

Politecnico di Torino

Master's Degree in Biomedical Engineering



**Politecnico
di Torino**

Master's Thesis

Digital Phenotyping of Bipolar Disorder Using Wearable Sensors: From Device Comparability to Mood State Recognition

Supervisor:

Prof. Luca Mesin

Co-Supervisor:

Prof. Jordi Fonollosa Magrinya

Candidate:

Valerio Chimisso

July 2026

A.A. 2025/2026

Abstract

Bipolar Disorder (BD) is a chronic psychiatric condition characterized by recurring manic and depressive episodes that severely impair a person’s social, occupational and interpersonal functioning. Traditional psychiatric assessment relies almost exclusively on clinical interviews and subjective patient reports, which are highly susceptible to recall bias and often lead to diagnostic delays of several years. Digital phenotyping offers a different approach by utilizing wearable sensors for the continuous, passive, and objective quantification of physiological and behavioral patterns in every day settings to support faster diagnosis.

This thesis, aims to address these challenges through two primary research tracks. The first objective is to evaluate the comparability and interchangeability of the Empatica E4 and EmbracePlus wearable devices both at the raw signal and extracted features levels, with the purpose of seamlessly employing data recorded with both devices. The second objective is to identify viable biomarkers and develop machine learning models capable of distinguishing between acute mood states and remission baselines in manic and depressive episodes.

The project employs two independent datasets: a cohort of 30 healthy controls wearing both devices simultaneously to assess device agreement, and a clinical cohort of patients with BD monitored during acute and remission phases. Clinical analyses focused on sleep-period data, and patients were grouped into Mania and Depression cohorts based on episode type. The physiological data was processed through a unified pipeline and features were extracted using the FLIRT framework, a standardized tool for computing statistical descriptors from raw wearable signals. Device comparability was assessed through signal and feature-level agreement analyses. Statistical modeling was employed to identify candidate biomarkers, while machine learning classifiers were developed to distinguish acute and remission states.

The device comparability analysis at the raw signal level showed very good agreement for heart rate ($r = 0.86$) and temperature ($r = 0.80$), while electrodermal activity and acceleration exhibited only moderate comparability. At the descriptor level, accelerometry features such as entropy and standard deviation demonstrated

the strongest stability, though most other physiological features were limited to moderate comparability. High variability between patients was observed in all the results. In the clinical study, the Mania cohort yielded several universal biomarkers; specifically, acute episodes were characterized by a significant decrease in HRV and a significant increase in skin temperature. On the contrary, the Depression cohort lacked universal biomarkers. While no depressive biomarkers were observed, the analysis revealed two distinct subtypes characterized by opposite responses of HRV, suggesting that depressive episodes may not follow a single physiological profile. Predictive modeling for the Mania cohort achieved a balanced accuracy of 68.7% using the identified universal biomarkers. The model reached an overall session recognition rate of 77% at 60% sample threshold, successfully distinguishing between acute and remission states. These findings support the feasibility of wearable-based digital phenotyping for mood-state monitoring in BD while highlighting the importance of accounting for physiological heterogeneity across patients.

Contents

Abstract	i
1 Introduction	1
1.1 Bipolar Disorder	1
1.1.1 Manic Episodes	2
1.1.2 Depressive Episodes	4
1.1.3 Clinical Heterogeneity	6
1.2 Current Clinical Assessment and Monitoring	7
1.2.1 Traditional Psychiatric Assessment	7
1.2.2 Limitations of Traditional Monitoring: Subjectivity and Recall Bias	7
1.2.3 The Need for Objective and Continuous Monitoring	7
1.3 Digital Phenotyping and Wearable Monitoring	8
1.3.1 Digital Phenotyping in Psychiatry	8
1.3.2 Wearable Monitoring	8
1.4 Machine Learning for Mood State Recognition	10
1.4.1 Machine Learning in Computational Psychiatry	10
1.4.2 Physiological Heterogeneity Across Patients and Generalizability Challenges	10
1.4.3 Interpretable Machine Learning Approaches	10
1.5 Research Motivation and Objectives	11
1.5.1 Context: The INTREPIBD Project	11
1.5.2 Thesis Contributions	11
1.5.3 Motivation of the Comparability Study	12
1.5.4 Research Hypotheses	12
1.5.5 Research Objectives	12
2 Materials and Methods	13
2.1 Wearable Devices	13

2.1.1	Empatica E4	13
2.1.2	EmbracePlus	14
2.1.3	Device Differences and Implications	14
2.2	Data Composition	15
2.2.1	Dataset 1	15
2.2.2	Dataset 2	15
2.3	Preprocessing	17
2.3.1	AVRO files to CSV	17
2.3.2	Signal cleaning and reviewing	17
2.3.3	Masking	19
2.3.4	Window Selection	21
2.4	Feature Extraction	22
2.4.1	The FLIRT Framework	23
2.4.2	Feature Extraction Pipeline	25
2.5	Device Comparability Analysis	26
2.5.1	Raw Signal Correlation	27
2.5.2	Descriptor Consistency Analysis	28
2.6	Bipolar Disorder Study	32
2.6.1	Dataset Construction and Normalization	33
2.6.2	Biomarker Discovery	34
2.6.3	Predictive Modeling Pipeline	40
3	Results and Discussion	46
3.1	Device Comparability Analysis	46
3.1.1	Raw Signal Correlation Analysis	46
3.1.2	Sleep-Period Comparability Analysis	55
3.1.3	Descriptor Consistency Analysis	58
3.2	BD Biomarkers and Prediction Models	65
3.2.1	Dataset Construction and Cohort Overview	65
3.2.2	Biomarker Discovery via Linear Mixed Models	65
3.2.3	Biological Phenotype Discovery	70
3.2.4	Mania Cohort Predictive Models	74
3.2.5	Depression Cohort Predictive Modeling Results	77
4	Conclusions and Future Work	81
4.1	Device Comparability	81
4.2	BD Biomarkers and Prediction Models	82
4.3	Future Work	83

A	Complete Feature List	88
A.1	Standard Statistical and Complexity Descriptors	88
A.2	Domain-Specific and Custom Features	89
B	Additional Results	92
B.1	Raw Signal Correlation Results	92
B.2	Predictions Visualizations	94

List of Figures

1.1	Progression of bipolar disorder and recommended pharmacological treatments, Figure extracted from [1].	2
2.1	Sleep Mask over ACC magnitude plot. Where the blue signal is the ACC magnitude and the red colored parts represent sleep mask = 1 .	21
2.2	Example of plot used for visual comparison of IBI signal. The figure overlays the E4 and EP IBI signals over the entire recording, with time on the x-axis and IBI values on the y-axis.	28
2.3	Samples and Bland-Altman representative plot (magnitude variance)	32
2.4	Correlation matrix of the Depression Dataset's feature set. Heatmap with feature names on both axes, where each value (color from blue to red) represents the level of correlation between the crossing features.	36
3.1	Distribution of EDA Pearson correlations across subjects	47
3.2	Distribution of ACC Pearson correlation across subjects	48
3.3	Distribution of TEMP Pearson correlation across subjects	50
3.4	Distribution of IBI Pearson correlation across subjects	51
3.5	EDA, ACC, TEMP and IBI overlays for a representative high agreement subject. Shows plots of 48-hour signals with the E4 recording in orange and the aligned EP recording overlaid in blue	52
3.6	EDA, ACC, TEMP and IBI overlays for a representative low agreement subject. Shows plots of 48-hour signals with the E4 recording in orange and the aligned EP recording overlaid in blue	54

3.7	Distribution of heartbeat detection for E4 relative to EP during the full recordings and when restricting the analysis to sleep periods. Heartbeat detection rate is expressed as the percentage of IBI samples detected by the Empatica E4 with respect to the EmbracePlus for each participant. Violin plots represent the distribution across participants, boxplots indicate the interquartile range and median, individual points correspond to single participants. The dashed horizontal line indicates perfect agreement.	57
3.8	Percentage of missing HRV feature windows across subjects for EP and E4.	58
3.9	Feature correlation heatmap across subjects and features	62
3.10	Feature comparability scatter plot for representative subject. Each feature is plotted with its Spearman correlation on the x axis and LOA on the y axis, color coded based on the assigned tier	63
3.11	Scatter plot of the feature (dots representing the single values, EP value on the x axis and corresponding E4 value on the y axis); and Bland-Altman plot for ACC magnitude std (dots plotted with mean (x axis) and difference (y axis) between the devices, gray lines representing LOA boundaries and red line marking the bias.	64
3.12	Scatter plot of the feature (dots representing the single values, EP value on the x axis and corresponding E4 value on the y axis); and Bland-Altman plot for EDA tonic mean (dots plotted with mean (x axis) and difference (y axis) between the devices, gray lines representing LOA boundaries and red line marking the bias.	64
3.13	Volcano plot of Mania cohort biomarkers. Volcano plot of Mania features, color coded for feature profile, with effect size on the x axis and significance on y axis.	66
3.14	Volcano plot of the Depression cohort biomarkers. Volcano plot of Depression features, color coded for feature profile, with effect size on the x axis and significance on y axis.	69
3.15	Forest plot of hrv_mean response for each patient. Average HRV response magnitude plotted for each patient with uncertainty intervals. Diamond shape at the bottom representing the averaged response of the cohort.	71
3.16	Balanced Accuracy across LOSO folds.	75

3.17	Representative plots of sample predictions over time for two folds of the Mania model. For each subplot the predictions for a single session (acute or remission) are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.	77
3.18	Balanced accuracy across folds for the three depression datasets: full cohort, droppers sub-cohort and increasers sub-cohort	78
3.19	Representative plots of session predictions over time for the two Depression sub-cohorts. For each subplot the predictions for a single session (acute or remission) are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.	80
B.1	Plots of sample predictions over time for the Mania cohort model. For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.	95
B.2	Plots of sample predictions over time for the Depression Increaser sub-cohort model. For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.	96

B.3 Plots of sample predictions over time for the Depression Dropper sub-cohort model. For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions. 97

List of Tables

1.1	Episode severity from YMRS score	4
1.2	Episode severity from HDRS score	6
2.1	Comparison between the E4 and EP sensors.	14
2.2	Metadata associated with each recording session.	16
2.3	Summary of features extracted per FLIRT module and additional descriptors over each 5-minute sliding window.	25
2.4	Depression Cohort Sample Distribution	34
2.5	Mania Cohort Sample Distribution	35
2.6	LMM output example	39
3.1	Summary statistics of signal comparability metrics.	46
3.2	Descriptive statistics for EDA correlation, RMSE and lag	47
3.3	Descriptive statistics for ACC correlation, RMSE and lag	48
3.4	Descriptive statistics for TEMP correlation, RMSE, and lag	49
3.5	Descriptive statistics for IBI correlation, RMSE, and lag	50
3.6	Summary statistics of signal comparability metrics, restricted to sleep periods.	56
3.7	Global descriptor comparability results between devices. The table reports the chosen metrics for each feature, along with an associated tier that classifies the agreement between devices in poor, moderate and strong.	60
3.8	Representative physiological biomarkers identified in the Mania cohort	66
3.9	Universal Biomarkers identified in the Mania cohort	67
3.10	Representative physiological biomarkers identified in the Depression cohort	69
3.11	Universal Biomarkers identified in the Depression cohort	70
3.12	Universal Biomarkers identified in the HRV-Dropper sub-cohort . . .	72
3.13	Universal Biomarkers identified in the HRV-Increaser sub-cohort . . .	73

3.14	Performance comparison between the evaluated feature configurations in the Mania cohort using LOSO cross-validation	74
3.15	Session Recognition Performance Across Thresholds	76
3.16	Predictive performance comparison between the Global Depression model and the phenotype-stratified sub-cohort models using LOGO cross-validation	78
3.17	Session recognition rates across depression configurations at 0.50 / 0.55 / 0.60 evaluation thresholds	79
A.1	Standard statistical descriptors applied to EDA, ACC, TEMP, and IBI sequences.	89
A.2	Domain-specific and custom engineered features extracted per window.	90
B.1	Complete subject-level raw signal comparison results across all modal- ities.	93
B.2	Complete subject-level raw signal comparison results across all modal- ities, restricted to signal recorded during sleep.	94

Chapter 1

Introduction

1.1 Bipolar Disorder

Bipolar Disorder (BD) is a chronic psychiatric condition characterized by repeated Manic (or hypomanic) episodes, alternating with Major Depressive episodes, as shown in Figure 1.1. These episodes are highly debilitating pathological states that last for weeks or months, impairing a person's ability to function across social, occupational, and interpersonal contexts. The complexity and unpredictability of these episodes make the identification of biomarkers critical for timely clinical intervention [2].

BD has a substantial global impact, in 2021 BD was affected approximately 37 million people worldwide, roughly 0.5% of the global population [3]. It is also one of the main causes of disability-adjusted life years among mental disorders, particularly because the onset usually occurs during late adolescence or early adulthood, a critical period for development. While the prevalence between men and women is roughly equivalent, type and frequency of episodes often differ between genders: men typically experience a higher frequency of manic episodes, whereas women are more affected by depressive or mixed states [2]. Furthermore, the mortality rate associated with BD is alarmingly high, mostly driven by a significant risk of suicide; lifetime suicide attempt rates are estimated between 20% and 60% [4].

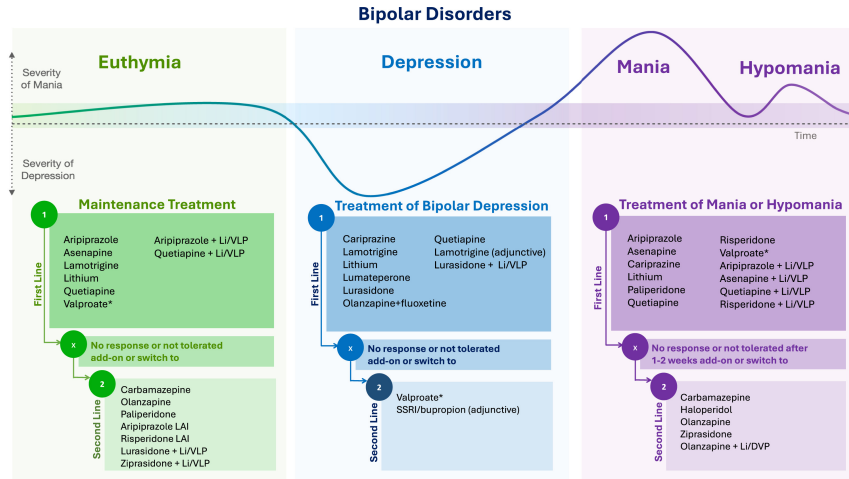


Figure 1.1: Progression of bipolar disorder and recommended pharmacological treatments, Figure extracted from [1].

1.1.1 Manic Episodes

To understand the longitudinal course of BD, it is essential to define Mania and Depression, as they represent two very different physiological and behavioral states characteristic of the condition.

A Manic Episode is defined as "a distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased activity or energy, lasting at least 1 week and present most of the day, nearly every day" [2]. During these periods, at least three of the following symptoms (four if the mood is only irritable) must be present:

1. Inflated self-esteem or grandiosity.
2. Decreased need for sleep (e.g. feeling rested after only 3 hours of sleep).
3. More talkative than usual or pressure to keep talking.
4. Flight of ideas or the subjective experience that thoughts are racing.
5. Distractibility (attention too easily drawn to unimportant or irrelevant external stimuli).
6. Increase in goal-directed activity (socially, at work, or sexually) or psychomotor agitation.
7. Excessive involvement in activities with a high potential for painful consequences (e.g. unrestrained buying sprees, sexual indiscretions, or foolish business investments).

The main symptoms, heightened self-esteem, increased motor activity and pressured speech, are often so pronounced that manic episodes are rarely overlooked by clinicians. The severity of these symptoms often requires hospitalization to prevent self-harm or damage to the patient's social and financial situation. Hypomanic episodes involve similar symptoms but are characterized by reduced severity and shorter duration (at least 4 days), without causing the same level of impairment or requiring hospitalization [2].

Young Mania Rating Scale

In clinical research and this study, the severity of manic symptoms is quantitatively assessed using the Young Mania Rating Scale (YMRS). This 11-item questionnaire conducted by the clinician rates symptoms based on the patient's report of their condition over the previous 48 hours and the clinician's observations during the interview [5].

The items are:

1. Elevated Mood
2. Increased Motor Activity-Energy
3. Sexual Interest
4. Sleep Decrease
5. Irritability (scored 0-8)
6. Speech (rate and amount) (scored 0-8)
7. Language-Thought Disorder
8. Content (scored 0-8)
9. Disruptive-Aggressive Behavior (scored 0-8)
10. Appearance
11. Insight

The items scored on a 0-8 scale (Irritability, Speech, Content, and Disruptive Behavior) are weighted more heavily to account for the frequent lack of cooperation or insight found in patients with severe symptoms. A total score is calculated to categorize the patient's state, ranging from euthymia (baseline state) to severe mania ($YMRS \geq 26$), with a maximum of 60, as shown in Table 1.1.

Table 1.1: Episode severity from YMRS score

Clinical State	YMRS Score
Severe	≥ 26
Moderate	15–25
Subsyndromal	8–14
Remission	0–7 (after an episode)
Euthymia	0–7 (for ≥ 8 weeks)

1.1.2 Depressive Episodes

Depressive episodes in BD are often more extended and frequent than manic ones, representing the phase where patients experience the greatest subjective distress and functional disability. A Major Depressive Episode is characterized by a period of at least 2 weeks during which there is either a depressed mood or a loss of interest or pleasure (anhedonia) in nearly all activities [2].

Patients must exhibit at least five of the following symptoms:

1. Depressed mood most of the day, nearly every day.
2. Markedly diminished interest or pleasure in all, or almost all, activities.
3. Significant weight loss or gain, or decrease or increase in appetite.
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation.
6. Fatigue or loss of energy.
7. Feelings of worthlessness or excessive/inappropriate guilt.
8. Diminished ability to think or concentrate, or indecisiveness.
9. Recurrent thoughts of death or suicidal ideation.

Hamilton Depression Rating Scale

The severity of depressive symptoms is measured via the Hamilton Depression Rating Scale (HDRS), specifically in the case of this study the 17-item version (HDRS-17) was used. Developed by Max Hamilton in 1960, it remains the gold standard for assessing depression severity in clinical settings. The scale emphasizes somatic symptoms such as weight loss, insomnia, and physical anxiety, as well as cognitive and affective symptoms. [6]

Items are scored 0-2 or 0-4, the higher scores in the 0-4 items are weighted to reflect the clinical significance of specific behavioral observations and somatic complaints. The items are:

1. Depressed Mood (0-4)
2. Feelings of Guilt (0-4)
3. Suicide (0-4)
4. Insomnia: Early (0-2)
5. Insomnia: Middle (0-2)
6. Insomnia: Late (0-2)
7. Work and Activities (0-4)
8. Retardation (slowness of thought/speech, impaired concentration) (0-4)
9. Agitation (0-4)
10. Anxiety: Psychic (0-4)
11. Anxiety: Somatic (0-2)
12. Somatic Symptoms: Gastrointestinal (0-2)
13. Somatic Symptoms: General (0-2)
14. Genital Symptoms (0-2)
15. Hypochondriasis (0-4)
16. Loss of Weight (0-2)
17. Insight (0-2)

Total scores are interpreted as in Table 1.2 to categorize episode severity, with a maximum total score of 50 [6].

Table 1.2: Episode severity from HDRS score

Clinical State	HDRS-17 Score
Severe	≥ 24
Moderate	15–23
Subsyndromal	8–14
Remission	0–7 (after an episode)
Euthymia	0–7 (≥ 8 weeks)

1.1.3 Clinical Heterogeneity

A persistent challenge in the monitoring of BD is the heterogeneity of its symptoms'. BD, and its two polarities, do not manifest as easily predictable symptoms, but as a wide spectrum of behavioral and physiological phenotypes, even within the same episode type.

A patient can also experience a mixed state, where symptoms of the opposite polarity occur simultaneously. For example a patient in a Depressive Episode may present some symptoms characteristic of mania such as racing thoughts, pressured speech and motor agitation, despite maintaining a low mood [2].

Individual responses to the same episode type can also vary drastically, especially in depression:

- **Psychomotor Variation:** In depression one patient may exhibit psychomotor retardation, characterized by a complete "stillness," reduced speech and decreased physical activity. While another may present with agitated depression, experiencing excitability and restlessness.
- **Vegetative Symptoms:** While some patients experience insomnia and weight loss, others show features such as hypersomnia (excessive sleep) and hyperphagia (increased eating) [7].

This variability is an obstacle to the identification of global biomarkers, if one patient's depressive state is marked by immobility and another's by constant movement, a single prediction model based on activity features fails to categorize both correctly. However, this challenge may be circumvented by finding specific endophenotypes or subgroups using digital phenotyping and machine learning (ML). Identifying these subgroups would allow for the development of personalized prediction models that account for the specific symptomatic "signature" of a subgroup of patients, leading to more precise and effective clinical interventions [7].

1.2 Current Clinical Assessment and Monitoring

1.2.1 Traditional Psychiatric Assessment

Standard diagnosis for BD currently relies on clinical observation and interviews, a practice that stayed mostly unchanged for decades. This methods cause significant diagnosis delay, with patients frequently waiting up to 10 years from the first symptoms to receive an accurate diagnosis. This delay is also caused by high rates of misdiagnosis, with approximately 40% to 70% of BD patients initially identified as having Unipolar Depression [8] [9].

Historically, this field has been characterized by an absence of objective, biological markers, such as validated laboratory tests or neuroimaging protocols, meaning that clinicians must navigate unclear diagnostic boundaries without the support of strong physiological evidence [9].

1.2.2 Limitations of Traditional Monitoring: Subjectivity and Recall Bias

During current monitoring and consultations, patients are asked to summarize their mood and behavior over the preceding weeks, subjecting the assessments to recall bias and subjectivity of the patient. This retrospection is often compromised by the episode itself: depressive episodes may produce a pessimistic exaggeration of symptoms, while manic phases are accompanied by a lack of clinical insight, causing patients to minimize them. Additionally, self-reports collected at fixed intervals separated by long periods of time fail to capture the fluctuations of mood cycling [10].

1.2.3 The Need for Objective and Continuous Monitoring

To bridge the gap between periodic clinical visits, there is a need for passive and continuous monitoring solutions that can quantify a patient's current state in real-world settings. Consumer and medical grade wearables capture real-time features, such as alterations in sleep architecture, physical activity and autonomic nervous system regulation, that could help in the detection of manic and depressive states. These devices provide an objective secondary data stream that avoids the limitations of a patient's perception of his condition. Recent evidence suggests that such continuous monitoring can detect "Early Warning Signals" of relapse approximately 13 days before of an episode onset, providing a critical window for interventions that could prevent the worst consequences of the disorder [11] [12].

1.3 Digital Phenotyping and Wearable Monitoring

1.3.1 Digital Phenotyping in Psychiatry

Digital phenotyping is defined as the continuous, passive quantification of the human phenotype using data from personal digital devices, such as smartphones and advanced wearables. Instead of relying on traditional psychiatric assessments, this approach captures objective behavioral and physiological indicators directly within the patient’s daily environment [10].

By tracking the evolution of these continuous patterns in time, ML algorithms can establish a personalized baseline for each individual. This baseline makes it possible to detect subtle deviations in behavior, sleep architecture and autonomic activity that typically manifest days or weeks before a full clinical relapse into a Manic or major Depressive state [10].

1.3.2 Wearable Monitoring

To implement digital phenotyping in clinical cohorts, researchers utilize specialized hardware capable of continuous, non-invasive data acquisition. Research grade devices are the gold standard for clinical investigations requiring access to raw signals. Within this domain, the Empatica E4 wristband is historically the most used and validated tool in academic research, now currently in the process of being replaced by the next-gen bracelet EmbracePlus (EP) [13, 14]. The E4 provides synchronous, high-resolution physiological signals, including blood volume pulse, electrodermal activity, and accelerometry, in a comfortable bracelet form that makes it usable over extended periods of time [15].

Physiological Signals Acquired From Wearables

Current multi-sensor wearables typically track three core physiological signals:

Heart Rate and Heart Rate Variability Photoplethysmography (PPG) sensors project light into the skin and measure differences in light absorption to derive changes in blood volume pulse, through this method wearable devices are able to track both instantaneous heart rate (HR) and Heart Rate Variability (HRV). HRV is a useful indicator of the autonomic nervous system’s activities: A decrease in HRV parameters indicates a suppression of the parasympathetic system and an activation of the sympathetic nervous system, a physiological shift characteristic of Manic states [13, 15].

Electrodermal Activity Electrodermal Activity (EDA) measures the electrical conductance of the skin, subject to fluctuation based on the filling of sweat glands, innervated by the sympathetic nervous system and susceptible to changes in emotional arousal and stress. In the context of bipolar disorder, patients in deep Depressive states often exhibit slow EDA responses, whereas transitions toward mania is characterized by elevated baseline conductance and frequent EDA spikes [14, 15].

Accelerometry and Activity Patterns Tri-axial accelerometers capture physical movement across the three spatial axes, quantifying variations in movement activity, which is fundamental in both polarities of the disorder. It can identify the psychomotor retardation (physical stillness and lethargy) that characterizes major depression, as well as the psychomotor agitation, and disrupted sleep-wake cycles typical of a manic episode [10, 15].

