Original Paper

Passive Monitoring for Detecting Hypomanic Symptoms in Bipolar Disorder Using Explainable Machine Learning: Longitudinal Observational Study

Ramzi Halabi[1], BEng, MSc, PhD; Martin Alda[2,3], MD, FRCP; Mirkamal Tolend[4], PhD; Claire O'Donovan[2], MD, FRCP; Arend Hintze[5], PhD; Benoit H Mulsant[4,6,7], MS, MD, FRCP, FRCS; Abigail Ortiz[1,4,7], MS, MD, FRCP

[1]Department of Psychiatry, The University of Texas Southwestern Medical Center, Dallas, TX, United States
[2]Department of Psychiatry, Dalhousie University, Halifax, NS, Canada
[3]Scientific Council, National Institute of Mental Health, Klecany, Czech Republic
[4]Campbell Family Research Institute, Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada
[5]Department of MicroData Analytics, Dalarna University, Falun, Sweden
[6]Department of Psychiatry, Royal Ottawa Mental Health Centre, Ottawa, ON, Canada
[7]Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

Corresponding Author:

Abigail Ortiz, MS, MD, FRCP
Department of Psychiatry
The University of Texas Southwestern Medical Center
1440 Empire Central Dr LD4.244
Dallas, TX 75247
United States
Phone: 1 214-968-0742
Email: Tania.OrtizDominguez@utsouthwestern.edu

Abstract

Background: When trying to intervene early to prevent manic episodes, clinicians are hindered by patients underreporting their symptoms as they are becoming hypomanic or manic (ie, emerging symptoms). Continuous passive monitoring using wearable technology offers a potential solution to address this problem by capturing behavioral signals that do not depend on self-awareness. However, its clinical usefulness requires validation before implementation.

Objective: This study aimed to evaluate whether passively collected wearable data can detect hypomanic symptoms using an explainable machine learning framework.

Methods: We applied explainable machine learning to densely sampled multimodal data from a 2-year observational study of digital monitoring in 178 patients with a diagnosis of bipolar disorder (mean follow-up 540 days; range 120-706 days). Data on activity and sleep were collected continuously using wearable technology. Self-reported data were collected daily on their mood, energy, and anxiety, and weekly using the Altman Self-Rating Mania Scale (ASRM). To identify robust digital markers of hypomanic symptoms, we analyzed 252 time- and frequency-domain features from 43 participants who experienced at least 1 ASRM-defined symptomatic interval.

Results: Passively collected data achieved high accuracy, comparable with daily self-reported data in detecting symptomatic intervals (area under the receiver operating characteristic curve 0.91 vs 0.93, respectively). Combined daily passive wearable data features achieved higher sensitivity (92.0% vs 85.4%), balanced accuracy (0.90 vs 0.89), and stronger rule-out usefulness (negative predictive value 97.8% vs 96.1%) than self-report data, with daily sleep features alone driving most of this signal (area under the receiver operating characteristic curve 0.91, sensitivity 91.8%). The high classification performance using wearable data was explained by reductions in spectral entropy and flux across domains, reflecting a decreased complexity and variability in mood, activity, and sleep patterns, consistent with a shift toward pathologically stable states.

Conclusions: Passive monitoring using wearable technology can detect hypomanic symptoms with high accuracy, supporting its potential as a clinically meaningful tool for continuous, objective monitoring in patients who underreport hypomanic symptoms. These findings are associative and concurrent in nature; they do not establish predictive validity. Prospective studies are needed to evaluate predictive validity for early detection of ASRM-defined symptomatic intervals.

JMIR Mhealth Uhealth 2026;14:e105647; doi: 10.2196/105647

Keywords: bipolar disorder; mania; machine learning; wearables; mood; activity; sleep

Introduction

Clinical assessment typically relies on retrospective self-reports during clinical visits, which are prone to recall bias. A key limitation of symptom self-report in bipolar disorder (BD) is that insight into illness is frequently impaired, particularly during hypomanic or manic episodes: fewer than a quarter of patients recognized having ever experienced a hypomanic or manic episode, compared with over 80% who recognized past depressive episodes [1]. Because diagnosis and ongoing monitoring rely heavily on patients reporting their own symptoms [2,3], this lack of insight contributes directly to the underdetection and undertreatment of hypomanic or manic episodes [4].

Passive monitoring using wearable technology offers a way around this problem: physiological (eg, sleep architecture) and behavioral (eg, activity levels) signals are captured continuously in the patient's natural environment, moving beyond retrospective clinical assessments toward continuous, objective monitoring [5,6]. This approach, often termed digital phenotyping, can capture subtle yet significant changes in patterns that precede the onset of a mood episode and has shown promise in mental health [7]. In BD specifically, disruptions in activity and circadian rhythms have shown their potential for episode prediction [8-13]. Furthermore, a major insight in digital phenotyping is that day-to-day variability in behavioral and physiological metrics may be more informative than static measures for detecting episode onset. We have demonstrated that variability in activity levels precedes mood changes by approximately 7 days in depressive episodes [14]. Similarly, we identified variability in sleep stages and activity levels as the earliest indicators of hypomanic episode onset, achieving high sensitivity (0.94) and specificity (0.80) [8]. While these findings underscore variability as a fundamental concept for the early detection of episodes, spectral or frequency-domain features that reveal rhythmic dynamics remain underexplored in the context of BD.

Additionally, the optimal integration of objective, passively collected wearable data with subjective reports of a participant's state (eg, electronic visual analog scales [e-VASs] and self-rating scales) remains underexplored [15]. In this context, while machine learning (ML) is essential for detecting subtle patterns in high-dimensional data, its clinical translation is often hindered by limited interpretability and the need for personalized approaches that account for heterogeneity [16-18]. Hence, robust, multimodal digital markers that combine passive and active data, validated through repeated within- and across-participant testing, are needed to develop personalized, clinically actionable early-warning systems.

Beyond methodological considerations, several clinical and technical challenges further complicate the development of digital markers for hypomanic or manic episodes. Clinically, hypomanic or manic episodes occur less frequently and are shorter in duration than depressive episodes [19]. This imbalance results in fewer training examples for predictive models and reduced statistical power to detect changes. From a technical perspective, many digital phenotyping studies in BD are characterized by short follow-up durations, limiting the capture of hypomanic or manic symptoms and hindering the development of robust, generalizable predictive models for detecting these types of episodes.

To address all these gaps, this manuscript describes a longitudinal observational study designed to identify and validate robust digital markers of hypomanic or manic symptoms. Our approach integrates 2 key data sources: objective (ie, continuous, passively collected activity and sleep data from a consumer-grade wearable), and subjective (ie, daily self-reports of mood, energy, and anxiety collected via e-VAS, and weekly self-reports of symptom severity collected via Altman Self-Rating Mania Scale [ASRM], a validated self-report measure [20] used to operationalize mood state classification). By applying explainable ML techniques to a diverse set of time- and frequency-domain features extracted from a longitudinal multimodal dataset, we aimed to (1) identify the most robust and accurate digital markers of hypomanic symptoms and (2) compare the discriminative accuracy of passively collected wearable data against self-reported measures.

Methods

Participants and Data Collection

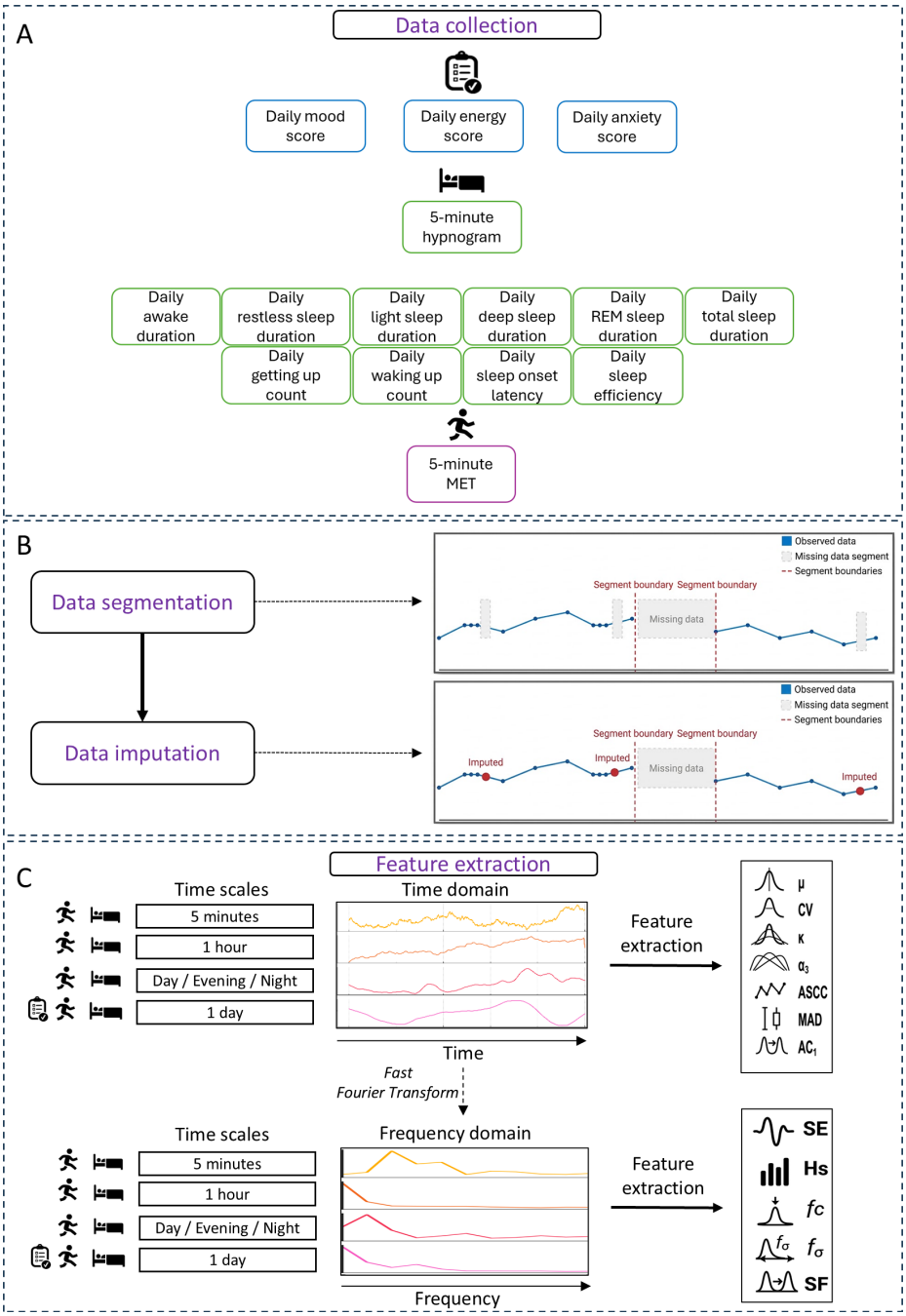
The study was approved by the Research Ethics Board (REB) at the Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada, and at the Queen Elizabeth II Health Sciences Centre (QEII Health Care Center), Halifax, Nova Scotia, Canada, in accordance with the Declaration of Helsinki, REB number 059-2019. All participants signed an informed consent form approved by the local REB before any research procedure was initiated.

