that feature group over participants. Annotations above each bar indicate the average model R^2 across all test folds and included participants, together with the number of participants contributing to the aggregate (p). ACR_frag: activity fragmentation; ACR_intens: activity intensity; ACR_rhythm: activity rhythmicity; ADHERENCE: adherence indicator features; APP_TRAFFIC: app traffic features; APP_USAGE (COM): app usage (communication apps); APP_USAGE (SM): app usage (social media apps); AUDIO: speech and acoustic features; EN: elastic net; HRV: heart rate variability; RF: random forest; SLEEP: sleep duration feature group; TIME: time-related features (eg, day of week, month of year, and elapsed time).



Discussion

Principal Findings

This study evaluated whether multimodal smartphone and wearable data provide predictive information beyond a simple person-specific baseline under strictly prospective,

time-aware evaluation and, where predictive value was observed, which sensor domains contributed most to these predictions. In line with the primary objective (RQ1), participant-specific models identified incremental predictive signal in a subset of individuals, with 8 of 11 (73%) participants (95% Wilson CI 43%-90%) showing significant improvement over baseline for each model type. Although the

overall number of participants showing significant improvement was identical, the sets of eligible participants were not completely overlapping, further emphasizing the participant-specific nature of predictive performance. Consistent with RQ2, no single sensor domain consistently dominated prediction across participants; instead, feature-group importance profiles were highly heterogeneous. Regarding RQ3, RF models exhibited statistically greater aggregated feature-group contributions than EN models, suggesting differences in how the 2 model classes used the available sensor information during prediction. However, this did not translate into a higher overall proportion of participants showing significant improvement over the person-specific baseline, indicating that nonlinear models appear to exploit behavioral information differently rather than uniformly improving predictive performance. Together, these findings indicate that both predictive performance and the behavioral domains underlying prediction are strongly participant-specific.

Interpretation of the Main Findings

This study found that multimodal smartphone and wearable data provided incremental predictive information beyond a simple person-specific baseline for a majority of participants. Approximately three-quarters of participants showed statistically significant improvement over baseline for both EN and RF models. Although these proportions suggest that participant-specific predictive signal was present in many individuals, the CIs remain wide because of the limited number of evaluable participants. Moreover, predictive performance varied considerably across participants, indicating that both the strength and nature of the detectable signal were highly individualized. The wide Wilson CIs further emphasize that the observed proportion of participants showing predictive improvement should be interpreted cautiously until replicated in larger cohorts.

Importantly, improvement over a simple person-specific baseline should not be equated with clinically useful prediction accuracy. Likewise, statistically significant improvements over baseline should be interpreted alongside their effect magnitude, as some participant-model combinations exhibited only small absolute reductions in prediction error despite meeting the predefined statistical criterion. Accordingly, the present findings support the presence of incremental predictive signal in some individuals rather than establishing passive sensing as a reliable stand-alone tool for symptom monitoring. Consistent with this idiographic perspective, by focusing on individual time series rather than group-level averages (eg, Zhang et al [56]), the present study further contributes to the growing literature on idiographic MS in depression.

The observation that predictive performance was highly individualized is broadly consistent with previous personalized mobile-sensing studies, although important methodological differences should be considered when interpreting the findings. Balliu et al [4] developed participant-specific EN models to predict interpolated Computer Adaptive Test Depression Inventory scores and evaluated performance

using mean absolute percentage error and R^2 . Price et al [52] used a stacked ensemble incorporating RFs to predict 12-month Patient Health Questionnaire-9 symptom variability and reported correlation coefficients together with normalized MAE. These substantial differences in prediction targets, outcome measures, feature representations, and evaluation procedures preclude direct numerical comparison. At the same time, they reinforce observations from recent reviews highlighting considerable methodological heterogeneity across mobile-sensing studies [12-14] and underline the importance of rigorous prospective evaluation and model-agnostic, domain-level explainability for improving comparability between studies. Against this background, this study extends previous work by combining strictly prospective evaluation with comparison against a person-specific baseline and model-agnostic feature-group explainability, thereby contributing a structured framework for evaluating individualized mobile-sensing models under prospective, time-aware conditions.

