

Bipolar disorder relapse detection and prediction using smartwatches. A pilot study for general and personalized machine learning models using anomaly detection methods.

Arnaud Pouchon, Saifeddine Aloui, Luca Mayer Dalverny, Lydia Djennaoui, Claire Guyon-Gardeux, Christelle Godin, Mircea Polosan

Submitted to: Journal of Medical Internet Research
on: October 13, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 39

 Multimedia Appendixes 40

 Multimedia Appendix 0..... 40

Bipolar disorder relapse detection and prediction using smartwatches. A pilot study for general and personalized machine learning models using anomaly detection methods.

Arnaud Pouchon^{1, 2, 3}; Saifeddine Aloui⁴; Luca Mayer Dalverny⁴; Lydia Djennaoui¹; Claire Guyon-Gardeux⁴; Christelle Godin⁴; Mircea Polosan^{1, 2, 3}

¹ Université Grenoble Alpes Grenoble FR

² Centre Hospitalier Universitaire de Grenoble Grenoble FR

³ Grenoble Institute of Neurosciences Grenoble FR

⁴ CEA Grenoble Grenoble FR

Corresponding Author:

Arnaud Pouchon

Université Grenoble Alpes
boulevard de la Chantourne
Grenoble
FR

Abstract

Background: Bipolar disorder (BD) is a chronic and recurrent mental illness characterized by alternating mood episodes, with relapse affecting up to 70% of patients within two years. Early identification of prodromal symptoms preceding mood episodes is critical for timely intervention, yet remains challenging, especially in patients with impaired alexithymia, self-awareness or early trauma. Wearable technologies and digital phenotyping offer novel avenues to monitor physiological and behavioral changes in real time, potentially enabling the prediction of relapses.

Objective: This pilot study aimed to assess the feasibility and performance of general and personalized machine learning models using smartwatch-derived physiological data for detecting and predicting mood relapses and prodromal phases in individuals with BD.

Methods: Ten patients with BD were monitored over six months using Fitbit smartwatches continuously recording heart rate (and heart rate variability - HRstd), sleep (light sleep, REM sleep, deep sleep), and physical activity (represented by the number of steps). Monthly clinical assessments and psychometric evaluations (e.g., MADRS, YMRS) were used to label periods as healthy, prodromal, or relapse. Anomaly detection models were trained on healthy data only and evaluated on their ability to detect (relapse) and predict (prodromal) episodes using several time windows (1h, 4h, 24h) and features. Both general and personalized modeling strategies were explored using unsupervised methods such as Isolation Forest, Local Outlier Factor, and One-Class SVM.

Results: Across 1,395 days of usable data (79% of total), 15 depressive and 8 hypomanic relapses were observed. Significant physiological changes were detected during prodromal and relapse phases. In general models, the best-performing feature combinations for detecting depression were HRstd and steps (Precision Recall Area Under the Curve - PR AUC = 0.67), while light sleep and mean HR were best for hypomania (PR AUC = 0.33). Prodromal depression and hypomania could be predicted with lower but above-chance performance (e.g., PR AUC = 0.34 for prodromal depression using light sleep). Personalized models showed improved performance for relapse detection (e.g., PR AUC = 0.99 for one patient), but not for prodromal prediction, likely due to limited data (median of 2–4 prodromal days per patient).

Conclusions: Smartwatch-derived physiological signals combined with unsupervised anomaly detection methods can effectively detect mood relapses and, to a lesser extent, predict prodromal states in BD. General models are more suitable for predicting prodromes, whereas relapse detection benefits from personalized approaches. These findings support the integration of wearable monitoring and AI into clinical practice to anticipate relapses and enable timely intervention. Clinical Trial: ClinicalTrials.gov Identifier: NCT04703972

(JMIR Preprints 13/10/2025:85815)

DOI: <https://doi.org/10.2196/preprints.85815>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>

Original Manuscript

Bipolar disorder relapse detection and prediction using smartwatches. A pilot study for general and personalized machine learning models using anomaly detection methods.

Arnaud Pouchon^{1,2*}, Saifeddine Aloui³, Luca Mayer Dalverny³, Lydia Djennaoui⁴, Claire Guyon-Gardeux³, Christelle Godin³ and Mircea Polosan^{1,2}

¹ Univ. Grenoble Alpes, Inserm, CHU Grenoble-Alpes, GIN, 38000 Grenoble, France

² Department of Psychiatry, Grenoble Alpes University Hospital, 38000 Grenoble, France

³ Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France

⁴ Univ. Grenoble Alpes, CHU Grenoble-Alpes, 38000 Grenoble, France

* Corresponding authors: apouchon@chu-grenoble.fr

Abstract

Background: Bipolar disorder (BD) is a chronic and recurrent mental illness characterized by alternating mood episodes, with relapse affecting up to 70% of patients within two years. Early identification of prodromal symptoms preceding mood episodes is critical for timely intervention, yet remains challenging, especially in patients with impaired alexithymia, self-awareness or early trauma. Wearable technologies and digital phenotyping offer novel avenues to monitor physiological and behavioral changes in real time, potentially enabling the prediction of relapses.

Objective: This pilot study aimed to assess the feasibility and performance of general and personalized machine learning models using smartwatch-derived physiological data for detecting and predicting mood relapses and prodromal phases in individuals with BD.

Methods: Ten patients with BD were monitored over six months using Fitbit smartwatches continuously recording heart rate (and heart rate variability - HRstd), sleep (light sleep, REM sleep, deep sleep), and physical activity (represented by the number of steps). Monthly clinical assessments and psychometric evaluations (e.g., MADRS, YMRS) were used to label periods as healthy, prodromal, or relapse. Anomaly detection models were trained on healthy data only and evaluated on their ability to detect (relapse) and predict (prodromal) episodes using several time windows (1h, 4h, 24h) and features. Both general and personalized modeling strategies were explored using unsupervised methods such as Isolation Forest, Local Outlier Factor, and One-Class SVM.

Results: Across 1,395 days of usable data (79% of total), 15 depressive and 8 hypomanic relapses were observed. Significant physiological changes were detected during prodromal and relapse phases. In general models, the best-performing feature combinations for detecting depression were HRstd and steps (Precision-Recall Area Under the Curve - PR AUC = 0.67), while light sleep and mean HR were best for hypomania (PR AUC = 0.33). Prodromal depression and hypomania could be predicted with lower but above-chance performance (e.g., PR AUC = 0.34 for prodromal depression using light sleep). Personalized models showed improved performance for relapse detection (e.g., PR AUC = 0.99 for one patient), but not for prodromal prediction, likely due to limited data (median of

2–4 prodromal days per patient).

Conclusions: Smartwatch-derived physiological signals combined with unsupervised anomaly detection methods can effectively detect mood relapses and, to a lesser extent, predict prodromal states in BD. General models are more suitable for predicting prodromes, whereas relapse detection benefits from personalized approaches. These findings support the integration of wearable monitoring and AI into clinical practice to anticipate relapses and enable timely intervention.

Trial Registration: ClinicalTrials.gov Identifier: NCT04703972

Keywords: bipolar disorder; relapse; prediction; detection; personalization; smartwatches; anomaly detection; digital phenotyping; biomarker detection.