We have previously described the overall methods of this ongoing longitudinal study [21]. In summary, between December 12, 2020, and January 1, 2025, we recruited and followed 178 adults diagnosed with BD type I or II in the outpatient clinics of 2 academic psychiatric hospitals in Canada: the CAMH and the QEII.

Primary diagnoses of BD I or II were confirmed with the SCID-5 (Structured Clinical Interview for the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*) [22] based on the diagnostic criteria of the DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*) [23]. Baseline assessments included sociodemographic and clinical data. Clinical polarity at study entry was defined using Young Mania Rating Scale (YMRS) [24] and Montgomery-Asberg Depression Rating Scale [25] thresholds.

We instructed each participant to wear an Oura Ring (Oura Health Oy; Generation 2) continuously throughout the study course. The ring collected multiple activity and sleep data variables, of which we selected a comprehensive subset [26] relevant to the goals of this study (Figure 1A). Additionally, participants received a daily email with a link to an e-VAS to self-rate their mood, energy, and anxiety levels over the previous 24 hours. Reminders were sent to participants when e-VASs were not completed for 3 consecutive days. Participants also received a weekly email with a link to an electronic version of the ASRM [20] to rate their symptoms for the previous week. Weekly symptom self-reports were stored in a REDCap (Research Electronic Data Capture) [27] secure database.

Figure 1. Data collection, processing, and feature extraction pipeline. This figure illustrates the workflow applied to data from a longitudinal observational study of 178 outpatients with bipolar disorder I or II recruited at the Centre for Addiction and Mental Health, Toronto, Ontario, and the Queen Elizabeth II Health Sciences Centre, Halifax, Nova Scotia, Canada (December 2020 to January 2025). Panel (A) shows the selected objective (activity and sleep) and subjective (electronic visual analog scale) variables collected via Oura Ring Generation 2 and daily/weekly self-reports. Panel (B) depicts the segmentation and feature extraction process. Panel (C) displays the 12 time- and frequency-domain features computed per variable per segment. See Table S1 in Multimedia Appendix 1 for the full variable list and Table S2 in Multimedia Appendix 1 for feature definitions. AC_1: autocorrelation at lag 1; ASCC: absolute sum of consecutive changes; CV: coefficient of variation; MAD: mean absolute differences; MET: metabolic equivalent of task; REM: rapid eye movement (sleep stage); SE: spectral energy; SF: spectral flux; μ: mean; ϰ: kurtosis; α_3: skewness; H_s: spectral entropy; f_c: spectral centroid; f_σ: spectral spread.



We evaluated participant adherence to device-wear protocols by quantifying nonwear time and modality-specific missing data ratios. We obtained nonwear time directly from the Oura Ring activity dataset and calculated missing data ratios separately for each modality (activity, sleep, and e-VAS) [28,29].

Data Generation and Processing

Data Generation

Figure 1 presents an overview of the clinical course and data generation time scales. Using the longitudinal ASRM scores of each participant, we applied a 1-week rolling window with 50% overlap to identify ASRM-defined symptomatic intervals, defined as at least 1 weekly rating with an ASRM score of ≥10. The ASRM is a self-rating scale comprising 5 items designed to evaluate the existence and severity of 5 hypomanic or manic symptoms: elation, overconfidence, decreased need for sleep, talkativeness, and increased energy levels, with good validity and reliability [30] when compared with the clinician-administered YMRS [20]. Each item is rated on a scale from 0 to 4, with 0 representing the absence of symptoms and 4 indicating sustained and pronounced symptoms.

Symptomatic Interval Labeling

Because the ASRM has shown high sensitivity and low specificity [31], our operationalized definition of a score of ≥10 typically indicates the presence of 3 or more symptoms nearly every day. By requiring a score of ≥10 for a week, we also captured the DSM-5 four days duration requirement for hypomania and the 1-week duration requirement for mania. The standard ASRM threshold of ≥6 has been recommended by the scale developers for screening purposes [31]. However, we used a more conservative threshold of ≥10 for the following reasons: (1) the ASRM has been shown to have high sensitivity but low specificity [31], and a higher threshold reduces false positives, which is critical when weekly ratings are used to define interval-level labels for ML classification; (2) a score of ≥10 typically requires that a participant endorse at least 3 of 5 symptoms at a moderate-to-severe level (ie, ≥2 on at least 3 items), providing higher clinical specificity for a symptomatic interval; and (3) by requiring this threshold to be sustained over a rolling 1-week window, we further reduce noise from transient elevations ("symptomatic blips"). Moreover, we defined a symptomatic interval as 1 or more consecutive weeks in which the ASRM overall score was ≥10 based on previous studies defining "symptomatic remission" as an ASRM score of <10 [32,33].

To assess the robustness of our findings to this threshold choice, we conducted a sensitivity analysis repeating the full analytical pipeline (feature extraction, classification, and feature importance) using the standard ASRM threshold of ≥6. Some participants presented with an ASRM score of ≥10 on 1 occasion. Because our dataset uses weekly symptom ratings, we could not determine whether these elevated ASRM scores lasted 4 to 6 days (ie, the minimum duration for a hypomanic episode according to DSM-5). Thus, these elevations in ASRM scores that did not span at least 2 weekly ratings were not considered symptomatic intervals in the primary analysis.

Because elevations in symptoms were not independently confirmed by a clinician or a clinician-rated scale, we refer to them as "symptomatic intervals," rather than "episodes." Weeks for which the ASRM was not completed were not assigned a label and were excluded from the definition of symptomatic intervals and the euthymic comparator. Depressive and mixed symptomatic intervals were identified using a Patient Health Questionnaire (9-item; PHQ-9) threshold of ≥10; such intervals were excluded from the binary classification analysis (see euthymic comparator definition).

Euthymic Comparator Labeling

Euthymic segments were defined as data segments during which the participants' weekly ASRM score was <10 and their weekly PHQ-9 score was <10. The use of both ASRM and PHQ-9 thresholds ensured that euthymic segments fell below the standard screening threshold for hypomanic or manic symptoms (ASRM score of <10) and depressive symptoms (PHQ-9 <10), although we cannot ensure that subsyndromal presentations were present. Segments that did not meet criteria for either ASRM-defined symptomatic intervals or euthymia (eg, intervals with elevated PHQ-9 but not ASRM, or transitional periods) were excluded from the binary classification analysis.

Data Variables

Among the 39 data variables generated by the wearable, we selected a smaller set of 12 objective variables (1 of 16 related to activity and 11 of 23 related to sleep). We derived the activity variables included in the analysis from (1) the 5-minute-sampled metabolic equivalent of task (MET), produced by the wearable; and (2) its aggregated variants over 1-hour periods, as well as daytime, evening, and nighttime. Similarly, we derived the sleep variables from (1) the 5-minute hypnograms produced by the wearable and (2) a set of 10 daily data variables provided by the wearable's proprietary algorithms (eg, wake-up count). Moreover, 3 daily e-VAS variables were collected to describe self-reported mood, energy, and anxiety dynamics (Figure 1A and Table S1 in Multimedia Appendix 1). Overall, based on objective and subjective data across the various time periods, we created a set of 21 variables.

Sleep stages in the 5-minute hypnogram were encoded categorically into 4 stages: awake, light sleep (N1/N2), deep sleep (N3/slow-wave sleep), and rapid eye movement (REM) sleep using the Oura Ring's proprietary stage labels. This encoding preserves the physiological ordering of sleep stages and allows the same time series feature extraction pipeline to be applied to both activity and sleep-stage data. Irregular sampling within 5-minute windows (eg, brief gaps from device removal) was handled by forward-filling missing epochs from the immediately preceding nonmissing epoch, capped at a maximum of 2 consecutive missing epochs (10 minutes); isolated stage flickers that are typical of automated staging outputs were filtered via moving-window

mode smoothing (categorical "majority vote"); days with more than 60 consecutive minutes of missing epochs were classified as nonwear and excluded from analysis.

Data Inclusion Criteria

Data inclusion criteria required at least 80% completeness per data segment; segments failing this threshold were excluded from analysis. For the 5-minute MET activity data, nonwear periods were identified using the Oura Ring's proprietary wear detection algorithm and confirmed by sustained zero-MET readings of ≥60 minutes. The 5-minute hypnograms were filtered to remove artefactual sleep-stage transitions (ie, rapid cycling between stages within <5 minutes inconsistent with physiological sleep architecture). Daily sleep variables provided by the wearable's proprietary algorithms (eg, wake-up count and sleep efficiency) were used as is without additional filtering, as these are already validated outputs of the device's sleep pipeline.