Challenges of Wearable Physiological Data

One technical issue of using wearable physiological data in clinical deployment is its low signal-to-noise ratio. Because data is captured passively during daily life, simple physical actions, such as walking, typing, or washing hands, introduce massive motion artifacts into PPG and EDA signal streams. These artifacts can mask true physiological responses or mimic autonomic arousal, leading to false positives in prediction models if not aggressively filtered [16].

Longitudinal monitoring with wearable devices also suffers from frequent data gaps due to technical faults, sensor displacement or patients sometimes removing the device to shower or sleep. The presence of these data gaps require robust imputation frameworks and appropriate windowing techniques that preserve signal integrity [16].

Wearables in Bipolar Disorder

A growing literature demonstrates that these physiological signals can predict mood alterations with effective statistical power. Recent architectures utilizing deep learning networks on combined actigraphy and EDA streams have achieved classification accuracies exceeding 91% in differentiating manic phases from euthymia [15]. The state of the art trend is the passive collection of data via smartphones and wearables to quantify patient behavior in real-world settings [11]. This objective data has the purpose of identifying behavioral changes during mood states in physical activity, sleep architecture and autonomic nervous system regulation, without requiring active patient effort [10].

1.4 Machine Learning for Mood State Recognition

1.4.1 Machine Learning in Computational Psychiatry

ML has become an essential tool in computational psychiatry because it provides the means to exploit the massive amount of data generated by wearable sensors. While traditional diagnosis relies on subjective interviews, ML algorithms can use objective measurements, like heart rate and movement, to predict a patient’s clinical state. These models act as a bridge between daily behavior and neurobiology, offering a way to identify mood patterns and helping compensate for the current lack of reliable biological markers [8, 10].

1.4.2 Physiological Heterogeneity Across Patients and Generalizability Challenges

A significant challenge for ML models is that BD is highly heterogeneous, what is normal for one patient may be a sign of illness for another. Because of this variability, generalized models often fail to categorize unseen patients correctly. Recent research highlights that personalized models, which establish a patient-specific baseline to detect deviations from an individual’s own norm, are significantly more accurate than general models, preventing the algorithm from misinterpreting a patient’s natural habits as a symptom of relapse. An intermediate approach to this problem could be to find subtypes of patients within the same condition, grouped by similarities in acute phase symptoms and behaviors [17, 18].

1.4.3 Interpretable Machine Learning Approaches

To be useful in clinical practice, machine learning models have to be interpretable: Clinicians and patients need to understand the reasoning behind a prediction to trust and act upon it. To address this, researchers are adopting methods such as SHAP (Shapley Additive Explanations), which reveal the specific features guiding a prediction. For example a model might show that it predicted a depressive episode because it detected a combination of an increased resting HR and a significant drop in sleep duration, this level of transparency makes digital biomarkers usable in a clinical setting, allowing for faster and more precise interventions while maintaining reliability [8, 10].

1.5 Research Motivation and Objectives

1.5.1 Context: The INTREPIBD Project

The INTREPIBD project is an international multicenter research initiative focused on improving the monitoring and management of BD through digital phenotyping and machine learning. The project aims to address one of the major limitations of current psychiatric practice: the lack of objective biomarkers for assessing mood states and tracking their evolution over time.

To achieve this goal INTREPIBD combines continuous physiological monitoring with standard clinical assessments. Participants are monitored using wearable devices that collect physiological signals including EDA, HRV, skin temperature and accelerometry, while clinical evaluations are performed regularly using traditional psychiatric rating scales. These data are then integrated to identify physiological patterns associated with different mood states and treatment responses.

The project has established a large longitudinal dataset of patients with BD monitored across multiple clinical centers. This dataset provides a valuable foundation for investigating digital biomarkers of mood episodes and developing machine learning models capable of supporting objective and personalized psychiatric care.

Several studies produced within the INTREPIBD project have already demonstrated the potential of wearable signals for characterizing mood states, autonomic nervous system alterations, sleep disturbances, and treatment response in BD. These findings support the growing role of digital phenotyping as a tool for continuous and objective mental health monitoring [19, 20].

1.5.2 Thesis Contributions

This thesis is developed within the INTREPIBD framework and focuses on the analysis of physiological data collected through wearable devices. In particular it investigates data acquired from the Empatica E4 and Empatica EP platforms, which provide continuous measurements of EDA, PPG, skin temperature, and accelerometry.

The work addresses two complementary objectives. The first is the evaluation of device comparability between the E4 and EP platforms, an important step for ensuring continuity between historical and future datasets collected within the project. The second is the exploration of viable biomarkers and the development of machine learning models for mood state recognition using wearable-derived physiological features. Through this investigations, this thesis aims to contribute to the project's

objective of developing objective, data-driven tools for the monitoring of BD.

1.5.3 Motivation of the Comparability Study

The recent introduction of the Empatica EP as a substitute for the old-gen E4 creates a challenge for longitudinal research. If physiological signals and derived features are not comparable across devices, datasets collected with the E4 previously and EP more recently cannot be analyzed and processed as if registered with the same device. Furthermore, existing feature extraction pipelines, including those based on the FLIRT framework, were originally developed around the E4 data format and require adaptation to EP’s output format. For these reasons, a comparability study between the two devices was performed to evaluate the degree of agreement between the two devices at both the signal and feature levels.

1.5.4 Research Hypotheses

- **H1:** Physiological raw signals acquired by the Empatica EP are comparable to those acquired by the Empatica E4.
- **H2:** Most physiological FLIRT features will demonstrate agreement between the two devices.
- **H3:** Wearable-derived physiological features contain information capable of discriminating between mood states in Mania and Depression episodes.
- **H4:** Machine learning models trained on multimodal physiological features will outperform the baseline in state classification.

1.5.5 Research Objectives

- Evaluate signal agreement between Empatica E4 and EP recordings.
- Evaluate feature agreement after standardized preprocessing and feature extraction.
- Identify Universal Biomarkers capable of distinguishing acute and remission mood states.
- Develop and evaluate machine learning models for mood-state classification using wearable-derived physiological features.

Chapter 2

Materials and Methods

2.1 Wearable Devices

Physiological data was acquired using two bracelet devices developed by Empatica: the E4 and the EP. Both devices are designed for clinical use and provide continuous recordings of multiple physiological signals used in clinical research.

2.1.1 Empatica E4

The E4 is a wearable device that has been widely used between 2016 and 2022 for research, especially with the purpose of physiological monitoring. It provides access to raw sensor data, allowing full control over signal processing.

The device includes the following sensors:

- **Electrodermal Activity:** measures skin conductance as an indicator of sympathetic nervous system activity (sampling rate: 4 Hz)
- **Photoplethysmography:** used to evaluate heart rate and inter-beat intervals (IBI)
- **Accelerometer (ACC):** 3-axis motion sensor (sampling rate: 32 Hz)
- **Skin Temperature (TEMP):** temperature sensor (sampling rate: 4 Hz)

The E4 allows both Bluetooth streaming and autonomous recording, with a storage capacity of over 48 hours of continuous data. It has been extensively validated in research contexts and is considered a reference device for wearable signals acquisition.

2.1.2 EmbracePlus

The EP is the next generation wearable developed by Empatica, designed to extend the capabilities of the E4 with improved hardware and advanced cloud data management. The device includes the same type of sensors of E4, with some differences. The EP provides:

- **EDA sensor** with ventral wrist placement
- **Advanced PPG sensor** with 4 multi-wavelength channels for improved cardiovascular signal quality
- **Accelerometer** at higher sampling frequency (64 Hz)
- **Digital skin temperature sensor**
- Extended battery life (>2 days)
- Additional derived digital biomarkers (e.g., sleep, respiratory rate)

2.1.3 Device Differences and Implications

Although both devices measure the same core physiological modalities (EDA, ACC, IBI, and TEMP), they present major differences in hardware implementation and signal acquisition, as in Table 2.1. These differences introduce variability in signal

Table 2.1: Comparison between the E4 and EP sensors.

Feature	Empatica E4	EmbracePlus
Accelerometer sampling rate	32 Hz	64 Hz
PPG technology	Single-channel PPG	Four-channel multi-wavelength PPG
Temperature sensor	Thermistor-based sensor	Digital temperature sensor
EDA sensor	Standard ventral electrodes	Improved electrode geometry and skin contact for increased signal stability
Signal processing	Direct access to raw physiological data	Cloud-based processing with derived biomarkers and digital health insights

quality, scaling, and noise characteristics, particularly for EDA and features derived

from IBI. For example, wearable EDA measurements are known to be sensitive to electrode design, skin contact and motion artifacts, which may differ across devices.

2.2 Data Composition

Two datasets were used in this project, both provided by Hospital Clinic of Barcelona. The first dataset was used exclusively for the comparison between the two generations of bracelets, containing only healthy control subjects. The second includes BD patients in various stages of the disorder, recorded using either the E4 or EP bracelet, and was used for biomarker extraction and the development of prediction models. The two datasets are independent and involve different subjects

2.2.1 Dataset 1

The dataset consists of 30 control subjects, each wearing both bracelets for a continuous period of 48 hours, during the recording the E4 and EP bracelets were worn respectively in proximal and distal position on the left wrist. The subjects are aged between 25 and 40 and do not present any comorbidities relevant to the purpose of the study. Clinical evaluation was performed by expert clinicians using the YMRS and HDRS described in 1.1.1 and 1.1.2 ensuring that all participants are below the thresholds for mania and depression.

2.2.2 Dataset 2

This is the main dataset used for biomarker extraction and prediction. It consists of several recordings (one recording is about 48 hours of continuous signal) of BD patients both hospitalized and home-based, which are approximately 70% E4 and 30% EP bracelets. The dataset is also paired to a metadata table which contains all relevant information on the patients and the recording sessions (Table 2.2).

In order to describe the clinical state of the patient there are three different **Episode Types** and three **Episode Times**. All patients are diagnosed with BD, the episode they are going through can be: Mania, Depression or Euthymia. Episode Time describes the phase of the episode for that specific recording session, it can be either acute, response or remission, based on the rating scales' score seen in 1.1; An **acute** phase corresponds to an initial diagnosis with a score above the clinical threshold (score > 7), response indicates a reduction from the initial score without reaching remission, and **remission** is defined as a score below the threshold

following a previously acute state. Euthymia is not associated to an episode time since it stands for normal state, when a patient is neither in mania nor in depression.

Raw scores and clinical states associated to the recording are assigned by clinicians in a hospital setting and patients are instructed to wear the bracelet for the following two days after a diagnosis or check-up, either in the hospital if they are to be hospitalized or at home.

The dataset was collected between september 2021 and october 2025, including around 400 recordings of 280 patients aged between 20 and 79 years old, with an average of 44 years old and a standard deviation of 14.5. Only the patients with both acute and remission recordings of the same episode were selected to be part of the study, enabling the evaluation of episode responses within subjects and baseline removals. Of the 280 patients only 51 fulfilled this requirement, after applying further signal quality and recording length requirements this selection resulted in 13 patients with mania and 19 patients affected by depression.

Table 2.2: Metadata associated with each recording session.

Category	Variables
Recording information	Recording ID, visit, inclusion date, recording start time, session code, device type (E4/EmbracePlus), device serial number, recording duration.
Patient demographics	Patient/control label, sex, year of birth, age.
Clinical information	Episode type (mania, depression, euthymia), episode time (acute, response, remission), clinical setting (hospital/home).
Clinical assessment	YMRS item scores and total score; HDRS item scores and total score; Positive and Negative Syndrome Scale (PANSS) positive, negative, general psychopathology, and total scores.
Diagnosis	Bipolar disorder subtype (I, II, NOS), comorbid psychiatric diagnosis, age at onset, years since onset, seasonal pattern, rapid cycling, suicide attempts.
Comorbidities	Cardiovascular, neurological, endocrine, respiratory, oncological, autoimmune disorders, chronic pain.
Medication	Current pharmacological treatment at the time of recording, including lithium, antidepressants (SSRIs, SNRIs, tricyclics, MAOIs, others), antipsychotics (first- and second-generation), anticonvulsants, benzodiazepines, beta-blockers, opioids, antihistamines, antiarrhythmics, and other medications.

2.3 Preprocessing

2.3.1 AVRO files to CSV

A conversion pipeline was developed to convert the fragmented AVRO recordings into CSV format. Since the E4 stores physiological signals as continuous CSV files, whereas the EP records data in segmented AVRO files (approximately 30-minute blocks), the pipeline collects all the AVRO and transforms them into a compatible file system.

The pipeline was implemented in Python (version 3.10.12) and consists of two main stages. First, AVRO files were extracted using the PySpark (version 3.3.4) framework to handle nested data structures, loading AVRO segments and storing the raw data in an intermediate structure. Second, the segmented data were reconstructed into continuous time series for each signal (EDA, TEMP, IBI and ACC). This process included ordering segments in chronological order, attaching consecutive blocks and handling gaps and overlaps based on timestamps. When time gaps between AVRO files exceeded 3 minutes the CSV is split into two, to avoid long Not a Number (NaN) segments that could interfere with the preprocessing. Signal specific processing was applied where necessary, for example, IBI were derived from systolic peak timestamps, while accelerometer data were scaled with device specific factor and multiplied by 64 to match the E4 format, where 1g equals 64 ACC units.

To ensure compatibility with the E4 format, signals were resampled to match the corresponding sampling frequencies (e.g., 4 Hz for temperature and 32 Hz for accelerometer). When downsampling was required, a simple anti-aliasing strategy based on moving average filtering was applied before resampling. Finally, the signals were exported in CSV files in order to replicate the E4 structure (initial timestamp, sampling frequency, and sequential samples).

2.3.2 Signal cleaning and reviewing

Basic signal cleaning was applied to EDA, ACC, IBI and TEMP to reduce noise and artifacts while preserving relevant signal. This pipeline was designed with two main objectives: to enable a comparison of signals across the two devices and to prepare the data for the FLIRT feature extraction package.

EDA preprocessing

EDA signals were preprocessed using a minimal pipeline designed to remove major artifacts while preserving the structure of the signal for feature extraction.

A fourth-order **Butterworth low-pass filter** with a cutoff frequency of **0.2 Hz** was applied to reduce high frequency noise and motion artifacts while retaining the slow dynamics of the EDA signal.

Large non physiological **spikes were identified and removed** using a threshold method. Spikes were defined as rapid signal variations exceeding a threshold of 3 μS within a window of 12 samples (3 seconds) and returning back to the previous value within 5 min. To avoid interpolating meaningful data only short events were considered, while longer deviations were preserved. Detected artifacts were linearly interpolated to ensure signal continuity.

ACC preprocessing

ACC signals were preprocessed using an approach that reduced high frequency noise while preserving movement relevant for agitation analysis.

A fourth-order **Butterworth low-pass filter** with a cutoff frequency of **10 Hz** was applied to each acceleration axis, to attenuate high frequency sensor noise while preserving rapid motion components necessary for mood prediction.

No additional transformations, such as gravity removal, were applied in order to preserve the original E4 signal characteristics needed from FLIRT to properly compute features.

TEMP preprocessing

TEMP signal preprocessing was handled similarly to EDA preprocessing. Fourth-order **Butterworth low-pass filter** with a cutoff frequency of **0.1 Hz** was applied, being TEMP a slow moving signal and less sensitive to artifact than EDA.

The same spike removal function described in EDA preprocessing was used, with the difference that, since temperature moves in both directions, also rapid temperature drops were included as possible artifacts. Parameters were also adjusted to account for the different characteristics of the signal, the spike detection threshold was lowered to an increase or a decrease of 2 $^{\circ}\text{C}$ inside a window of 16 samples. This approach accounts for the fact that temperature artifacts may be either sudden heating or cooling (e.g. loss of skin contact).

IBI preprocessing

IBI signals were preprocessed using a conservative approach that removed non physiological values and outliers while preserving the temporal structure of the signal.

First physiologically implausible values were removed by applying **fixed thresh-**

olds, excluding intervals shorter than 0.3 seconds or longer than 3.0 seconds (above 200 bpm or below 20 bpm).

To further reduce the impact of noise and detection errors, in case of artifacts in the physiological range, a **Hampel filter** was applied to the IBI series. This robust outlier detection method identifies local deviations based on the median and median absolute deviation within a sliding window, replacing outliers with the local median. No interpolation or resampling was applied to preserve the original timing of cardiac events and avoid introducing artificial variability. This is important for downstream feature extraction, which relies on accurate temporal relationships between beats.

2.3.3 Masking

Two masks are produced, the first identifies whether the bracelet is being worn or not (On/Off mask), while the second one detects sleep periods (Sleep mask). The On/Off mask is then applied to the clean signals in order to exclude the segments where the bracelet is not in contact with the skin.

On/Off mask

The On/Off mask was computed using a rule based approach combining information from EDA, TEMP and ACC activity.

To reduce computational weight and focus on relevant signal, all signals were first downsampled to 1 Hz (ACC magnitude variance was first computed at native frequency (32 Hz) and then downsampled).

Off-body conditions were identified based on three main criteria:

- **Temperature-based detection:** Off-body periods were associated with low absolute temperature values and sudden drops in temperature. A rolling median over a 5-minute window was used to detect abrupt decreases, capturing events such as device removal from the skin.
- **EDA-based detection:** Low or absent EDA values were considered indicative of poor skin contact or device removal.
- **Motion-based detection:** Periods of very low accelerometer variance (computed over a 60 second window) were interpreted as inactivity.

Thresholds for temperature and EDA were defined using mean and standard deviation to account for variability across subjects, at the same time absolute bounds were also employed at minimum or maximum physiological thresholds.

These criteria were combined into a decision rule, where a segment was classified as off-body when temperature and EDA conditions indicated invalid signals or one of the two if the low movement condition is True.

$$\text{Off-body} = (T_{\text{off}} \wedge E_{\text{off}}) \vee ((T_{\text{off}} \vee E_{\text{off}}) \wedge A_{\text{still}})$$

T_{off} : temperature based off-body condition E_{off} : EDA based off-body condition A_{still} : low activity condition based on accelerometer variance

To improve temporal consistency and reduce frequent transitions, the resulting binary mask was postprocessed by setting minimum segment durations. Short off-body segments (<5 minutes) were reclassified as on-body, while short on-body segments (<90 seconds) were removed.

The final output is a binary 1 Hz time series indicating valid on-body (1) and invalid off-body (0) periods, which is then applied to all physiological signals by filling the segments corresponding to off-body with NaNs.

Sleep mask

Sleep periods were identified using an activity based approach derived from the van Hees method [21]. Raw tri-axial acceleration signals were first used to compute the wrist orientation angle relative to gravity, the median angle was estimated over a short window to reduce noise, and its temporal variation was used to quantify movement. In parallel, the Euclidean Norm Minus One (ENMO) was computed as a measure of movement intensity.

Periods of inactivity were defined as segments where both the variation in wrist angle and the ENMO are below predefined thresholds. Specifically, low movement was identified when:

$$\text{InactiveSample} = |\Delta\theta(t)| < \theta_{\text{tol}} \wedge \text{ENMO}(t) < \text{ENMO}_{\text{tol}}$$

To ensure robustness, inactivity was evaluated over a sliding window of 5 minutes and a segment was classified as inactive when at least 95% of the samples within the window satisfied the inactivity criteria. Detected inactivity onsets were then retroactively extended backward to account for the detection delay. Lastly, short inactivity periods (<30 minutes) were discarded to avoid the flagging of brief resting periods as sleep.

The result is a binary mask that identifies sleep periods (1) and wake periods (0), sampled at 1 Hz and it is not applied to the signal at this step but kept for later analyses.

In Figure 2.1 it is possible to observe an example of sleep mask detection over a 40 hrs recording, plotted above the acceleration magnitude signal.

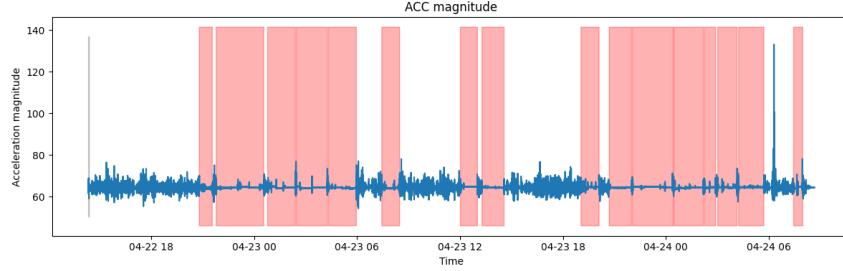


Figure 2.1: Sleep Mask over ACC magnitude plot. Where the blue signal is the ACC magnitude and the red colored parts represent sleep mask = 1

2.3.4 Window Selection

This step was used to identify recording segments suitable for feature extraction with FLIRT. Candidate windows were selected according to three criteria: **data quality**, **signal continuity**, and consistency with **behavioral state** (sleep vs wake). Because FLIRT computes statistical and temporal descriptors over fixed windows, segments containing excessive missing data or fragmented signals were excluded to improve feature reliability.

Preprocessed EDA, ACC, TEMP and IBI signals were loaded and temporally aligned. Continuous signals (EDA, ACC, TEMP) were resampled to 1 Hz, using linear interpolation only for short gaps of up to 2 seconds. Longer missing periods were preserved as missing values. This reduced computational weight while maintaining the temporal resolution required for window quality assessments.

The common valid interval across all signals and masks was identified. Fixed windows of **60 minutes duration**, a **1 minute step**, and a **60 minute stride** after acceptance were evaluated, avoiding overlap between segments. The following rules were applied to determine window quality:

Continuity Constraints For each candidate window, signal completeness was evaluated on EDA, ACC, and TEMP. Let C_s be the proportion of available samples for each signal s . Overall continuity was defined as:

$$C = \frac{C_{EDA} + C_{ACC} + C_{TEMP}}{3}$$

Windows were retained only if $C \geq 0.90$, meaning at least 90% usable data on average.

Cardiac Validity The number of valid inter-beat intervals inside each window was counted. A minimum of $N_{beats} \geq 60$ was required for a 60-minute window. This ensured at least minimal heart rate coverage and excluded windows where cardiac data were absent or heavily corrupted.

Sleep-State Consistency Using the previously generated sleep mask, each window was classified based on the fraction of sleep-labeled seconds (p_{sleep}):

- **Awake** if $p_{sleep} < 0.05$
- **Sleep** if $p_{sleep} > 0.97$
- **Mixed** otherwise (Discarded)

This avoids extracting features from windows containing transitions between states. The thresholds allow a small portion of the window to be not consistent in order to avoid discarding a selection in case of few misclassified seconds.

Quality Scoring of Valid Windows Windows passing all filters were ranked using a quality score combining **continuity**, **number of beats**, **accelerometer variability**, and **EDA variability**. The goal was to prefer moderate and informative variability while penalizing flat or erratic segments. The final score was:

$$Q = w_C S_{cont} + w_B S_{beat} + w_A S_{acc} + w_E S_{eda}$$

With weights set at $w_C = 0.30$, $w_B = 0.10$, $w_A = 0.30$, and $w_E = 0.30$. This score will be useful in the Device Compatibility Analysis to choose the best windows.

For each recording, all accepted windows' data such as start/end timestamps, continuity scores, beat counts, sleep/wake labels, and quality scores (Q) was saved in a CSV file, to be retrieved when extracting the features with FLIRT.

2.4 Feature Extraction

In order to transform continuous wearable signals into structured numerical descriptors, the FLIRT package [16] was employed. In this study FLIRT is used as both a secondary signal processing layer and a feature extraction tool, selected for its compatibility with E4 data and the ability to generate a wide array of features across the signals of interest.

2.4.1 The FLIRT Framework

FLIRT is an open source Python package designed to standardize the extraction of features from raw wearable data.

The package is designed to be robust against noise and artifacts, it incorporates signal quality checks and filters that handle the irregularities found in long real-world recordings, providing a smooth transition from raw sensor recordings to processed data.

The specific features extracted are divided into modules based on the input signal.

EDA module

To process the EDA signal FLIRT implements filtering models like the Extended Kalman Filter and the Particle Filter to get the actual skin conductance response in the presence of severe artifacts. After cleaning, the module extracts features describing the sympathetic activation of the subject, including:

- **Tonic/phasic decomposition**, to separate the slow background signal (tonic level) from the rapid responses (phasic components).
- **Statistical indicators**, capturing the general distribution of the values.
- **Peak descriptors**, specifically related to skin conductance responses, which identify the frequency and amplitude of rapid arousal peaks.
- **Temporal dynamics**, which describe how the signal changes over the given time window.

ACC module

The ACC module processes raw tri-axial movement data to capture different activity and agitation states. The extracted features include:

- **Axis and magnitude statistics**, capturing the fundamental metrics of movement across all three axis and their combined vector.
- **Movement intensity**, providing an aggregate measure of physical effort.
- **Variability**, indicating how movement patterns change within the given window.
- **Entropy and activity descriptors**, to quantify the complexity and randomness of the subject's motor activity.

HRV module

Heart Rate Variability (HRV) feature extraction from wearable devices is harder to process and to get stable features from, due to frequent measurement gaps and misplaced beats. FLIRT first applies a series of artifact rejection rules to discard implausible IBIs, then it extracts features across multiple domains:

- **Time domain:** Metrics such as RMSSD and pNN50 (percentage of successive intervals that differ by more than 50 milliseconds), which represent the variability of the IBIs over the time period.
- **Frequency domain:** Estimations of the power spectral density, including Low Frequency (LF), High Frequency (HF), and the LF/HF ratio. FLIRT uses the Lomb-Scargle periodogram method to perform spectral analysis directly on heart intervals, avoiding the artifacts introduced by methods that employ beat interpolation.