1. Introduction

Bipolar disorder (BD) is a chronic mental illness defined by alternating episodes of depression and mania or hypomania [1]. It affects approximately 1 to 2% of the population [2] according to the most restrictive definitions, falling to around 5% with the broader definitions [3], and frequently leads to chronic functional and cognitive disorders, associated with a reduced quality of life [4]. Despite treatment advancements, relapses remain prevalent, affecting up to 70% of patients within two years of initiating therapy [5]. Preventing relapses is vital to mitigate acute complications, cognitive decline, treatment resistance, and long-term disease progression [6]. One of the most promising strategies for relapse prevention is the timely identification of early warning signs (EWS) that often precede mood episodes by days or even weeks. Psychoeducation has demonstrated its effectiveness in delaying relapses by helping patients recognize and manage these prodromal symptoms, thereby improving overall functioning and reducing hospitalizations [7,8]. However, EWS are frequently subtle, heterogeneous, and often missed—especially in individuals with impaired insight or emotional awareness—highlighting the urgent need for more objective, reliable, and continuous methods to detect these early indicators.

However, detecting EWS remains a major clinical challenge due to their subtle, non-specific, and highly variable presentation across individuals and episodes. These signs often emerge gradually and may go unnoticed by both patients and clinicians, especially in individuals with limited emotional awareness (e.g., high alexithymia) or histories of early life adversity, which are both prevalent in BD and associated with reduced interoceptive accuracy [9–11]. Moreover, EWS may be internal (e.g., changes in thought patterns or energy levels) and difficult to verbalize or quantify, further complicating early detection. Stressful life events are among the most consistent and powerful triggers of relapse, mediated by alterations in the hypothalamic-pituitary-adrenal axis and impaired stress reactivity [12–14]. Even during remission, BD patients exhibit elevated baseline

cortisol levels and reduced diurnal variation, suggesting a persistent state of allostatic load that may lower the threshold for future mood episodes [15,16].

In this context, continuous, objective monitoring of physiological markers of stress and arousal—such as heart rate, heart rate variability, skin temperature, and electrodermal activity—offers a promising alternative to traditional self-reported approaches. These biomarkers are now accessible through consumer-grade wearable devices like smartwatches, which enable unobtrusive and scalable long-term monitoring in real-world environments [17–19]. Digital phenotyping, which involves the continuous collection of behavioral and physiological data via wearables and smartphones, offers a promising and personalized avenue to anticipate relapse trajectories in BD by capturing subtle changes in real time [20]. This approach facilitates the early identification of prodromal symptoms—often elusive in clinical settings—enabling timely, proactive interventions that may prevent full-blown episodes and reduce the burden of reactive treatment strategies [21,22]. When integrated with artificial intelligence, digital phenotyping can go beyond mere monitoring: combining passive sensor data with clinical, demographic, and stress-related variables holds the potential to markedly enhance the prediction of individual relapse risk [23,24]. In particular, wearable technologies such as smartwatches provide continuous, non-invasive tracking of physiological markers like heart rate, sleep architecture, and physical activity—key domains known to fluctuate prior to mood episodes [25,26]. Yet, despite the growing body of evidence supporting this approach, a major unmet need remains: the ability to detect prodromal changes not just concurrently, but days in advance—allowing for genuinely preventive, personalized care. Most existing models focus on relapse detection rather than prediction, and few studies have systematically evaluated the performance of general versus individualized algorithms in capturing early warning signals.

To address this gap, the present pilot study introduces a novel framework that leverages smartwatch-derived physiological and behavioral data to develop both general and personalized

machine learning models. Using unsupervised anomaly detection techniques, we aim to identify patterns that reliably signal emerging prodromal states and imminent relapses in BD. By comparing different time windows, feature sets, and modeling strategies, our goal is to assess the feasibility and accuracy of this approach, and ultimately to pave the way for scalable, real-time prediction tools tailored to the unique trajectories of individuals with BD.

2. Method

2.1.Participants

BD patients were recruited at the Grenoble-Alpes University Hospital. Inclusion criteria were a BD diagnosis defined by DSM-5 [1] for patients and diagnosed by a psychiatrist, an age between 18 and 75 years old. Exclusions criteria were other principal psychiatric diagnoses than BD, lack of fluency in French. 10 patients were included (6 females, 4 males). This study was approved by the French Institutional Ethic Committee (CPP Ile-de-France VII: ID-RCB 2020-A00637-32) and was compliant with international standards for tests involving human participants (Helsinki Declaration of 1975, as revised in 2008). The experiment was registered at ClinicalTrials.gov (NCT04703972).

2.2.Study design

Participants were monitored over six months, with data collected through smartwatches, self-reported questionnaires, and monthly clinical evaluations. At the first visit, participants were provided with Fitbit Inspire HR smartwatches to continuously record heart rate, activity, and sleep metrics. Psychometric evaluations were conducted using clinician-administered scales, including Young Mania Rating Scale (YMRS) and Montgomery Asberg Depression Rating Scale (MADRS). The devices were worn continuously by the patients, who also attended monthly follow-ups for data evaluation and a clinical interview with a psychiatrist to determine whether any mood episodes had occurred in the previous month.

2.3.Data management and analysis

We adapted the methodology developed by our group in a prior study [27] to focus more specifically on prodromal phase detection. The novelty here is that we focus more on prodromal phases of the mood episodes. A model is trained using unsupervised anomaly detection methods using feature vectors extracted from only non-relapse days. An “anomaly score” is used to assess if a new feature vector is from a relapse day or not. The higher this score is the more the data is considered as a relapse. We systematically compared multiple features, anomaly detection methods, and time windows within two modeling approaches: a general framework, trained on data pooled across all participants, and a personalized framework, in which a separate model was developed for each individual patient. The models are evaluated on their ability to “detect” the relapses periods but also to “predict” them (i.e. detect prodromal phases).

2.3.1. Participants and Labeling

Clinical phases—including acute episodes (relapses), euthymic states, and periods with prodromal or residual symptoms—were identified during monthly visits based on psychiatrist evaluations, patient self-reports, and daily mood diaries. Relapse states were categorized, with specific temporal boundaries, as minor depression, major depression, hypomania, mania. For our study, we will consider minor and major depression together (labeled D) and as no mania relapse occurred, only hypomania is considered (labeled M). One week (seven days) before the relapse, we label data as prodromal period (Dp for depression, Mp for mania) and one week after, we label data as residual period (Dr for depression, Mr for mania) due to potential residual symptoms. Remaining days labels are defined as “healthy” (H), intercritical, mood-stabilized periods.

2.3.2. Feature Extraction

Behavioral data were continuously collected through Fitbit devices, providing information about (i) daily activity, (ii) sleep, and (iii) heart rate patterns. These data were split into 5-minutes time windows for each day.

Not used as an input feature for model training, a binary indicator wearing watch (WW) was computed to determine whether the participant was wearing the Fitbit during each time window. This measure was crucial for assessing data quality, as missing or unreliable data were often linked to periods when the device was not worn.

The data were preprocessed such that any missing data points (e.g., no recorded steps or heart rate during a time window) were handled appropriately. When no data was available, the corresponding feature values were set to “NaN” (Not a Number), a standard placeholder indicating missing or undefined numerical data. Additionally, if more than three HR measurements were available within the time window, HR variability (HRstd) could be calculated and supplementary “wearing watch” was set to one, indicating active participation; otherwise, “wearing watch” was set to zero for the corresponding time window. For time windows where no steps or sleep data were recorded but “wearing watch” was set to one due to existing HR values, the participant was considered not sleeping, nor walking and the sleep and step features were set to zero for the corresponding time window. In total, seven features were computed per time window:

- i. **Steps:** (1) The total number of steps detected during the 5-minutes time window. This provides a measure of physical activity.
- ii. **Sleep Metrics:** The proportion of time spent in each of the following states: (2) light sleep, (3) REM sleep, (4) deep sleep (5). The total sleep proportion (sum of light, REM, and deep sleeps) is also calculated to represent overall sleep activity during the time window.
- iii. **Heart Rate (HR):** (6) The mean heart rate (HRmean) during the 5-minutes window, providing insight into cardiovascular activity. Additionally, (7) the standard deviation of heart rate (HRstd) was computed as a measure of variability.