Data Completeness

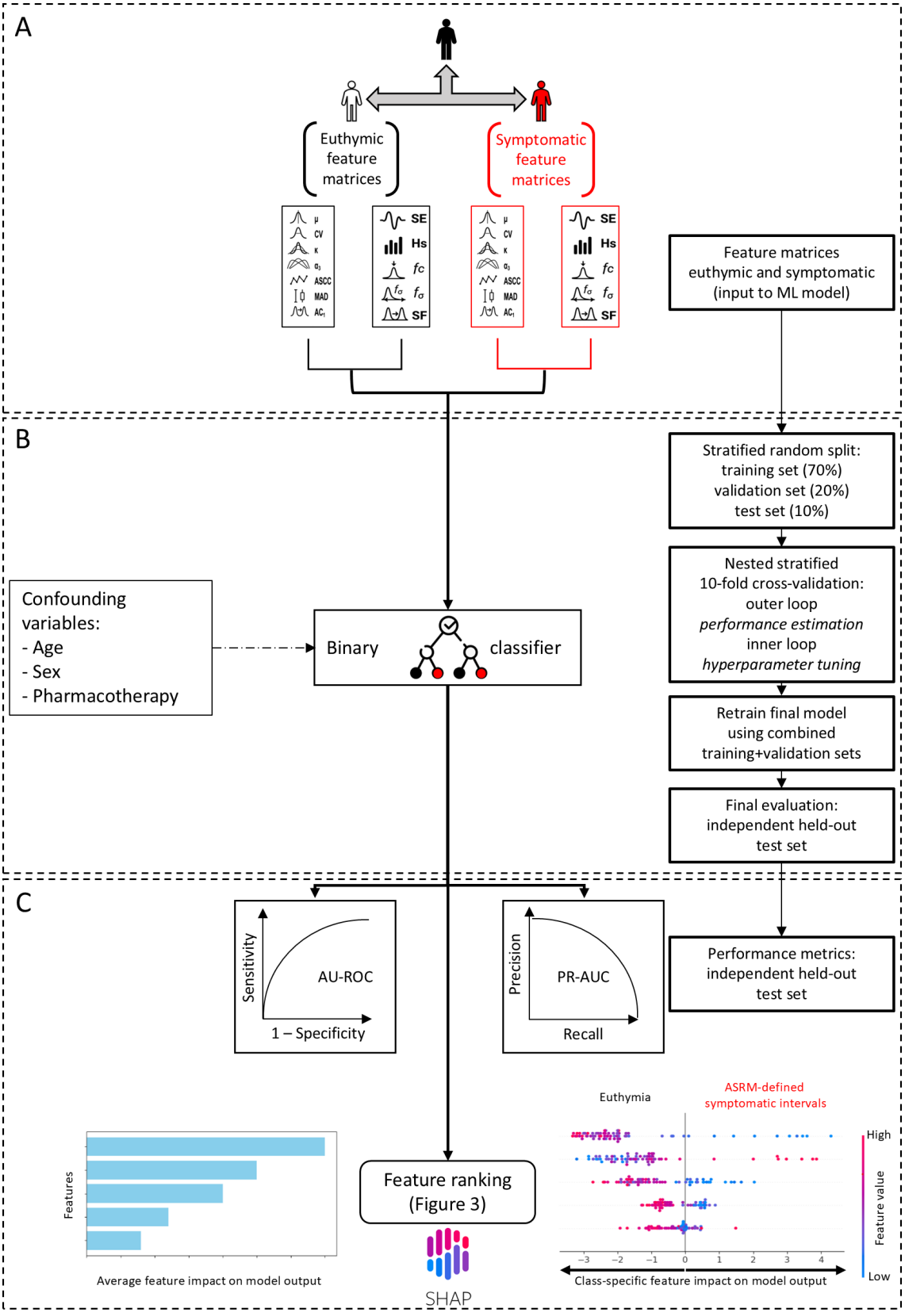
Given our previous findings showing that data in our sample were missing not at random [29], we imputed the daily sleep variables and the 5-minute hypnograms using completeness-based adaptive segmentation (Figure 1B). The k-nearest neighbors (KNNs) method [34] was performed with k=5 neighbors using Euclidean distance, with the imputation model fit within each participant's data and within each cross-validation training fold to prevent information leakage. The adaptive segmentation approach identified contiguous data segments of >7 days with ≥80% completeness; segments shorter than 8 days or with >20% missingness were excluded. To systematically characterize missingness, we report the proportion of missing data for each modality (activity, sleep, e-VAS, and ASRM) both overall and stratified by clinical state (euthymic, ASRM-defined symptomatic interval). We assessed whether data were missing completely at random (MCAR) using Little MCAR test, reporting the test statistic and associated *P* value [35]. Based on our previous work [29], we hypothesized that Little MCAR test would be rejected; we therefore report sensitivity analyses comparing classifier performance, selected features, and importance rankings with and without imputation.

Data Features

For each participant and each of the 21 variables' data segments, we calculated 12 features in the time and frequency domains. Seven time-domain features (ie, using the raw data) were calculated to characterize the statistical properties of the data segments over the entire duration of participation while euthymic and during ASRM-defined symptomatic intervals (where applicable): the mean (μ), coefficient of variation, kurtosis ($\varkappa$), skewness (α_3), absolute sum of consecutive changes, mean absolute differences, and autocorrelation at lag 1 (Figure 1C; Table S2 in Multimedia Appendix 1). Moreover, we transformed the segmented time-domain variables to the frequency domain (ie, fluctuation rate) using the Fast Fourier Transform [36] and then calculated 5 frequency-domain features: spectral energy (SE), spectral entropy (H_s), spectral centroid (f_c), spectral spread (f_σ), and spectral flux (SF) (Figure 2A). These frequency-domain (ie, spectral) features were calculated to assess the overall and instantaneous fluctuation rate's central tendency, variability, and complexity. All features were normalized using minimum-maximum scaling to the range (0, 1), with scaling parameters fit only on the training data within each cross-validation fold. Outliers were clipped to the 1st and 99th percentiles of the training distribution within each fold. Feature extraction and preprocessing were implemented in Python (version 3.11; Python Software Foundation) using NumPy (version 1.24; NumPy Developers), SciPy (version 1.10; SciPy Developers), and scikit-learn (version 1.3; scikit-learn Developers).

See Table S1 in Multimedia Appendix 1 for the full list of data, variables, and features, and Table S2 in Multimedia Appendix 1 for the mathematical description and interpretation of the features.

Figure 2. Machine learning pipeline. Panels (A) model input: binary feature matrices labeled by clinical state (euthymic vs ASRM-defined symptomatic interval) from 43 participants who experienced ≥1 symptomatic interval, including 252 features and clinical and demographic covariates (age, sex, and pharmacotherapy at study entry) used for adjustment; (B) model training, testing, and validation: nested cross-validation (10 outer×10 inner folds) with group k-fold at the participant level, 70/20/10 train/validation/test split; and (C) model explainability and feature importance: SHAP values and permutation importance aggregated across outer folds and 100 bootstraps. AC_1: autocorrelation at lag 1; ASCC: absolute sum of consecutive changes; ASRM: Altman Self-Rating Mania Scale; AUROC: area under the receiver operating characteristic curve; CV: coefficient of variation; MAD: mean absolute differences; ML: machine learning; PR-AUC: precision-recall–area under the curve; SE: spectral energy; SF: spectral flux; SHAP: Shapley additive explanations; μ: mean; ϰ: kurtosis; α_3: skewness; H_s: spectral entropy; f_c: spectral centroid; f_σ: spectral spread.



Classification Model Development and Performance Evaluation

ML Model Development and Training

The present analysis was designed to assess the association between specific features and concurrently assessed clinical polarities: ML models were used to distinguish ASRM-defined symptomatic intervals from euthymic states (rather than to forecast future episodes) (Figure 2B-C). We constructed binary feature matrices labeled by clinical polarity (euthymic vs ASRM-defined symptomatic interval) and trained an ensemble of tree-based classifiers, specifically Random Forest, Extra Trees, and XGBoost (Extreme Gradient Boosting), to discriminate states within concurrent data segments. Features were computed from the same 7-day window used to assign the ASRM-based clinical state label; accordingly, this is a retrospective state classification analysis and not a predictive or forecasting model. Data were split into training (70%), validation (20%), and test (10%) sets with stratified sampling. Nested cross-validation with 10 outer folds and 10 inner folds was used for unbiased performance estimation, hyperparameter tuning, and feature selection. To prevent information leakage from nonindependent observations, all cross-validation folds were grouped at the participant level (group k-fold), so that every data segment from a given participant was assigned entirely to either the training or the testing partition within each fold.

To evaluate the independent contribution of passive wearable data relative to active self-reports, we trained and evaluated 5 model variants on the following feature subsets: (1) passive wearable features (daily activity+covariates), (2) passive wearable features only (daily sleep features+covariates), (3) passive wearable features only (activity+sleep features combined+covariates), (4) active self-report features only (daily e-VAS mood, anxiety, and energy features +covariates), and (5) the full combined model (passive wearable+active self-report+covariates). All variants used the same nested cross-validation procedure, class-weighted training, and out-of-fold (OOF) evaluation applied to the primary analysis.

Hyperparameter tuning was performed via randomized search within the inner cross-validation folds. The search spaces were as follows: Random Forest (n_estimators: {100, 200, 500, 1000}, max_depth: {3, 5, 8, 12, None}, min_samples_split: {2, 5, 10}, min_samples_leaf: {1, 2, 4}, max_features: {"sqrt", "log2", None}); Extra Trees (n_estimators: {100, 200, 500, 1000}, max_depth: {3, 5, 8, 12, None}, min_samples_split: {2, 5, 10}); and XGBoost (n_estimators: {100, 200, 500}, max_depth: {3, 5, 7, 9}, learning_rate: {0.01, 0.05, 0.1, 0.3}, subsample: {0.7, 0.8, 1.0}, colsample_bytree: {0.7, 0.8, 1.0}, reg_alpha: {0, 0.1, 1.0}, reg_lambda: {0.5, 1.0, 5.0}) (Table S3 in Multimedia Appendix 1). The best-performing model and hyperparameter configuration were selected based on inner-fold area under the receiver operating characteristic curve (AUROC) and refit on the full outer training fold.