The explainability analyses further support the interpretation that predictive relationships are highly individualized. No single sensor domain consistently dominated prediction across participants, and feature-group importance profiles differed markedly between individuals. Although accelerometry, smartphone interaction, HRV, adherence indicators, and audio features contributed meaningfully for some participants, no common sensing profile emerged. Instead, different participants relied on different combinations of behavioral domains. This pattern is consistent with previous evidence that depression follows highly individualized behavioral trajectories and evolving within-person dynamics rather than stable group-level patterns [17,20]. The comparatively greater contribution of the TIME feature group in EN models likely reflects the ability of linear models to exploit approximately monotonic temporal trends over the observation period. In particular, elapsed study time likely captures gradual longitudinal changes in depressive symptom severity rather than psychologically meaningful behavioral variation, although day of week and month of year may additionally account for calendar-related temporal structure. Consequently, the TIME feature group should be interpreted as representing temporal adjustment within the observation period rather than a passive sensing modality. As permutation importance reflects model reliance rather than causation, the identified feature groups should be interpreted as predictive correlates of symptom variation rather than mechanistic drivers of depression.

Finally, the comparison between EN and RF suggests that the 2 modeling approaches may exploit available sensor information differently rather than consistently improving prediction. RF models showed larger aggregated permutation-importance values and, for some participants, relied on a broader range of domains. Nevertheless, RF did not identify predictive improvement in more participants than EN. As the aggregate comparison was based on only 7 participants with eligible models under both approaches, these findings should be interpreted cautiously. Given the exploratory comparison involving 7 participants, these findings do not establish a

general advantage of nonlinear modeling but suggest that model suitability may depend on the individual symptom-behavior relationship.

Limitations

Several limitations should be considered when interpreting the present findings.

First, participant-specific modeling of 130 predictors from relatively short individual time series creates a risk of overfitting, despite nested time-series validation and regularization. In addition, unmeasured time-varying factors such as treatment changes, life events, or contextual changes may have influenced both sensor features and symptom ratings. Second, as a proof-of-concept study on individualized, long-term depression monitoring, the current work is based on a small sample. Given the heterogeneity of depression, the symptoms and patterns identified here may not represent the broader population of patients with depressive disorders. Consequently, the generalizability of these findings is limited and will require validation in larger samples.

Third, although the target variable was assessed daily, the rolling evaluation and explainability analyses were based on 30-day test windows and 30-day training extensions. This design supports stable prospective assessment but may smooth over shorter-lived day-to-day fluctuations and transient temporal relationships that could also be clinically relevant. Consequently, short-term temporal relationships and how they may fluctuate across days remain unexamined. Future work should build on the present framework by investigating prediction errors across different levels of symptom severity, symptom transitions, and phases of longitudinal follow-up. Such analyses could help determine under which conditions passive sensing models provide reliable predictions and where their predictive performance deteriorates, thereby improving the clinical interpretability of individualized mobile-sensing models.

Fourth, although each participant contributed intensive repeated measurements, the number of independent individuals available for aggregate inference was only 11 participants, and only 7 participants contributed to the paired model-class comparison. Consequently, statistical power for participant-level inference remained limited despite the large number of repeated measurements. In addition, participant-level inference involved multiple comparisons across participants, model classes, and performance metrics.

To reduce the expected proportion of false-positive findings, bootstrap-based inference was combined with Benjamini-Hochberg FDR adjustment. While this represents an appropriate strategy for exploratory analyses, it cannot fully compensate for the limited number of independent participants. Consequently, both false-positive and false-negative findings remain possible, and the present results should be interpreted as hypothesis-generating until replicated in larger independent cohorts. In addition, although participant I exhibited markedly larger improvements relative to the baseline model than the remaining participants, a sensitivity analysis excluding this participant yielded qualitatively unchanged aggregate results, supporting the robustness of the principal conclusions.