Statistical module (TEMP)

For other generic signals such as body temperature, FLIRT provides a standardized statistical module. This captures the signal’s dynamics through basic metrics. The extracted features include:

- **Mean and standard deviation (std)**, indicating the average temperature and its variability.
- **Percentiles**, representing the distribution of temperature values.
- **Minimum and maximum (min/max)**, tracking the extreme values inside the window.
- **Skewness and kurtosis**, detailing the asymmetry and the tail characteristics of the signal’s distribution.

Summary of Extracted Features

Table 2.3 represents the number of features produced by each module/custom process.

A comprehensive table of every single feature and its mathematical description can be found in **Appendix A**.

Source	Module / Category	Number of Features
FLIRT	EDA	44
FLIRT	ACC	88
FLIRT	HRV	46
FLIRT	TEMP	22
Custom	Autoregressive Dynamics (AR)	6
Custom	Circadian Encoding	2
Custom	Rolling Sleep History	4
Total		212

Table 2.3: Summary of features extracted per FLIRT module and additional descriptors over each 5-minute sliding window.

2.4.2 Feature Extraction Pipeline

To actually extract the features from the valid data segments for each signal, a custom Python pipeline was developed. This workflow ingests the masked signals, coordinates the FLIRT modules over the selected signal segments, adds custom features and aligns the outputs based on the timestamps.

Windowing and FLIRT Application

For each patient, the clean EDA, ACC, TEMP, and IBI signals were loaded into continuous time series and each valid recording segment was processed separately. The segments of each signal were processed by their respective FLIRT modules using non-overlapping sliding windows of **5 minutes**, resulting in a matrix equal to the number of 5 min segments multiplied by the number of feature for specific module, see table 2.3.

Supplementary Custom Features

Three further sets of custom features were engineered and added to the output vectors for each 5-minute window:

- **Autoregressive (AR) Dynamics:** To further describe the temporal evolution of movement, a third order autoregressive model ($AR(3)$) was calculated for the ACC magnitude and the EDA signals:

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \varepsilon_t \quad (2.1)$$

The coefficients ϕ_1, ϕ_2, ϕ_3 were used as additional descriptors of short-term temporal dependence to capture signal predictability.

- **Rolling Sleep History:** Using the sleep mask generated in the previous steps, a sleep time evaluation metric was calculated, evaluating for each window how many hours the patient was sleeping in the last 24 hours. To address the first 24 hours of the recordings (where previous sleep data is unavailable) the first hours of available sleep data were duplicated and shifted backward in time, creating a synthetic historical baseline that allows the rolling sum to be computed consistently without missing data.

Alignment and Export

To align output vectors in case of missing data or different initial timestamp of the modalities and create a unified feature matrix, an asynchronous merging strategy was developed. The timestamp index of the EDA features was used as the temporal anchor, while the ACC, TEMP, and HRV feature matrices were sequentially merged onto it using a nearest neighbor matching algorithm. To prevent aligning features from distant windows just because the nearest available, a maximum tolerance of 2 minutes was applied for ACC and TEMP and 2.5 minutes (50% of the window) for HRV, dropping the row if no match was found.

Finally, a data quality threshold was applied to the unified matrix, dropping any rows where an excessive number of features failed to compute. The resulting synchronized feature datasets were exported as csv files for each recording session, ready for further analysis and ML applications.

2.5 Device Comparability Analysis

In this part of the study Dataset 1 is used to compare E4 and EP data, in order to evaluate comparability and interchangeability of the two devices' output. 27 out of the 30 available control subjects were employed in this analysis, due to the removal of three subjects for severe signal corruption.

The analysis is structured in two different parts. First a **Raw Signal Correlation** was carried out, to quantify the temporal overlap and morphological similarity between the time series signals. Secondly a **Descriptor Consistency Analysis** was performed on features extracted from selected synchronized segments, evaluating the degree of agreement between the devices once the data is in a structured numerical descriptor format.

2.5.1 Raw Signal Correlation

Continuous Signal Alignment (EDA, ACC, TEMP)

For signals with a constant sampling rate, the objective was to quantify waveform similarity after correcting for clock desynchronization. This analysis is done on signals that have already been preprocessed in the preprocessing pipeline described in 2.3.

Clipping and Downsampling Both signals were clipped to their common temporal window and resampled to a frequency of 1 Hz to handle minor sampling variability. This process transforms the raw signals into a comparable vector format $x, y \in \mathbb{R}^n$ (in the case of ACC, acceleration magnitude is used as the main signal to compare, to account for non aligned axis).

Temporal Synchronization via Cross-Correlation To find the optimal temporal alignment between the recordings of the two devices, the cross-correlation sequence $R_{xy}[m]$ was computed using the Fast Fourier Transform:

$$R_{xy}[m] = \sum_n (x[n] - \bar{x})(y[n - m] - \bar{y}) \quad (2.2)$$

The optimal lag τ was determined by computing the sample shift m that maximizes the correlation:

$$\tau_{opt} = \arg \max_m R_{xy}[m] \quad (2.3)$$

The EP signal was then shifted by τ_{opt} seconds to ensure maximum alignment before evaluating signal comparability.

Statistical Similarity Metrics Two metrics were used to prove interchangeability: Pearson Correlation Coefficient (r), which measures the linear correlation of the morphological relationship and the Root Mean Square Error (RMSE), representing the absolute distance between the signals in their respective units (e.g., μS for EDA).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (2.4)$$

Discrete Event Alignment (IBI)

IBI signals depend on the precise timing of detected heartbeats, making it an event based signal, this requires a different comparing logic from the pipeline applied to

continuous signal.

Beat Matching Because the two devices could detect the same beat at slightly different timestamps, a temporal tolerance window was used: A beat pair (IBI_{E4}, IBI_{EP}) was considered a match only if the difference in their absolute timestamps Δt was within a relative tolerance $\Delta t \leq 0.20 \cdot IBI_{E4}$ and the difference in the IBI values themselves was within a ratio tolerance of 15%. This threshold reduces the chances of pairing physiological beats with artifacts.

Optimal Lag Search Because of its nature, the discrete signal is less susceptible to misalignment during the recording, however a small clock offset could still reduce the number of matches. Therefore IBI alignment was actuated by iteratively testing lags between -1 and +1 seconds (in 50ms increments) to compensate for this. The lag that presented the highest number of valid matches and the highest Pearson correlation was selected.

All signals were also plotted for visual comparison, the figures were realized by plotting the shifted EP signal over the E4 time series. For IBIs visual comparison a dot plot was used, each dot representing the time between two beats, see Figure 2.2 .

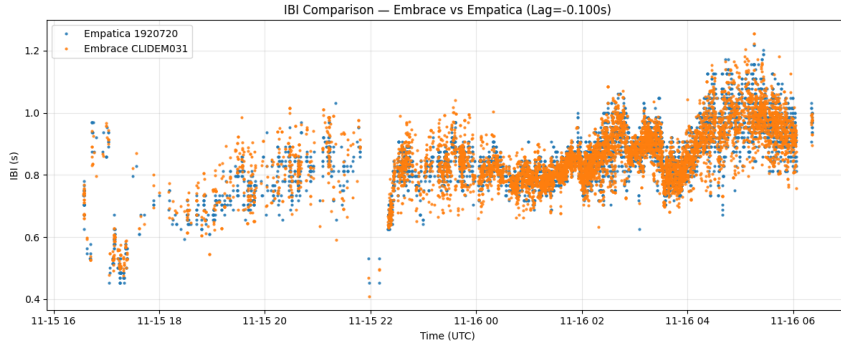


Figure 2.2: Example of plot used for visual comparison of IBI signal. The figure overlays the E4 and EP IBI signals over the entire recording, with time on the x-axis and IBI values on the y-axis.

2.5.2 Descriptor Consistency Analysis

To evaluate the comparability of the devices at the feature level, a window selection pipeline was used, in order to select a suitable window from where to extract the features. This window would then be passed on to FLIRT for feature computation and submitted to evaluation.

Cross-Device Window Selection

The window selection pipeline, previously defined in Section 2.3.4, was adapted for the comparative analysis. The concept of evaluating signal continuity, cardiac validity and overall signal quality (Q) remained the same, while the parameters were modified specifically for the task. The main differences in the selection rules were:

- **Window Dimensions:** The evaluation window was extended from 60 minutes to **120 minutes** to capture a more physiological variance and the stride was reduced to 1 minute (evaluating overlapping windows), to capture every valid window and increase the possibility of matches between the devices.
- **Continuity Constraints:** The threshold for the overall continuity score across EDA, ACC and TEMP was raised from 0.90 to 0.95, requiring at least 95% valid data.
- **Cardiac Validity:** The minimum threshold for acceptable IBIs was increased to ensure basic cardiac data coverage. A minimum average of 25 beats per minute was required, corresponding to $N_{beats} \geq 3000$ for the 120 minutes window, this conservative threshold was set to exclude implausibly low numbers of detected beats while accounting for the limitations of the E4 PPG sensor.
- **Behavioral State:** Because the objective of this specific analysis is technical sensor comparison rather than behavioral monitoring, the sleep/wake state filter was not taken into account.

Temporal Intersection and Best Segment Extraction The modified window selection algorithm was executed independently on the E4 and EP datasets for each subject, generating two separate lists of possible valid segments and their Quality Scores (Q_{e4} and Q_{ep}). To find a common timeframe, the start timestamps of the candidate windows were rounded to the nearest minute and matched by time, keeping only the 120 minute segments that satisfied the quality criteria on both devices simultaneously. Finally, the segment with the highest combined quality score was selected and the raw data corresponding to this timeframe was then clipped and exported to be fed into the FLIRT extraction pipeline.

Feature Extraction

To evaluate the comparability of the devices at the descriptor level, the previously identified, 120-minute segments were processed using the FLIRT pipeline. Features were extracted using overlapping 5-minute sliding windows, with a 1-minute stride

to maximize window matches across devices.

The integration logic remained the same as the main extraction pipeline: the timestamp index of the EDA features served as the reference, and the ACC, TEMP, and HRV feature matrices were merged using a nearest-neighbor algorithm, with a temporal tolerance to prevent misalignment caused by data gaps (2 minutes for ACC and TEMP, 2.5 minutes for HRV).

The output of this extraction was two aligned matrices (one for the E4, one for the EP), where each row represented the descriptors of a 5-minute segment.

Feature Subset While the FLIRT library generates over 200 features, conducting a comparative analysis across all of them would be heavy and redundant. A subset of 32 features was selected to perform the device comparison.

- **Temperature (TEMP) Features:** `temp_mean`, `temp_std`, `temp_pct_5`, `temp_pct_95`, `temp_ptp`

These features summarize the thermal baseline and its dispersion. Instead of minimum and maximum values, the 5th and 95th percentiles (`pct_5`, `pct_95`), combined with the peak-to-peak range (`ptp`), were chosen to provide a representation of the temperature trend more robust to outliers and artifacts.

- **ACC Features:** `l2_mean`, `l2_std`, `l2_energy`, `l2_entropy`, `l2_skewness`, `l2_pct_5`, `l2_pct_95`

The selected features aim to quantify both movement intensity and structural complexity. The Euclidean magnitude (L2) of the acceleration is represented by its central tendency, dispersion, total physical energy and entropy. These descriptors capture the overall volume and randomness of the motor activity. Additionally, the standard deviations of the individual spatial axes (`x`, `y`, `z`) were also kept to account for directional movement variance, which can differ if the devices sit differently on the wrist.

- **HRV Features:** `hrv_mean_hr`, `hrv_std_hr`, `hrv_median_nni`, `hrv_rmssd`, `hrv_sddn`, `hrv_pnni_50`, `hrv_lf`, `hrv_hf`, `hrv_lf_hf_ratio`, `num_ibis`

These metrics represent the gold standard in cardiac and autonomic nervous system monitoring. Basic descriptors establish the fundamental cardiac baseline (e.g., `mean_hr`). Time-domain HRV measures, specifically `rmssd`, `sddn`, and `pnni50`, are widely used indicators for cardiac activity and variability. Some features evaluated in the frequency domain (`lf`, `hf`, `lf_hf_ratio`) are included to capture spectral consistency. Finally, the raw count of valid intervals

(num_ibis) was selected to have an easily readable metric representing beat detection efficiency of the two different PPG sensors.

- **EDA Features:** tonic_mean, tonic_std, tonic_ptp, tonic_rms, tonic_iqr, phasic_std, phasic_energy.

This selection comprehends both the slow background state (tonic) and the rapid responses (phasic).

Statistical Comparison of Descriptors

The final step of the descriptor analysis quantifies the degree of comparability between the E4 and the EP across the 32 selected features, specifically relative consistency and absolute agreement are evaluated.

For each subject, paired feature vectors (x for EP, y for E4) were aligned by matching their 5-minute window timestamps with a 30 seconds tolerance. Segments containing missing values in at least one device due to signal loss or failure to compute were discarded for both devices; features with less paired samples than the minimum amount were also excluded from the analysis.

Relative Consistency (Spearman Correlation) Because physiological features can have not normal distributions or not linear relationships, the Spearman’s rank correlation coefficient (ρ) was chosen instead of Pearson’s r . Spearman’s ρ is a non parametric measure that evaluates the monotonic relationship between the two devices, determining if an increase in a physiological metric on the E4 corresponds to an increase on the EP, regardless of absolute scaling.

Absolute Agreement (Bland-Altman Analysis) To evaluate the absolute numerical discrepancy and systematic bias, a Bland-Altman analysis was conducted for each feature [22]. The mean difference (bias) evaluates systematic offset, while the 95% limits of agreement define the variability range of most measurements. The Mean Difference (Bias, $\bar{d}$) between the devices was calculated as:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) \quad (2.5)$$

The 95% Limits of Agreement (LOA) were computed using the standard deviation of the differences (SD_{diff}):

$$LOA_{low,high} = \bar{d} \pm 1.96 \cdot SD_{diff} \quad (2.6)$$

These limits define the range within which 95% of the differences between the devices measurements are expected. Bland Altman output is represented in a plot such as the one on the right in figure 2.3, showing the mean of the two samples along the x axis and the difference on the y .

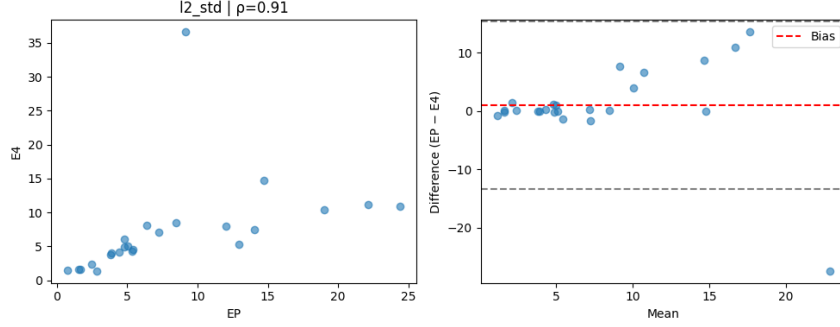


Figure 2.3: Samples and Bland-Altman representative plot (magnitude variance)

LOA Normalization and Feature Classification A LOA of 50 is negligible for heart rate variance (ms^2) but catastrophic for temperature ($^{\circ}C$). To allow for an easily readable metric of agreement, the LOA width was normalized against the range of the feature:

$$LOA_{norm} = \frac{LOA_{high} - LOA_{low}}{\max(X \cup Y) - \min(X \cup Y)} \quad (2.7)$$

Where X and Y are the complete sets of feature values from both devices. This yields a dimensionless metric bounded between 0 and 1, representing device noise as a percentage of the total observed physiological variance.

Using the combination of relative consistency and normalized agreement, each feature was classified into one of three reliability classes:

- **Strong:** $\rho \geq 0.70$ and $LOA_{norm} \leq 0.30$. (High consistency, tight agreement).
- **Moderate:** $0.50 \leq \rho \leq 0.70$ (Good consistency, potential scaling differences or wider noise bands).
- **Poor:** $\rho < 0.50$. (Unreliable translation between devices).

2.6 Bipolar Disorder Study

Data from patients affected by Bipolar Disorder (Dataset 2) was employed to identify biomarkers and developing machine learning predictive models. This analysis was executed separately on two cohorts, containing either manic or depressive episodes,

comparing their respective acute states to their remission recordings.

The complete dataset was first processed using the preprocessing and feature extraction pipelines already described in Sections 2.3 and 2.4, resulting in a matrix of 212 physiological features.

2.6.1 Dataset Construction and Normalization

To construct the predictive modeling dataset, the 5 minute FLIRT feature vectors corresponding to the acute and remission periods were loaded for each subject. A strict data density threshold was applied: in order to have enough data for each subject, patients were kept only if they possessed a minimum of 60 valid 5 minutes samples (equal to 5 hours of data) in both their acute and remission phases. For patients meeting this requirement, their sessions were concatenated into a unified dataframe and each 5-minute sample was labeled to distinguish acute (Class 1) from remission (Class 0).

Within-Subject Normalization

To isolate physiological variations due to the bipolar state rather than inter individual baseline differences, a within-subject standardization strategy was implemented. Any infinite values generated during the mathematical feature extraction phase were replaced with missing values (NaNs). Every continuous feature was then transformed into a Z-score calculated using the specific patient’s mean (μ_i) and standard deviation (σ_i) across their combined recordings:

$$z = \frac{x - \mu_i}{\sigma_i}$$

Data Cleaning and Imputation

Following normalization, a missing data filter was applied, removing any features with a missingness rate higher than 15%. For the remaining features, missing values were replaced with the median value of the descriptor.

Final Dataset Composition

Because of the application of the data density threshold (< 60 samples in either phase), several patients were dropped to ensure a minimum representation of the recording. For the depression cohort, two patients were excluded, while for mania, six patients were dropped.

The final datasets counted 18 patients for the depression cohort and 13 patients for

the mania cohort. The distribution of the 5-minute recording segments across the acute and remission states, along with the resulting class imbalance ratios for each subject in both cohorts, is in the following Tables 2.4, 2.5.

Table 2.4: Depression Cohort Sample Distribution

Patient ID	Acute	Remission	Total Samples	Imbalance Ratio
56	142	137	279	1.04
143	207	215	422	1.04
MAL049	179	190	369	1.06
112	197	183	380	1.08
99	302	280	582	1.08
MAL043	265	242	507	1.10
136	208	189	397	1.10
MAL032	223	202	425	1.10
MAL009	162	191	353	1.18
182	192	140	332	1.37
IPM009	78	119	197	1.53
MAL013	127	204	331	1.61
106	106	195	301	1.84
IPM001	81	155	236	1.91
157	89	176	265	1.98
MAL007	276	118	394	2.34
151	198	74	272	2.68
126	276	67	343	4.12

2.6.2 Biomarker Discovery

Dimensionality Reduction and Feature Pruning

Before moving to the predictive models, it was necessary to cut redundant signals from the 212 feature space in order to prevent overfitting and reduce multicollinearity. FLIRT output contains redundant descriptors that can be highly correlated with each other (e.g. the mean, median, rms, mean_nni and median_nni of heart rate).

Correlation Filtering A Spearman rank correlation matrix was computed across all extracted features to identify overlapping signals and a list of feature pairs with an absolute correlation greater than **0.90** was identified. This threshold was chosen with a conservative approach in mind, targeting dimensionality reduction rather than complete mathematical independence. In Figure 2.4 an example of correlation

Table 2.5: Mania Cohort Sample Distribution

Patient ID	Acute	Remission	Total Samples	Imbalance Ratio
82	170	164	334	1.04
79	185	170	355	1.09
40	147	131	278	1.12
MAL040	209	173	382	1.21
107	135	110	245	1.23
100	134	196	330	1.46
IPM002	152	94	246	1.62
MAL002	119	73	192	1.63
IPM012	153	84	237	1.82
110	76	173	249	2.28
MAL075	247	108	355	2.29
41	70	180	250	2.57
183	212	72	284	2.94

matrix is displayed, color coded for correlation between each pair of features.

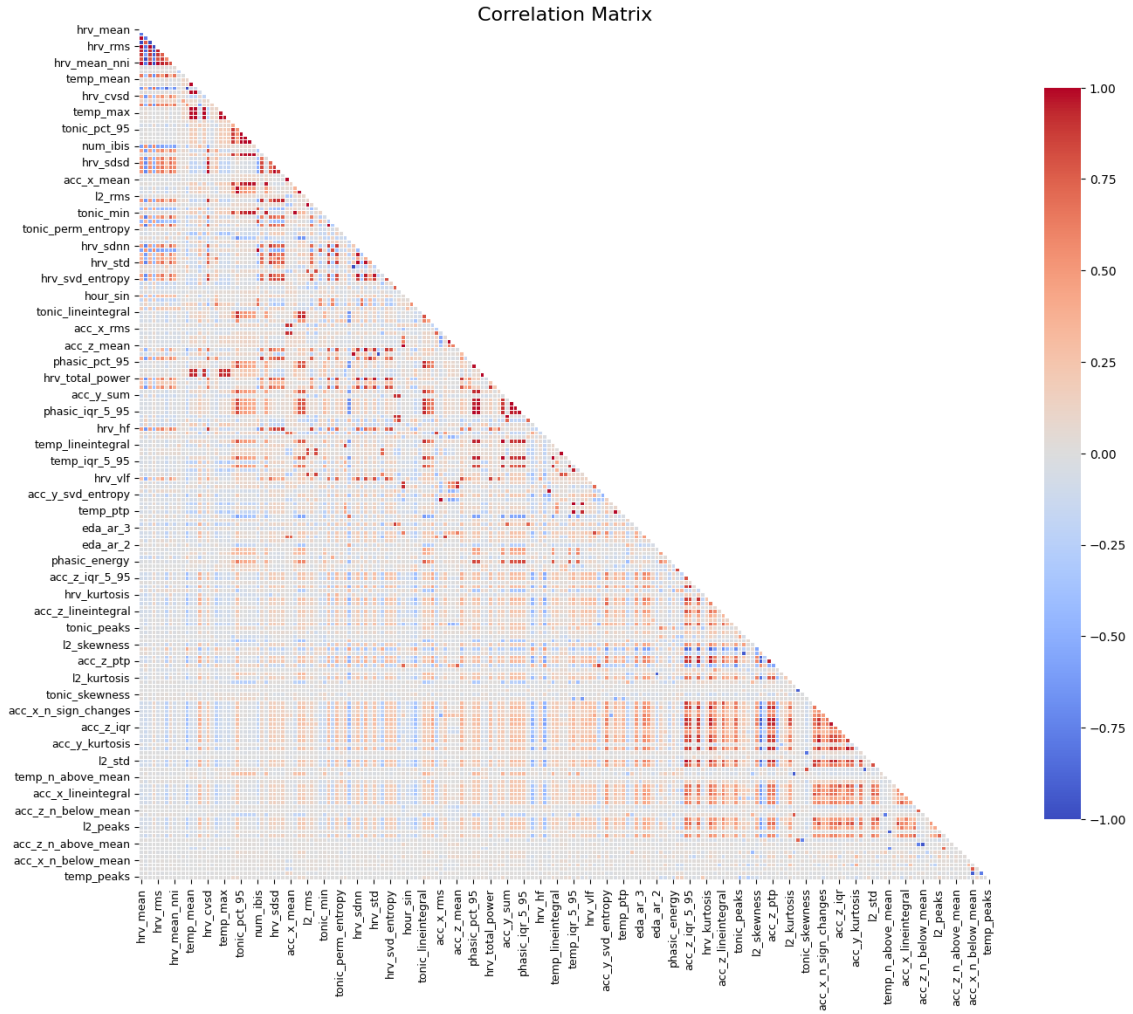


Figure 2.4: Correlation matrix of the Depression Dataset’s feature set. Heatmap with feature names on both axes, where each value (color from blue to red) represents the level of correlation between the crossing features.

When a highly correlated pair was detected, only one feature was retained, discarding the other, and this was executed iteratively until all remaining features had a correlation lower than the threshold. This process significantly reduced the dataset’s dimensionality

In the case of the Depression Dataset 75 features were dropped, bringing the number of features from 205 to 130.

While for the Mania Dataset 66 features were dropped, bringing the number of features from 203 to 137.

Biomarker Identification via Linear Mixed-Effects Models

After the dimensionality reduction process, the surviving features were subjected to statistical evaluation to identify any physiological biomarkers that could distinguish

acute phases from remission. To achieve this, a Linear Mixed-Effects Model (LMM) was implemented using the statsmodel package ([23]) in Python.

LMM Introduction Continuous wearable datasets contain hundreds of repeated measures taken from the same individuals over time, standard statistical methods operate on the assumption that all observations are independent, applying these standard methods to repeated measures violates this assumption, as samples of the same recording/patient are highly correlated.

Linear Mixed-Effects Models resolve this by dividing the variance into two components:

1. **Fixed Effects:** The global relationships between the predictors and the target variable, the universal "rule" describing how a descriptor changes between the two classes across all subjects.
2. **Random Effects:** The subject specific deviations from the population average, considering that samples within the same patient share a common behavior.