Class Imbalance Handling and ML Model Performance

Given the unequal number of euthymic and ASRM-defined symptomatic interval segments, we addressed class imbalance through 2 complementary approaches. First, we applied class weighting (ie, assigning weights inversely proportional to class frequency) during model training so that the minority class (symptomatic intervals) received proportionally greater influence during learning. We did not use oversampling (eg, Synthetic Minority Over-sampling Technique) or undersampling techniques, as class weighting preserves the full data distribution while adjusting for imbalance. In addition to AUROC and AUPRC (area under the precision-recall curve), we report sensitivity, specificity, positive predictive value, and negative predictive value computed at the threshold that maximizes the F_1-score across outer-fold held-out test sets, aggregated across 100 bootstrap iterations. Class prevalence (proportion of ASRM-defined symptomatic interval segments) is reported alongside AU-PRC to facilitate interpretation relative to the no-information baseline. Calibration was assessed using the Brier score and a reliability diagram (calibration curve) constructed on OOF predicted probabilities; the no-information Brier score equals the product of class prevalence and (1 – class prevalence). All bootstrapped confidence intervals were computed by resampling at the participant level (ie, all observations from a given participant were sampled as a block in each bootstrap replicate) to account for the nonindependence of repeated observations within participants. Held-out test performance is referred to as internal validation throughout, as the held-out fold was drawn from the same cohort.

Model Explainability: Permutation Feature Importance and Robustness

To interpret the contribution of individual features to the classification model, we computed SHAP (Shapley Additive Explanations) [37] values on the held-out test sets, providing model-agnostic, local explanations of feature effects (Figure 2C). We aggregated SHAP values across all outer cross-validation folds to derive an overall feature importance ranking, thereby characterizing how each feature influenced the model's outputs. SHAP values were computed using TreeSHAP applied to held-out (OOF) test observations only, ensuring that feature importance reflects model behavior on unseen data and is not inflated by in-sample fitting. SHAP values quantify each feature's marginal contribution to the model's predictions and should be interpreted as characterizing how the fitted model used the features to discriminate ASRM-defined symptomatic intervals from euthymia; they do not establish that these features causally influence mood state or constitute validated biomarkers.

To further evaluate feature robustness and importance stability, we performed permutation importance (PI) analyses on the held-out test sets, derived 3 stability metrics for each feature, and aggregated feature importance and ranking statistics across bootstraps and outer folds to quantify robustness, thereby identifying features that consistently contributed to model performance across

resampling procedures. The details of our variable selection and processing, feature extraction, and ML model's robust explainability pipeline are described thoroughly in our previous manuscript [26]. Model explainability was conducted using TreeSHAP (for tree ensembles) via the SHAP library (version 0.43; SHAP Developers) using Random Forest classifiers. All models were implemented in Python (version 3.11).

Statistical Analysis

We performed label permutation testing (n=100) to generate a null distribution of group differences between euthymia and ASRM-defined symptomatic intervals and applied bootstrapping (n=1000) to estimate variability of the test statistics, thereby addressing limitations imposed by the relatively small sample size. Group differences were assessed using the Mann-Whitney *U* test across the resampled distributions. To ensure reliable inference, we included only clinical polarities with at least 5 observations per group, a threshold chosen to maintain adequate statistical power in each bootstrap iteration. Multiple comparisons across top 5 features in Table 1 were controlled using the Benjamini-Hochberg false discovery rate correction method; statistical significance was assumed for false discovery rate–corrected *P* value <.05. We also computed Cohen *d* effect sizes to quantify differences between the euthymic and symptomatic intervals. All models were adjusted for clinical and demographic covariates (age, sex, and pharmacotherapy) at study entry (Table 2), which we included as additional input features in the classification models. Additionally, we conducted a series of sensitivity analyses by repeating the full analytical pipeline with and without imputation of missing data. We compared classifier performance, selected features, and permutation-based importance rankings across the 2 datasets to assess whether imputation significantly influenced the results.

All statistical tests are reported with the test name, test statistic, and exact *P* value (reported as *P*<.001 when the computed value is below this threshold). Effect sizes are reported as Cohen *d* with 95% CIs derived from bootstrap resampling (n=1000). Sample sizes (n) for each comparison are reported. Nonwear time was also expressed as a percentage of each participant's total follow-up duration to enable comparison across participants with varying enrollment lengths: % nonwear = (nonwear days/total follow-up days) ×100.

Table 1. Predictive performance and clinical operating characteristics of machine learning models across time scales and data modalities for detecting symptomatic intervals in 43 participants with bipolar disorder (I or II) who experienced ≥1 Altman Self-Rating Mania Scale (ASRM)–defined symptomatic interval (ASRM score of ≥10) and were included in the machine learning analysis, recruited from outpatient clinics at Centre for Addiction and Mental Health (Toronto, Ontario) and Queen Elizabeth II Health Sciences Centre (Halifax, Nova Scotia), Canada (December 2020 to January 2025)[a].

Time scales and data types	AUROC[b] (95% CI)	AUPRC[c] (95% CI)	Balanced accuracy (95% CI)	F_1-score optimal threshold (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV[d], % (95% CI)	NPV[e], % (95% CI)	Brier score (95% CI)
Within day									
Passive wearable features: activity+covariates	0.74 (0.64-0.81)	0.43 (0.30-0.54)	0.73 (0.67-0.82)	0.23 (0.17-0.32)	0.72 (0.40-0.93)	0.75 (0.57-0.96)	0.44 (0.29-0.74)	0.92 (0.86-0.98)	0.14 (0.12-0.16)
Within night									
Passive wearable features: sleep+covariates	0.74 (0.66-0.81)	0.50 (0.37-0.63)	0.71 (0.66-0.78)	0.33 (0.21-0.41)	0.57 (0.39-0.80)	0.86 (0.65-0.96)	0.53 (0.29-0.73)	0.890 (0.848-0.944)	0.14 (0.11-0.16)
Daily									
Passive wearable features: activity+covariates	0.90 (0.84-0.95)	0.73 (0.61-0.85)	0.86 (0.80-0.91)	0.39 (0.20-0.53)	0.83 (0.70-0.96)	0.89 (0.82-0.95)	0.66 (0.50-0.81)	0.95 (0.92-0.99)	0.09 (0.07-0.12)
Passive wearable features: sleep+covariates	0.91 (0.86-0.95)	0.71 (0.58-0.85)	0.90 (0.85-0.94)	0.29 (0.26-0.50)	0.92 (0.83-0.99)	0.88 (0.81-0.94)	0.66 (0.51-0.78)	0.98 (0.95-0.99)	0.09 (0.06-0.11)
Active self-report features only: e-VAS[f] mood, energy, and anxiety features+covariates	0.93 (0.88-0.96)	0.75 (0.62-0.87)	0.89 (0.83-0.93)	0.44 (0.25-0.61)	0.85 (0.72-0.95)	0.92 (0.88-0.97)	0.73 (0.61-0.89)	0.96 (0.93-0.99)	0.08 (0.06-0.11)
Passive wearable features only: activity+sleep features combined+covariates	0.91 (0.86-0.95)	0.72 (0.58-0.85)	0.90 (0.85-0.94)	0.29 (0.18-0.52)	0.92 (0.80-1.00)	0.88 (0.82-0.94)	0.66 (0.52-0.80)	0.98 (0.95-1.00)	0.09 (0.06-0.11)
Full combined model: passive wearable+active self-report+covariates	0.92 (0.87-0.96)	0.75 (0.63-0.85)	0.89 (0.82-0.93)	0.34 (0.15-0.57)	0.89 (0.69-0.99)	0.88 (0.82-0.97)	0.67 (0.54-0.85)	0.97 (0.93-0.99)	0.08 (0.06-0.10)

[a]Values represent pooled out-of-fold point estimates along with 95% CIs derived from 100 iterations of participant-level block bootstrapping to account for repeated within-subject observations. Models were evaluated using nested, participant-grouped cross-validation (*StratifiedGroupKFold*) to ensure internal validation on participants unseen during training and feature preprocessing. All models included baseline demographic and clinical covariates (age, sex, and pharmacotherapy at study entry). Decision thresholds were determined per bootstrap iteration by identifying the cutoff that maximized the F_1-score on held-out test predictions. Sensitivity, specificity, positive predictive value, and negative predictive value are presented as percentages. Calibration was evaluated using the Brier score, where lower scores indicate superior probabilistic accuracy.

[b]AUROC: area under the receiver operating characteristic curve.

[c]AUPRC: area under the precision-recall curve.

[d]PPV: positive predictive value.

[e]NPV: negative predictive value.

[f]e-VAS: electronic visual analog scale.

(page number not for citation purposes)

Table 2. Sociodemographic and clinical characteristics of the 43 participants with bipolar disorder (I or II) who experienced ≥1 Altman Self-Rating Mania Scale (ASRM)–defined symptomatic interval (ASRM score of ≥10) and were included in the machine learning analysis, recruited from outpatient clinics at Centre for Addiction and Mental Health (Toronto, Ontario) and Queen Elizabeth II Health Sciences Centre (Halifax, Nova Scotia), Canada (December 2020 to January 2025).

Characteristics	Participants (N=43)
Age (years), mean (SD)	33.7 (9.4)
Sex assigned at birth, n (%)	
Male	14 (32.6)
Female	29 (67.4)
Self-reported gender, n (%)	
Man	13 (30.2)
Woman	22 (51.1)
Queer/gender nonconforming	3 (7)
Prefer not to disclose	5 (11.6)
Education, n (%)	
Completed high school or less	7 (16.3)
College diploma or some university	23 (53.5)
University degree	13 (30.2)
Race, n (%)	
White	20 (46.5)
Asian	10 (23.3)
Black or African Canadian	2 (4.7)
More than 1 race	3 (7)
Unknown	8 (18.6)
Marital state, n (%)	
Single	27 (62.8)
Married/partnered	14 (32.6)
Separated/divorced	1 (2.3)
Widowed	1 (2.3)
Socioeconomic status, n (%)	
Student	5 (11.6)
Full-time work	21 (48.8)
Part-time work	5 (11.6)
Social or disability assistance	2 (4.7)
Unemployed and other	10 (23.3)
Primary diagnosis, n (%)	
Bipolar disorder I	29 (67.4)
Bipolar disorder II	14 (32.6)
Rapid cycling, n (%)	15 (34.9)
Clinical state at study entry, n (%)	
Euthymic	21 (48.8)
In a depressive episode	20 (46.5)
In a hypomanic episode	2 (4.7)
Pharmacotherapy at study entry, n (%)	
None	4 (9.4)
Lithium monotherapy	2 (4.7)
Anticonvulsant monotherapy	5 (11.6)
Antipsychotic monotherapy	5 (11.6)
Combination treatment	27 (62.8)

Sensitivity Analysis for ASRM Threshold

To evaluate whether the choice of ASRM threshold (≥10 vs the standard ≥6) substantially influenced the results, we repeated the complete classification pipeline, including feature extraction, model training with nested cross-validation, SHAP-based explainability, and PI, using ASRM score

of ≥6 as the alternative threshold for defining symptomatic intervals. We report the number of participants and intervals identified under each threshold, the AUROC and AUPRC for each modality, and the correlation of the performance metrics between the 2 thresholds (Spearman ϱ).