2.3.3. Feature Aggregation

For each participant, daily feature vectors were constructed by aggregating the data from the 5-minutes windows into larger time intervals: 1 hour, 4 hours, and 24 hours. This process involved summing the steps over each larger time window, while the sleep metrics and heart rate features and the wearing watch indicator were averaged over the same period. The aggregation was performed for each of the three-time windows (1-hour, 4-hour, and 24-hour), resulting in three distinct datasets for each participant:

- **For 1-hour windows:** 7 features were summed (Steps) or averaged (Sleep, HR): 1 steps, 4 sleep metrics and 2 HR metrics.
- **For 4-hour windows:** the aggregation process was repeated with a similar approach.
- **For 24-hour windows:** a daily summary of the features was produced.

The wearing watch entry is used only for handling missing data and not for training the models, it is not considered as a feature. The final features vector size varied depending on the time window chosen:

- 168 features for the 1-hour time window (24 hours $\times$ 7 features).
- 28 features for the 4-hour time window (6 segments of 4 hours $\times$ 7 features).
- 7 features for the 24-hour time window (1 full day $\times$ 7 features).

2.3.3.1. Handling Missing Data

As wearable data can be incomplete, missing information was a critical issue to address in the dataset construction. Instances where no data or labels were available for an entire day were discarded from the analysis. Additionally, days where the "wearing watch" feature represented less than 90% of the total day (indicating the participant did not wear the device consistently) were excluded to ensure

sufficient data quality.

For the remaining days, where individual features were missing within a given time window, we imputed the missing values with the mean of the corresponding feature values from the “healthy” state (labeled H). This approach was chosen because it preserves the distribution of healthy data and minimizes the bias introduced by extreme values. This approach, following the methodology in previous studies [27], helps ensure that the dataset remains reliable while accounting for occasional missing data without introducing significant bias.

2.3.3.2. Normalization

Standard preprocessing for machine learning was applied to normalize the feature vectors. For each participant, features were normalized using z-score normalization, where the feature value was subtracted by the mean and divided by the standard deviation computed over the “healthy” state (H) data. The normalization parameters (mean and standard deviation) are calculated only on the training set and applied unchanged to the validation and test sets. This normalization ensures that features are on a comparable scale and removes individual patient-specific biases, allowing the model to generalize across subjects.

2.3.4. Data Partitioning

For effective machine learning model evaluation, appropriate data partitioning is crucial. Data were randomly split into training, validation, and test sets. The test set was selected once, while training and validation sets were split 10 times, with results averaged over the 10 iterations.

For all partitioning schemes, only Healthy data (labeled H) were used to train the models, as anomaly detection requires data representing the “normal” state for training purposes. Training data comprised half of the Healthy data, while the remaining data (healthy and non-healthy) were split separately for each label: 40% for validation and 60% for testing. This setup ensured that validation

and test sets were representative of the different clinical states.

2.3.5. Anomaly Detection Methods

Several unsupervised anomaly detection methods were tested in this study, including One-Class Support Vector Machine (OCSVM), Isolation Forest (IF), and Elliptic Envelope (EE), Local outlier Factor (LOF). Autoencoders were not included, as previous work has shown that classical anomaly detection techniques perform better when dealing with small sample sizes and the computational cost of autoencoders is considerably higher. See **supplementary materials** for more details.

Each method produces an anomaly score indicating how much a given feature vector deviates from the healthy baseline. These anomaly scores were then used to classify the data into either normal (Healthy) or anomalous (Relapse, Prodromal). The models were trained either in a general manner (combining data from all participants to train a generic model) or in a personalized manner (training a separate model for each patient).

2.3.6. Evaluation Metrics

The performance of the anomaly detection methods was evaluated using several metrics, including the Area Under the Receiver Operating Characteristic Curve (ROC AUC), the Precision-Recall Area Under the Curve (PR AUC), and Mean AUC (the average of ROC AUC and PR AUC). Additionally, sensitivity ($TP/(TP+FN)$) and specificity ($TN/(TN+FP)$) were computed at the optimal threshold determined using the ROC curve as the point maximizing TP/FP, providing insights into the models' ability to correctly classify both healthy and anomalous states, TP corresponding to the number of cases correctly identified as patient, FP the number of cases incorrectly identified as patient, TN the number of cases correctly identified as healthy, and FN the number of cases incorrectly identified as healthy. **Figure 1** summarizes the methodology used for this study.