Finally, although adherence-related indicators derived from the original sensor availability patterns were included as auxiliary predictors to partially capture informative missingness, they cannot fully account for MNAR mechanisms. Consequently, if missingness depended on unobserved factors beyond those reflected by the adherence indicators, some bias in the imputed predictor values may remain.

Conclusions

This study demonstrates that multimodal smartphone and wearable data can provide incremental predictive information beyond a simple person-specific baseline for daily depressive symptom severity in a subset of individuals when evaluated under strictly prospective, time-aware conditions. However, predictive performance remained highly participant-specific, and the observed improvements alone do not support immediate clinical application.

These findings suggest that future passive-sensing approaches for depression are unlikely to benefit from a universal sensing profile or a single modeling strategy. Instead, predictive modeling and interpretation may need to be individualized, reflecting the heterogeneous nature of depressive symptom dynamics. More broadly, these findings suggest that previously reported associations in mobile-sensing research, often derived from heterogeneous study designs and feature representations, may not consistently translate into prospective prediction. Future research may therefore benefit from evaluation frameworks that assess predictive signal under rigorous prospective, individual-level conditions while considering individualized combinations of behavioral domains rather than assuming a universal sensing profile.

Acknowledgments

The authors would like to express their gratitude and acknowledge the contributions of all members of the MONDY (secure and open platform for AI-based health care apps) Consortium:

Mohamed Alhaskir^{ab}, Rebeka Amin^{cd}, Angela Carell^e, Tobias Dunker^e, Kismet Ekinici^e, Florian Fischer^a, Enrico Lohmann^f, Elisabeth Schriewer^a, Yvonne Weber^a, Stefan Wolking^a, Jil Zippelius^{cd}.

Affiliations:

^aSection of Epileptology, Department of Neurology, RWTH Aachen University, Aachen, Germany

^bInstitute for Medical Informatics, RWTH Aachen University, Aachen, Germany

^cGerman Foundation for Depression and Suicide Prevention, Leipzig, Germany

^dResearch Centre of the German Foundation for Depression and Suicide Prevention, Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital, Goethe University Frankfurt, Frankfurt am Main, Germany

^e adesso SE, Dortmund, Germany

^f Institute for Applied Informatics, Leipzig University, Leipzig, Germany

The authors used generative AI (ChatGPT, OpenAI) for language editing, formatting assistance, and code refactoring or optimization. All content was critically reviewed and approved by the authors, who take full responsibility for the integrity of this work.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF, 13GW0576D, 13GW0576A, 13GW0576B, 13GW0576C, date awarded: November 1, 2021). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The data of individual participants underlying the results reported in this paper are not publicly available due to privacy considerations but are available upon reasonable request. The complete analysis pipeline and visualization code are publicly available [55].

Authors' Contributions

Conceptualization: HR, UH

Data curation: SS, JL, SL

Formal analysis: SS, JL

Funding acquisition: HR, UH, MONDY consortium

Investigation: SS

Methodology: SS, JL, HR, MONDY consortium

Project administration: SS, MP, HR, MONDY consortium

Resources: SL

Software: SS, JL, SL

Supervision: AD, DH, UH

Validation: SS, JL, SL

Visualization: SS, JL

Writing – original draft: SS, JL, MP, HR

Writing – review & editing: SS, JL, MP, HR, SL, AD, DH, UH, MONDY consortium

Conflicts of Interest

None declared.

Collaborators

Consortium: Mohamed Alhaskir, Rebeka Amin, Angela Carell, Tobias Dunker, Kismet Ekinci, Florian Fischer, Enrico Lohmann, Elisabeth Schriewer, Yvonne Weber, Stefan Wolking, and Jil Zippelius.

Multimedia Appendix 1

Overview of all features used, their feature-group assignments, and the preprocessing scaler applied to each feature.