Mathematical Formulation and Z-Score Integration The standard formula of an LMM is defined as:

$$y = X\beta + Zu + \epsilon$$

Where $X\beta$ represents the fixed effects matrix and coefficients, Zu represents the random effects matrix and coefficients, and ϵ is the residual error.

In this specific pipeline, the model architecture was modified to fit with the within subject Z-scoring applied during the dataset construction. Because every patient's feature data was already standardized to have a mean of 0, fitting a traditional "random intercept" (allowing each patient to start at a different baseline) was redundant. To accommodate for this the LMM was modified to drop the random intercept and exclusively model a random slope for the clinical state, so instead of accounting for baseline differences between patients the Random Slope captures the different ways patients react to a state change. For a given physiological feature y , the model is expressed as:

$$y_{ij} = \beta_0 + \beta_1 \cdot State_{ij} + u_{1j} \cdot State_{ij} + \epsilon_{ij}$$

Where:

- y_{ij} is the normalized physiological feature for patient j at time i .

- β_0 is the fixed intercept (near zero due to the prior normalization).
- β_1 is the Fixed Effect (the global average shift caused by the acute state).
- u_{1j} is the Random Slope for patient j , capturing their specific deviation from the global average shift.
- ϵ_{ij} is the residual error.
- $State_{ij}$ is the independent variable representing the phase of the clinical condition.

Implementation and Output Interpretation For each evaluated feature, the LMM produced three critical metrics:

- Global Average Shift (β_1): because the input variable y is Z-scored, the fixed effect coefficient β_1 is adimensional and acts as a normalized effect size, quantifying the expected universal shift in the physiological feature between remission and acute phase of the episode, expressed in standard deviations.
- Global P-Value: the statistical significance of the fixed effect β_1 , measuring whether the observed global shift is reliably different from zero across the cohort.
- Individual Reaction SD ($\sqrt{\text{Var}(u_{1j})}$): as the square root of the random slope variance (σ_{u1}^2), this metric quantifies the degree of heterogeneity between patients. A high Individual Reaction SD indicates that while there may be a global trend, individual physiological responses to an episode are highly volatile and patient-specific.

By extracting these three parameters, the pipeline isolated features that not only shifted significantly at the population level during an episode but also demonstrated stability across individual patient profiles, establishing them as robust clinical biomarkers.

In Table 2.6 it is possible to observe an example of the adapted LMM output.

Visualization of Biomarkers For an easier interpretation, a custom metric was calculated from the raw variance components, the **Individuality Ratio**, defined as defined as the Individual Reaction standard deviation divided by the absolute Global Average Shift. :

$$\text{Individuality Ratio} = \sqrt{\text{Var}(u_{1j})}/\beta_1$$

Table 2.6: LMM output example

Feature	Global Average Shift	Global P-Value	Individual Reaction SD
hrv_min_hr	0.944892	9.621929e-10	0.335254
hrv_pnni_20	-1.035466	5.633752e-07	0.478565
hrv_pct_5	-0.856125	2.184099e-05	0.438594
hrv_sdn	-0.738057	5.175779e-05	0.315960
hrv_ptp	-0.646987	9.766110e-05	0.300299

This ratio is a quantification of physiological heterogeneity, based on this metric and the global p-value, each feature was classified into one of three clinical profiles:

- **Universal Biomarker:** Features with a statistically significant global shift ($p < 0.05$) and an Individuality Ratio < 1 . For these metrics, the universal response to an episode outweighs inter-patient variability, making them suitable to be biomarkers at the global level.
- **Personalized (High Variance):** Features with an Individuality Ratio ≥ 1 and an Individual Reaction SD > 0.2 . Even if a global trend exists, the individual variance is so bigger then that, signaling that the feature's behavior is more individual then global. The threshold on Individual Reaction leaves out of this category those features with negligible change between the two states. These features could be useful for personalized machine learning models or subanalyses.
- **Non-Predictive Noise:** Features that are not significantly impacted by BD episodes both on a global scale and at a subject specific level.

To visually synthesize these results, a **Volcano Plot** mapping the normalized Effect Size against the $-\log_{10}(\text{P-Value})$ was produced. This visualization highlights features that are both highly significant and possess a large magnitude of change, differentiated by color based on their respective clinical profile.

Sub-Phenotype Discovery via Cluster Analysis

For features identified in the "Personalized" zone of the biomarker landscape, where the Individual Reaction SD was high relative to the Global Average, a secondary sub-analysis was performed to identify possible biological phenotypes. The goal of this stage was to determine if the high variance was driven by distinct subgroups of patients having opposing physiological responses to an episode.

To isolate individual physiological trajectories, a Linear Mixed-Effects Model was

fitted to the raw (non-normalized) feature data, raw data retains a lot more information on the differences between patients. For each patient j , the Total Reaction Slope (S_j) was calculated by summing the fixed effect (β_1) and the random slope deviation (u_{1j}):

$$S_j = \beta_1 + u_{1j}$$

This value represents the direction and magnitude of the response for a specific individual when going from a state of remission to an acute episode.

To objectively group these trajectories, a K-Means clustering algorithm ($k = 2$) was applied to the vector of individual reaction slopes. This mathematical partition allows for the separation of the cohort into two hypothesized phenotypes:

- "Increasers": Patients who increase during an acute episode.
- "Droppers": Patients who decrease during an acute episode.

By re-running the feature analysis on these discovered clusters independently, the pipeline can identify biomarkers that are highly predictive for specific patient subtypes, which would otherwise be masked in a whole-cohort analysis. The eventual sub cohorts generated by the subanalysis will also be analyzed to look for a metadata correlation that could explain the different phenotypical response to the condition.

2.6.3 Predictive Modeling Pipeline

The final section of the methodology is the predictive pipeline designed to classify clinical states and assess the accuracy of those classifications at both the sample and session levels. The architecture utilizes ensemble learning and subject-independent validation to ensure that the findings are generalizable to a broader clinical population.

Feature set and Dimensionality Reduction

A combination of different feature sets was evaluated during model training in order to assess the impact of feature selection and dimensionality reduction on classification performance. Three configurations were considered:

- **Full FLIRT Feature Set:** the complete set of features extracted by the FLIRT library. This configuration maximizes the amount of physiological information available to the classifier, including redundant or correlated variables.

- **Uncorrelated Feature Set:** a reduced subset of features obtained after the pruning described in Section 2.6.2. Highly correlated variables were removed to reduce redundancy and limit overfitting.
- **Universal Biomarkers Set:** a compact subset of features identified through the LMM analysis as statistically significant across subjects and conditions.

The comparison evaluated of the tradeoff between amount of data, interpretability, and classification performance.

Machine Learning Framework: Random Forest ([24])

The core estimator chosen for this study is a **Random Forest (RF)** classifier. RF is an ensemble learning method based on the combination of multiple decision trees, where the final prediction is obtained through majority voting across all trees.

Each decision tree is independently trained using a bootstrap sample of the training set, meaning that for every tree a new dataset is generated by randomly substituting samples with replacements from the original dataset. As a consequence, some samples may appear multiple times while others may not be included at all. This process introduces variability between trees and increasing the trees reduces the risk of overfitting to specific observations, for this reason a forest of 500 trees was employed in all the models used in this study.

A decision tree is composed of a sequence of binary decision rules. At each internal node, the algorithm selects a feature and a threshold value that best separate the classes according to an impurity criterion, commonly the Gini impurity. The data is recursively partitioned at the nodes until a termination rule is reached, after which the terminal nodes assign the final class prediction. A key component of RF is that, in addition to bootstrap sampling, each split in a tree is computed using only a random subset of the available features rather than the entire feature space. In this work, the parameter `max_features="sqrt"` was chosen, meaning that each split considered only $\sqrt{tot_features}$ randomly selected features. This mechanism decorrelates the trees, preventing the ensemble from repeatedly relying on the same dominant predictors and improving generalization performance.

RF was selected due to its inherent versatility and suitability for physiological wearable data:

- **Non-linear Relationships:** Bipolar symptom manifestations in physiological signals often involve complex interactions (e.g., the relationship between skin conductance and heart rate variability) that are non-linear. RF captures these interactions without requiring manual feature transformations.

- **High-Dimensional Robustness:** The extracted physiological feature set is high-dimensional and contains partially redundant information. By randomly selecting only a subset of features at each split, RF reduces overfitting and avoids excessive dependence on individual variables.
- **Resistance to Noise and Outliers:** Physiological signals collected from wearable devices are frequently affected by motion artifacts, missing values, and inter-subject variability. Since the final prediction results from the aggregation of many trees trained on different subsets of data and features, RF is relatively robust against noisy samples and unstable predictors.
- **Subject Generalization:** To reduce even more the risk of the model memorizing physiological patterns of single patients, a minimum of five samples per leaf (`min_samples_leaf=5`) was set. This constraint makes the model learn patterns shared across multiple observations rather than isolated individual cases, resulting in a more generalized learning.

Validation Strategy: Leave One Subject Out

To test the clinical usability of the model, a **Leave One Subject Out (LOSO)** cross-validation strategy was implemented. In this approach, for each iteration the model is trained on the data of all but one patient, and evaluation is performed exclusively on the subject left out. This process is repeated until every subject has been used once as the test set, resulting in a number of folds equal to the number of subjects.

This protocol is good practice for validation of models based on wearable physiological data because it prevents subject leakage between training and testing sets. Since physiological patterns can contain strong subject specific characteristics, random splitting would inflate performance by allowing the model to learn individual dynamics instead of clinically relevant patterns. Therefore LOSO provides a more realistic estimate of how the model would generalize to an unseen patient in a clinical setting.

Dual Layer Sample Weighting

Given the longitudinal nature of the data the dataset is imbalanced regarding both the number of samples per patient and the duration of clinical phases (Acute vs. Remission). To prevent the model from biasing toward more active patients or more frequent clinical states, a **combined weighting strategy** was applied during the fitting process:

1. **Subject Weighting:** Scales the importance of each sample so that every patient contributes equally to the model’s learning, regardless of how many days of data they provided.
2. **Class Weighting:** Accounts for imbalances between the Acute and Remission labels.

Tiered Classification and Threshold Optimization

The pipeline produces both a classification at the sample level and at the session level, providing two different evaluation methods.

Sample-Level Classification

The model first generates a probability $p \in [0, 1]$ for every 5-minute window. Rather than using an arbitrary 0.5 cutoff, the optimal threshold is determined dynamically for each fold using Youden’s J statistic ($J = \text{Sensitivity} + \text{Specificity} - 1$) calculated on the training set. This identifies the probability threshold that best balances the detection of acute episodes avoiding false positives.

Session-Level Recognition

To determine if a clinical session (the entire recorded period of an Acute or Remission state) has been successfully "recognized", an additional evaluative layer was implemented. A session is marked as "Recognized" only if the percentage of its samples correctly predicted exceeds a specific recognition threshold.

For this secondary threshold an array of four distinct levels were tested: **0.5, 0.55, 0.60 and 0.65**. By evaluating the amount of correct predictions required to identify a clinical state, the framework provides a more realistic approximation of how such a system could operate in a clinical monitoring scenario, where robustness and temporal consistency are essential.

Prediction Visualization

To provide a transparent view of the model’s behavior, a custom visualization framework was developed to plot the prediction trajectory for each test patient. These plots integrate:

- **Confidence Mapping:** Scatter points where the Y-axis is the probability of an acute state, and the size of the point is scaled by model confidence ($20 + [\text{confidence} \times 100]$).
- **Success Visualization:** Green points indicate correct classification, while red indicates errors relative to the clinical ground truth.

- **Contextual Shading:** Night-time periods (22:00 - 08:00) are shaded to observe model stability during circadian transitions. Considering that the samples are all from detected sleep periods, highlighting the night regions is helpful to distinguish predictions on night-sleep or day-sleep.
- **Integrated Metadata:** Each timeline is annotated with the patient’s specific HDRS/YMRS scores and current medications, facilitating a qualitative assessment on how pharmacological treatments or symptom severity impact the ability of the model to predict the correct class.

Evaluation Metrics

To quantify the model’s predictive power across both clinical states, different metrics focused on class-balanced performance were employed.

1. **Balanced Accuracy:** Standard accuracy is often misleading in datasets with unbalanced classes. Therefore, Balanced Accuracy was used, defined as the arithmetic mean of Sensitivity and Specificity:

$$\text{Balanced Accuracy} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

Where TP is defined as True Positive, TN is True Negative, FN is False Negative and FP is False Positive

2. **Macro F1-score:** The F1-score represents the harmonic mean of Precision and Recall (*Sensitivity*). The Macro F1-score was calculated by taking the unweighted average of the F1-scores for both the acute and remission classes:

$$\text{Precision} = \frac{TP}{TP + FP} \quad , \quad F1_{class} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F1_{macro} = \frac{F1_{Acute} + F1_{Remission}}{2}$$

A high Macro F1-score therefore indicates that the model maintains a balanced trade-off between false positives and false negatives across both acute and remission conditions.

3. **ROC-AUC:** The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was used to measure the model’s discriminative capability across all possible thresholds. This provides a single scalar value where 1.0 represents perfect separation and 0.5 represents a model no better than a random coin flip.

ROC-AUC quantifies the probability that the classifier assigns a higher prediction score to a randomly selected positive sample than to a randomly selected negative

sample. Therefore the metric evaluates how well the model separates the two classes independently of the final decision boundary.

In the context of wearable physiological data, ROC-AUC is particularly informative because it evaluates the ranking quality of the probabilistic outputs even when the optimal operating threshold may vary depending on the intended clinical application.

Chapter 3

Results and Discussion

3.1 Device Comparability Analysis

3.1.1 Raw Signal Correlation Analysis

Table 3.1 presents the summary of agreement metrics across all sensor modalities. The comparability of each modality on the 48 hour recordings will be examined in depth in the following paragraphs. For each type of signal a table and figures were

Table 3.1: Summary statistics of signal comparability metrics.

Metric	EDA	ACC	IBI	TEMP
Correlation Mean	0.636	0.532	0.859	0.796
Correlation Median	0.693	0.594	0.870	0.867
Correlation SD	0.187	0.189	0.074	0.190
Lag Mean	-42	-123	0	-95
Lag Median	-61	-123	0	-91
Lag SD	168	62	0	164
RMSE Mean	3.161	0.453	0.059	1.906
RMSE Median	2.368	0.433	0.061	1.792
RMSE SD	2.629	0.115	0.012	0.813

produced. The tables describe the behavior of the sensors when compared across the two devices in the matter of: optimal lag found, Pearson’s correlation and RMSE. The figures represent the visual overlay of the two recordings, clipped to a common timeframe and shifted by the best lag.

EDA As shown in Table 3.2, the analysis of the raw EDA time series revealed moderate and highly variable correlation between the two devices, with a mean

correlation of 0.64 and a median correlation of 0.69 ($SD = 0.19$). Average RMSE is massive at $3.16\mu S$ showing incompatibility in absolute values. These raw amplitude differences are to be expected considering the hardware differences between the two devices. Sensor gains, positioning and electrode geometry differences have a direct impact on absolute EDA value and in EP they were redesigned to render the readings more accurate and responsive.

Table 3.2: Descriptive statistics for EDA correlation, RMSE and lag

Metric	Value
Mean Correlation	0.64
Median Correlation	0.69
SD Correlation	0.19
Mean RMSE(μS)	3.16
Mean Lag (s)	-42

The distribution of EDA correlations across all recordings is shown in Figure 3.1. Most recordings clustered between correlation values of 0.60 and 0.90, while a smaller subset exhibited substantially weaker agreement.

Overall, more than 70% of the recordings reached a correlation coefficient bigger than 0.60, indicating consistent morphological similarity between the EDA signals in the majority of subjects.

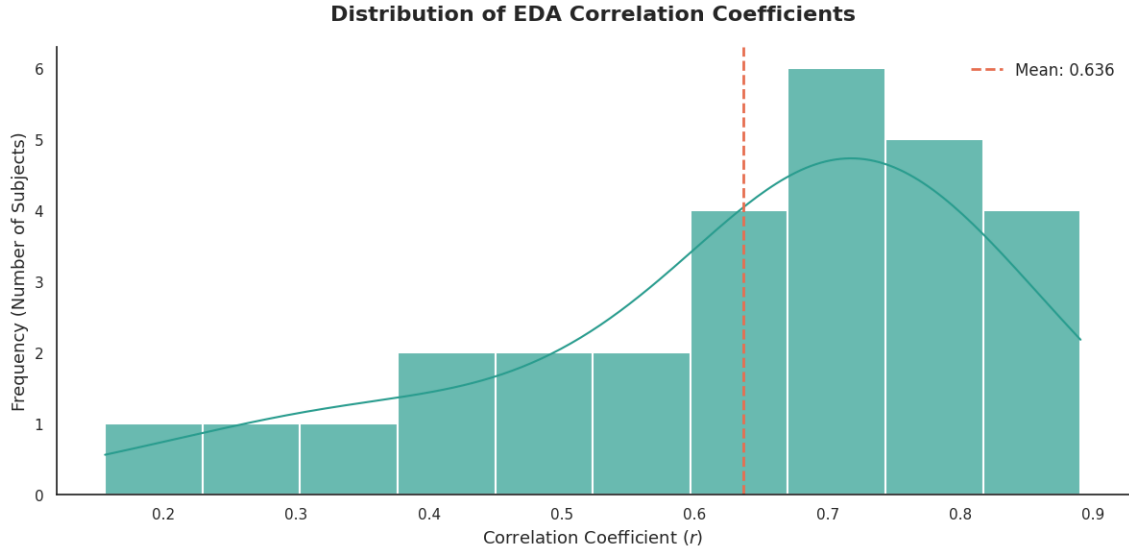


Figure 3.1: Distribution of EDA Pearson correlations across subjects

ACC As shown in Table 3.3, the accelerometer magnitude signals demonstrated moderate agreement after temporal synchronization, with a mean correlation of 0.53

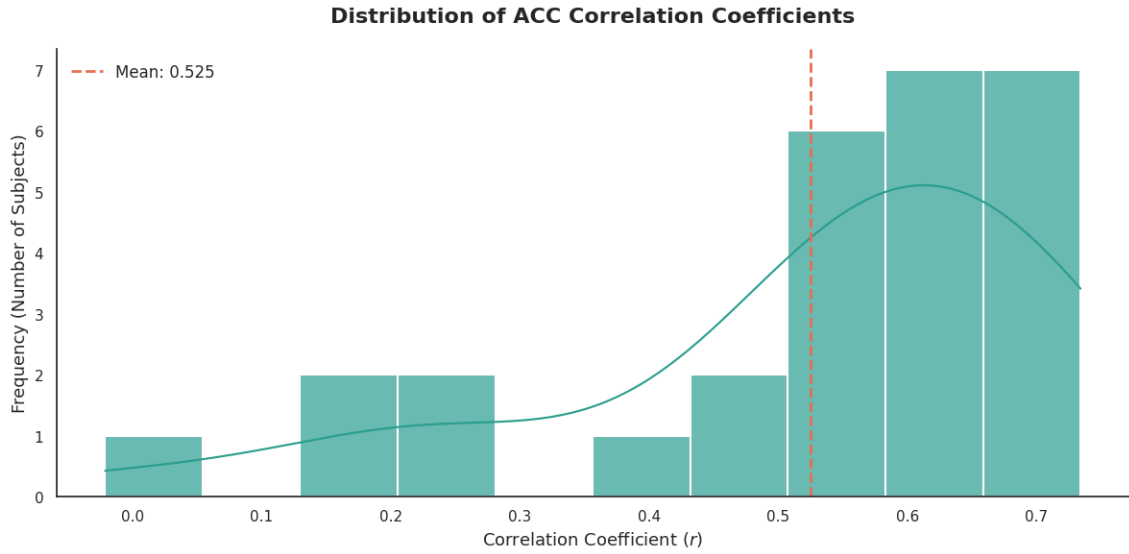


Figure 3.2: Distribution of ACC Pearson correlation across subjects

and a standard deviation of 0.19 (median = 0.59). The average lag identified through cross-correlation optimization was -123.57 seconds, indicating a substantial offset between the two recordings before alignment.

Table 3.3: Descriptive statistics for ACC correlation, RMSE and lag

Metric	Value
Mean Correlation	0.53
Median Correlation	0.59
SD Correlation	0.19
Mean RMSE(g)	0.44
Mean Lag (s)	-124

Figure 3.2 shows the distribution of correlation coefficients obtained across all recordings. Overall, only 43% of the recordings achieved a correlation coefficient greater than 0.60, with a 22% of subjects falling in the 0.5 to 0.6 range. This result is the lowest between the four modalities, mainly due to the highly volatile nature of the signal. Unlike the slow moving EDA and TEMP signals, ACC is a less forgiving signal that requires perfect temporal alignment and sensor placement conditions to be valuated accurately. Since accelerometers are highly sensitive to sensor orientation and placement, small differences in wrist position, device rotation or strap tension may produce different signals despite being the recording of the same arm movement. The different positioning of the two bracelets (E4 proximal, EP distal) also introduces slight differences between the two signals, much more in ACC than in any other signal. These variables could also explain the high variability of

ACC agreement between subject, as is can depend heavily on the bracelets’ relative position. As a result, lower cross-device correlation may partly reflect biomechanical rather than technological differences.

TEMP As shown in Table 3.4, the temperature signals present the strongest agreement among the continuous sensors, with a mean correlation of 0.80 and a median of 0.87 ($SD = 0.19$). The average temporal offset evaluated through cross-correlation alignment was -85.89 seconds, indicating a relevant lag between the two devices, consistently with the findings in ACC and EDA signals. The high RMSE can be explained by the position of the bracelet on the body, as observable in the overlay plots the E4 signal is persistently higher. This is probably because the skin has a lower temperature in distal positions, causing the E4 (proximal) to be warmer than EP. Further differences in amplitude can be explained by the hardware changes between the two devices, in this case the E4’s termistor was replaced with a digital temperature sensor so accuracy, sensitivity and response time are expected to differ. The distribution of temperature correlations across all recordings is shown in

Table 3.4: Descriptive statistics for TEMP correlation, RMSE, and lag

Metric	Value
Mean Correlation	0.80
Median Correlation	0.87
SD Correlation	0.19
Mean RMSE($^{\circ}$ C)	1.90
Mean Lag (s)	-86

Figure 3.3. Most recordings clustered at high correlation values, with the 71% of subjects exhibiting coefficients above 0.80 indicating strong morphological similarity between the synchronized temperature signals. Only a limited number of recordings presented substantially weaker agreement, possibly due to device placement or other obstacles that prevented the correct detection of the signal.

It is important to noticing that the lag estimates exhibited substantial variability, often differing considerably between modalities of the same recording. This observation suggests that the offsets calculated do not reflect an actual temporal displacement between devices, as a genuine clock offset should affect all modalities similarly. Rather, they likely derive from the cross-correlation optimization and signal characteristics , including differences in filtering, smoothing, sampling rate variations and sensor response dynamics. As a result, the lag values should be interpreted as alignment parameters that maximize signal similarity rather than direct

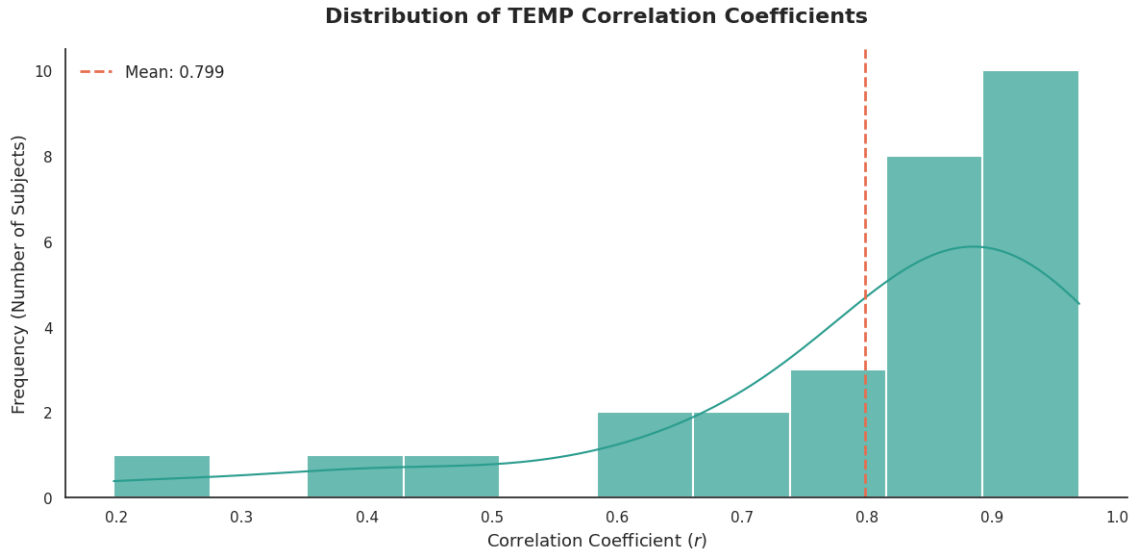


Figure 3.3: Distribution of TEMP Pearson correlation across subjects

estimates of timestamp error between the devices.

IBI As shown in Table 3.5, the IBI signals demonstrated high agreement between the two devices, with a mean correlation of 0.86 and a standard deviation $SD = 0.07$.