Segments with ASRM score of ≥10 satisfy both the ≥6 and ≥10 threshold criteria and therefore represent matched observations that are classified as symptomatic under both thresholds. On the other hand, segments with ASRM scores of 6 to 9 satisfy only the ≥6 threshold and constitute the additional unmatched observations unique to the sensitivity analysis and are therefore not classified as symptomatic under the primary ≥10 criterion.

Identification of Top Descriptors of ASRM-Defined Symptomatic Intervals

We used a supervised ML approach within a 10-fold nested cross-validation framework to classify clinical polarity (euthymia vs ASRM-defined symptomatic intervals) using activity, sleep, and e-VAS–derived features. We evaluated model performance on held-out test sets and quantified feature contributions using SHAP impact scores, complemented by a feature robustness analysis based on PI (top-k selection frequency) across 100 bootstraps with 100 permutation repeats (Figure 2).

As shown in Figures 1C and 2B, we directly compared features across multiple temporal scales within a unified analytical framework, including corresponding classification performance metrics and SHAP analyses for each time scale.

Ethical Considerations

The study was approved by the REB at the CAMH, Toronto, Ontario, Canada, and at the QEII Health Care Center, Halifax, Nova Scotia, Canada, in accordance with the Declaration of Helsinki REB number 059-2019. All participants consented in writing to participate in the study. All participants signed an informed consent form approved by the local REB before any research procedure was initiated. As an incentive, participants received CAD $180 (CAD $1=US $0.72 as of May 28, 2026) upon completing their first year. For the second year, they could choose between an additional $180 payment or ownership of the Oura Ring.

Results

Participant Characteristics

Of the 178 participants enrolled in the study, 43 (24.2%) experienced at least 1 ASRM-defined symptomatic interval during the study period, totaling 133 intervals. Participants were followed for a mean of 540.1 (SD 170.9; range 120-706) days; of these 43 participants, 21 (48.8%) entered the study euthymic, 20 (46.5%) in a depressive episode, and only 2 (4.7%) in a hypomanic episode. Using the operationalized euthymia definition (ASRM score of <10 and PHQ-9<10 on 2 consecutive ratings separated by at least 7 days), 4774 euthymic segments were identified across participants; 1216 met the duration (>7 days) and wearable completeness (≥80%) criteria for inclusion as comparator euthymic segments. The 43 participants who experienced at least 1 symptomatic interval experienced a mean of 3.1 (SD 2.6; range 1-10) ASRM-defined symptomatic intervals, lasting a mean of 10.7 (SD 11.1) days. Table 2 summarizes the sociodemographic and clinical characteristics of these 43 participants. The distribution of ASRM-defined symptomatic intervals (ASRM score of ≥10) per participant is presented in Figure S1 in Multimedia Appendix 1.

Using objective and subjective data, we extracted 252 features characterizing variability in objective (ie, wearable-derived) and subjective (eg, mood) variables during euthymic periods versus symptomatic intervals (see Methods section and Table S1 in Multimedia Appendix 1).

Adherence and Data Completeness

Across all participants, nonwear accounted for a median of 19.9% (IQR 6.5-42.0) of the follow-up duration, corresponding to a median wearable adherence of 80.1% (IQR 58.0-93.5). Median daily self-reported e-VAS adherence was 82.6% (IQR 60.2-94.1), and median weekly self-reported ASRM adherence was 81.2% (IQR 58.1-91.9) (Table S4 in Multimedia Appendix 1).

Across pooled follow-up observations, missingness was 15.0% for daily activity, 23.8% for sleep, 23.9% for daily mood ratings, and 26.2% for weekly ASRM assessments. At the participant level, median missingness was 8.1% (IQR 1.1%-28.0%) for daily activity, 20.4% (IQR 6.8%-42.8%) for sleep, 17.4% (IQR 5.9%-39.8%) for daily mood ratings, and 18.8% (IQR 8.1%-41.9%) for weekly ASRM assessments (Table S5 in Multimedia Appendix 1). When stratified by clinical state, missingness was lower during ASRM-defined symptomatic intervals compared with euthymic periods (activity: 11.4% vs 15.1%, sleep: 19.4% vs 23.9%, and e-VAS: 7.8% vs 24.2%). Missingness in activity and e-VAS differed significantly between clinical states at the participant level (Wilcoxon signed-rank test: activity, P=.016; e-VAS, P<.001), whereas the difference in sleep missingness was not significant (P=.472) (Table S6 in Multimedia Appendix 1). Little MCAR test rejected the null hypothesis of MCAR for both the daily passive/active stream matrix (χ^2_{1031}=20696.27, P<.001) and the weekly assessment matrix (χ^2_{15}=29.70, P=.013), confirming that the data are not MCAR and that missingness is associated with clinical state (Table S7 in Multimedia Appendix 1).

Identification of Top Descriptors of ASRM-Defined Symptomatic Intervals

Daily features consistently outperformed features derived at other temporal scales. Accordingly, we present results for daily features in the main text and report results for other time scales (Figures S2 and S3 in Multimedia Appendix 1).

The daily activity feature (Figure 3A-C) that best described an ASRM-defined symptomatic interval was lower activity spectral entropy (ie, consistently more predictable daily activity patterns: PI=0.25; P<.001; Cohen d=1.92, 95%

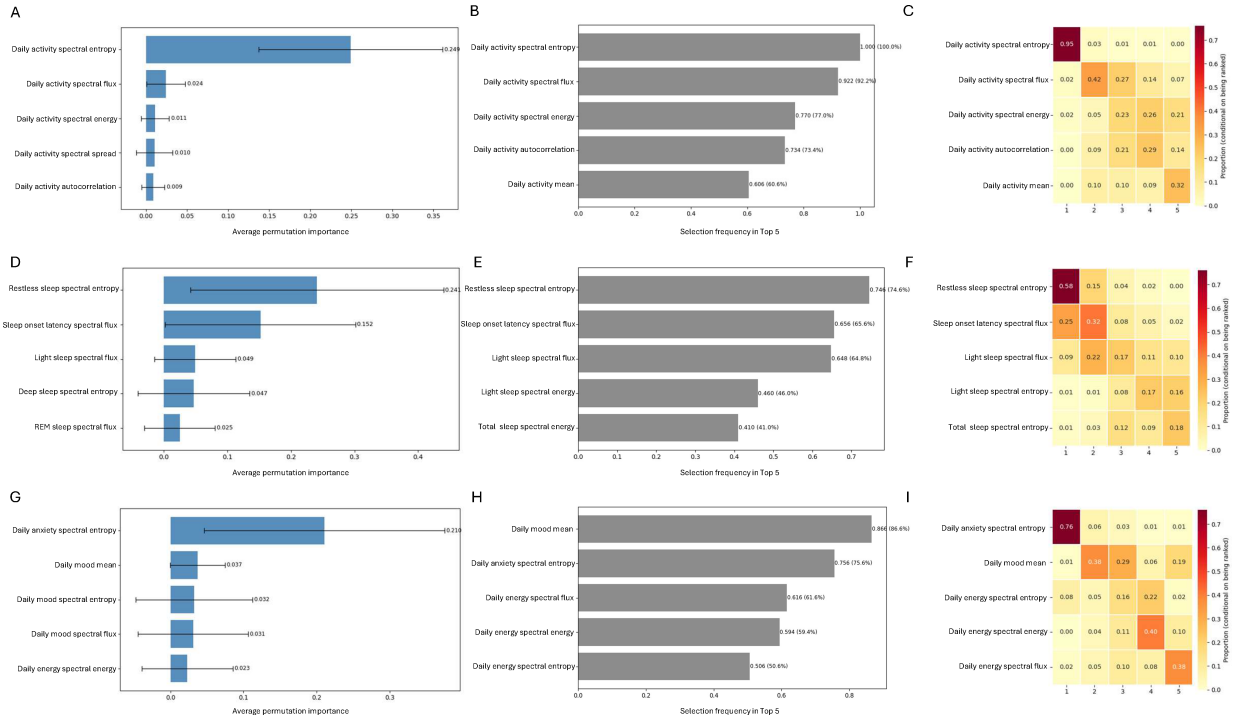
CI 1.56-2.36) (Table 1). Moreover, symptomatic intervals showed significantly lower daily activity SF (ie, lower day-to-day change in behavioral rhythms: PI=0.02; *P*<.001; Cohen *d*=1.46, 95% CI 1.24-1.67) and significantly lower activity SE (ie, weaker rest-activity cycle rhythmicity: PI=0.01; *P*<.001; Cohen *d*=1.01, 95% CI 0.89-1.15).

The daily sleep features (Figure 3D-F) that best described an ASRM-defined symptomatic interval were lower restless sleep spectral entropy (ie, consistently higher restlessness during sleep; PI=0.24; *P*<.001; Cohen *d*=2.24, 95% CI 1.79-2.74) and lower deep sleep spectral entropy (ie, reduced complexity and variability of deep sleep duration; PI=0.05; *P*<.001; Cohen *d*=2.19, 95% CI 1.73-2.76) (Table 1). In addition, ASRM-defined symptomatic intervals were characterized by significantly lower sleep onset latency SF (ie, reduced night-to-night variability in sleep onset; PI=0.15; *P*<.001; Cohen *d*=1.33, 95% CI 1.17-1.49), light sleep SF (ie, consistently longer and more predictable light sleep durations; PI=0.05; *P*<.001; Cohen *d*=1.43, 95% CI 1.22-1.65), and REM sleep SF (ie, consistently shorter and more predictable REM sleep durations; PI=0.03; *P*<.001; Cohen *d*=1.29, 95% CI 1.10-1.47), indicating consistently lower variability across sleep stages.