[\[DOCX File \(Microsoft Word File\), 28 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Table of feature missingness per subject.

[\[DOCX File \(Microsoft Word File\), 22 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Table of hyperparameter grids used for elastic net and random forest models.

[\[DOCX File \(Microsoft Word File\), 21 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Per-participant plots of eligibility and explainability results.

[\[DOCX File \(Microsoft Word File\), 4491 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Sensitivity analysis of aggregated feature-group contributions after exclusion of participant I.

[\[DOCX File \(Microsoft Word File\), 218 KB-Multimedia Appendix 5\]](#)

References

1. Nemesure MD, Collins AC, Price GD, et al. Depressive symptoms as a heterogeneous and constantly evolving dynamical system: idiographic depressive symptom networks of rapid symptom changes among persons with major depressive disorder. *J Psychopathol Clin Sci*. Feb 2024;133(2):155-166. [doi: [10.1037/abn0000884](https://doi.org/10.1037/abn0000884)] [Medline: [38271054](https://pubmed.ncbi.nlm.nih.gov/38271054/)]
2. Bauer AM, Baldwin SA, Anguera JA, Areán PA, Atkins DC. Comparing approaches to mobile depression assessment for measurement-based care: prospective study. *J Med Internet Res*. Jun 19, 2018;20(6):e10001. [doi: [10.2196/10001](https://doi.org/10.2196/10001)] [Medline: [29921564](https://pubmed.ncbi.nlm.nih.gov/29921564/)]
3. Stone AA, Schneider S, Smyth JM. Evaluation of pressing issues in ecological momentary assessment. *Annu Rev Clin Psychol*. May 9, 2023;19(1):107-131. [doi: [10.1146/annurev-clinpsy-080921-083128](https://doi.org/10.1146/annurev-clinpsy-080921-083128)] [Medline: [36475718](https://pubmed.ncbi.nlm.nih.gov/36475718/)]
4. Balliu B, Douglas C, Seok D, et al. Personalized mood prediction from patterns of behavior collected with smartphones. *npj Digital Med*. 2024;7(1):49. [doi: [10.1038/s41746-024-01035-6](https://doi.org/10.1038/s41746-024-01035-6)]
5. Choi J, Lee S, Kim S, Kim D, Kim H. Depressed mood prediction of elderly people with a wearable band. *Sensors*. Jan 2022;22(11):4174. [doi: [10.3390/s22114174](https://doi.org/10.3390/s22114174)]
6. Corponi F, Li BM, Anmella G, et al. Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. *Transl Psychiatry*. Mar 26, 2024;14(1):161. [doi: [10.1038/s41398-024-02876-1](https://doi.org/10.1038/s41398-024-02876-1)] [Medline: [38531865](https://pubmed.ncbi.nlm.nih.gov/38531865/)]
7. Ricka N, Pellegrin G, Fompeyrine DA, Lahutte B, Geoffroy PA. Predictive biosignature of major depressive disorder derived from physiological measurements of outpatients using machine learning. *Sci Rep*. Apr 25, 2023;13(1):6332. [doi: [10.1038/s41598-023-33359-w](https://doi.org/10.1038/s41598-023-33359-w)] [Medline: [37185788](https://pubmed.ncbi.nlm.nih.gov/37185788/)]
8. Zhuparris A, Maleki G, van Londen L, et al. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. *Sci Rep*. Nov 1, 2023;13(1):18844. [doi: [10.1038/s41598-023-46075-2](https://doi.org/10.1038/s41598-023-46075-2)] [Medline: [37914808](https://pubmed.ncbi.nlm.nih.gov/37914808/)]
9. Shah RV, Grennan G, Zafar-Khan M, et al. Personalized machine learning of depressed mood using wearables. *Transl Psychiatry*. Jun 9, 2021;11(1):1-18. [doi: [10.1038/s41398-021-01445-0](https://doi.org/10.1038/s41398-021-01445-0)] [Medline: [34103481](https://pubmed.ncbi.nlm.nih.gov/34103481/)]