Table 3.5: Descriptive statistics for IBI correlation, RMSE, and lag

Metric	Value
Mean Correlation	0.86
Median Correlation	0.87
SD Correlation	0.07
Mean RMSE	0.06
Mean Lag (s)	0.1

The distribution of IBI correlations across all recordings is shown in Figure 3.4. Most recordings exhibited consistently high agreement, with correlation coefficients concentrated between 0.80 and 0.95. Only a small number of subjects presented moderate reductions in similarity. Overall, approximately 79% of the recordings achieved a correlation coefficient greater than 0.80, indicating very strong agreement between the beat-to-beat interval sequences extracted from the two devices.

Representative Signal Overlays

To visually evaluate the effectiveness of the synchronization pipeline and the degree of agreement, representative aligned signal overlays are reported for two subjects

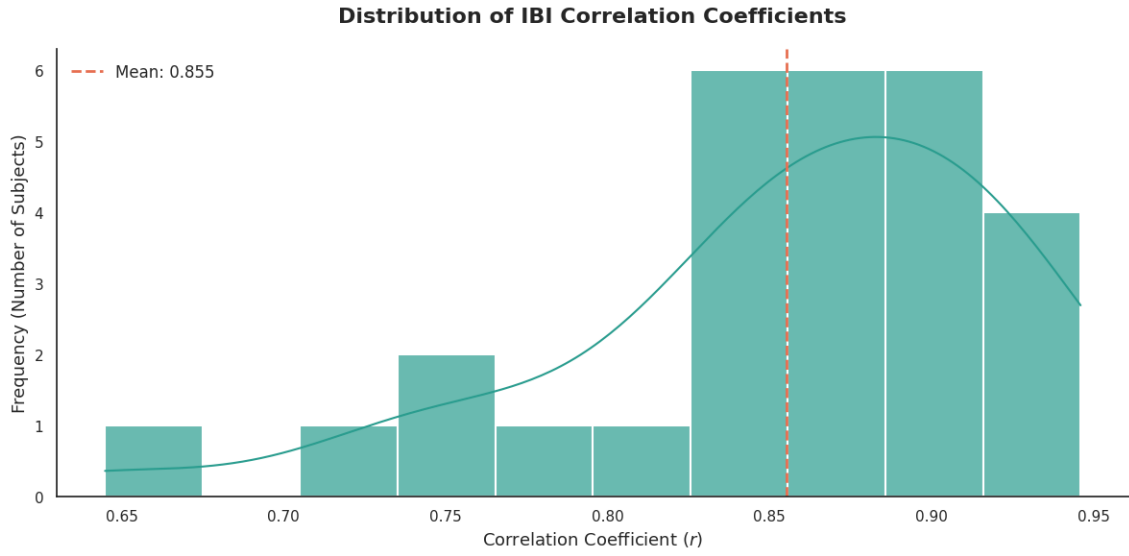


Figure 3.4: Distribution of IBI Pearson correlation across subjects

that exhibit different levels of similarity. For each subject, separate visualizations are shown for EDA, ACC, TEMP, and IBI signals after temporal alignment.

High Agreement Example: CLIDEM15 Figure 3.5 shows the overlaid recordings of the two devices for a subject characterized by high agreement across all modalities.

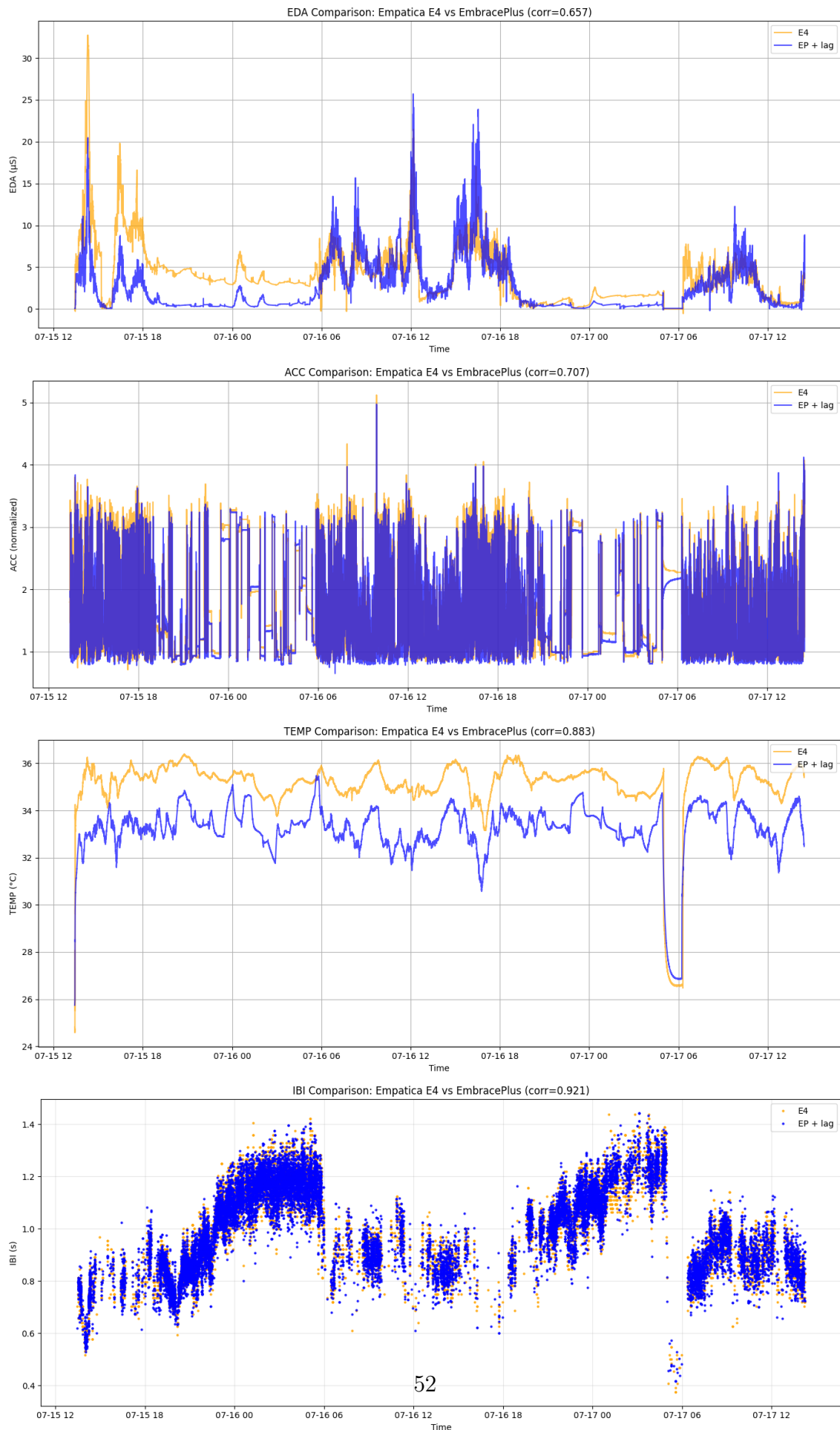


Figure 3.5: EDA, ACC, TEMP and IBI overlays for a representative high agreement subject

Lower Agreement Example: CLIDEM08 Figure 3.6 present recordings characterized by weaker agreement between the two devices. Noticeable inconsistencies in signal morphology and amplitude are visible in multiple modalities, specifically very low correlations for EDA and ACC signals, IBI and TEMP proved more

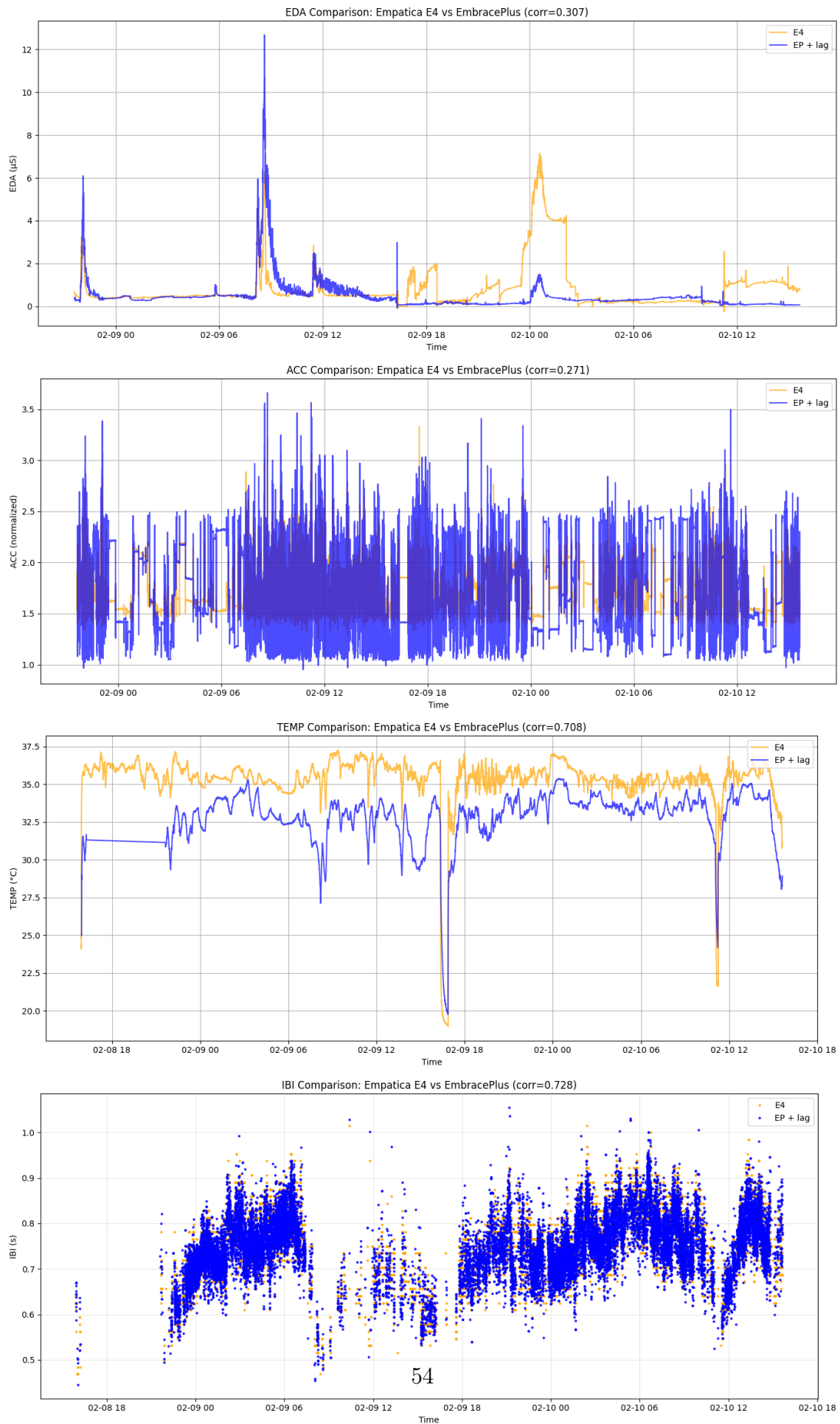


Figure 3.6: EDA, ACC, TEMP and IBI overlays for a representative low agreement subject

The visual inspection of the overlaid signals was consistent with the quantitative findings of the previous sections. Subjects characterized by high correlation coefficients showed strong morphological , while recordings with lower statistical agreement showed visible differences. The disagreement in absolute values is instead noticeable in both subjects, especially in EDA and TEMP signals, confirming that it is a consistent behavior, likely caused by hardware and sensor positioning differences between devices.

The complete results at the patient level, including individual lag values, correlation coefficients and RMSE metrics for all modalities, are reported in Appendix B.1. In the complete table it is possible to appreciate the heterogeneity of the cohort, with some patients such as CLIDEM09, presenting very low agreement all over the four modalities, indicating bracelets misplacement or other severe failures of the experimental setup that could invalidate the analysis for that specific patient. These patients were kept because of the impossibility to prove such claims.

3.1.2 Sleep-Period Comparability Analysis

To investigate if periods characterized by reduced movement and stable conditions improve agreement, the comparability analysis was repeated considering only the sleep periods of each recording, utilizing the sleep_mask created before. Reducing the influence of motion artifacts, changes in device position and other external factors that may affect signal quality, sleep periods are expected to provide better conditions for stable recordings.

The aggregated results are reported in Table 3.6. The analysis restricted to sleep time confirmed the theorized improvement in device agreement for most modalities compared to the full-recording analysis. Correlation values increased, indicating that the two devices captured more similar physiological dynamics during sleep. This result is consistent with the more stable physiological parameters and the lower amount of artifacts caused by movements.

Some subjects still exhibited low correlation values across one or more modalities. These recordings generally correspond to subjects that already presented poor agreement in the previous analysis or patients with not enough consequent sleep time to be properly analyzed (e.g. CLIDEM13, CLIDEM06).

The lag estimates remained inconsistent across modalities, similarly to the observations made in the full-recording analysis. This supports the interpretation that the lags are to be considered more as alignment parameters rather than estimates of timestamp differences between the devices.

In order to compare heart beat detection performance across devices, the propor-

Table 3.6: Summary statistics of signal comparability metrics, restricted to sleep periods.

Metric	EDA	ACC	IBI	TEMP
Correlation Mean	0.687	0.651	0.700	0.603
Correlation Median	0.786	0.845	0.688	0.641
Correlation SD	0.327	0.461	0.115	0.257
Lag Mean	-55.607	-46.107	-0.034	-33.571
Lag SD	173.081	178.161	0.690	85.301
RMSE Mean	3.176	0.366	0.058	1.368
RMSE Median	1.444	0.285	0.061	1.106
RMSE SD	4.146	0.227	0.014	1.524

tion of IBI detected by E4 compared to EP was computed for all subjects. For the analysis restricted to sleep periods, E4 detected on average 76% (median 84%) of the IBI samples detected by EP. As shown in Figure 3.7, this is a massive improvement compared to the full recordings, where IBI detection rates of E4 compared to EP were on average 37% (median 34%). The lower overall performance of E4 beat detection is to be expected considering the inferior PPG sensor. This results suggests that E4 achieves a more reliable heart beat detection during sleep, due to less movement and improved skin contact.

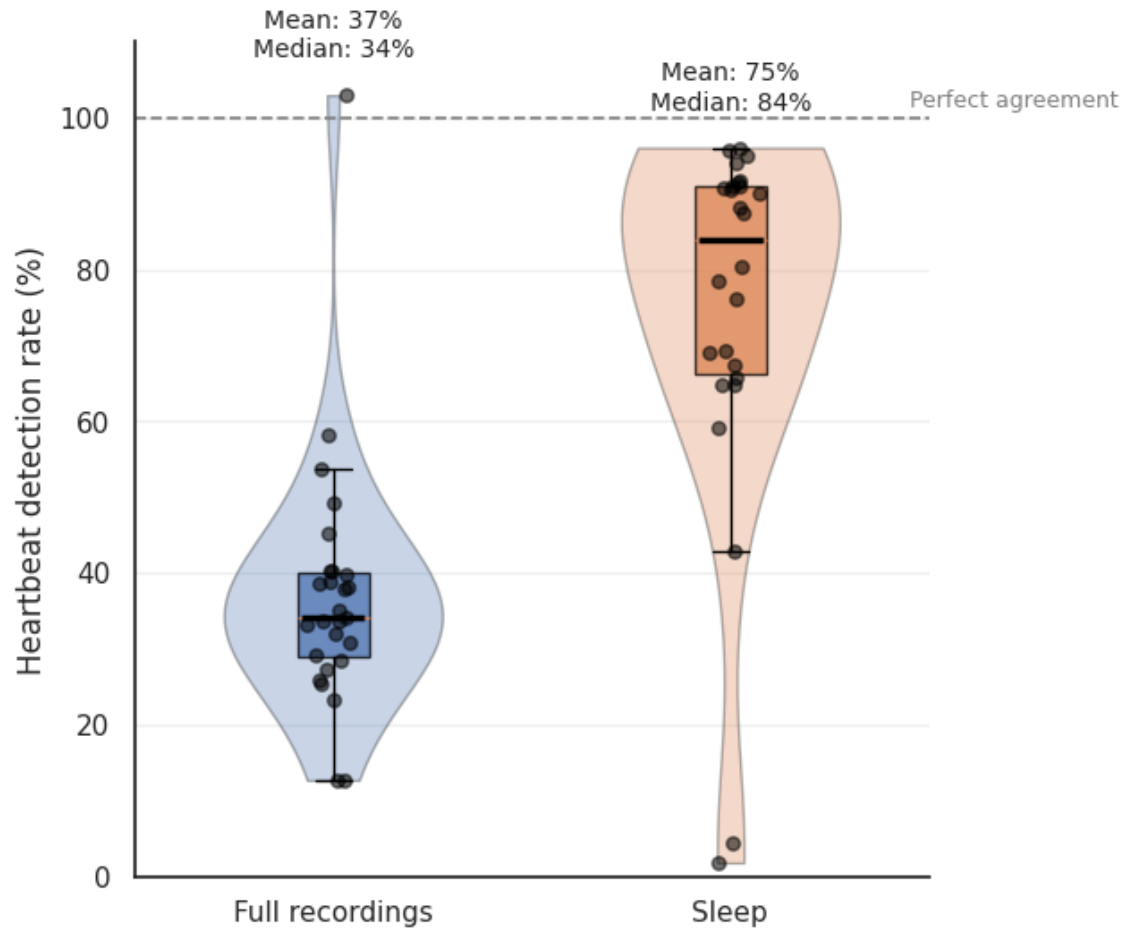


Figure 3.7: Distribution of heartbeat detection for E4 relative to EP during the full recordings and when restricting the analysis to sleep periods. Heartbeat detection rate is expressed as the percentage of IBI samples detected by the Empatica E4 with respect to the EmbracePlus for each participant. Violin plots represent the distribution across participants, boxplots indicate the interquartile range and median, individual points correspond to single participants. The dashed horizontal line indicates perfect agreement.

Overall the results indicate that recordings restricted to sleep time are better suited for device comparability. The signals recorded during sleep time have a higher agreement, lower variability and much more consistent IBI detection than the full 48h recordings. These findings show that the two devices are more comparable when monitoring sleeping subjects rather than during the day, and allow to employ HRV features from E4 that would be too unreliable to use when the patient is awake.

3.1.3 Descriptor Consistency Analysis

Feature Availability

The amount of missing windows in the extracted feature was evaluated, showing significant signal loss in HRV feature windows. FLIRT extracts the HRV descriptors based on the IBI detections, if there are not enough heart beats detected in the 5 minute window the algorithm is unable to return a value for that specific time span. The percentage of missing windows across patients for both devices was grouped in Figure 3.8.

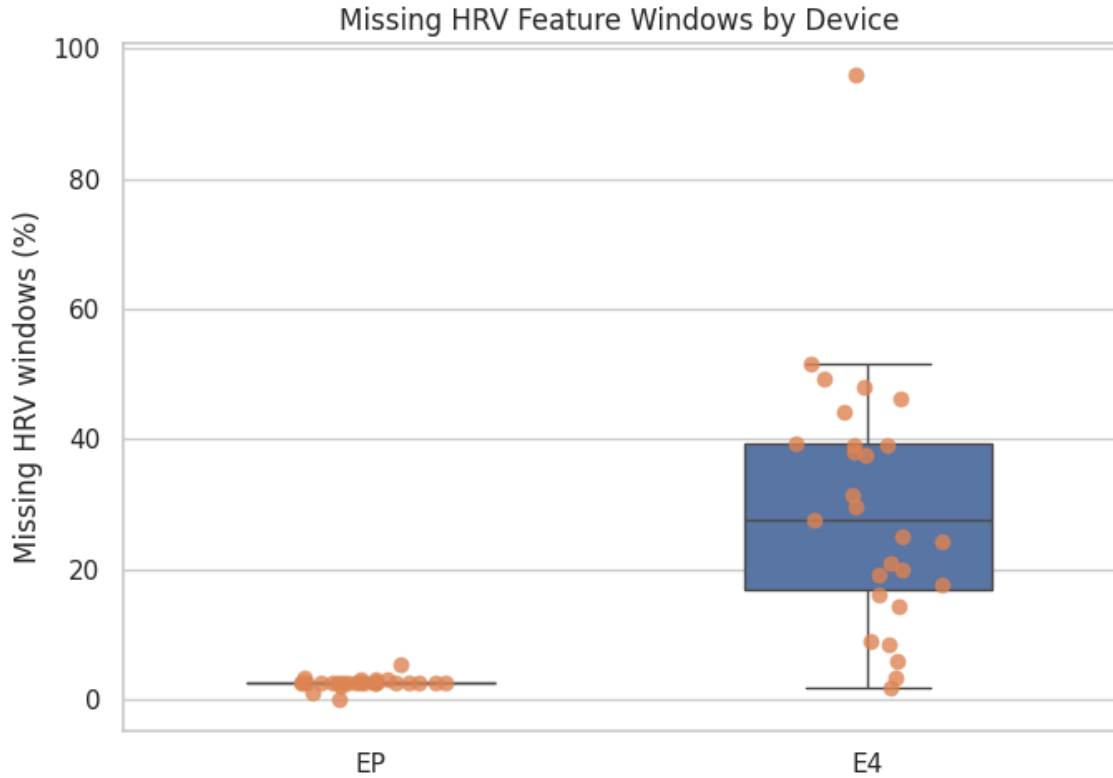


Figure 3.8: Percentage of missing HRV feature windows across subjects for EP and E4.

EP shows stable behavior with very low signal loss, consistent at around 2-3% of total windows. In contrast, E4 shows heavy and erratic window loss, with a median of 27.5% missing rows and several recordings exceeding 40% invalid HRV windows. This behavior is a direct consequence of the high beat missingness rate found in the previous section and is easily explained by the improvements made to EP's PPG sensor. Because of its four multi-wavelength channels, the next gen device demonstrates significantly higher beat detection sensitivity than the single channel mounted on the E4.

Global Feature Comparability

To evaluate the descriptor agreement between E4 and EP data the selected features were extracted from the 2 hour windows. For each patient Spearman's correlation ρ , Bland-Altman bias and normalized Limits of Agreement (LOA_{norm}) were computed.

The resulting metrics provide an assessment of:

- **Relative consistency** (Spearman), describing whether both devices register the same physiological trends in time.
- **Absolute agreement** (Bland-Altman), evaluating whether the extracted feature values are numerically similar.

Table 3.7 summarizes the global descriptor comparability results for all evaluated features.

Table 3.7: Global descriptor comparability results between devices. The table reports the chosen metrics for each feature, along with an associated tier that classifies the agreement between devices in poor, moderate and strong.

Feature	Mean ρ	Median ρ	SD ρ	Median Bias	Median LOA_{norm}	Tier
HRV Features						
hrv_hf	0.392	0.414	0.277	954.98	0.618	Poor
hrv_lf	0.399	0.405	0.238	1038.51	0.629	Poor
hrv_mean_hr	0.619	0.810	0.435	0.49	0.473	Moderate
hrv_median_nni	0.638	0.818	0.431	-7.24	0.453	Moderate
hrv_pnni_50	0.520	0.515	0.250	12.26	0.688	Moderate
hrv_rmssd	0.432	0.505	0.280	39.67	0.743	Moderate
hrv_sdn	0.399	0.430	0.261	36.80	0.781	Poor
hrv_std_hr	0.441	0.510	0.314	3.22	0.751	Moderate
num_ibis	0.553	0.667	0.370	135.47	0.692	Moderate
ACC Features						
12_energy	0.566	0.690	0.317	1.37e6	0.247	Moderate
12_entropy	0.675	0.679	0.227	0.00004	0.121	Moderate
12_mean	0.461	0.606	0.367	1.31	0.410	Moderate
12_skewness	0.480	0.498	0.261	-0.06	0.600	Poor
12_std	0.758	0.799	0.292	0.07	0.477	Moderate
EDA Features						
phasic_energy	0.521	0.691	0.407	-9.98	0.761	Moderate
phasic_std	0.501	0.676	0.407	-0.02	0.807	Moderate
tonic_iqr	0.524	0.640	0.319	-0.07	0.624	Moderate
tonic_mean	0.612	0.725	0.307	-0.54	0.657	Moderate
tonic_rms	0.621	0.731	0.304	-0.54	0.675	Moderate
tonic_std	0.538	0.694	0.370	-0.05	0.719	Moderate
TEMP Features						
temp_mean	0.545	0.632	0.411	-1.17	0.611	Moderate
temp_pct_5	0.553	0.646	0.408	-1.22	0.618	Moderate
temp_pct_95	0.527	0.624	0.399	-1.44	0.622	Moderate
temp_ptp	0.346	0.418	0.346	-0.08	0.801	Poor
temp_std	0.311	0.338	0.337	-0.01	0.810	Poor

The analysis revealed moderate agreement between the two wearable generations, with substantial variability across modalities and feature classes. Contrarily to the time series analysis, ACC descriptors showed the strongest consistency (**12_std** and **12_entropy**), showing that global movement intensity and activity complexity are stable across devices, despite the low agreement when comparing the whole signals. While fine differences at the single-value level are lost, overall movement energy and activity present very high agreement, making the devices highly comparable in a mood or stress detection context.

On the contrary, HRV descriptors (**hrv_lf** and **hrv_hf**) analyzing the frequency domain demonstrated poor comparability, characterized by low correlation and wide limits of agreement. Cardiac metrics in the time domain such as **hrv_mean_hr** and

`hrv_median_nni` achieved substantially stronger consistency, although their high standard deviations indicate significant variability among subjects. This outcome is likely caused by the missingness rate of the E4 that skews the results and produces unreliable readings, especially in periods characterized by high movement.