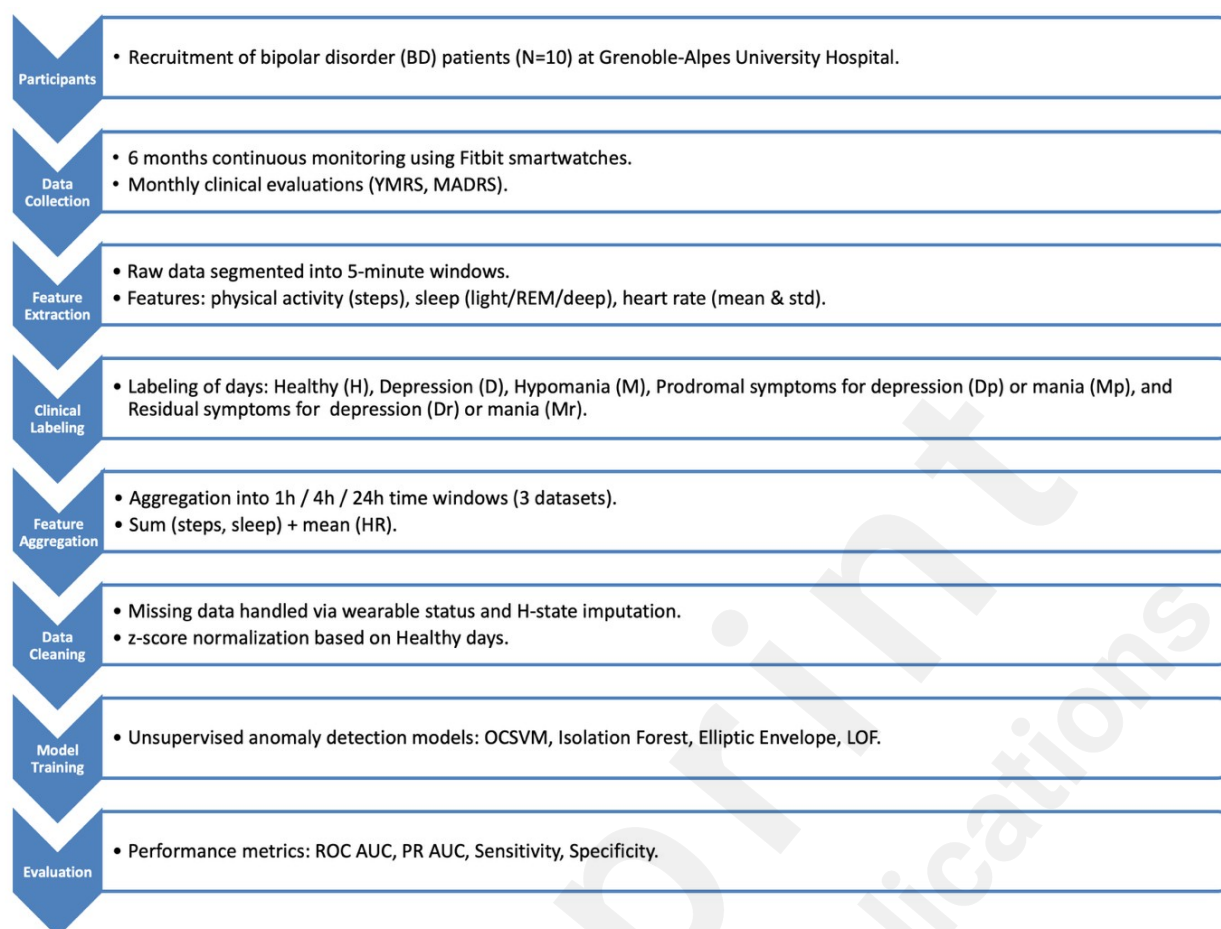


Figure 1. Overview of the study pipeline. After inclusion, participants wore smartwatches continuously and underwent monthly clinical assessments over 6 months. Raw data (steps, sleep, heart rate) were preprocessed and segmented into 5-minute time windows, then labeled according to clinical status (healthy, relapse, prodromal, residual). Features were aggregated over three time windows (1h, 4h, 24h), normalized, and used to train and evaluate unsupervised anomaly detection models aiming to detect and predict mood episodes. BD = bipolar disorder; HR = heart rate; OCSVM = One-Class Support Vector Machine; LOF = Local outlier Factor; PR AUC = Precision-Recall Area Under the Curve; REM = rapid eye movement; ROC AUC = Area Under the Receiver Operating Characteristic Curve.

3. Results

3.1. Participants

The **Table 1** gives the age, gender, and number of relapses for each patient. Among the 10 recruited patients, two discontinued participations in the study before six months, the 6th subject (S06) and the 9th subject (S09).

Subject	Age	Gender	BD subtype	Number of depressive relapses	Number of hypomanic relapses
S01	47	F	Type I	0	2
S02	44	M	Type II	2	1
S03	53	F	Type I	1	1
S04	34	M	Type I	1	1
S05	43	F	Type II	5	4
S06	49	F	Type I	1	0
S07	56	F	Type II	2	0
S08	36	F	Type II	2	0
S09	28	M	Type II	0	0
S10	50	M	Type II	1	0
Total				15	8

Table 1: Patients socio-demographic data, follow-up duration, BD subtype and number of mood relapses. BD = bipolar disorder; F = Female; M = male; S = Subject

3.2.Data availability

The **Table 2** summarizes the number of days available for statistical analysis and machine learning methods across 1767 days of experimentation, all patients included. Of these, 1395 days (79%) were deemed relevant, with the Fitbit worn at least 90% of the time. The low proportion of missing days (21%) highlights strong adherence to the protocol. Exclusions were mainly due to interruptions caused by Fitbit recharging.

Patients generally had more healthy days than relapse days, except for S03 and S05. S03 experienced a prolonged depressive phase followed by persistent hypomania until the study's end. S05 alternated between depression and relapses throughout the study, with few remission days.

	Duration (days)	Wearing watch (90% of time)	H	D	D p	D r	M	M p	M r
S01	215	133	96	0	0	0	20	11	6
S02	208	191	143	6	12	7	9	7	7
S03	217	177	37	108	7	0	25	0	0
S04	182	142	98	16	2	6	13	7	0
S05	187	141	12	54	26	4	23	12	10
S06*	105	89	66	10	6	7	0	0	0
S07	201	184	103	57	13	11	0	0	0
S08	180	178	93	57	14	14	0	0	0
S09*	85	52	52	0	0	0	0	0	0
S10	187	108	91	17	0	0	0	0	0
Total	1767	1395	791	325	80	49	90	37	23

Table 2: Number of Valid Days per Clinical State and Wearable Compliance for Each Participant.

For each patient, the total number of days the watch has been worn (duration in days), the number of days kept for the analysis (wearing fitbit (90%)) and kept for the analysis for each label (H, D, Dp, Dr, M, Mp, Mr). D = minor and major depression together; Dp = prodromal depression; Dr = residual depression; H = healthy or remitted period; M = hypomania; Mp = prodromal mania; Mr = residual mania; S = subject.

3.3.Comparative analysis of relapse and healthy states

The results present statistical analyses comparing relapse states to the healthy state. For all patients combined, mean values and standard deviations of features are compared to normalized values (mean = 0, standard deviation = 1) for the healthy state. Several features show significant differences, such as standard deviation of heart rate variability (HRstd), sleep (total duration), and light sleep, which are significant in both depression and prodromal depression. An increase in sleep proportion is observed during depression and prodromal depression, indicating potential EWS. Light sleep follows a similar trend, while deep and REM sleeps decrease during hypomania and hypomania prodromal phases. Standard deviation of heart rate variability (HRstd) decreases during prodromal phases but increases during depression (**Table 3**).

All patients	HR mean	HRstd	Steps	Sleep (total duration)	Light sleep	Deep sleep	REM
D	-0.16 (1.4)	0.37 (1.9)*	0.36 (1.8)	0.23 (1.3)*	0.5 (1.7)***	-0.06 (1.1)	0.064 (1)
Dp	-0.39 (1.2)**	-0.36 (1.1)**	-0.076 (1)	0.38 (1)**	0.47 (1.3)**	0.26 (1.1)*	0.17 (0.97)
Dr	-0.82 (1.3)***	-0.071 (1.4)	-0.19 (0.89)	-0.059 (0.9)	-0.0027 (1)	-0.063 (1)	0.05 (0.88)
M	0.032 (1.5)	0.49 (1.5)**	0.26 (1.6)	-0.37 (1.6)**	-0.15 (1.9)	-0.27 (1.2)*	-0.36 (1.3)**
Mp	0.53 (1.6)	0.49 (1.2)**	0.59 (1.8)	-0.45 (0.96)**	-0.2 (0.97)	-0.39 (1)*	-0.5 (1)**
Mr	0.38 (1.2)	-0.13 (0.88)	1e-16 (0.99)	-0.098 (1.9)	-0.06 (2)	-0.19 (1.7)	0.1 (1.6)

Table 3: Mean and standard deviation and significance of each feature for each state. D = minor and major depression together; Dp = prodromal depression; Dr = residual depression; M = hypomania; H = healthy or remitted period; HRmean = mean heart rate; HRstd = standard deviation of heart rate variability; Mp = prodromal mania; Mr = residual mania. Light sleep, deep sleep and REM corresponding to sleep stages.

Individual analyses revealed significant variability between patients. For instance, mean HR decreased during depressive states for some patients (S04, S05, S06, S10) but increased during hypomania for S01 and during prodromal hypomanic phases for S02 and S04. Sleep patterns also varied: S02 exhibited decreased sleep during depression, while S05 showed a notable increase. These findings emphasize the need for personalized approaches to model patient-specific variations. See **supplementary materials** for more details.

3.4. General approach: anomaly detection using all subjects

3.4.1. Relapse detection

All combinations of features, models, and time windows were tested, and the best model was selected based on the highest mean of ROC AUC and PR AUC scores. The tables in **supplementary materials** summarize the selected models, their features, time windows, and performance metrics, including mean ROC AUC, PR AUC, sensitivity, and specificity over 10 runs, along with 95% confidence intervals (calculated as 1.96 times the standard deviation). Sensitivity and specificity are reported for the threshold on the anomaly score that maximizes the true positive samples (TP) to false positive samples (FP) ratio.