10. Makhmutova M, Kainkaryam R, Ferreira M, Min J, Jaggi M, Clay I. Predicting changes in depression severity using the PSYCHE-D (Prediction of Severity Change-Depression) model involving person-generated health data: longitudinal case-control observational study. *JMIR mHealth uHealth*. Mar 25, 2022;10(3):e34148. [doi: [10.2196/34148](https://doi.org/10.2196/34148)] [Medline: [35333186](https://pubmed.ncbi.nlm.nih.gov/35333186/)]
11. Huang Y, Ma Y, Xiao J, Liu W, Zhang G. Identification of depression state based on multi-scale acoustic features in interrogation environment. *IET Signal Process*. Apr 2023;17(4):e12207. [doi: [10.1049/si2.12207](https://doi.org/10.1049/si2.12207)]
12. Shen S, Qi W, Zeng J, et al. Passive sensing for mental health monitoring using machine learning with wearables and smartphones: scoping review. *J Med Internet Res*. Aug 14, 2025;27:e77066. [doi: [10.2196/77066](https://doi.org/10.2196/77066)] [Medline: [40811794](https://pubmed.ncbi.nlm.nih.gov/40811794/)]
13. Amin R, Schreynemackers S, Oppenheimer H, Petrovic M, Hegerl U, Reich H. Use of mobile sensing data for longitudinal monitoring and prediction of depression severity: systematic review. *J Med Internet Res*. Aug 21, 2025;27:e57418. [doi: [10.2196/57418](https://doi.org/10.2196/57418)] [Medline: [40839863](https://pubmed.ncbi.nlm.nih.gov/40839863/)]
14. Leimhofer J, Petrovic M, Dominik A, Heider D, Hegerl U. Cross-platform availability of smartphone sensors for depression indication systems: mixed-methods umbrella review. *Interact J Med Res*. Aug 7, 2025;14:e69686. [doi: [10.2196/69686](https://doi.org/10.2196/69686)] [Medline: [40773748](https://pubmed.ncbi.nlm.nih.gov/40773748/)]
15. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J*. Sep 2015;38(3):218-226. [doi: [10.1037/prj0000130](https://doi.org/10.1037/prj0000130)] [Medline: [25844912](https://pubmed.ncbi.nlm.nih.gov/25844912/)]
16. Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. *npj Digital Med*. Feb 2019;2(1):3. [doi: [10.1038/s41746-019-0078-0](https://doi.org/10.1038/s41746-019-0078-0)]
17. Kathan A, Harrer M, Küster L, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Front Digit Health*. 2022;4:964582. [doi: [10.3389/fdgth.2022.964582](https://doi.org/10.3389/fdgth.2022.964582)] [Medline: [36465087](https://pubmed.ncbi.nlm.nih.gov/36465087/)]
18. Lin Y, Liyanage BN, Sun Y, et al. A deep learning-based model for detecting depression in senior population. *Front Psychiatry*. 2022;13:1016676. [doi: [10.3389/fpsy.2022.1016676](https://doi.org/10.3389/fpsy.2022.1016676)] [Medline: [36419976](https://pubmed.ncbi.nlm.nih.gov/36419976/)]
19. Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. *Knowl Based Syst*. Jun 2022;245:108632. [doi: [10.1016/j.knosys.2022.108632](https://doi.org/10.1016/j.knosys.2022.108632)]
20. Siepe BS, Sander C, Schultze M, et al. Time-varying network models for the temporal dynamics of depressive symptomatology in patients with depressive disorders: secondary analysis of longitudinal observational data. *JMIR Ment Health*. Apr 18, 2024;11(1):e50136. [doi: [10.2196/50136](https://doi.org/10.2196/50136)] [Medline: [38635978](https://pubmed.ncbi.nlm.nih.gov/38635978/)]
21. Liu D, Liu B, Lin T, et al. Measuring depression severity based on facial expression and body movement using deep convolutional neural network. *Front Psychiatry*. 2022;13:1017064. [doi: [10.3389/fpsy.2022.1017064](https://doi.org/10.3389/fpsy.2022.1017064)] [Medline: [36620657](https://pubmed.ncbi.nlm.nih.gov/36620657/)]

22. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *Artif Intell.* Sep 2021;298:103502. [doi: [10.1016/j.artint.2021.103502](https://doi.org/10.1016/j.artint.2021.103502)]
23. Au Q, Herbringer J, Stachl C, Bischl B, Casalicchio G. Grouped feature importance and combined features effect plot. *Data Min Knowl Discovery.* Jul 2022;36(4):1401-1450. [doi: [10.1007/s10618-022-00840-5](https://doi.org/10.1007/s10618-022-00840-5)]
24. Reich H, Schreynemackers S, Amin R, et al. Links between self-monitoring data collected through smartphones and smartwatches and the individual disease trajectories of adult patients with depressive disorders: study protocol of a one-year observational trial. *Contemp Clin Trials Commun.* Jun 2025;45:101492. [doi: [10.1016/j.conctc.2025.101492](https://doi.org/10.1016/j.conctc.2025.101492)] [Medline: [40491661](https://pubmed.ncbi.nlm.nih.gov/40491661/)]
25. WMA Declaration of Helsinki – ethical principles for medical research involving human participants. World Medical Association. URL: <https://www.wma.net/policies-post/wma-declaration-of-helsinki/> [Accessed 2026-08-08]
26. Arrieta J, Aguerreberre M, Raviola G, et al. Validity and utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for screening and diagnosis of depression in rural Chiapas, Mexico: a cross-sectional study. *J Clin Psychol.* Sep 2017;73(9):1076-1090. [doi: [10.1002/jclp.22390](https://doi.org/10.1002/jclp.22390)] [Medline: [28195649](https://pubmed.ncbi.nlm.nih.gov/28195649/)]
27. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann.* Sep 2002;32(9):509-515. [doi: [10.3928/0048-5713-20020901-06](https://doi.org/10.3928/0048-5713-20020901-06)]
28. Sun S, Folarin AA, Zhang Y, et al. Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: retrospective analysis. *J Med Internet Res.* Aug 14, 2023;25:e45233. [doi: [10.2196/45233](https://doi.org/10.2196/45233)] [Medline: [37578823](https://pubmed.ncbi.nlm.nih.gov/37578823/)]
29. Wang X, Pathiravasan CH, Zhang Y, et al. Association of depressive symptom trajectory with physical activity collected by mHealth devices in the electronic Framingham Heart Study: cohort study. *JMIR Ment Health.* Jul 14, 2023;10:e44529. [doi: [10.2196/44529](https://doi.org/10.2196/44529)] [Medline: [37450333](https://pubmed.ncbi.nlm.nih.gov/37450333/)]
30. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography.* Jan 2013;36(1):27-46. [doi: [10.1111/j.1600-0587.2012.07348.x](https://doi.org/10.1111/j.1600-0587.2012.07348.x)]
31. O'brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.* Sep 11, 2007;41(5):673-690. [doi: [10.1007/s11135-006-9018-6](https://doi.org/10.1007/s11135-006-9018-6)]
32. Adamowicz L, Christakis Y, Czech MD, Adamusiak T. SciKit digital health: Python package for streamlined wearable inertial sensor data processing. *JMIR mHealth uHealth.* Apr 21, 2022;10(4):e36762. [doi: [10.2196/36762](https://doi.org/10.2196/36762)] [Medline: [35353039](https://pubmed.ncbi.nlm.nih.gov/35353039/)]
33. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. *Behav Res.* Aug 2021;53(4):1689-1696. [doi: [10.3758/s13428-020-01516-y](https://doi.org/10.3758/s13428-020-01516-y)]
34. Lorenz N, Sander C, Ivanova G, Hegerl U. Temporal associations of daily changes in sleep and depression core symptoms in patients suffering from major depressive disorder: idiographic time-series analysis. *JMIR Ment Health.* Apr 23, 2020;7(4):e17071. [doi: [10.2196/17071](https://doi.org/10.2196/17071)] [Medline: [32324147](https://pubmed.ncbi.nlm.nih.gov/32324147/)]
35. Eyben F, Wöllmer M, Schuller B. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. Presented at: MM '10: Proceedings of the 18th ACM International Conference on Multimedia; Oct 25-29, 2010:1459-1462; Firenze, Italy. [doi: [10.1145/1873951.1874246](https://doi.org/10.1145/1873951.1874246)]
36. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* Jul 2015;71:10-49. [doi: [10.1016/j.specom.2015.03.004](https://doi.org/10.1016/j.specom.2015.03.004)]
37. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun.* Dec 2015;75:27-49. [doi: [10.1016/j.specom.2015.09.003](https://doi.org/10.1016/j.specom.2015.09.003)]
38. Dhelim S, Chen L, Ning H, Nugent C. Artificial intelligence for suicide assessment using audiovisual cues: a review. *Artif Intell Rev.* Jun 2023;56(6):5591-5618. [doi: [10.1007/s10462-022-10290-6](https://doi.org/10.1007/s10462-022-10290-6)]
39. Moore E, Clements MA, Peifer JW, Weisser L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Trans Biomed Eng.* Jan 2008;55(1):96-107. [doi: [10.1109/TBME.2007.900562](https://doi.org/10.1109/TBME.2007.900562)] [Medline: [18232351](https://pubmed.ncbi.nlm.nih.gov/18232351/)]
40. Quatieri TF, Malyska N. Vocal-source biomarkers for depression: a link to psychomotor activity. Presented at: Interspeech 2012: 13th Annual Conference of the International Speech Communication Association; Sep 9-13, 2012:1059-1062; Portland, OR. [doi: [10.21437/Interspeech.2012-311](https://doi.org/10.21437/Interspeech.2012-311)]
41. Stasak B, Epps J, Schatten HT, Miller IW, Provost EM, Armey MF. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Commun.* Sep 2021;132:10-20. [doi: [10.1016/j.specom.2021.05.004](https://doi.org/10.1016/j.specom.2021.05.004)] [Medline: [42079368](https://pubmed.ncbi.nlm.nih.gov/42079368/)]
42. Choi A, Ooi A, Lottridge D. Digital phenotyping for stress, anxiety, and mild depression: systematic literature review. *JMIR mHealth uHealth.* May 23, 2024;12:e40689. [doi: [10.2196/40689](https://doi.org/10.2196/40689)] [Medline: [38780995](https://pubmed.ncbi.nlm.nih.gov/38780995/)]
43. Tian Y, Zhou K, Pelleg D. What and how long: prediction of mobile app engagement. *ACM Trans Inf Syst.* Jan 31, 2022;40(1):1-38. [doi: [10.1145/3464301](https://doi.org/10.1145/3464301)]