EDA descriptors showed intermediate behavior, with tonic features showing slightly better performance than phasic components. Similarly TEMP descriptors (`temp_mean`, `temp_pct_5` and `temp_pct_95`) demonstrated moderate consistency characterized by high correlation and wide LOAs, confirming the findings of the raw signal comparison section that indicated coherent trends but different absolute values.

No descriptor satisfied the predefined criteria for the "Strong" reliability tier. This tier was purposely set with very high thresholds, to indicate features that match almost perfectly across all subjects, with high correlation and tight LOA. The heterogeneous nature of the results doesn't allow for these criteria to be met, as there is a small group of patients that leans heavily on near zero correlation in some or all of the modalities. This is to be expected, as a real-world everyday setting involving a double device setup, can often present sensor malfunctions or suboptimal recording conditions.

Subject Specific Performance

To observe the behavior of feature comparison at the subject level an heatmap was produced, with the purpose of observing differences between patients. The heatmap in Figure 3.9 ranks spearman's ρ across all subjects and features analyzed.

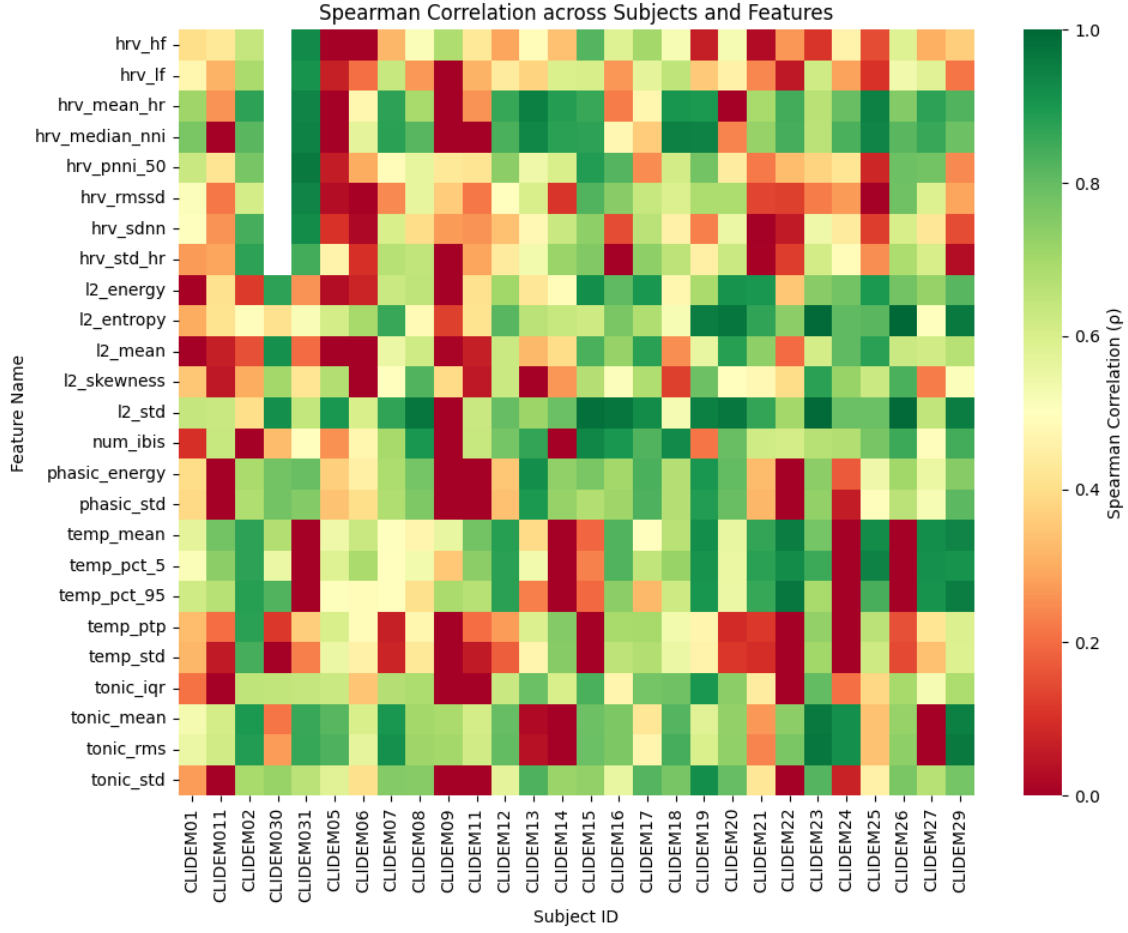


Figure 3.9: Feature correlation heatmap across subjects and features

Some patterns are identifiable in the subject specific correlations, demonstrating heterogeneous behavior of the sensors across recordings. Observable recurring patterns:

- TEMP descriptors showed consistent disagreement in subjects *CLIDEM24* and *CLIDEM26*
- HRV descriptors disagreed strongly in subjects *CLIDEM05* and *CLIDEM06*
- Overall inconsistencies are strongly present in subjects *CLIDEM09* and *CLIDEM11*

These findings suggest that some recordings are affected by factors such as movement artifacts, device placement, unreliable skin contact or occasional faulty sensors.

Representative Subject-Level Comparisons

To appreciate the variability observed across recordings, a representative subject was selected for detailed visualization using the feature comparability plots and

Bland-Altman representations introduced in the methodology section. The same representative subject previously introduced in the raw signal correlation analysis was retained for the descriptor-level evaluation.

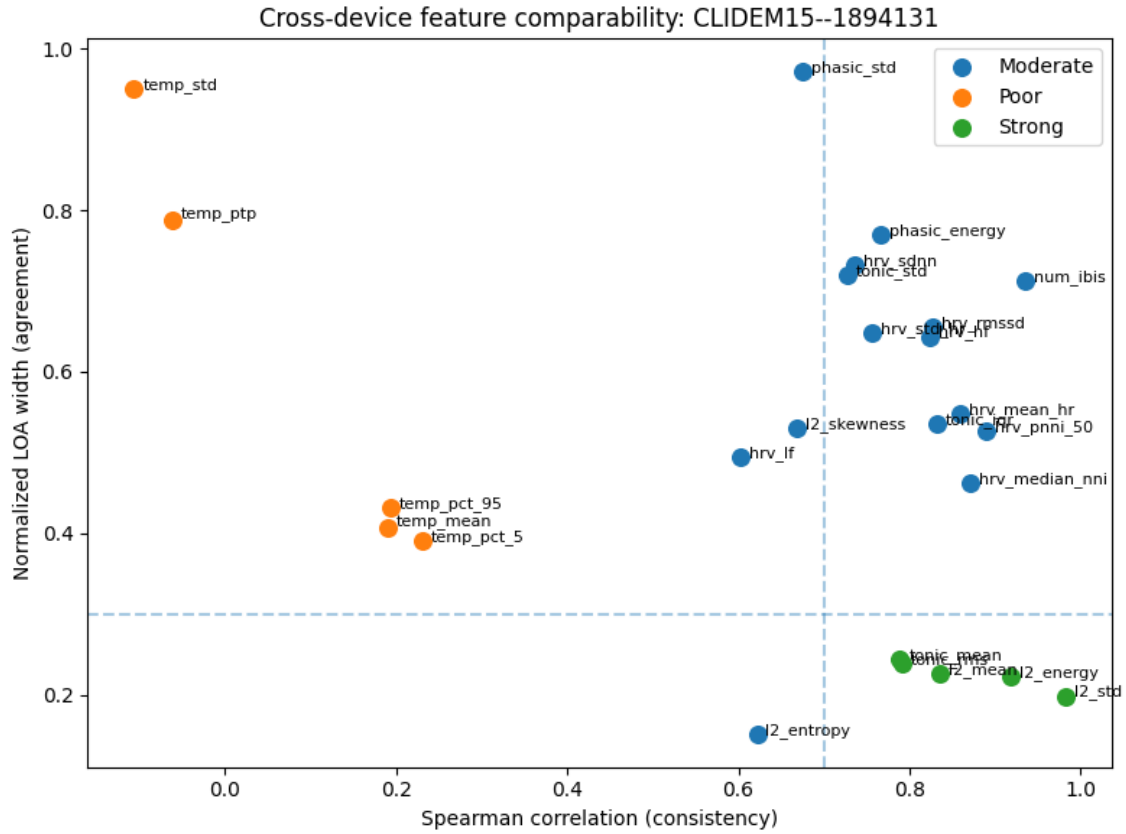


Figure 3.10: Feature comparability scatter plot for representative subject. Each feature is plotted with its Spearman correlation on the x axis and LOA on the y axis, color coded based on the assigned tier

Example: CLIDEM15 As shown in Figure 3.10, in this recording most descriptors cluster within the Moderate or Strong agreement regions, with high Spearman correlations on most features and particularly narrow LOA on ACC and tonic EDA features.

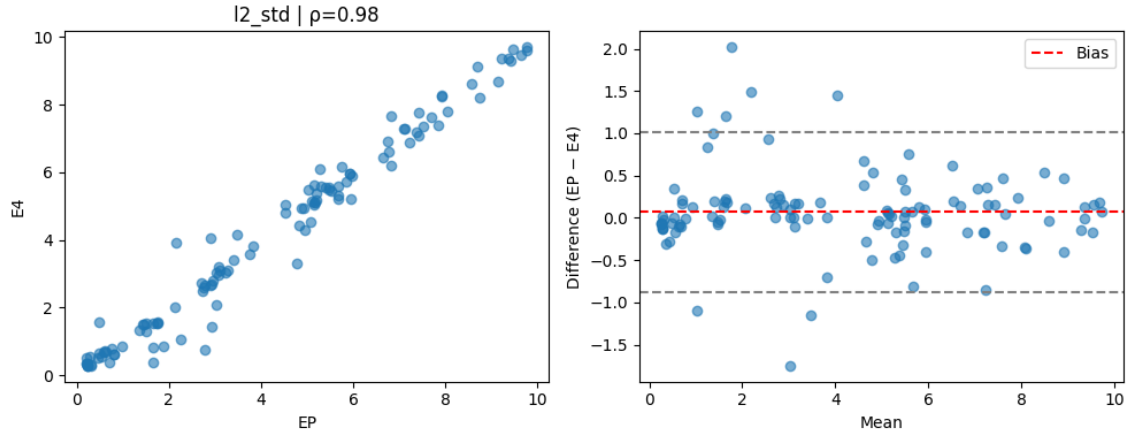


Figure 3.11: Scatter plot of the feature (dots representing the single values, EP value on the x axis and corresponding E4 value on the y axis); and Bland-Altman plot for ACC magnitude std (dots plotted with mean (x axis) and difference (y axis) between the devices, gray lines representing LOA boundaries and red line marking the bias.

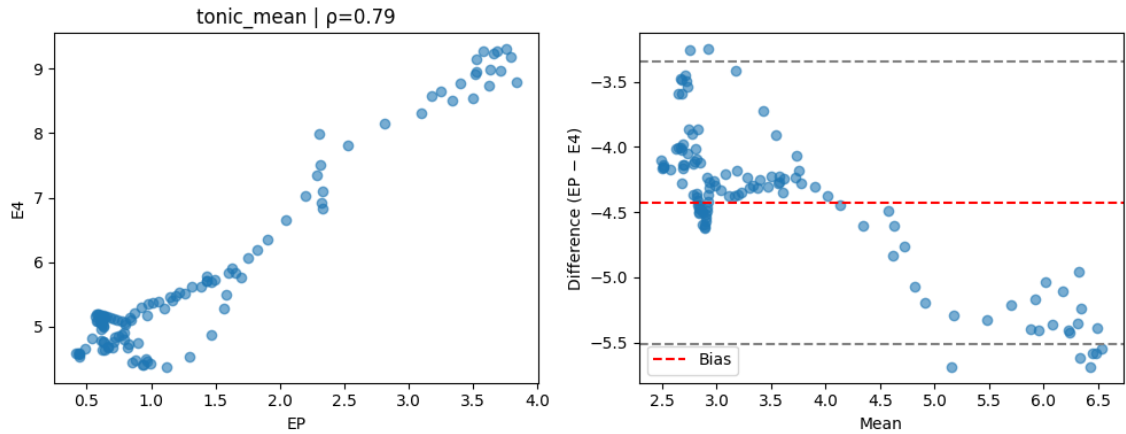


Figure 3.12: Scatter plot of the feature (dots representing the single values, EP value on the x axis and corresponding E4 value on the y axis); and Bland-Altman plot for EDA tonic mean (dots plotted with mean (x axis) and difference (y axis) between the devices, gray lines representing LOA boundaries and red line marking the bias.

The Bland-Altman plots in Figures 3.11-3.12 are brought as an example of high consistency between devices. Although EDA tonic mean exhibited strong consistency ($\rho = 0.79$), Bland-Altman analysis revealed evidence of proportional bias, with disagreement increasing at higher tonic values. This finding suggests that the devices capture similar physiological variation while differing in measurement scaling, limiting their direct interchangeability for analyses relying on absolute tonic EDA amplitudes. On the other hand ACC magnitude std resulted highly compa-

rable with a correlation near to 1 and bias near zero demonstrating that `12_std` is a highly robust and interchangeable feature for measuring activity intensity across these wearable devices. This indicates that despite underlying sensor hardware differences, both devices reliably capture the same aggregate movement trends without significant systematic measurement error.

3.2 BD Biomarkers and Prediction Models

3.2.1 Dataset Construction and Cohort Overview

As described in Section 2.6.1, the final datasets employed in this study consisted of 13 patients for the Mania cohort and 18 patients for the Depression cohort. The total number of extracted 5-minute physiological segments amounted to 3737 samples for the Mania dataset and 6385 samples for the Depression dataset, considering Acute and Remission labels.

Starting from the original FLIRT feature space of 212 extracted descriptors (Section 2.4.1), an initial filtering stage was applied during dataset construction to remove features with excessive missing samples. Specifically, any feature exhibiting more than 15% missing values across the cohort was discarded. Following this procedure, 6 features were removed from the Mania dataset and 4 from the Depression dataset, resulting in effective starting dimensionalities of 204 and 206 features respectively before correlation filtering.

As described in Section 2.6.2 the feature set is further reduced with the pruning of highly correlated features, resulting in 130 features for the Depression dataset and 137 for the Mania Dataset. To lighten computational weight and avoid using highly correlated features the biomarker analysis will be based only on these features.

Following on the findings illustrated in Section 3.1.2, the whole clinical analysis will be focused exclusively on signals recorded during sleep time. This choice allows for the use of HRV descriptors (otherwise impossible to use reliably because of the high missingness rate of E4’s PPG sensor) and much more stable signals in general.

3.2.2 Biomarker Discovery via Linear Mixed Models

Mania Cohort Table 3.8 displays 8 representative examples of the biomarker analysis across features. The Mania cohort was mostly characterized by a reduction in HRV descriptors during acute episodes, including decreases in `hrv_mean`, `hrv_pct_5`, and `hrv_max`. On the other hand HR descriptors such as `max_hr` and `num_ibis` (number of detected beats in a window) displayed significant increase

during the acute phase. Temperature features also exhibited significant increases, suggesting altered thermoregulatory dynamics during manic episodes.

Table 3.8: Representative physiological biomarkers identified in the Mania cohort

Feature	Profile	The Effect Size	Indiv. Ratio
hrv_mean	UB	-0.888	0.401
hrv_pct_5	UB	-0.831	0.418
hrv_max	UB	-0.696	0.424
num_ibis	UB	0.567	0.654
hrv_max_hr	UB	0.540	0.427
temp_pct_5	UB	0.519	0.755
temp_energy	Pers.	0.301	1.105
sleep_past_24h_hours	Pers.	0.161	2.772

While several descriptors demonstrated stable behavior across subjects and were classified as Universal Biomarkers (UB), other variables such as `temp_energy` and `sleep_past_24h_hours` showed high inter-subject variability (Personalized markers), indicating the presence of heterogeneous physiological responses across patients.

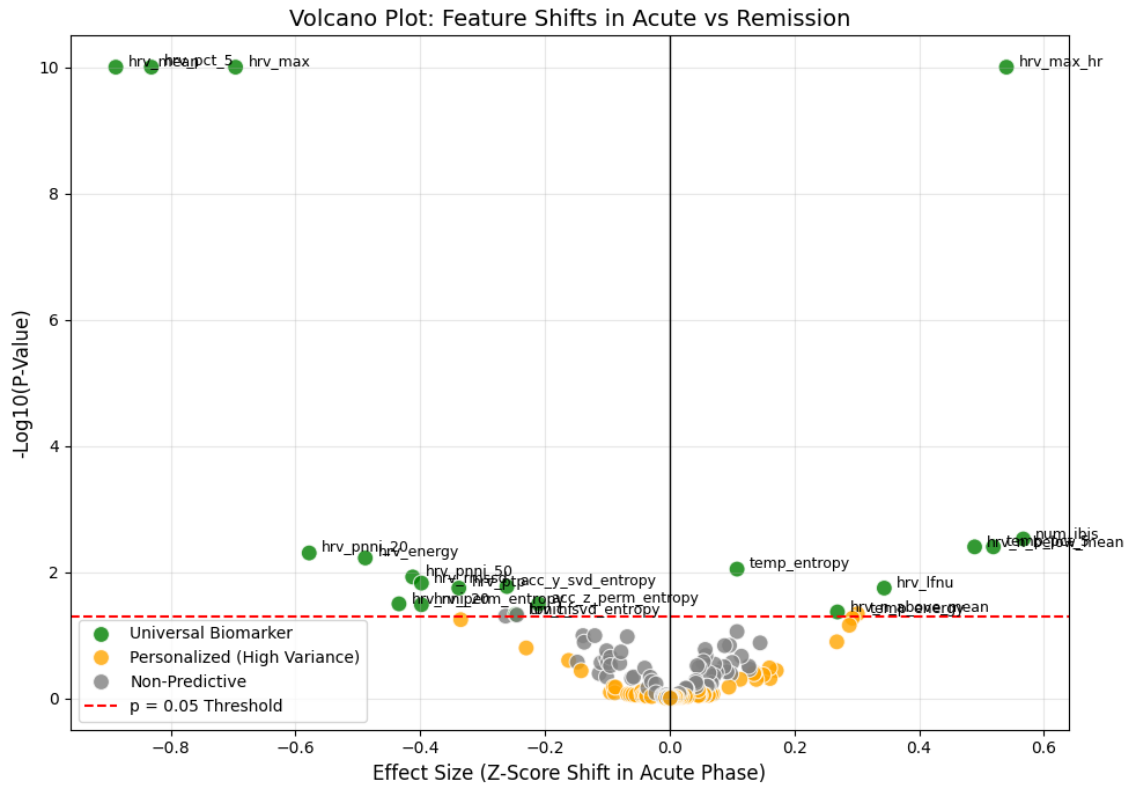


Figure 3.13: Volcano plot of Mania cohort biomarkers. Volcano plot of Mania features, color coded for feature profile, with effect size on the x axis and significance on y axis.

Figure 3.13 displays a color-coded volcano plot representing the distribution

of all extracted features within the Effect Size – P-Value space. Feature colors indicate their classification into the three previously defined categories: Universal Biomarkers, Personalized Indicators, and Non-Predictive Descriptors.

The most stable UBs, positioned in the upper corners of the plot, were mostly composed of HRV descriptors, which exhibited both the largest effect sizes and the lowest p-values across the cohort, accompanied by the 5th percentile of temperature `temp_pct_5`.

A complete overview of all identified UB is reported in Table 3.9. Overall, the Mania cohort analysis identified 24 UB, 53 Personalized Indicators, and 64 Non-Predictive Descriptors. The physiological biomarkers identified within the Mania

Table 3.9: Universal Biomarkers identified in the Mania cohort

Feature	Profile	Effect Size	Indiv. Ratio
<code>hrv_mean</code>	UB	-0.847	0.417
<code>hrv_pct_5</code>	UB	-0.775	0.427
<code>hrv_pnni_20</code>	UB	-0.674	0.554
<code>temp_pct_5</code>	UB	0.654	0.418
<code>hrv_max</code>	UB	-0.651	0.440
<code>hrv_nni_20</code>	UB	-0.534	0.652
<code>hrv_max_hr</code>	UB	0.503	0.440
<code>num_ibis</code>	UB	0.503	0.707
<code>hrv_pnni_50</code>	UB	-0.488	0.615
<code>hrv_rmssd</code>	UB	-0.457	0.656
<code>hrv_energy</code>	UB	-0.417	0.931
<code>hrv_perm_entropy</code>	UB	-0.416	0.727
<code>temp_energy</code>	UB	0.416	0.521
<code>hrv_nni_50</code>	UB	-0.409	0.811
<code>hrv_n_below_mean</code>	UB	0.399	0.693
<code>hrv_lfnu</code>	UB	0.325	0.734
<code>hrv_iqr_5_95</code>	UB	-0.314	0.863
<code>hrv_ptp</code>	UB	-0.311	0.877
<code>hrv_n_above_mean</code>	UB	0.301	0.793
<code>acc_y_sum</code>	UB	0.258	0.832
<code>acc_y_min</code>	UB	0.252	0.930
<code>acc_y_svd_entropy</code>	UB	-0.190	0.797
<code>tonic_skewness</code>	UB	0.164	0.024
<code>acc_y_max</code>	UB	0.163	0.899

cohort are for the most part highly consistent with the established literature. The pronounced reduction in autonomic regulation, indicated by the drop in `hrv_mean`, `hrv_pct_5`, and `hrv_max`, combined with the acceleration of HR metrics (`max_hr` and `num_ibis`) aligns with the most recent studies regarding sympathetic activation and

diminished heartbeat complexity during acute manic phases [9,17]. The increase in `temp_pct_5` and `temp_energy` similarly fits within the recognized pattern of altered thermoregulatory dynamics and heightened autonomic arousal as an indicator of acute mania [25].

In contrast, the increase observed in `sleep_past_24h_hours` (classified as a personalized marker but presenting a relevant effect size) stands in contradiction to the clinical framework, which associates manic episodes with restricted sleep duration [7,17]. This divergence can be accounted for by both methodological and environmental confounding factors. Computationally, extracting sleep metrics from a 48-hour recording window introduces significant statistical instability into the feature. At the same time, acute manic episodes are recorded during hospitalization and administration of intensive pharmacological therapy, such as sedatives or antipsychotics, which can artificially induce sleep. This contrasts with the remission phase recording conditions, typically recorded in a home environment under symptom maintenance therapy. Within the LMM framework, these confounders explain why `sleep_past_24h_hours` is classified as a highly volatile Personalized Indicator rather than a global trend, reinforcing the conclusion that robust physiological metrics like HRV and temperature remain the most reliable Universal Biomarkers for generalized acute state detection.

Depression Cohort Table 3.10 displays 8 representative examples of the biomarker analysis across features, reflecting the distribution of the feature classification obtained through LMM in the Depression dataset. Unlike the highly synchronized shifts seen in mania, the only Biomarkers found in the Depression cohort were subtle reductions in ACC descriptors during acute episodes, such as `acc_x_max` and `acc_ar_1`. All other features were classified as personalized or non-predictive profiles; variables like `waso_mins` (wake time after sleep onset) and `hrv_cvstd` (coefficient of variation of successive differences) displayed downward trends that were highly volatile across subjects, while `temp_mean` exhibited a general increasing trend but an individuality ratio close to 2, identifying it as an highly volatile descriptor.

While a small subset of descriptors demonstrated stable behavior across subjects and were classified as Universal Biomarkers, the majority of variables, such as `waso_mins`, `hrv_min_hr`, and `hrv_perm_entropy`, showed high variability (high individuality ratio) or low statistical significance that indicates a highly heterogeneous and individualized response across patients.