When relapse (depression and hypomania) is considered as an anomaly, the random PR AUC baseline is 0.52. The best single feature is HRstd, and combining Steps and Light (sleep) improves performance. The 24-hour time window is most effective, indicating global patterns. Isolation Forest and Local Outlier Factor are the best models for one and two features, respectively, with significant performance improvements when using two features ($p=0.0004$) and three features ($p=0.03$), but adding a fourth feature offers no benefit ($p=0.2$). For depression as an anomaly, the baseline PR AUC is 0.45. HRstd is the best feature, with Steps improving performance when combined with it ($p=0.009$). Adding a third feature offers no significant improvement ($p=0.3$). For hypomania as an anomaly, the baseline is 0.19. Light is the best single feature, and combining HRmean and Steps performs better ($p=0.0004$). A 1-hour time window is optimal, and adding a third feature (REM) improves performance ($p=0.006$), but a fourth feature does not ($p=0.1$).

3.4.2. Prodromal phases detection

For detecting prodromal phases, the best features selected for depression and hypomania anomalies were applied to prodromal depression and prodromal hypomania, respectively. The random PR AUC baseline is 0.16 for prodromal depression and 0.07 for prodromal hypomania, with results exceeding these baselines, suggesting the potential for detecting prodromal phases using the selected features. For prodromal depression, the model with two features performs slightly worse than the single-

feature model, though not statistically significantly ($p=0.1$). However, the three-feature model, which was not chosen for depression, performs significantly better ($p=0.004$). For prodromal hypomania, the models perform well, with performance improving as more features are added. The improvements are statistically significant, with $p=0.03$ when increasing from one to two features, and $p=0.001$ from two to three features.

Specific models for prodromal phase detection were selected based on their performance in detecting prodromal rather than relapse phases. While the results are slightly better than those using relapse data, they are less robust due to the smaller sample size. For prodromal depression as an anomaly, the two-feature model does not significantly outperform the one-feature model ($p=0.06$) and shows lower performance. The best feature for prodromal depression is Light, whereas HRstd was the best for depression. The best combinations involve a 24-hour time window, with HRstd still performing well. The two-feature model is not significantly better than the one-feature model ($p=0.064$). For prodromal hypomania, the two-feature model performs better ($p=0.01$), and adding a third feature improves accuracy, though not significantly ($p=0.2$). The optimal model includes a 1-hour time window, with HRmean and REM being the best features, even when used individually. See **supplementary materials** for more information.

3.5. Personalized models

In this section, a model is trained for each patient. The best models are selected based on validation performance, evaluated using the mean AUCs (ROC and PR AUC). Due to the low number of samples and the high computation requirements, not all feature combinations were tested. Instead, features were selected by modality: HR (combining HRmean and HRstd), Steps, Sleep (total), and SleepD (combining Light, Deep, and REM sleep features). Given the small sample size, most of the selected models use only one feature, with two-feature models presented only when they were more relevant.

3.5.1. Relapse detection

The results are presented separately for depression and hypomania, as separating these conditions was shown to yield better performance. For depression as an anomaly, when patients are considered individually, the results are significantly higher, with feature selection varying by patient. The two-feature model is only selected for S03 ($p=0.001$), while one-feature models are chosen for the other patients. The 24-hour time window is relevant for most, but smaller time windows are beneficial for S04, S07, and S08. These improved results should be viewed cautiously due to the small sample size for validation. For hypomania as an anomaly, performance is again much higher than with general models. A 1-hour time window is selected for S02, while 4-hour time windows are chosen for S04. For other patients, the 24-hour time window is used.

3.5.2. Prodromal phases detection with relapse models

The models selected using corresponding relapse data for depression prodromal and hypomania prodromal are evaluated here. For prodromal depression as an anomaly, results are higher than those from general models. The ROC AUC is above random performance (0.5) for most patients, except for S06, which has a very small validation sample (only 2 prodromal depression samples), making the result unreliable. S08 also shows poor performance, possibly due to the use of small time windows that may not be robust. For prodromal hypomania as an anomaly, results are generally poor, with the exception of S01. Only S01 and S05 have 4 validation samples for hypomania, while other patients have only 2, which likely contributes to the lower performance.

3.5.3. Prodromal phases detection with specific models

The performance is higher when we select specific features, time windows and models for the considered task. However, it is important to notice, that with such small set of samples, this could be

due to overfitting. We will evaluate that in the final section with the test set.

3.6. Testing the final models

Here we evaluate the performances of the selected models on the test set. First, we use rather the general models, than the personalized ones. For the general models, the **Table 4** gives the results. We can see that, even bellow the results from validation, the results are above random guess even for prodromal states. We can also observe that prodromal states benefit from a dedicated model choice even with a small number of samples. This should be nuanced by the fact that we do not have many samples for prodromal states.

State M	State	Features	TW	Mode l	mean AUCs	ROC AUC	PR AUC	sensitivit y	specificit y
D	D	HRstd Steps	24h r	LOF	0.67	0.67	0.67	0.05	1.00
M	M	HRmean REM	1hr	LOF	0.47	0.61	0.33	0.10	0.99
D	Dp	HRstd Steps	24h r	LOF	0.37	0.56	0.18	0.24	0.84
Dp	Dp	Light	24h r	LOF	0.50	0.66	0.34	0.02	1.00
M	Mp	Steps HRmean REM	1hr	LOF	0.37	0.61	0.13	0.10	0.99
Mp	Mp	REM Hrmean	1hr	IF	0.39	0.63	0.16	0.05	1.00

Table 4: Performance of general models on the test set by clinical state and selected features.

StateM is the state used to choose the model. State is the state used for evaluation with the test set. D = minor and major depression together; Dp = prodromal depression; Dr = residual depression; EE = elliptic envelope; IF = isolation forest; M = hypomania; H = healthy or remitted period; Hrmean = mean heart rate variability; HRstd = standard deviation of heart rate variability; Mp = prodromal mania; Mr = residual mania; OCSVM = one-class support vector machine; LOF = Local Outlier

Factor; PR AUC = precision-recall area under the curve; ROC AUC = area under the receiver operating characteristic curve (ROC AUC); TW = time window. Light, deep and REM corresponding to sleep stages.

For personalized models, the problem is amplified, more particularly for prodromal states. Personalized models show superior performance for relapse detection, but their benefit for prodromal prediction is limited. This is likely due to the too small amount of data used to select models and to evaluate performance in the validation and in the testing sets. The **Table 5** summarizes the results on the testing set for depression as anomaly, the **Table 6** for hypomania as anomaly, the **Table 7** Depression prodromal as anomaly, and the **Table 8** for Hypomania prodromal as anomaly.

Subject	Feature	TW	Model	mean	ROC	PR	sensitivity	specificity
t	s			AUCs	AUC	AUC	y	y
S02	SleepD	24hr	EE	0,42	0,74	0,11	1	0,62
S03	Steps	24hr	EE	0,86	0,77	0,95	0,39	1
S03	HR	24hr	OCSV	0,93	0,88	0,98	0,58	1
	Steps		M					
S04	Sleep	4hr	EE	0,59	0,75	0,42	0,56	0,9
S05	HR	24hr	LOF	0,95	0,91	0,99	0,81	1
S06	HR	24hr	OCSV	0,53	0,71	0,36	0,67	0,79
			M					
S07	SleepD	4hr	OCSV	0,49	0,47	0,51	0,15	0,9
			M					
S08	HR	1hr	OCSV	0,63	0,57	0,68	0,09	1
			M					
S10	Sleep	24hr	LOF	0,47	0,53	0,42	0,4	0,81

Table 5: Personalized models per subject for depression as anomaly. EE = elliptic envelope; Hrmean = mean heart rate variability; HRstd = standard deviation of heart rate variability; IF = isolation forest; LOF = Local Outlier Factor; Mp = prodromal mania; Mr = residual mania; OCSVM = one-class support vector machine; PR AUC = precision-recall area under the curve; ROC AUC = area under the receiver operating characteristic curve (ROC AUC); S = subject; SleepD = combining

of Light, Deep, and REM features; TW = time window.