The daily e-VAS features (Figure 3G-I) that best described ASRM-defined symptomatic intervals were anxiety spectral entropy (PI=0.21; *P*<.001; Cohen *d*=1.92, 95% CI 1.53-2.38) and mood spectral entropy (PI=0.03; *P*<.001; Cohen *d*=1.89, 95% CI 1.48-2.36), both significantly lower during symptomatic intervals, associated with reduced day-to-day complexity in self-reported anxiety and mood (Table 1). Symptomatic intervals were also characterized by significantly lower mood SF (ie, lower day-to-day change in mood rhythms; PI=0.03; *P*<.001; Cohen *d*=1.38, 95% CI 1.13-1.60), alongside significantly higher mean mood levels (PI=0.04; *P*<.001; Cohen *d*=–0.76, 95% CI –1.16 to –0.35). Similarly, energy spectral entropy was significantly lower during symptomatic intervals (PI=0.02; *P*<.001; Cohen *d*=1.92, 95% CI 1.50-2.40), indicating consistently higher and less complex energy levels during symptomatic intervals.

The results of the statistical comparison of the top 5 daily features of activity, sleep, and e-VAS, during euthymia or an ASRM-derived symptomatic interval, and their average PI are presented in Table S7 in Multimedia Appendix 1. The absolute and clinical state-specific permutation feature importance of the aggregated passive data features (ie, sleep and activity) is presented in Figure S5 in Multimedia Appendix 1, and those of the combined active self-report and passive data features are presented in Figure S6 in Multimedia Appendix 1. The ASRM threshold sensitivity analysis (ASRM score of ≥6 vs ASRM score of ≥10) revealed moderate overlap in the most robust descriptors of ASRM-defined symptomatic intervals, namely, in light sleep features (Figure S7 in Multimedia Appendix 1), which is expected upon relaxing the ASRM symptom severity cutoff and including less symptomatic intervals in the analysis (Figure S8 in Multimedia Appendix 1).

Figure 3. Feature importance results for daily activity, sleep, and electronic visual analog scale features in a longitudinal cohort of 43 outpatients with bipolar disorder who experienced ≥1 Altman Self-Rating Mania Scale–defined symptomatic interval (N=133 intervals). Panels (A, D, and G): feature importance plot showing average permutation importance values with error bars representing 95% CI of bootstrapped samples. Panels (B, E, and H): feature selection frequency bar plot showing the percentage of times each feature was selected among the top predictors across models. Panels (C, F, and I): feature rank distribution heat maps showing feature ranking probabilities, with darker colors indicating higher likelihood of being selected in the top rank positions. (A-C): activity features, (D-F): sleep features, and (G-I): electronic visual analog scale features. Results are from nested cross-validation with group k-fold at the participant level.



Classification Performance

Primary Classification: ASRM Score of ≥10

Classification performance increased significantly when aggregating sensor data to daily 24-hour summaries compared with fine-grained sub–daily intervals (within-day activity: AUROC 0.74, 95% CI 0.64-0.81 and Brier score of 0.14, 95% CI 0.12-0.16; within-night sleep: AUROC 0.74, 95% CI 0.66-0.81 and Brier score of 0.14, 95% CI 0.11-0.16). The prevalence of ASRM-defined symptomatic intervals in the daily analytic sample was 20.4%, yielding a no-information AUPRC of 0.20. Among daily data modalities, active daily self-report data (e-VAS mood, energy, and anxiety) yielded the highest point estimates for overall discrimination (AUROC 0.93, 95% CI 0.88-0.96; AUPRC 0.75, 95% CI 0.62-0.87; specificity 91.9%, 95% CI 0.88-0.97; positive predictive value 72.9%, 95% CI 0.61-0.89; Brier score of 0.08, 95% CI 0.06-0.11). However, passive wearable data models achieved broadly comparable discrimination with substantially overlapping confidence intervals, while demonstrating distinct advantages for continuous detection. Specifically, combined daily passive wearable data features (activity and sleep) achieved higher point-estimate sensitivity (92.0%, 95% CI 0.80-1.00 vs 85.4%, 95% CI 0.72-0.95), higher balanced accuracy (0.90, 95% CI 0.85-0.94 vs 0.89, 95% CI 0.83-0.93), stronger rule-out usefulness (negative predictive value 97.8%, 95% CI 0.95-1.00 vs 96.1%, 95% CI 0.93-0.99), and similar risk prediction error (Brier score of 0.09, 95% CI 0.061-0.108) to active self-report data, with daily sleep features alone driving most of this passive signal (AUROC 0.91, 95% CI 0.86-0.95; sensitivity 91.8%, 95% CI 0.83-0.99; Brier score of 0.09, 95% CI 0.06-0.11). Integrating active self-reports and passive sensor metrics into a full multimodal model (AUROC 0.92, 95% CI 0.87-0.96; AUPRC 0.75, 95% CI 0.63-0.85; Brier score of 0.08, 95% CI 0.06-0.10) yielded no meaningful incremental gain over either modality alone (Table 1; Figure S9 and Table S10 in Multimedia Appendix 1).

Class weighting (models were trained with weights inversely proportional to class frequency) modestly improved minority-class performance: across nested coefficient of variation, the AUPRC increased by a median of 0.07, balanced accuracy by 0.06 with weighting, while AUROC increased the least by a median of 0.01, indicating weighting boosted sensitivity/precision for ASRM-defined symptomatic intervals without substantially altering overall discrimination.

For all top-performing daily models, AUROC values were significantly higher than those obtained under random label assignment, as demonstrated by permutation testing (all $P<.01$). In contrast, models based on within-day features alone achieved lower performance than daily features while consistently outperforming random-chance models (within-night sleep: AUROC, mean 0.72, SD 0.12, $P<.05$; within-day activity: mean 0.64, SD 0.11, $P<.05$). Statistical comparison of the top 5 daily features of activity, sleep, and e-VAS during euthymia and during symptomatic intervals are reported in Table S8 in Multimedia Appendix 1, and permutation-based performance comparisons are shown in Figure S4 in Multimedia Appendix 1.

Sensitivity Analysis: ASRM Score of ≥6

Using the threshold of ASRM score of ≥6, 93 participants (vs 43 when using a threshold of ASRM score of ≥10) experienced at least 1 symptomatic interval, totaling 392 intervals (vs 133 with ASRM score of ≥10). Segments with ASRM score of ≥10 (n=133) represent matched observations classified as symptomatic under both thresholds; the remaining 259 segments with ASRM scores of 6 to 9 are classified as symptomatic exclusively under the ≥6 threshold. In the sleep-only (selected data modality) ASRM-threshold sensitivity analysis using primary KNN imputation, model discrimination was comparable with the ASRM score of ≥10 and ASRM score of ≥6 thresholds (pooled OOF AUROC 0.91 vs 0.89; AUPRC 0.66 vs 0.81), with a paired difference (ASRM score of ≥6 minus ASRM score of ≥10) of 0.01 for AUROC (95% CI –0.05 to 0.08) and 0.18 for AUPRC (95% CI 0.01-0.35). However, the OOF-predicted probabilities from the 2 threshold-specific models showed expected negligible correlation at the individual-observation level (Spearman $\varrho=0.016$; $P=.81$; 211 matched observations), indicating that although both thresholds yielded similarly separable classifications, the specific observations identified as high-risk differed substantially between thresholds. This suggests that the ASRM score of ≥10 and ASRM score of ≥6 thresholds capture a different set of symptomatic segments rather than a shared signal at differing sensitivity levels. This interpretation is based on the OOF-predicted probability correlation ($\varrho=0.016$), a prevalence-independent measure of model agreement at the individual-observation level, rather than on the AUPRC threshold comparison (0.66 vs 0.81), which is susceptible to between-threshold differences in positive class prevalence and no-information AUPRC baseline.

Sensitivity Analysis: Missing Data Imputation

Using daily sleep features as sample representative inputs to a Random Forest classifier, OOF-predicted probabilities were highly correlated between KNN and mean imputation (Spearman $\varrho=0.984$; $P<.001$; 211 matched observations) and remained strongly correlated between KNN imputation and complete-case analysis ($\varrho=0.91$; $P<.001$; 186 matched observations), supporting the robustness of model performance to the missing data strategy. Using KNN imputation (k=5), the model achieved mean AUROC 0.91 (SD 0.07) across folds and pooled AUROC 0.90 (95% CI 0.84-0.95); pooled AUPRC was 0.66 (95% CI 0.52-0.82) versus a no-information AUPRC of 0.20. Mean imputation produced similar discrimination (AUROC mean 0.92, SD 0.05; pooled AUROC 0.91, 95% CI 0.86-0.95; pooled AUPRC 0.64, 95% CI 0.51-0.81). Complete-case analysis gave comparable AUROC (mean 0.90, SD 0.12; pooled 0.90, 95% CI 0.83-0.96) but a wider AUPRC CI and lower pooled AUPRC point estimate (0.59, 95% CI 0.42-0.81) (Table S9 in Multimedia Appendix 1). Moreover, no changes were observed across the top 5 sleep features best characterizing ASRM-defined symptomatic intervals from euthymia, further proving the

robustness of the extracted digital markers to imputation strategies.

Discussion

Principal Findings

The primary aim of this study was to identify the most accurate and robust digital markers for detecting ASRM-defined symptomatic intervals. We systematically analyzed multivariate, densely sampled data that combined continuously collected objective data (ie, from a wearable) and subjective data (ie, daily and weekly self-reported measures). We analyzed more than 250 features extracted from 43 participants who experienced at least 1 ASRM-defined symptomatic interval during the study period, totaling 133 intervals.

Our analysis revealed a distinctive pattern: ASRM-defined symptomatic intervals are characterized by reduced variability across all measured domains (ie, activity, sleep, and e-VAS), rather than simply elevated activity levels. In other words, not only were mood and energy levels elevated during ASRM-defined symptomatic intervals but their day-to-day patterns also became more stationary (ie, less variable). Moreover, our analysis revealed that complexity measures derived from frequency-domain analysis outperform traditional time-domain statistics, and spectral entropy emerged as the most robust discriminator across mood, energy, anxiety, sleep, and activity domains.