Table 3.10: Representative physiological biomarkers identified in the Depression cohort

Feature	Profile	Effect Size	Indiv. Ratio
acc_x_max	UB	-0.153	0.901
acc_ar_1	UB	-0.140	0.666
waso_mins	Pers.	-0.381	1.757
hrv_min_hr	Pers.	0.034	15.675
temp_mean	Pers.	0.213	1.968
hrv_sdn	Pers.	-0.151	2.631
hrv_perm_entropy	Non-Pred.	-0.199	1.486
acc_y_pct_95	Non-Pred.	-0.165	1.020

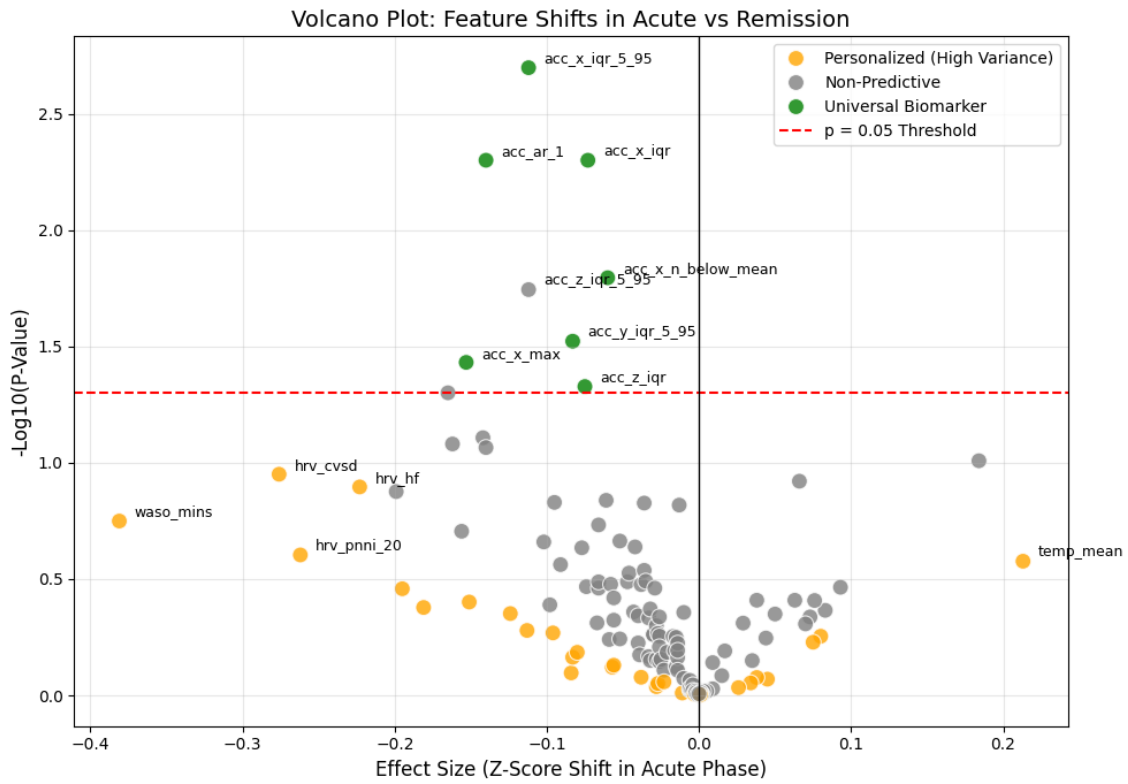


Figure 3.14: Volcano plot of the Depression cohort biomarkers. Volcano plot of Depression features, color coded for feature profile, with effect size on the x axis and significance on y axis.

Figure 3.14 shows the volcano plot representing the distribution of all extracted features across the three classes in the Effect Size – P-Value space. Noticeably, the 7 biomarkers barely pass the p-value requirements for significant and present low effect sizes, marking them as not very strong biomarkers

In contrast to the highly synchronized features observed in the Mania cohort, the analysis of the Depression dataset revealed an absence of uniform global biomarkers.

As documented in Table 3.10, autonomic, thermoregulatory and sleep parameters failed to achieve significance at the cohort level, instead being classified as personalized or non-predictive. The only features surviving the universal classification threshold were a subset of 7 ACC descriptors, as reported in Table 3.11, all of which presented subtle effect sizes (ranging from -0.060 to -0.153).

Overall, the Depression cohort analysis identified 7 Universal Biomarkers, 28 Personalized Indicators, and 95 Non-Predictive Descriptors. The absence of strong

Table 3.11: Universal Biomarkers identified in the Depression cohort

Feature	Profile	Effect Size	Indiv. Ratio
acc_x_max	UB	-0.153	0.901
acc_ar_1	UB	-0.140	0.666
acc_x_iqr_5_95	UB	-0.112	0.605
acc_y_iqr_5_95	UB	-0.083	0.841
acc_z_iqr	UB	-0.075	0.968
acc_x_iqr	UB	-0.073	0.076
acc_x_n_below_mean	UB	-0.060	0.012

global biomarkers is backed by contemporary literature that highlights the individualized nature of depression, where highly diverse individual expressions are often found. The exclusive survival of a sparse subset of seven actigraphy descriptors (such as `acc_x_max` and `acc_ar_1`) aligns directly with established findings documenting blunted daily motor activity and psychomotor deceleration in depressed patients. [7]

The classification of sleep and thermoregulatory metrics into personalized or non-predictive features contextualizes this variance. Features such as `waso_mins` and `temp_mean` failed to reach uniformity at the global level precisely because of the opposing responses characteristic of depressive episodes. For instance, the clinical variation between insomnia and hypersomnia neutralizes globalized trends in sleep duration and metrics deviates from sleep. This mixed directionality in sleep time variations during depressive episodes is widely supported in literature [7, 11], suggesting that sleep tracking should be used only as a personalized marker. Similarly, the highly individual shifts in skin temperature align with recent wearable sensor studies, which note that nocturnal temperature variations are inconsistent, reinforcing the validity of the LMM’s personalized classification [7].

3.2.3 Biological Phenotype Discovery

To explore this diversity in responses a targeted sub-analysis was conducted on the Personalized Indicators in order to look for clusters of patients that respond

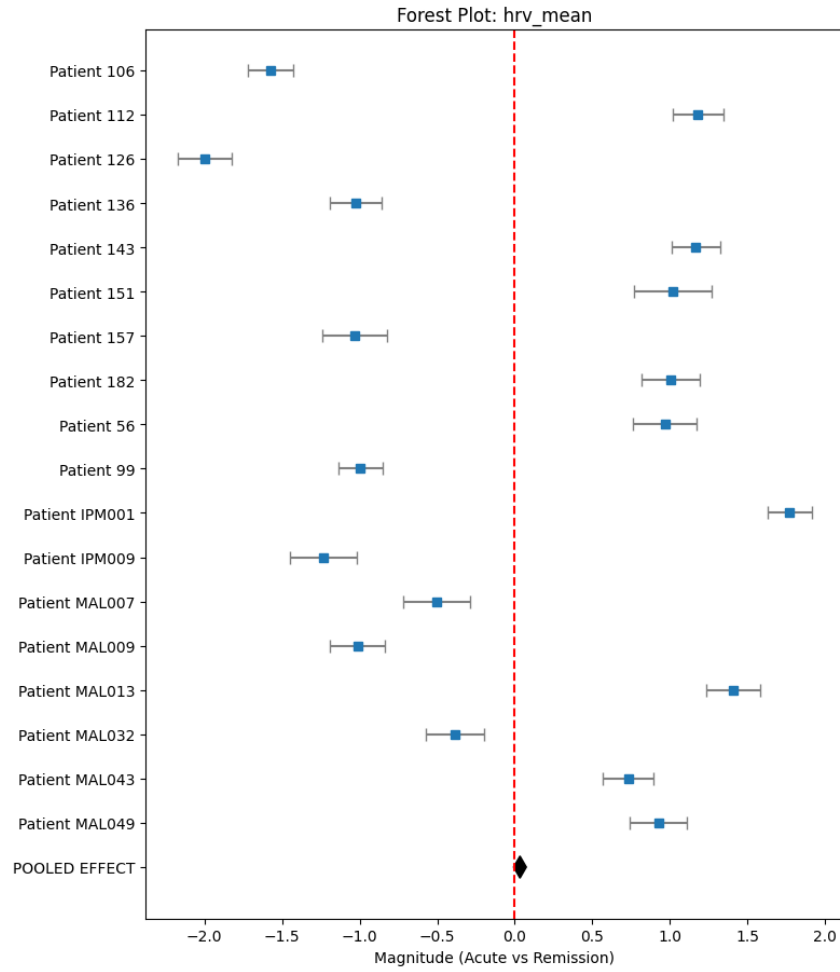


Figure 3.15: Forest plot of hrv_mean response for each patient. Average HRV response magnitude plotted for each patient with uncertainty intervals. Diamond shape at the bottom representing the averaged response of the cohort.

similarly to depression. A stratified response was found in the variations of average HRV, as shown in the forest plot of Figure 3.15, where two well defined clusters were identified, one group of patients with strongly decreasing hrv_mean and one where it increased sharply.

Phenotype-Specific Biomarkers

Following the stratification analysis, the Depression dataset was split into two distinct phenotypes based on hrv_mean trajectories: HRV-Increasers and HRV-Droppers ($n = 9$ patients each). To evaluate the clinical utility of this stratification, the LMM was re-executed independently on both groups.

HRV-Dropper sub-cohort Contrarily to the highly heterogeneous response of the Depression cohort, the HRV-Dropper subgroup demonstrated to be much more aligned in patient’s responses. As displayed with the Universal Biomarkers examples in Table 3.12, this phenotype is characterized by a uniform reduction in parasympathetic tone and Heart Rate Variability during acute depressive episodes. This is marked by significant decrease in metrics such as `hrv_pnni_20`, `hrv_pct_5`, and `hrv_sdn`. Cardiac activation indicators like mean and minimum heart rate (`hr_mean_hr`, `hrv_min_hr`) increased sharply, while baseline actigraphy metrics showed an uniform decreasing trend coherent to the findings of the whole depression cohort. The discovery of 24 Universal Biomarkers demonstrate that this specific

Table 3.12: Universal Biomarkers identified in the HRV-Dropper sub-cohort

Feature	Profile	Effect Size	Indiv. Ratio
<code>hrv_pnni_20</code>	UB	-1.035	0.462
<code>mean_hr</code>	UB	1.043	0.341
<code>min_hr</code>	UB	0.945	0.355
<code>hrv_pct_5</code>	UB	-0.856	0.512
<code>hrv_nni_50</code>	UB	-0.755	0.685
<code>hrv_cvsd</code>	UB	-0.749	0.623
<code>hrv_nni_20</code>	UB	-0.740	0.756
<code>hrv_sdn</code>	UB	-0.738	0.428
<code>hrv_iqr</code>	UB	-0.683	0.478
<code>hrv_energy</code>	UB	-0.663	0.435
<code>hrv_ptp</code>	UB	-0.647	0.464
<code>hrv_hf</code>	UB	-0.645	0.675
<code>hrv_min</code>	UB	-0.572	0.776
<code>hrv_lf</code>	UB	-0.504	0.523
<code>hrv_perm_entropy</code>	UB	-0.448	0.608
<code>hrv_vlf</code>	UB	-0.226	0.943
<code>hrv_skewness</code>	UB	0.157	0.533
<code>acc_x_iqr_5_95</code>	UB	-0.137	0.019
<code>acc_z_iqr_5_95</code>	UB	-0.118	0.530
<code>acc_ar_1</code>	UB	-0.116	0.544
<code>acc_x_iqr</code>	UB	-0.112	0.012
<code>acc_ar_2</code>	UB	-0.107	0.153
<code>acc_z_iqr</code>	UB	-0.104	0.024
<code>acc_z_n_below_mean</code>	UB	-0.092	0.016
<code>acc_z_skewness</code>	UB	-0.083	0.054

sub-cohort experiences a uniform response during acute episodes. The large negative effect sizes across all main HRV markers (e.g., `hrv_pnni_20` and `hrv_cvsd`) coupled with suppressed actigraphy variables indicate that the HRV-dropper sub-

cohort represents a cluster of patients characterized by sympathetic overactivation and slight psychomotor deceleration.

HRV-Increaser sub-cohort The HRV-Increaser subgroup demonstrated a diametrically opposite response. As displayed with the Universal Biomarkers examples in Table 3.13, this phenotype is characterized by a uniform increase in parasympathetic tone and overall HRV during acute depressive episodes, marked by positive effect sizes in metrics such as `hrv_pct_5`, `hrv_energy` and `hrv_sdn`. At the same time, cardiac activation indicators like mean and min heart rate, as well as specific actigraphy metrics, showed a uniform decreasing trend.

Table 3.13: Universal Biomarkers identified in the HRV-Increaser sub-cohort

Feature	Profile	Effect Size	Indiv. Ratio
<code>mean_hr</code>	UB	-1.086	0.062
<code>hrv_pct_5</code>	UB	0.932	0.096
<code>min_hr</code>	UB	-0.865	0.166
<code>hrv_energy</code>	UB	0.652	0.304
<code>hrv_min</code>	UB	0.636	0.185
<code>num_ibis</code>	UB	-0.545	0.441
<code>hrv_iqr</code>	UB	0.505	0.334
<code>hrv_n_below_mean</code>	UB	-0.500	0.328
<code>hrv_pnni_20</code>	UB	0.490	0.450
<code>hrv_sdn</code>	UB	0.412	0.458
<code>hrv_ptp</code>	UB	0.380	0.541
<code>hrv_n_above_mean</code>	UB	-0.374	0.570
<code>hrv_nni_20</code>	UB	0.360	0.647
<code>acc_y_pct_95</code>	UB	-0.350	0.460
<code>acc_y_max</code>	UB	-0.311	0.491
<code>acc_y_pct_5</code>	UB	-0.241	0.499
<code>hrv_skewness</code>	UB	-0.210	0.744
<code>hrv_vlf</code>	UB	0.176	0.527
<code>hrv_kurtosis</code>	UB	-0.135	0.413
<code>acc_y_iqr_5_95</code>	UB	-0.128	0.623

By successfully isolating these biomarkers in each group, the analysis demonstrated that while depression doesn't allow clear global tracking, it can be divided into predictable biological subtypes.

Evaluation of Clinical and Demographic Confounders

To determine if the discovered HRV sub-cohorts were driven by external clinical or demographic factors, the two groups were compared across age, sex, clinical

episode severity (HDRS/YMRS baselines) and current pharmacological treatment (including antidepressants, antipsychotics, mood stabilizers). Fisher’s tests and independent t-tests revealed no significant differences between the HRV-Droppers and HRV-Increasers across any evaluated clinical or demographic variable.

3.2.4 Mania Cohort Predictive Models

Feature Set Comparison

Table 3.14: Performance comparison between the evaluated feature configurations in the Mania cohort using LOSO cross-validation

Feature Config.	N	Balanced Accuracy	Macro F1-score	ROC-AUC
Full FLIRT	204	0.689 ± 0.108	0.668 ± 0.122	0.767 ± 0.128
Uncorrelated	137	0.680 ± 0.129	0.663 ± 0.142	0.753 ± 0.146
UB Only	24	0.687 ± 0.095	0.666 ± 0.108	0.747 ± 0.109

Table 3.14 reports the predictive performance obtained using the three evaluated feature configurations within the Mania cohort. The Full FLIRT configuration achieved the highest ROC-AUC (0.767 ± 0.128), suggesting that the complete 206 features retain some additional information useful for class separation. However, despite this reduction in dimensionality, the classification performance remained mostly stable across all configurations. Eventually the 24 Universal Biomarkers model achieved a Balanced Accuracy of 0.687 ± 0.095 , which is comparable to the 0.689 ± 0.108 produced by the full feature set, with reduced fold variability.

This stable performance proves that the biomarker discovery pipeline lowered redundant noise while preserving the main predictive signal. By isolating a subset of meaningful descriptors that retain the same classification power of the original space, this configuration proves to be the most optimal of the ones tested. In practice this reduction in feature dimensionality is highly convenient for digital health applications outside of hospitals, minimizing computational weight and partially protecting the model from overfitting on unseen patients. For these reasons the 24 features Universal Biomarker set was selected as the main configuration for all the following analysis on the Mania Dataset.

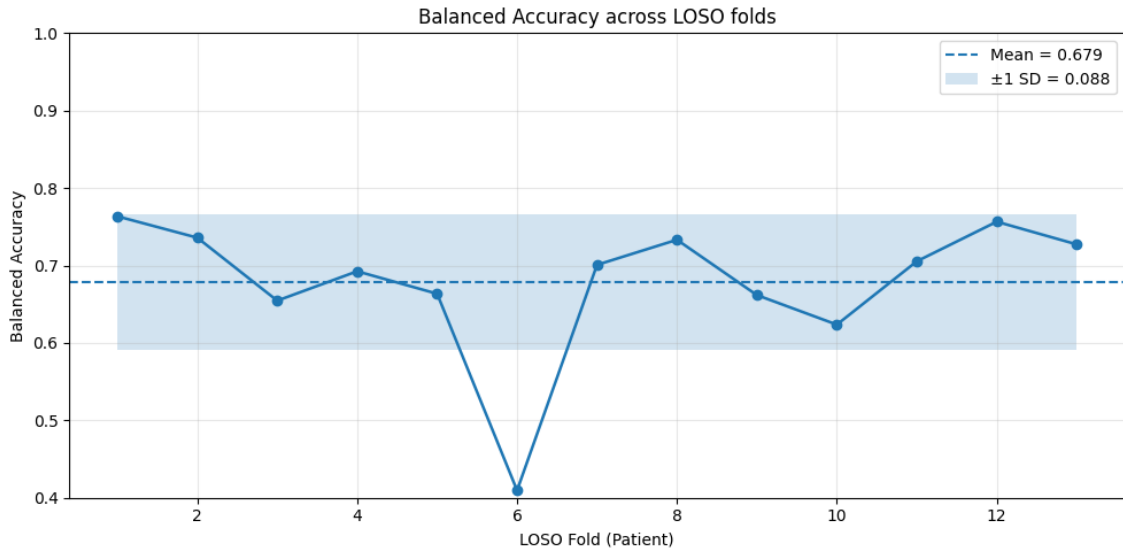


Figure 3.16: Balanced Accuracy across LOSO folds.

Figure 3.16 represents the performance on the model across LOSO folds. In this image each dot is a LOSO fold, where the model is tested exclusively on one patient, demonstrating a Balanced Accuracy mostly confined between 0.6 and 0.8. Fold 6 is the only one performing much below average, demonstrating an inability of the model to adapt to this patient’s specific response to an acute manic phase.

Session-Level Recognition Analysis

Table 3.15 summarizes the performance of session recognition across different evaluation thresholds, ranging from a baseline of 0.50 to a stricter requirement of 0.65. As expected, recognition rates declined as the thresholds became stricter. The analysis also revealed a notable asymmetry between the performances of the two clinical states.

At the pre-defined threshold of 0.60 Balanced Accuracy, the model successfully recognized 85% of the remission sessions, compared to 69% of the acute manic sessions. This unbalance becomes more evident at the 0.65 threshold. This asymmetry suggests that the physiological signature of remission is more stable compared to acute manic episodes, which is characterized by higher intra and inter-patient variability. This volatility is also noticeable in the model’s failure on Fold 6, where individual variation deviated too much from the generalized cohort trend, forbidding the model from correctly predicting both classes.

Table 3.15: Session Recognition Performance Across Thresholds

Threshold	Acute	Remission	Overall
0.50	0.92	1.00	0.96
0.55	0.77	0.92	0.85
0.60	0.69	0.85	0.77
0.65	0.46	0.77	0.62

Temporal Prediction Dynamics

Figure 3.17 shows some examples of prediction visualization. This timeline plot, representing the predictions of the model for each sample over time, allows an assessment of the model’s behavior during different times of the day. Since all the samples are taken during sleep periods it is interesting to observe how the quality of the prediction changes during the circadian cycle.

For example in the first Remission plot the prediction confidence increases as the samples go further into the night period, likely corresponding with deep sleep onset, which minimizes external behavioral noise that could impair the capability of the model to identify the class of the sample. On the contrary the mania predictions appear to be more consistent throughout the circadian cycle, this indicates that manic arousal might be visible to the model independently of confounding factors and regardless of sleep stage transitions.

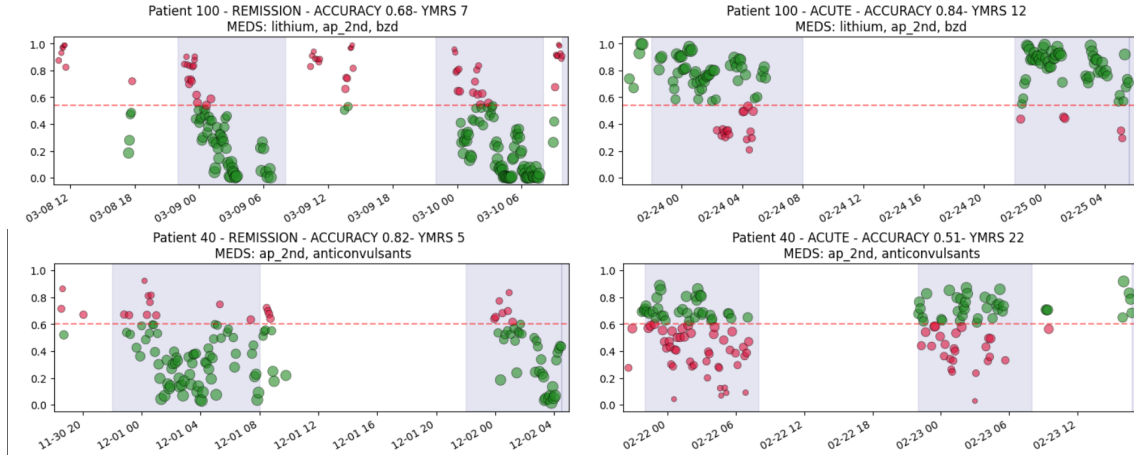


Figure 3.17: Representative plots of sample predictions over time for two folds of the Mania model.

For each subplot the predictions for a single session (acute or remission) are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.

The prediction plots of all the folds in the Mania cohort model can be found in Appendix B.2, Figure B.1.

3.2.5 Depression Cohort Predictive Modeling Results

Full Cohort vs. Stratification

Three independent predictive models were generated using only the features classified as Universal Biomarkers during the LMM analysis. The baseline model was trained on the complete, unstratified Depression dataset, while the other two models were restricted to the HRV-Dropper and HRV-Increaser sub-cohorts, utilizing only the universal biomarkers found in the LMM analysis. This pipeline resulted in a input dimensionality of 7 features for the full cohort, 25 features for the HRV-Dropper sub-cohort, and 20 features for the HRV-Increaser sub-cohort.

Table 3.16 aggregates the cross-validation performance metrics across these three distinct configurations.

The Global Depression model fails to generalize the prediction, reporting a performance below chance with a Balanced Accuracy of 0.493 and a ROC-AUC of 0.494 . This is due to the lack of common direction of the features across patient, that doesn't allow the model to find a common behavior to separate acute mania from

Table 3.16: Predictive performance comparison between the Global Depression model and the phenotype-stratified sub-cohort models using LOGO cross-validation

Model	N Features	Balanced Acc.	F1-score	ROC-AUC
Depression Cohort	7	0.493 ± 0.038	0.466 ± 0.050	0.494 ± 0.065
HRV-Dropper	25	0.653 ± 0.110	0.622 ± 0.082	0.742 ± 0.126
HRV-Increaser	20	0.696 ± 0.048	0.682 ± 0.044	0.793 ± 0.067

remission, as seen for most HRV features.

On the other hand separating the sub-cohorts unlocks better performance across all evaluation metrics. The HRV-Dropper model achieves a ROC-AUC of 0.742 and a Balanced Accuracy of 0.653. The HRV-Increaser configuration, which reaches the highest performance with a Balanced Accuracy of 0.696, Macro F1-score of 0.682 and ROC-AUC of 0.793. The rise in both F1 and ROC-AUC scores proves that the identified sub-cohort biomarker features allow the model to distinguish acute and remission phases.

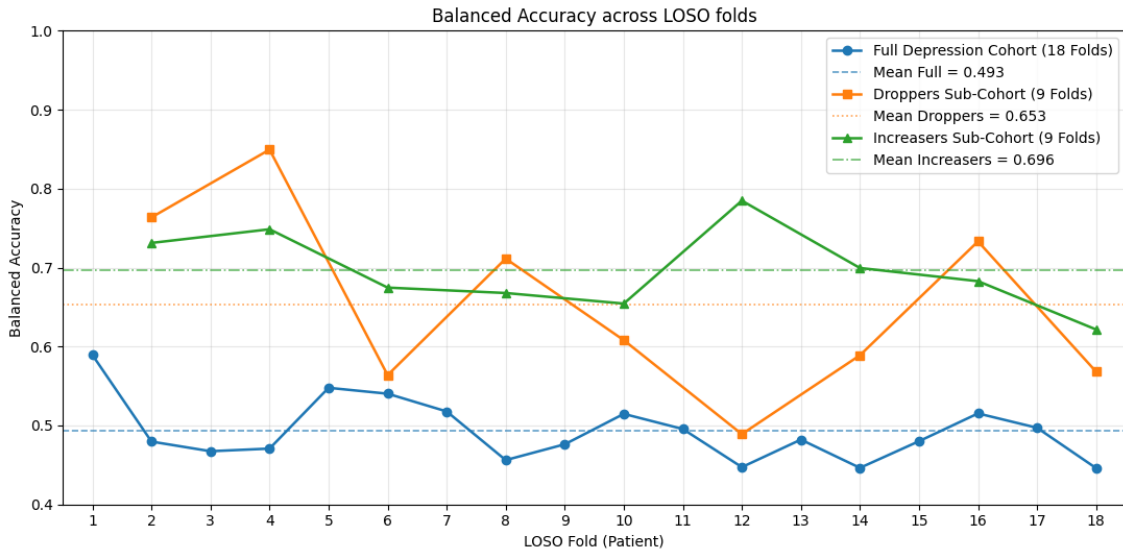


Figure 3.18: Balanced accuracy across folds for the three depression datasets: full cohort, droppers sub-cohort and increasers sub-cohort

Figure 3.18 reports the Balanced Accuracy metrics for each of the individual LOSO cross-validation folds for each of the three cohorts. The full cohort performance profile remains mostly flat and concentrated around 0.5 across evaluation folds, demonstrating chance classification and inability to predict the two classes.

On the contrary the specialized models are mostly able to correctly classify the samples in the two classes. The HRV-Dropper model accuracy averages at around 0.65, exhibiting high volatility, showing drops in performance at folds 3 and 6 which indicate that these patients might have a stronger variability in responses direction compared to the others. The HRV-Increaser model exhibits the best performance curve achieving more stability and a higher mean accuracy at 0.7, with boundaries at 0.79 and 0.62 for folds 6 and 9.