Subje ct	Featur es	T W	Mode l	mean AUCs	ROC AUC	PR AUC	sensitivit y	specifici ty
S01	Sleep	24hr	EE	0,67	0,71	0,64	0,33	1
S02	Sleep	1hr	IF	0,31	0,52	0,11	0,6	0,67
S03	Steps	24hr	EE	0,82	0,79	0,84	0,4	1
S03	HR Steps	24hr	OCSV M	0,84	0,81	0,87	0,4	1
S04	SleepD	4hr	OCSV M	0,38	0,54	0,22	0,14	0,97
S05	Steps	24hr	OCSV M	0,54	0,36	0,73	0,54	0,67

Tables 6: Personalized models per subject for hypomania as anomaly. EE = elliptic envelope; Hrmean = mean heart rate variability; HRstd = standard deviation of heart rate variability; IF = isolation forest; Mp = prodromal mania; Mr = residual mania; OCSVM = one-class support vector machine; PR AUC = precision-recall area under the curve; ROC AUC = area under the receiver operating characteristic curve (ROC AUC); S = subject; SleepD = combining of Light, Deep, and REM features; TW = time window.

Subjec t	Feature s	TW	Model	mean AUCs	ROC AUC	PR AUC	sensitivit y	specificit y
S02	Sleep	24hr	IF	0,31	0,38	0,24	0,14	1
S03	Steps	24hr	EE	0,42	0,57	0,28	0,25	0,91
S05	SleepD	1hr	IF	0,87	0,79	0,96	0,67	1
S06	SleepD	4hr	IF	0,28	0,44	0,12	0,33	0,79
S07	HR	1hr	OCSV M	0,32	0,47	0,17	1	0,23
S08	Steps	24hr	IF	0,3	0,41	0,18	1	0,19

Tables 7: Personalized models per subject for depression prodromal as anomaly. EE = elliptic envelope; Hrmean = mean heart rate variability; HRstd = standard deviation of heart rate variability;

IF = isolation forest; Mp = prodromal mania; Mr = residual mania; OCSVM = one-class support vector machine; PR AUC = precision-recall area under the curve; ROC AUC = area under the receiver operating characteristic curve (ROC AUC); S = subject; SleepD = combining of Light, Deep, and REM features; TW = time window.

Subject	Features	TW	Model	mean AUCs	ROC AUC	PR AUC	sensitivity	specificity
S01	Sleep	24hr	EE	0,21	0,3	0,12	1	0,11
S02	Sleep	1hr	IF	0,17	0,29	0,06	0,25	0,83
S04	SleepD	4hr	OCSVM	0,25	0,41	0,09	1	0,21
S05	Steps	24hr	OCSVM	0,59	0,52	0,65	0,71	0,67

Table 8: Personalized models per subject for hypomania prodromal as anomaly. Hrmean = mean heart rate variability; HRstd = standard deviation of heart rate variability; Mp = prodromal mania; Mr = residual mania; OCSVM = one-class support vector machine; PR AUC = precision-recall area under the curve; ROC AUC = area under the receiver operating characteristic curve (ROC AUC); S = subject; SleepD = combining of Light, Deep, and REM features; TW = time window.

4. Discussion

This study explores the feasibility of using smartwatch-derived physiological and behavioral data to detect and predict relapses in BD, including the identification of prodromal phases. By employing a novel approach using unsupervised anomaly detection methods, we highlight key findings that contribute to understanding BD progression and potential intervention strategies.

Our findings suggest that wearable devices can be useful tools for monitoring BD symptoms and may aid in relapse detection; however, predictive performance for prodromal phases remains modest. The models developed demonstrated reliable performances in detecting both depressive and hypomanic episodes. For example, heart rate variability (HRstd) and sleep metrics emerged as

critical features, aligning with previous research on the physiological underpinnings of mood disorders [17,18,28].

4.1.Main results

Generalized models incorporating features like HRstd, steps, and sleep data showed significant predictive capability. Depression states exhibited increased sleep proportions during and before the relapse and decreased HRstd before the relapse and increased HRstd during it, indicative of physiological modification before a depressive episode. Hypomanic states, conversely, were marked by reduced sleep quality and elevated HRstd during and before episodes, reflecting heightened autonomic arousal. These findings support existing literature on sleep disturbances and autonomic dysregulation in BD [13,29].

Importantly, these results echo recent findings showing that sleep and HR variability may act as digital biomarkers of mood instability, especially when derived from passive, real-world monitoring [30,31]. In particular, variability in sleep duration and efficiency has been associated with increased risk of relapse in BD, even in the absence of overt mood symptoms [32]. Moreover, we observed a dynamic pattern in HR variability across phases: HRstd tended to increase during depressive relapses, likely reflecting heightened sympathetic nervous system activity linked to internal stress and ruminative processes [33], whereas its reduction during prodromal phases may signal decreased autonomic flexibility or a hypo-aroused physiological state that precedes episode onset. This interpretation is supported by longitudinal data in adults showing that lower resting HRV predicts the emergence of depressive symptoms over time [34].

Detecting prodromal phases posed a greater challenge due to limited data availability. Nevertheless, models trained specifically for prodromal states demonstrated better performance than those adapted from relapse detection models when leveraging general models' frameworks. Recent

literature highlights the complexities of digital phenotyping in BD relapse prediction. For instance, Chen and Liu (2025) discuss paradoxical patterns in behavioral data, suggesting that subtle changes in activity or physiological signals may precede mood episodes, supporting the potential utility of personalized monitoring approaches [35].

Although the 24-hour time window often showed better performance, it provides the lowest temporal granularity, which may limit sensitivity to rapid changes occurring during prodromal phases. This supports prior studies advocating for real-time monitoring to capture EWS [21,22]. However, concerning personalized models, certainly due to the small amount of data in prodromal phases for each participant, they failed to generalize across the test sets. Nonetheless, continuous high-frequency monitoring may be particularly advantageous for capturing ultra-rapid or ultradian cycles, where subtle early warning signals could be missed by coarser temporal windows.

This limitation is consistent with other digital phenotyping studies, where detecting transitions between mood states—rather than sustained syndromic episodes—requires both high-resolution temporal data and sufficient episode frequency [36]. The limited performance in prodrome detection underlines the necessity for extended follow-up and a higher density of labeled transitions to train more robust models. Nevertheless, the fact that some prodromal periods could be detected above chance suggests a real physiological shift preceding full-blown episodes, supporting the clinical relevance of EWS and advocating for future studies to target these windows as actionable clinical moments [8,37].

4.2. Implications for clinical practice

The integration of wearable technology and anomaly detection algorithms holds promise for transforming BD management. Early detection of prodromal symptoms may allow for prompt interventions, which may lessen the burden of related medical care and prevent full-blown relapses. Personalized models, in particular, demonstrated superior performance for relapse detection,

underscoring the importance of tailoring monitoring systems to individual patient profiles. This aligns with recent findings indicating that behavioral and physiological markers may show complex, sometimes counterintuitive patterns prior to mood episodes, emphasizing the value of personalized and high-frequency monitoring [35]. However, for prediction (i.e. prodromal symptoms detection), general models were superior. Personalized ones failed to generalize on test sets certainly due to a number of prodromal days used to select models too small to guaranty sufficient generalization capabilities. This outcome could be investigated in future studies with longer recording periods including more relapse episodes.