44. Yue C, Ware S, Morillo R, et al. Automatic depression prediction using internet traffic characteristics on smartphones. *Smart Health*. Nov 2020;18:100137. [doi: [10.1016/j.smhl.2020.100137](https://doi.org/10.1016/j.smhl.2020.100137)]
45. Brunner C, Hofer F. SleepECG: a Python package for sleep staging based on heart rate. *J Open Source Software*. Jun 9, 2023;8(86):5411. [doi: [10.21105/joss.05411](https://doi.org/10.21105/joss.05411)]
46. Tukey JW. *Exploratory Data Analysis*. Addison-Wesley Publishing Company; 1977. ISBN: 978-0-201-07616-5
47. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. Mar 2011;20(1):40-49. [doi: [10.1002/mpr.329](https://doi.org/10.1002/mpr.329)] [Medline: [21499542](https://pubmed.ncbi.nlm.nih.gov/21499542/)]
48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page [Accessed 2026-08-08]
49. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol*. Jun 2019;110:63-73. [doi: [10.1016/j.jclinepi.2019.02.016](https://doi.org/10.1016/j.jclinepi.2019.02.016)] [Medline: [30878639](https://pubmed.ncbi.nlm.nih.gov/30878639/)]
50. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. Apr 1, 2005;67(2):301-320. [doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)]
51. Breiman L. Random forests. *Mach Learn*. Oct 2001;45(1):5-32. [doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)]
52. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Transl Psychiatry*. Dec 9, 2023;13(1):1-10. [doi: [10.1038/s41398-023-02669-y](https://doi.org/10.1038/s41398-023-02669-y)] [Medline: [38071317](https://pubmed.ncbi.nlm.nih.gov/38071317/)]
53. Hegerl U, Schreynemackers S, Petrovic M, et al. Methods protocol: can AI-enabled analyses of smartphones and wearable data collected over one year provide patients suffering from depression with an objective marker of disease severity. *OSF*. Preprint posted online on Jul 12, 2025. [doi: [10.17605/OSF.IO/XDU3P](https://doi.org/10.17605/OSF.IO/XDU3P)]
54. Hegerl U, Schreynemackers S, Leimhofer J, et al. Can 1-year smartphone and wearable data allow individual-level modeling of depressive symptom severity? *Research Square*. Preprint posted online on Apr 1, 2026. [doi: [10.21203/rs.3.rs-9083929/v1](https://doi.org/10.21203/rs.3.rs-9083929/v1)]
55. Leimhofer J. GermanDepressionFoundation/mondy-fi-publication: initial release – individual-level modeling of depressive symptom severity using smartphone and wearable data. *Zenodo*; Jul 30, 2026. [doi: [10.5281/zenodo.21704247](https://doi.org/10.5281/zenodo.21704247)]
56. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. *JMIR mHealth uHealth*. Apr 12, 2021;9(4):e24604. [doi: [10.2196/24604](https://doi.org/10.2196/24604)] [Medline: [33843591](https://pubmed.ncbi.nlm.nih.gov/33843591/)]