Our findings showed that the complexity of sleep patterns is disrupted during symptomatic intervals, in accordance with a recent systematic review discriminating unipolar from bipolar depression [38]. Our work extends these findings by demonstrating that frequency-domain features capturing temporal dynamics are more informative than simple duration measures (eg, total sleep time).

In parallel, reduced complexity in self-reported mood and anxiety emerged as highly discriminative features. Mean mood levels were elevated during symptomatic intervals, but with significantly reduced variability compared with euthymia, illustrating the phenomenon of elevated mood with reduced variability (ie, participants were consistently happier than usual). This finding aligns with recent studies demonstrating that entropy-based measures can effectively classify mood states [39-41]. On the other hand, reduced variability in anxiety during ASRM-defined symptomatic intervals is in keeping with clinical observations and consistent with theoretical models proposing hypomanic or manic episodes as a persistent drive to seek rewards, with reduced flexibility to disengage from these behaviors [42].

Similarly, ASRM-defined symptomatic intervals are characterized by more predictable and less complex daily activity patterns. This finding aligns with prior research showing dysregulation of 24-hour motor activity patterns in BD [43]. Furthermore, our findings suggest that activity becomes less adaptive during symptomatic intervals (ie, consistently higher and more predictable). SF and SE of activity were also significantly reduced during symptomatic intervals, further supporting that activity patterns become more rigid [15].

Finally, temporal scale analysis revealed that daily feature aggregation consistently outperformed finer temporal resolutions (ie, within-day features [AUROC 0.76]), suggesting that day-to-day variability captures the most discriminatory and clinically relevant patterns of polarity changes. These findings are consistent with prior research showing that appropriate temporal aggregation improves predictive performance [44,45].

Overall, the consistent pattern of reduced complexity across domains (ie, activity, sleep, and e-VAS) is associated with a pattern consistent with "pathological stability" in underlying regulatory systems during ASRM-defined symptomatic intervals, which may reflect a shared underlying process affecting emotional regulation broadly [46]. These findings are consistent with, but do not establish, a mechanistic explanation.

Strengths and Limitations

This study represents a rigorous study of digital markers for ASRM-defined symptomatic intervals in BD, jointly extracting signature patterns in activity, sleep, mood, energy, and anxiety from more than 100 symptomatic intervals. Several limitations in our study require consideration. First, our participants had relatively high educational levels and limited racial diversity, which constrains generalizability. Future studies should validate these findings in more diverse samples. Second, the current analysis is associative rather than predictive (ie, features are concurrent with ASRM-defined symptomatic intervals). Third, symptomatic intervals were defined by ASRM self-ratings rather than clinical adjudication, which are subject to recall bias, insight limitations during elevated mood states, and variable item endorsement across participants and time points. Clinician interviews or clinician-administered ratings (eg, YMRS) were not routinely performed within each interval; accordingly, the outcome should be interpreted as an "ASRM-defined symptomatic interval" rather than a confirmed hypomanic or manic episode, and findings may not directly generalize to clinician-confirmed diagnoses. Fourth, the euthymic comparator was operationalized using ASRM score of <10 and PHQ-9 <10 thresholds, which reduce but do not eliminate the possibility of residual subsyndromal symptoms. Fifth, sleep staging by the Oura Ring is produced by a proprietary algorithm trained primarily on nonclinical populations; device-derived sleep stages may deviate from polysomnographic standards, particularly during mood episodes when sleep architecture is disrupted. The 5-minute hypnogram should therefore be treated as an approximation rather than a validated measure of sleep staging in BD. However, a major strength of this analysis is the length of follow-up and the value that longitudinal studies offer to the field (ie, after 2 years of follow-up, 25% of the sample had at least 1 ASRM-defined symptomatic interval). Another major strength is the use of explainable ML with systematic feature selection and validation coupled with a rigorous statistical analysis. The

additional analysis of all the studied features using traditional statistics identified the features that statistically differentiated ASRM-defined symptomatic intervals from periods of euthymia. However, some features did not contribute as much to the explainable ML retrospective state classifications because they were redundant in their contribution to the models' outputs (ie, even after managing multicollinearity, their SHAP importance was distributed among correlated features; eg, activity spectral spread).

The choice of ASRM threshold warrants discussion. We used a threshold ≥10 rather than the standard ≥6 to enhance specificity, given the ASRM's known high sensitivity and low specificity. This choice may limit direct comparability with studies using ≥6, as we may have excluded milder or subsyndromal presentations. However, our sensitivity analysis using ≥6 showed comparable clinical state classification results, suggesting that the identified digital markers are robust to threshold choice and identifiable at comparable accuracy and precision, yet detected across different sets of symptomatic intervals. Future studies should systematically evaluate the optimal threshold for digital phenotyping applications (eg, the degree to which the ≥6 and ≥10 thresholds capture distinct symptom presentations), potentially using receiver operating characteristic analysis to derive data-driven thresholds.

In addition, we introduced and validated a novel set of frequency-domain features and demonstrated their robustness. The use of nested cross-validation, PI analysis, and bootstrapped feature ranking, complemented by data-quality sensitivity analyses, further provided confidence in the robustness of the identified markers. This within-participant analytical framework allowed each individual to serve as their own control, reducing between-participant heterogeneity.

Conclusions

This systematic analysis identified a small set of robust digital markers across activity, sleep, and e-VAS domains to detect ASRM-defined symptomatic intervals using objective and subjective data. Our main finding is that passive monitoring using wearable technology matched self-reported data in detecting ASRM-defined symptomatic intervals. This finding is particularly important during hypomanic or manic episodes, as symptoms tend to be underreported. Furthermore, the finding that symptomatic intervals are characterized by decreased variability is consistent with nonlinear dynamics of mood regulation and warrants further prospective investigation. Our study also highlights the importance of study designs with longer follow-ups to capture symptoms that would otherwise go undetected with the conventional weeks-long observation windows in digital phenotyping studies. Together, our findings point to a future in which passively collected digital markers, embedded within a strong clinical foundation, enable us to improve the detection of mood episodes.

Generative AI Use

Generative AI tools were not used for language editing or manuscript formatting assistance during the preparation of this manuscript, nor to generate any of the text, tables, or figures for this paper. All scientific content, data analysis, interpretation of results, and intellectual contributions were conducted solely by the authors, who reviewed and take full responsibility for all content in this manuscript. Generative AI tools were used to improve the formatting of the penultimate version, as well as that of the response to the reviewers.

Funding

This study was funded by the National Institute of Mental Health grant 1R21MH123849-01A1 (AO) and by the Canadian Institutes of Health Research grant 02010PJT-450770-BSB-CEAH-188794 (AO). AO has also received support from the John Sloan (Trey) Leonard III Fund for Bipolar Disorder Research endowment. The work of MA on the project has been supported by the Ministry of Health of the Czech Republic, grant number NU23-04-00534. The rest of the authors' work received no funding. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to privacy considerations. The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Authors' Contributions

RH contributed to data management, development and implementation of signal processing and machine learning algorithms, original draft writing, formal analysis, and data curation. MA contributed to clinical interpretation, validation of results, and manuscript reviewing and editing. MT contributed to statistical analysis, manuscript reviewing, and editing. CO contributed to manuscript reviewing and editing. AH contributed to methodology, validation, and manuscript reviewing and editing. BHM contributed to clinical interpretation, validation of results, and manuscript reviewing and editing. AO contributed to study conceptualization, manuscript writing, reviewing and editing, methodology and results validation, project supervision and management, resource provision, and funding acquisition.

Conflicts of Interest

BHM holds, and receives support from, the Labatt Family Chair in Biology of Depression in Late-Life Adults at the University of Toronto. He also receives compensation from the Department of Psychiatry, University of Toronto, Toronto, Ontario,

Canada; the Centre for Addiction and Mental Health (CAMH), Toronto, Ontario; and The Royal, Ottawa Hospital, Ottawa, Ontario, Canada. He is a member of the Board of Trustees of The Royal Ottawa Hospital. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Supplementary file containing figures and tables.
[DOCX File (Microsoft Word File), 1464 KB-Multimedia Appendix 1]

References

1. Regeer EJ, Kupka RW, Have MT, Vollebergh W, Nolen WA. Low self-recognition and awareness of past hypomanic and manic episodes in the general population. Int J Bipolar Disord. Dec 2015;3(1):22. [doi: 10.1186/s40345-015-0039-8] [Medline: 26440507]
2. Bauer M, Andreassen OA, Geddes JR, et al. Areas of uncertainties and unmet needs in bipolar disorders: clinical and research perspectives. Lancet Psychiatry. Nov 2018;5(11):930-939. [doi: 10.1016/S2215-0366(18)30253-0] [Medline: 30146246]
3. Tania G, Mariam AJ, Shivangi T, Haleemah A, Talen W. The complexity of delayed diagnosis in bipolar disorder. a systematic review of associated precursors, outcomes, and suggested avenues for improvement. medRxiv. Preprint posted online on Oct 14, 2025. [doi: 10.1101/2025.10.13.25337871]
4. Ghaemi SN, Stoll AL, Pope HG. Lack of insight in bipolar disorder. The acute manic episode. J Nerv Ment Dis. Jul 1995;183(7):464-467. [doi: 10.1097/00005053-199507000-00007] [Medline: 7623019]
5. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry. Jun 2025;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
6. Hirschtritt ME, Insel TR. Digital technologies in psychiatry: present and future. Focus (Am Psychiatr Publ). Jul 2018;16(3):251-258. [doi: 10.1176/appi.focus.20180001] [Medline: 31975919]
7. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry. Oct 2018;17(3):276-277. [doi: 10.1002/wps.20550] [Medline: 30192103]
8. Ortiz A, Halabi R, Alda M, et al. Day-to-day variability in sleep and activity predict the onset of a hypomanic episode in patients with bipolar disorder. J Affect Disord. Apr 1, 2025;374:75-83. [doi: 10.1016/j.jad.2025.01.026] [Medline: 39793618]
9. Song YM, Jeong J, de Los Reyes V AA 5th, et al. Causal dynamics of sleep, circadian rhythm, and mood symptoms in patients with major depression and bipolar disorder: insights from longitudinal wearable device data. eBioMedicine. May 2024;103:105094. [doi: 10.1016/j.ebiom.2024.105094]
10. Lipschitz JM, Lin S, Saghafian S, Pike CK, Burdick KE. Digital phenotyping in bipolar disorder: using longitudinal Fitbit data and personalized machine learning to predict mood symptomatology. Acta Psychiatr Scand. Mar 2025;151(3):434-447. [doi: 10.1111/acps.13765] [Medline: 39397313]
11. Lim D, Jeong J, Song YM, et al. Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features. NPJ Digit Med. Nov 18, 2024;7(1):324. [doi: 10.1038/s41746-024-01333-z] [Medline: 39557997]
12. 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. Sep 16, 2025;13:e66277. [doi: 10.2196/66277] [Medline: 40957006]
13. Cho CH, Lee T, Kim MG, In HP, Kim L, Lee HJ. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the Circadian rhythm: prospective observational cohort study. J Med Internet Res. Apr 17, 2019;21(4):e11029. [doi: 10.2196/11029] [Medline: 30994461]
14. Ortiz A, Halabi R, Alda M, et al. Day-to-day variability in activity levels detects transitions to depressive symptoms in bipolar disorder earlier than changes in sleep and mood. Int J Bipolar Disord. Apr 2, 2025;13(1):13. [doi: 10.1186/s40345-025-00379-6] [Medline: 40175826]
15. Aledavood T, Luong N, Baryshnikov I, et al. Multimodal digital phenotyping study in patients with major depressive episodes and healthy controls (mobile monitoring of mood): observational longitudinal study. JMIR Ment Health. Feb 21, 2025;12:e63622. [doi: 10.2196/63622] [Medline: 39984168]
16. Pan Y, Wang P, Xue B, et al. Machine learning for the diagnosis accuracy of bipolar disorder: a systematic review and meta-analysis. Front Psychiatry. 2024;15:1515549. [doi: 10.3389/fpsyt.2024.1515549] [Medline: 39935623]
17. Sun J, Lu T, Shao X, et al. Practical AI application in psychiatry: historical review and future directions. Mol Psychiatry. Sep 2025;30(9):4399-4408. [doi: 10.1038/s41380-025-03072-3] [Medline: 40456953]