Such approaches could be particularly useful in complementing psychoeducation programs by providing patients and clinicians with real-time, objective feedback on subtle physiological changes indicative of vulnerability [7,8]. The combination of self-awareness strategies and passive digital monitoring may represent a hybrid model for next-generation relapse prevention tools. From a clinical standpoint, embedding such models into mobile apps or clinical dashboards could facilitate early warning alerts, prompting care teams or patients to initiate early interventions (e.g., medication adjustments, increased clinical monitoring, or psychological support). This aligns with the growing field of just-in-time adaptive interventions (JITAIs), which leverage digital data to deliver support at optimal times [38]. Additionally, as many BD patients report difficulty recognizing or verbalizing internal mood changes, such objective monitoring could serve as a compensatory tool to enhance self-monitoring and clinical communication.

4.3.Strengths and limitations

A notable strength of this study is its focus on prodromal phase prediction, which remains underexplored in BD research. The use of unsupervised methods allowed for anomaly detection without requiring extensive labeled datasets, enhancing the study's applicability to real-world scenarios. Furthermore, the combination of generalized and personalized modeling approaches

provided valuable insights into the trade-offs between scalability and specificity.

However, several limitations must be acknowledged. First, the small sample size, particularly for prodromal states, limits the generalizability of the findings. Only 10 participants were included, with two discontinuing the study prematurely, resulting in sparse data for certain analyses. Second, the reliance on self-reported data and clinician assessments for labeling phases introduces potential biases. Future studies should incorporate objective measures, such as ecological momentary assessments, to enhance labeling accuracy. Additionally, wearable device adherence and data quality posed challenges. Approximately 21% of days were excluded due to insufficient wearable usage, highlighting the need for improved device designs and patient compliance strategies. It is also worth noting that external contextual data—such as weather conditions, circadian rhythm disruptions, or social stressors—were not integrated into the current framework. Yet, such multimodal data could significantly improve the ecological validity and predictive power of relapse forecasting models [20,23,39].

4.4. Clinical Alert Management

The translation of anomaly detection into clinically meaningful alerts requires careful consideration. The optimal threshold for triggering an alert should not simply maximize sensitivity or specificity in isolation, but rather the point that balances true positive detection with minimizing false alarms – essentially maximizing the positive predictive value in a real-world clinical context. We propose a threshold determined by maximizing the trade-off between True Positives (TP) and False Positives (FP) on a validation dataset representative of the patient population.

To avoid overwhelming clinicians and ensure alert fatigue is minimized, alerts should not be continuous or overly frequent. We envision a system where a daily summary notification is sent to the clinician, indicating whether an anomaly indicative of potential mood destabilization has been detected. This notification would include a concise summary of the detected anomaly (e.g.,

“Significant increase in heart rate variability observed today”), along with relevant contextual information (e.g., recent activity levels, sleep patterns).

Critically, the system should incorporate a mechanism for clinicians to temporarily disable alerts for individual patients, allowing for clinical judgment to override the automated system when necessary. This “override” functionality is essential for managing situations where the algorithm generates false positives or when the patient is undergoing a planned intervention that might temporarily trigger an alert. Furthermore, a feedback loop should be established, allowing clinicians to flag false positives and contribute to the refinement of the anomaly detection threshold and algorithm over time. This iterative process of clinical validation and algorithm refinement is crucial for ensuring the long-term effectiveness and usability of the system.

4.5.Future directions

To advance this field, several avenues for future research should be explored. Expanding the study population to include larger and more diverse cohorts will be essential to improve the robustness and generalizability of findings. Prolonged monitoring periods are also needed to capture a broader range of relapse and prodromal episodes, enabling the development of more reliable and individualized models. Integrating multimodal data sources—such as GPS tracking, call and text logs, and environmental variables—could enhance model accuracy and deepen our understanding of bipolar disorder triggers [20,23]. Additionally, investigating advanced machine learning techniques, including deep learning and hybrid approaches that combine supervised and unsupervised learning, may reveal more subtle patterns in behavioral and physiological signals. Translating these models into real-world clinical settings will be crucial to assess their feasibility, scalability, and impact on patient care. Finally, future work must also address the ethical and psychological implications of continuous physiological monitoring, such as concerns related to data privacy, user autonomy, and the emotional toll of real-time feedback in vulnerable individuals [40,41]. Adopting participatory

design strategies that involve both patients and clinicians will be key to ensuring that these tools are both acceptable and clinically meaningful.

4.6. Conclusion

This study underscores the potential of wearable technology and machine learning for advancing BD management. By enabling early detection of prodromal symptoms and personalized relapse monitoring, these tools could significantly improve patient care and quality of life. While challenges remain, including data limitations and device adherence, the findings provide a strong foundation for future research aimed at integrating digital phenotyping into routine clinical practice. The next step will be to evaluate the algorithm in a prospective, real-time clinical trial and to integrate contextual data (GPS, stress questionnaires) to improve prodromal prediction.

5. Acknowledgment

We warmly thank Maud Veloso Da Fonseca for her valuable contribution on this study, particularly for coordinating patient scheduling, follow-up, and data collection logistics.

6. Fundings

This project was funded by Fondation d'Avenir (project AP-RM-19-029).

7. Conflicts of Interest

All authors declare no conflict of interest.

8. Data Availability (required)

The data supporting the findings of this study are not publicly available due to participant privacy restrictions and the conditions set by the ethics committee (CPP). Access to the data can be requested from the corresponding author, subject to approval and a formal data sharing agreement.

9. Authors' Contributions

Conceptualization: MP, CG

Data curation: AP, LMD, LD, CGG

Formal analysis: AP, AS, CG, CGG,

Funding acquisition: MP, CG

Investigation: AP, LMD, LD

Methodology: MP, CG, SA

Project administration: MP, CG

Resources: MP, CG

Supervision: MP, CG

Validation: AP, AS, LMD, LD, CGG, CG, MP

Visualization: AP, AS, CG

Writing – original draft: AP, CG

Writing – review & editing: AP, AS, LMD, LD, CGG, CG, MP

10. Abbreviations

BD: bipolar disorder

EE: elliptic envelope

EWS: early warning signs

HR: heart rate

HRstd: HR variability

IF: isolation forest

Mood labels:

- D: depression
- Dp: period with prodromal symptoms of depression
- Dr: period with residual symptoms of depression
- H: healthy period
- M: mania
- Mp: period with prodromal symptoms of mania
- Mr : period with residual symptoms of mania

REM: rapid eye movement

NaN: not a number

OCSVM: one-class support vector machine

PR AUC: precision-recall area under the curve

ROC AUC: area under the receiver operating characteristic curve (ROC AUC)