Abbreviations

APP_USAGE: app usage features
EN: elastic net
FDR: false discovery rate
GDPR: General Data Protection Regulation
HRV: heart rate variability
MAE: mean absolute error
MNAR: missing not at random
MONDY : secure and open platform for AI-based health care apps
MS: mobile sensing
PHQ-2: Patient Health Questionnaire 2
RF: random forest
RQ1: research question 1
RQ2: research question 2
RQ3: research question 3
R²: coefficient of determination
SHAP : Shapley Additive Explanations
TIME: time-related features (eg, day of week, month of year, and elapsed time)
ΔMAE: change in mean absolute error relative to baseline
ΔR²: change in coefficient of determination relative to baseline

Edited by Amaryllis Mavragani, Mamdooh Alzyood; peer-reviewed by Ariel Teles, Peter Tonn; submitted 11.May.2026; final revised version received 30.Jul.2026; accepted 30.Jul.2026; published 25.Aug.2026

Please cite as:

Schreynemackers S, Leimhofer J, Petrovic M, Reich H, Ludwig S, Dominik A, Heider D, Hegerl U, the MONDY Consortium
*Individual-Level Modeling of Depressive Symptom Severity Using Smartphone and Wearable Data: Time-Aware 1-Year
Study With Feature-Group Contributions*

JMIR Form Res 2026;10:e99204

URL: <https://formative.jmir.org/2026/1/e99204>

doi: [10.2196/99204](https://doi.org/10.2196/99204)

© Simon Schreynemackers, Johannes Leimhofer, Milica Petrovic, Hanna Reich, Sascha Ludwig, Andreas Dominik, Dominik Heider, Ulrich Hegerl, the MONDY Consortium. Originally published in JMIR Formative Research (<https://formative.jmir.org>), 25.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://formative.jmir.org>, as well as this copyright and license information must be included.