18. Kerz E, Zanwar S, Qiao Y, Wiechmann D. Toward explainable AI (XAI) for mental health detection based on language behavior. Front Psychiatry. 2023;14:1219479. [doi: 10.3389/fpsyt.2023.1219479] [Medline: 38144474]
19. Judd LL, Akiskal HS, Schettler PJ, et al. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry. Jun 2002;59(6):530-537. [doi: 10.1001/archpsyc.59.6.530] [Medline: 12044195]
20. Altman EG, Hedeker D, Peterson JL, Davis JM. The Altman Self-Rating Mania Scale. Biol Psychiatry. Nov 15, 1997;42(10):948-955. [doi: 10.1016/S0006-3223(96)00548-3] [Medline: 9359982]
21. Ortiz A, Hintze A, Burnett R, et al. Identifying patient-specific behaviors to understand illness trajectories and predict relapses in bipolar disorder using passive sensing and deep anomaly detection: protocol for a contactless cohort study. BMC Psychiatry. Apr 22, 2022;22(1):288. [doi: 10.1186/s12888-022-03923-1] [Medline: 35459150]
22. First MB. Structured clinical interview for the DSM (SCID). In: Cautin RL, Lilienfeld SO, editors. The Encyclopedia of Clinical Psychology. Wiley; 1-6. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118625392.wbecp351 [Accessed 2026-10-29]
23. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Publishing; 2013. [doi: 10.1176/appi.books.9780890425596]
24. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry. Nov 1978;133:429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
25. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. Apr 1979;134:382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
26. Halabi R, Mulsant BH, Tolend M, et al. A systematic exploration of digital biomarkers for the detection of depressive episodes in bipolar disorder. Npj Mental Health Res. 2026;5(1):13. [doi: 10.1038/s44184-026-00195-5]
27. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. Apr 2009;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
28. Halabi R, Selvarajan R, Lin Z, et al. Comparative assessment of multimodal sensor data quality collected using Android and iOS smartphones in real-world settings. Sensors. Jan 2024;24(19):6246. [doi: 10.3390/s24196246]
29. Halabi R, Mulsant BH, Alda M, et al. Not missing at random: missing data are associated with clinical status and trajectories in an electronic monitoring longitudinal study of bipolar disorder. J Psychiatr Res. Jun 2024;174:326-331. [doi: 10.1016/j.jpsychires.2024.04.036] [Medline: 38692162]
30. Meyer TD, Crist N, La Rosa N, Ye B, Soares JC, Bauer IE. Are existing self-ratings of acute manic symptoms in adults reliable and valid?—a systematic review. Bipolar Disord. Sep 2020;22(6):558-568. [doi: 10.1111/bdi.12906] [Medline: 32232950]
31. Altman E, Hedeker D, Peterson JL, Davis JM. A comparative evaluation of three self-rating scales for acute mania. Biol Psychiatry. Sep 15, 2001;50(6):468-471. [doi: 10.1016/s0006-3223(01)01065-4] [Medline: 11566165]
32. Vieta E, Suppes T, Eggens I, Persson I, Paulsson B, Brecher M. Efficacy and safety of quetiapine in combination with lithium or divalproex for maintenance of patients with bipolar I disorder (international trial 126). J Affect Disord. Aug 2008;109(3):251-263. [doi: 10.1016/j.jad.2008.06.001] [Medline: 18579216]
33. Tohen M, Calabrese JR, Sachs GS, et al. Randomized, placebo-controlled trial of olanzapine as maintenance therapy in patients with bipolar I disorder responding to acute treatment with olanzapine. Am J Psychiatry. Feb 2006;163(2):247-256. [doi: 10.1176/appi.ajp.163.2.247]
34. Mucherino A, Papajorgji PJ, Pardalos PM. K-nearest neighbor classification. In: Data Mining in Agriculture. Springer; 2009:83-106. [doi: 10.1007/978-0-387-88615-2_4]
35. Li C. Little's test of missing completely at random. Stata J. Dec 2013;13(4):795-809. [doi: 10.1177/1536867X1301300407]
36. Cooley JW, Tukey JW. An algorithm for the machine calculation of complex Fourier series. Math Comp. 1965;19(90):297-301. [doi: 10.1090/S0025-5718-1965-0178586-1]
37. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
38. Zhong R, Wu X, Chen J, Fang Y. Using digital phenotyping to discriminate unipolar depression and bipolar disorder: systematic review. J Med Internet Res. May 23, 2025;27:e72229. [doi: 10.2196/72229] [Medline: 40408762]
39. Langholm C, Breitinger S, Gray L, et al. Classifying and clustering mood disorder patients using smartphone data from a feasibility study. NPJ Digit Med. Dec 21, 2023;6(1):238. [doi: 10.1038/s41746-023-00977-7] [Medline: 38129571]
40. Ortiz A, Bradler K, Hintze A. Episode forecasting in bipolar disorder: Is energy better than mood? Bipolar Disord. Jan 22, 2018. [doi: 10.1111/bdi.12603] [Medline: 29356281]
41. Ortiz A, Bradler K, Garnham J, Slaney C, Alda M. Nonlinear dynamics of mood regulation in bipolar disorder. Bipolar Disord. Mar 2015;17(2):139-149. [doi: 10.1111/bdi.12246] [Medline: 25118155]

42. Johnson SL, Edge MD, Holmes MK, Carver CS. The behavioral activation system and mania. Annu Rev Clin Psychol. 2012;8:243-267. [doi: 10.1146/annurev-clinpsy-032511-143148] [Medline: 22077912]
43. Shou H, Cui L, Hickie I, et al. Dysregulation of objectively assessed 24-hour motor activity patterns as a potential marker for bipolar I disorder: results of a community-based family study. Transl Psychiatry. Aug 22, 2017;7(8):e1211. [doi: 10.1038/tp.2017.136] [Medline: 28892068]
44. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther. Feb 2022;149:104013. [doi: 10.1016/j.brat.2021.104013] [Medline: 35030442]
45. Pratap A, Atkins DC, Renn BN, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety. Jan 2019;36(1):72-81. [doi: 10.1002/da.22822] [Medline: 30129691]
46. Ortiz A, Alda M. The perils of being too stable: mood regulation in bipolar disorder. J Psychiatry Neurosci. Nov 1, 2018;43(6):363-365. [doi: 10.1503/jpn.180183] [Medline: 30371048]

Abbreviations

ASRM: Altman Self-Rating Mania Scale
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BD: bipolar disorder
CAMH: Centre for Addiction and Mental Health
DSM-5: *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
e-VAS: electronic visual analog scale
KNN: k-nearest neighbor
MCAR: missing completely at random
MET: metabolic equivalent of task
ML: machine learning
OOF: out-of-fold
PHQ-9: Patient Health Questionnaire (9-item)
PI: permutation importance
QEII: Queen Elizabeth II (Health Sciences Centre)
REB: research ethics board
REDCap: Research Electronic Data Capture
REM: rapid eye movement (sleep stage)
SCID-5: Structured Clinical Interview for the *Diagnostic and Statistical Manual of Mental Disorders*, *Fifth Edition*
SE: spectral energy
SF: spectral flux
SHAP: Shapley Additive Explanations
XGBoost: Extreme Gradient Boosting
YMRS: Young Mania Rating Scale

Edited by Alicia Stone; peer-reviewed by Babak Najand, Lauren Z Atkinson; submitted 26.Jun.2026; final revised version received 05.Oct.2026; accepted 12.Oct.2026; published 18.Nov.2026

Please cite as:
Halabi R, Alda M, Tolend M, O'Donovan C, Hintze A, Mulsant BH, Ortiz A
Passive Monitoring for Detecting Hypomanic Symptoms in Bipolar Disorder Using Explainable Machine Learning: Longitudinal Observational Study
JMIR Mhealth Uhealth 2026;14:e105647
URL: https://mhealth.jmir.org/2026/1/e105647
doi: 10.2196/105647

© Ramzi Halabi, Martin Alda, Mirkamal Tolend, Claire O'Donovan, Arend Hintze, Benoit H Mulsant, Abigail Ortiz. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 18.Nov.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 mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.