11. References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders [Internet]. DSM-5-TR. American Psychiatric Association Publishing; 2022 [cité 2025 août 5]. Available from: <https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425787>
2. Rowland TA, Marwaha S. Epidemiology and risk factors for bipolar disorder. *Therapeutic Advances in Psychopharmacology* 2018;8:251-69.
3. Ferrari AJ, Baxter AJ, Whiteford HA. A systematic review of the global distribution and availability of prevalence data for bipolar disorder. *Journal of Affective Disorders* 2011;134:1-13.
4. Grande I, Berk M, Birmaher B, Vieta E. Bipolar disorder. *The Lancet* 2016;387:1561-72.
5. Giménez-Palomo A, Andreu H, Olivier L, Ochandiano I, de Juan O, Fernández-Plaza T, et al. Clinical, sociodemographic and environmental predicting factors for relapse in bipolar disorder: A systematic review. *J Affect Disord* 2024;360:276-96.
6. López-Villarreal A, Sánchez-Morla EM, Jiménez-López E, Martínez-Vizcaíno V, Aparicio AI, Mateo-Sotos J, et al. Progression of the functional deficit in a group of patients with bipolar disorder: a cluster analysis based on longitudinal data. *Eur Arch Psychiatry Clin Neurosci* 2020;270:947-57.
7. Morriss RK, Faizal MA, Jones AP, Williamson PR, Bolton C, McCarthy JP. Interventions for helping people recognise early signs of recurrence in bipolar disorder. *Cochrane Database Syst Rev* 2007;CD004854.
8. Bos FM, Schreuder MJ, George SV, Doornbos B, Bruggeman R, van der Krieke L, et al. Anticipating manic and depressive transitions in patients with bipolar disorder using early warning signals. *Int J Bipolar Disord* 2022;10:12.
9. Ogrodniczuk JS, Piper WE, Joyce AS. Effect of alexithymia on the process and outcome of psychotherapy: A programmatic review. *Psychiatry Research* 2011;190:43-8.
10. Ospina LH, Shanahan M, Perez-Rodriguez MM, Chan CC, Clari R, Burdick KE. Alexithymia predicts poorer social and everyday functioning in schizophrenia and bipolar disorder. *Psychiatry Res* 2019;273:218-26.
11. Schäfer JÖ, Naumann E, Holmes EA, Tuschen-Caffier B, Samson AC. Emotion Regulation Strategies in Depressive and Anxiety Symptoms in Youth: A Meta-Analytic Review. *J Youth Adolescence* 2017;46:261-76.
12. Lex C, Bärner E, Meyer TD. Does stress play a significant role in bipolar disorder? A meta-analysis. *J Affect Disord* 2017;208:298-308.
13. Juruena MF, Cleare AJ, Young AH. Neuroendocrine Stress System in Bipolar Disorder. *Curr Top Behav Neurosci* 2021;48:149-71.
14. Belvederi Murri M, Prestia D, Mondelli V, Pariante C, Patti S, Olivieri B, et al. The HPA axis in bipolar disorder: Systematic review and meta-analysis. *Psychoneuroendocrinology* 2016;63:327-42.
15. Girshkin L, Matheson SL, Shepherd AM, Green MJ. Morning cortisol levels in schizophrenia and bipolar disorder: A meta-analysis. *Psychoneuroendocrinology* 2014;49:187-206.
16. Kapczinski F, Vieta E, Andreazza AC, Frey BN, Gomes FA, Tramontina J, et al. Allostatic load in bipolar disorder: Implications for pathophysiology and treatment. *Neuroscience & Biobehavioral Reviews* 2008;32:675-92.
17. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and Heart Rate Variability: A Meta-Analysis

and Review of the Literature. *Psychiatry Investig* 2018;15:235-45.

18. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart Devices and Wearable Technologies to Detect and Monitor Mental Health Conditions and Stress: A Systematic Review. *Sensors* 2021;21:3461.

19. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics* 2018;77:120-32.

20. Brietzke E, Hawken ER, Idzikowski M, Pong J, Kennedy SH, Soares CN. Integrating digital phenotyping in clinical characterization of individuals with mood disorders. *Neuroscience & Biobehavioral Reviews* 2019;104:223-30.

21. Ortiz A, Hintze A, Burnett R, Gonzalez-Torres C, Unger S, Yang D, 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* 2022;22:288.

22. Rotenberg L de S, Borges-Júnior RG, Lafer B, Salvini R, Dias R da S. Exploring machine learning to predict depressive relapses of bipolar disorder patients. *J Affect Disord* 2021;295:681-7.

23. Colombo F, Calesella F, Mazza MG, Melloni EMT, Morelli MJ, Scotti GM, et al. Machine learning approaches for prediction of bipolar disorder based on biological, clinical and neuropsychological markers: A systematic review and meta-analysis. *Neurosci Biobehav Rev* 2022;135:104552.

24. Passos IC, Ballester PL, Barros RC, Librenza-Garcia D, Mwangi B, Birmaher B, et al. Machine learning and big data analytics in bipolar disorder: A position paper from the International Society for Bipolar Disorders Big Data Task Force. *Bipolar Disord* 2019;21:582-94.

25. Moura PC, Ribeiro PA, Raposo M, Vassilenko V. The State of the Art on Graphene-Based Sensors for Human Health Monitoring through Breath Biomarkers. *Sensors* 2023;23:9271.

26. Masood W, Mukherjee D, Ali S. Machine Learning Tool: A Novel Complementary Method for Early Detection and Better Prognosis of Bipolar Disorder. *Psychiatr Danub* 2022;34:320.

27. Hamieh S, Heiries V, Al Osman H, Godin C. Relapse Detection in Patients with Psychotic Disorders Using Unsupervised Learning on Smartwatch Signals [Internet]. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE; 2023 [cité 2025 août 4]. page 1-2. Available from: <https://ieeexplore.ieee.org/document/10094686/>

28. Esaki Y, Obayashi K, Saeki K, Fujita K, Iwata N, Kitajima T. Circadian variability of objective sleep measures predicts the relapse of a mood episode in bipolar disorder: findings from the APPLE cohort. *Psychiatry Clin Neurosci* 2023;77:442-8.

29. Van Der Werf-Eldering MJ, Riemersma-van Der Lek RF, Burger H, Holthausen EAE, Aleman A, Nolen WA. Can Variation in Hypothalamic-Pituitary-Adrenal (HPA)-Axis Activity Explain the Relationship between Depression and Cognition in Bipolar Patients? *PLoS ONE* 2012;7:e37119.

30. Dagum P. Digital biomarkers of cognitive function. *npj Digital Med* 2018;1:10.

31. Reinertsen E, Clifford GD. A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses. *Physiol. Meas.* 2018;39:05TR01.

32. Krane-Gartiser K, Vaaler AE, Fasmer OB, Sørensen K, Morken G, Scott J. Variability of activity patterns across mood disorders and time of day. *BMC Psychiatry* 2017;17:404.

33. Ottaviani C, Thayer JF, Verkuil B, Lonigro A, Medea B, Couyoumdjian A, et al. Physiological concomitants of perseverative cognition: A systematic review and meta-analysis. *Psychological Bulletin* 2016;142:231-59.

34. Jandackova VK, Britton A, Malik M, Steptoe A. Heart rate variability and depressive symptoms: a cross-lagged analysis over a 10-year period in the Whitehall II study. *Psychol. Med.*

2016;46:2121-31.

35. Chen J, Liu J. Navigating the paradoxes and potential of digital phenotyping for bipolar relapse prediction. *Journal of Affective Disorders* 2026;392:120116.

36. Grunerbl A, Muaremi A, Osmani V, Bahle G, Ohler S, Troster G, et al. Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients. *IEEE J. Biomed. Health Inform.* 2015;19:140-8.

37. Faurholt-Jepsen M, Busk J, Frost M, Vinberg M, Christensen EM, Winther O, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry* 2016;6:e856-e856.

38. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* 2018;52:446-62.

39. Torous J, Onnela JP, Keshavan M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl Psychiatry* 2017;7:e1053-e1053.

40. Simons CJP, Hartmann JA, Kramer I, Menne-Lothmann C, Höhn P, Van Bommel AL, et al. Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. *Eur. psychiatr.* 2015;30:900-6.

41. Gooding P. Mapping the rise of digital mental health technologies: Emerging issues for law and society. *International Journal of Law and Psychiatry* 2019;67:101498.

Supplementary Files

Multimedia Appendixes

This file includes a methodological note regarding anomaly detection methods, as well as a detailed account of the results for each patient, both for the use of general approach and personalized models.

URL: <http://asset.jmir.pub/assets/50175de76037246cb2258dfb0e6bc9b8.docx>