Session-Level Recognition

Table 3.17: Session recognition rates across depression configurations at 0.50 / 0.55 / 0.60 evaluation thresholds

Model Config	Acute	Remission	Overall
Depression Cohort	0.72/0.50/0.39	0.22/0.17/0.17	0.47/0.33/0.28
HRV-Dropper	0.78/0.67/0.56	1.00/0.89/0.67	0.89/0.78/0.61
HRV-Increaser	1.00/1.00/0.89	1.00/1.00/0.89	1.00/1.00/0.89

Table 3.17 shows the session level recognition rates, evaluated at thresholds of 0.50, 0.55 and 0.60. The unstratified Global model was predictably the most unstable, maintaining an overall session recognition below 50% with a clear difficulty in recognizing remission samples.

The HRV-Dropper model maintained an overall recognition around 80%, with slightly better performance in remission samples, decreasing to 61% of sessions recognized at 0.60 evaluation threshold. The HRV-Increaser sub-cohort pipeline achieved much better performance with all sessions recognized at the 0.50 and 0.55 evaluation thresholds, dropping only to 90% overall recognition at a more strict threshold of 0.60.

This behavior shows exceptional stability and consistency in the Increaser phenotype model and moderate performance in the Dropper model.

Temporal Prediction Dynamics

Figure 3.19 shows three examples of prediction dynamics over time, two folds from the increasers sub-cohort and one from the droppers sub-cohort.

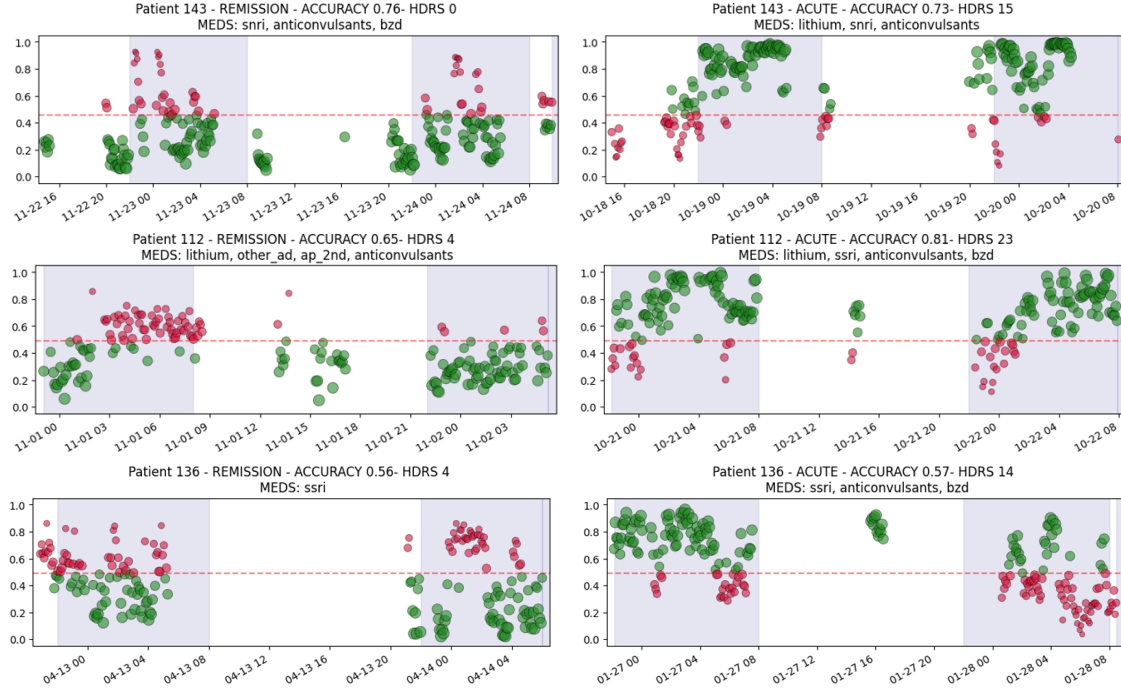


Figure 3.19: Representative plots of session predictions over time for the two Depression sub-cohorts.

For each subplot the predictions for a single session (acute or remission) are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.

The prediction plots of all the folds in the Depression sub-cohorts models can be found in Appendix B.2, Figures B.2 and B.3.

As in mania, also in the case of classifying depression the models struggled to accurately predict the classes when the samples were taken in close proximity to sleep onset or wake. Confirming that the best predictions were obtained some time after sleep onset, when the patient was most stable.

Chapter 4

Conclusions and Future Work

4.1 Device Comparability

The first phase of this study evaluated the comparability between the Empatica E4 and the Embrace Plus bracelets using a cohort of 30 healthy subjects.

Agreement at the signal level (Objective 1) : The hypothesis H1 was partially supported. While IBI and TEMP signals demonstrated strong morphological alignment, raw EDA and ACC magnitude signals were found have variable comparability across patients. Significant RMSE variations in EDA ($3.16 \mu S$) and TEMP ($1.90^{\circ}C$) showed that absolute values cannot be directly interchanged across device generations due to hardware differences, varying sensor geometries and proximal versus distal wrist placement. It was also found that the PPG sensor mounted on the E4 bracelet had detected on average 60% less beats than EP, making the use of IBI signal unreliable. Although the detected IBI sequences showed good morphological agreement when valid beats were available, the substantial beat loss observed in the E4 make the signal unreliable to be used in this analyses. The sub-analysis conducted on data restricted to sleep periods indicated that, despite few outliers, the devices were much more comparable and aligned while the patient was asleep. Producing a decisive increase in correlation in EDA and ACC signals. The beat detection capabilities of the E4 bracelet were also vastly improved during sleep, sampling just 12% less beats than EP.

Descriptor Consistency (Objective 2) : The hypothesis H2 was supported with limitations. Most features were found to have moderate agreement between devices, with high variability across subjects. The most comparable descriptors were found to be ACC magnitude std, energy and entropy features, balancing the

low agreement found in the full analysis of the ACC signal. Also in this case the E4 suffered from significant and highly variable data loss in HRV metrics (median 27.5% missing windows), whereas the EmbracePlus demonstrated highly stable performance (2-3% loss) thanks to its PPG sensor equipped with 4 multi-wavelength channels. Consequently, HRV features compared poorly in several subjects. It was concluded that the subsequent analysis were to be performed exclusively on data obtained from sleep periods. This decision was taken in order to employ the IBI signals from the E4 that would have to be excluded otherwise, and to use the signals where they are the most stable, excluding confounding factors and artifacts caused by daily activities.

4.2 BD Biomarkers and Prediction Models

The second phase consisted in the clinical application within the INTREPIBD project, developing ML models to differentiate acute phases of mania and depression from remission periods.

Biomarker Identification (Objective 3) : The hypothesis H3 was fully validated for Mania and partially validated for Depression. **Mania**: The LMM pipeline successfully isolated 24 Universal Biomarkers. Acute manic phases were characterized by a decrease of HRV descriptors and increase of HR and thermoregulatory metrics. **Depression**: The analysis on the full cohort failed to identify strong global biomarkers except for 7 actigraphy features reflecting slight motor deceleration. This absence of common direction highlights the clinical heterogeneity characteristic of depressive disorders (e.g., opposing patient expressions like insomnia vs. hypersomnia). However, clustering based on `hrv_mean` direction, two sub-cohorts were separated where patients share a similar behavior: HRV-Droppers (24 UBs showing reduced parasympathetic activity) and HRV-Increasers (20 UBs characterized by bradycardia and reduced motor activity).

Predictive Model Performance (Objective 4) : The hypothesis H4 was validated for mania and partially validated through stratification for depression. In the **Mania cohort**, a compact model trained on the 24 Universal Biomarkers matched the predictive power of the full 206-feature space, achieving a stable Balanced Accuracy of 68.7% utilizing LOSO cross-validation to predict the class of the single samples. When a further level was added in order to classify the whole recordings, the model was able to correctly predict 85% of unseen remission sessions and 69% of

acute. In the **Depression cohort**, the unstratified global model failed completely to classify the samples, due to the high heterogeneity of the patient’s responses. Separating the phenotypes improved the ability of the models to predict the state of the condition: the HRV-Dropper model achieved a Balanced Accuracy of 65.3% (ROC-AUC of 0.742), while the HRV-Increaser model demonstrated more stability, reaching a Balanced Accuracy of 69.6% (ROC-AUC of 0.793) and 100% recognition rate at the session level.

This thesis demonstrated the feasibility of using physiological biomarkers from wearable devices to predict the phases of mania. It confirmed that depression expresses itself with different phenotypes and it is highly heterogeneous, making it difficult to have a generalized model. Finally, sleep periods were discovered to be the highest quality window to utilize wearable data, improving recording stability and carrying enough information to distinguish acute phases from remission.

4.3 Future Work

To build on the contributions of this thesis, future research should focus on the following developments:

- **Phenotype Validation:** Validate the discovered HRV-Dropper and HRV-Increaser depressive sub-phenotypes on an independent, larger longitudinal cohort to confirm their stability. Investigate if these digital profiles can be linked to distinct clinical sub-categories based on other parameters.
- **Sleep Window Filtering:** Include sleep periods segmentation to explore the influence of different sleep stages (e.g. Deep, Light and REM) on classification accuracy. Validate or exclude the theorized impact of sleep-wake onset on the accuracy of the predictions. Such studies would improve the understanding of BD, enabling the development of more accurate detection methods.
- **From Classification to Forecasting:** Utilize the Biomarkers found in this study to predict an upcoming BD episode days before the manifestation of clinical symptoms. Rather than classifying the current mood state, the models should evaluate the risk of an episode onset in the near future. Such an instrument for early warning would be crucial to reduce episode severity and duration, allowing timely therapy interventions and possibly avoiding hospitalization.
- **Real-World Deployment:** Exploit the compact feature space of the models to design real-time passive monitoring concepts, such as smartphone or

wearable device applications. This work demonstrated that it is possible to classify mood states of BD patients from short wearable recordings using only a small number of features. This findings open the door to passive monitoring systems that detect episode onset exclusively relying on the processing power of a smartphone, alerting the clinician and reducing detection time massively.

Bibliography

- [1] V. Oliva, G. Fico, M. De Prisco, X. Gonda, A. R. Rosa, and E. Vieta, “Bipolar disorders: an update on critical aspects,” *The Lancet Regional Health – Europe*, vol. 48, p. 101135, Jan 2025.
- [2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. Washington, DC: American Psychiatric Association Publishing, 2022.
- [3] World Health Organization, “Bipolar disorder,” September 2025.
- [4] P. Dome, Z. Rihmer, and X. Gonda, “Suicide risk in bipolar disorder: A brief review,” *Medicina*, vol. 55, no. 8, p. 403, 2019.
- [5] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, “A rating scale for mania: Reliability, validity and sensitivity,” *British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.
- [6] M. Hamilton, “A rating scale for depression,” *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, pp. 56–62, 1960.
- [7] A. C. Tonon, A. Nexha, M. M. da Silva, F. A. Gomes, M. P. Hidalgo, and B. N. Frey, “Sleep and circadian disruption in bipolar disorders: From psychopathology to digital phenotyping in clinical practice,” *Psychiatry and Clinical Neurosciences*, vol. 78, no. 10, pp. 543–555, 2024.
- [8] R. Zhong, X. Wu, J. Chen, and Y. Fang, “Using digital phenotyping to discriminate unipolar depression and bipolar disorder: Systematic review,” *JMIR mHealth and uHealth*, 2024.
- [9] M. Nardelli, A. Lanata, G. Bertschy, E. P. Scilingo, and G. Valenza, “Heart-beat complexity modulation in bipolar disorder during daytime and nighttime,” *Scientific Reports*, vol. 7, no. 1, p. 17920, 2017.

- [10] S. Fedor, R. Lewis, P. Pedrelli, D. Mischoulon, J. Curtiss, and R. W. Picard, “Wearable technology in clinical practice for depressive disorder,” *New England Journal of Medicine*, vol. 389, no. 26, pp. 2457–2466, 2023.
- [11] V. De Angel, S. Lewis, K. White, C. Oetzmann, D. Leightley, E. Oprea, G. Lavelle, F. Matcham, A. Pace, D. C. Mohr, R. Dobson, and M. Hotopf, “Digital health tools for the passive monitoring of depression: a systematic review of methods,” *npj Digital Medicine*, vol. 5, no. 1, p. 3, 2022.
- [12] P. Jakobsen, U. Cote-Allard, M. A. Riegler, L. A. Stabell, A. Stautland, T. Nordgreen, J. Torresen, O. B. Fasmer, and K. J. Oedegaard, “Early warning signals observed in motor activity preceding mood state change in bipolar disorder,” *Bipolar Disorders*, vol. 26, no. 4, pp. 346–357, 2024.
- [13] A. A. T. Schuurmans, P. de Looft, K. S. Nijhof, C. Rosada, R. H. J. Scholte, A. Popma, and R. Otten, “Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: a comparison to electrocardiography (ecg),” *Journal of Medical Systems*, vol. 44, no. 11, p. 190, 2020.
- [14] N. Milstein and I. Gordon, “Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states,” *Frontiers in Behavioral Neuroscience*, vol. 14, p. 148, 2020.
- [15] U. Cote-Allard, P. Jakobsen, A. Stautland, T. Nordgreen, O. B. Fasmer, K. J. Oedegaard, and J. Torresen, “Long-short ensemble network for bipolar manic-euthymic state recognition based on wrist-worn sensors,” *arXiv preprint arXiv:2107.00710*, 2022.
- [16] S. Föll, M. Maritsch, F. Spinola, V. Mishra, F. Barata, T. Kowatsch, E. Fleisch, and F. Wortmann, “Flirt: A feature generation toolkit for wearable data,” *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106461, 2021.
- [17] C.-T. Wu, M. H. Hsieh, I.-M. Chen, L.-Y. Jhao, D.-S. Liu, S.-M. Wang, C.-T. Wu, and Y.-L. Chien, “Using wearable device and machine learning to predict mood symptoms in bipolar disorder: Development and usability study,” *JMIR Medical Informatics*, vol. 13, p. e66277, 2025.
- [18] J. M. Lipschitz, S. Lin, S. Saghaian, C. K. Pike, and K. E. Burdick, “Digital phenotyping in bipolar disorder: Using longitudinal fitbit data and personalized machine learning to predict mood symptomatology,” *Acta Psychiatrica Scandinavica*, vol. 150, no. 6, pp. 524–537, 2024.

- [19] G. Anmella, F. Corponi, B. M. Li, A. Mas, M. Sanabra, I. Pacchiarotti, G. Fico, *et al.*, “Exploring digital biomarkers of illness activity in mood episodes: Hypotheses generating and model development study,” *JMIR mHealth and uHealth*, vol. 11, p. e45405, 2023.
- [20] INTREPIBD Consortium, “The INTREPIBD project: Identifying digital biomarkers of illness activity and treatment response in bipolar disorder.” <https://intrepibd.github.io>, 2023. Accessed: June 5, 2026.
- [21] V. T. van Hees, S. Sabia, S. E. Jones, A. R. Wood, K. N. Anderson, M. Kivimäki, T. M. Frayling, A. I. Pack, M. Bucan, M. I. Trenell, D. R. Mazzotti, P. R. Gehrman, B. A. Singh-Manoux, and M. N. Weedon, “Estimating sleep parameters using an accelerometer without sleep diary,” *Scientific Reports*, vol. 8, no. 1, p. 12975, 2018.
- [22] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.
- [23] S. Seabold and J. Perktold, “Statsmodels: Econometric and statistical modeling with python,” in *Proceedings of the 9th Python in Science Conference*, pp. 92–96, SciPy, 2010.
- [24] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [25] A. Eto, K. Mochizuki, T. Fukami, W. Sakakibara, and K. Izumi, “Wearable-derived heart rate variability and sleep monitoring as predictors of mood episodes in bipolar disorder: a case report,” *Frontiers in Psychiatry*, vol. 16, p. 1695158, 2025.

Appendix A

Complete Feature List

The feature extraction pipeline produced a total of 212 descriptors for each 5-minute sliding window. For easier readability, standard statistical and non-linear descriptors applied across multiple signal modalities are grouped together in Table A.1. Domain-specific features (HRV, sleep, and custom engineered features) are detailed in Table A.2.

A.1 Standard Statistical and Complexity Descriptors

The FLIRT package calculates a standardized set of 22 descriptors. These were extracted for the following signals: EDA (Tonic and Phasic), ACC (X, Y, Z axes, and L2 magnitude), TEMP, and IBI (HRV).

Table A.1: Standard statistical descriptors applied to EDA, ACC, TEMP, and IBI sequences.

Feature Name	Domain	Description
<code>_mean, _std</code>	Time	Central tendency and standard deviation (dispersion) of the signal.
<code>_min, _max, _ptp</code>	Time	Minimum, maximum, and peak-to-peak (range) values.
<code>_sum, _energy</code>	Time	Sum of all values and total signal energy (sum of squared values).
<code>_skewness, _kurtosis</code>	Stat	Measure of distribution asymmetry and tail descriptor.
<code>_peaks</code>	Time	Number of detected local maxima within the window.
<code>_rms</code>	Time	Root mean square, indicating signal magnitude.
<code>_lineintegral</code>	Time	Sum of absolute differences between consecutive samples (waveform length).
<code>_n_above_mean, _n_below_mean</code>	Stat	Number of samples strictly above or below the window mean.
<code>_n_sign_changes</code>	Time	Number of times the signal crosses the zero-line (or mean-centered zero).
<code>_iqr, _iqr_5_95</code>	Stat	Interquartile range (25th-75th) and the 5th-to-95th percentile range.
<code>_pct_5, _pct_95</code>	Stat	Absolute values at the 5th and 95th percentiles.
<code>_entropy</code>	Non-lin	Shannon entropy, quantifying the information content/randomness.
<code>_perm_entropy</code>	Non-lin	Permutation entropy, measuring the complexity of a dynamic time series.
<code>_svd_entropy</code>	Non-lin	Singular Value Decomposition entropy, indicating dimensionality of the signal.

A.2 Domain-Specific and Custom Features

Specific physiological domains and custom engineered data require dedicated feature descriptors, detailed below.

Table A.2: Domain-specific and custom engineered features extracted per window.

Feature Base Name	Domain	Description
Heart Rate Variability (HRV) - Time Domain		
num_ibis	Time	Total count of valid Inter-Beat Intervals (IBIs).
hrv_mean_nni, _median_nni	Time	Mean and median of Normal-to-Normal (NN) intervals.
hrv_range_nni	Time	Range (max - min) of the NN intervals.
hrv_sdsd, hrv_rmssd	Time	Std. dev. of differences between adjacent NNs, and Root Mean Square of Successive Differences (parasympathetic proxy).
hrv_nni_50, hrv_pnni_50	Time	Number and percentage of successive NNs differing by > 50ms.
hrv_nni_20, hrv_pnni_20	Time	Number and percentage of successive NNs differing by > 20ms.
hrv_cvsd, hrv_cvnni	Time	Coefficient of variation of successive differences and of NN intervals.
hrv_sdn	Time	Standard deviation of all NN intervals.
hrv_mean_hr, _min_hr, _max_hr, _std_hr	Time	Mean, min, max, and standard deviation of the Heart Rate (converted to bpm).
Heart Rate Variability (HRV) - Frequency Domain		
hrv_total_power	Freq	Total spectral power of the IBI sequence.
hrv_vlf, hrv_lf, hrv_hf	Freq	Power in Very Low (VLF), Low (LF), and High Frequency (HF) bands calculated via Lomb-Scargle periodogram.
hrv_lf_hf_ratio	Freq	Ratio of LF to HF power (sympathovagal balance proxy).
hrv_lfnu, hrv_hfnu	Freq	LF and HF power measured in normalized units.
Autoregressive Dynamics (Custom)		
eda_ar_1, _2, _3	Time	1st, 2nd, and 3rd order coefficients of an AR(3) model fitted to the 1Hz EDA signal.

Continued on next page

Table A.2 – *Continued from previous page*

Feature Name	Domain	Description
acc_ar_1, _2, _3	Time	1st, 2nd, and 3rd order coefficients of an AR(3) model fitted to the 1Hz ACC magnitude.
Circadian Encoding (Custom)		
hour_sin, hour_cos	Time	Cyclical sine and cosine transformations of the decimal hour of the day.
Sleep History Metrics (Custom)		
sleep_past_24h_hours	Time	Sum of sleep hours in the strictly preceding 24 hours.
sleep_past_24h_padded_hours	Time	24h rolling sleep sum with historical padding to solve cold-start constraints.
waso_mins	Time	Wake After Sleep Onset (in minutes) during the current primary sleep period.
awakenings_count	Time	Number of distinct awakening events registered.

Appendix B

Additional Results

B.1 Raw Signal Correlation Results

This appendix reports the complete subject-level results obtained from the raw signal comparability analysis between the E4 and EP, both for the raw signal comparison on the full recording and for the same analysis when restricted to sleep time data

For each subject and modality, the optimal synchronization lag, the final correlation coefficient after lag application and the RMSE are reported.

Table B.1: Complete subject-level raw signal comparison results across all modalities.

Subject	EDA			ACC			IBI			TEMP		
	Lag	r	RMSE	Lag	r	RMSE	Lag	r	RMSE	Lag	r	RMSE
CLIDEM01	-56	0.275	0.687	-114	0.704	0.320	-0.600	0.645	0.064	-91	0.377	2.506
CLIDEM02	-194	0.382	0.417	-196	0.534	0.440	0.800	0.826	0.053	-125	0.960	1.504
CLIDEM030	188	0.875	6.246	-128	0.476	0.490	1.000	0.738	0.052	80	0.916	1.361
CLIDEM031	-24	0.709	0.980	-6	0.640	0.265	-0.100	0.830	0.054	-360	0.971	2.466
CLIDEM05	-167	0.890	4.570	-148	0.677	0.392	0.700	0.842	0.061	-289	0.906	1.998
CLIDEM06	-152	0.713	0.516	-128	0.710	0.342	0.700	0.764	0.068	-117	0.847	1.834
CLIDEM07	-71	0.771	1.352	-103	0.201	0.476	-0.050	0.782	0.062	-38	0.941	2.354
CLIDEM08	42	0.293	1.156	-102	0.253	0.392	-0.800	0.732	0.046	-154	0.708	3.030
CLIDEM09	5	0.156	2.576	-363	-0.015	0.623	-1.000	0.840	0.075	501	0.467	3.761
CLIDEM11	-182	0.650	1.054	-81	0.526	0.547	-0.050	0.908	0.066	-158	0.874	0.947
CLIDEM12	180	0.702	4.299	-127	0.488	0.433	0.400	0.856	0.043	-62	0.962	1.461
CLIDEM13	8	0.512	3.106	-88	0.597	0.595	-0.350	0.922	0.064	-80	0.784	0.852
CLIDEM14	-196	0.732	5.289	-125	0.615	0.419	-0.400	0.908	0.075	-36	0.824	1.479
CLIDEM15	-30	0.656	3.092	-115	0.707	0.518	-0.600	0.922	0.064	0	0.883	1.996
CLIDEM16	8	0.592	10.597	-113	0.574	0.416	0.900	0.946	0.048	-19	0.710	1.633
CLIDEM17	-69	0.753	7.067	-172	0.136	0.433	0.300	0.822	0.044	-164	0.656	1.790
CLIDEM18	-87	0.813	2.368	-134	0.397	0.654	0.950	0.941	0.065	-44	0.906	0.914
CLIDEM19	-128	0.683	1.580	-121	0.716	0.376	1.000	0.859	0.057	0	0.592	2.429
CLIDEM20	192	0.540	1.790	-53	0.602	0.280	0.850	0.873	0.050	-82	0.198	3.806
CLIDEM21	-128	0.833	1.659	-147	0.650	0.363	0.000	0.915	0.057	-210	0.813	2.122
CLIDEM22	-238	0.800	0.859	-123	0.596	0.550	-0.450	0.882	0.051	-244	0.888	1.402
CLIDEM23	-5	0.743	1.677	-143	0.542	0.511	-0.700	0.877	0.062	-105	0.852	1.736
CLIDEM24	-228	0.754	8.228	-138	0.647	0.520	0.500	0.911	0.069	-71	0.806	2.327
CLIDEM25	-2	0.518	1.301	-123	0.734	0.359	-0.100	0.855	0.064	-103	0.941	1.793
CLIDEM26	-127	0.404	5.292	-128	0.532	0.240	-0.950	0.906	0.043	0	0.970	1.270
CLIDEM27	299	0.714	2.899	-113	0.679	0.509	-0.350	0.871	0.043	-41	0.918	1.678
CLIDEM29	-61	0.628	7.062	-128	0.271	0.418	0.950	0.870	0.047	-93	0.857	1.893

Table B.2: Complete subject-level raw signal comparison results across all modalities, restricted to signal recorded during sleep.

Subject	EDA			ACC			IBI			TEMP		
	Lag	r	RMSE	Lag	r	RMSE	Lag	r	RMSE	Lag	r	RMSE
CLIDEM01	-80	0.432	0.334	-86	0.910	0.265	1.00	0.561	0.063	0	0.668	0.621
CLIDEM02	-130	0.829	0.542	-125	0.856	0.308	-1.00	0.630	0.053	-180	0.879	1.091
CLIDEM030	196	0.719	15.847	-69	0.527	0.698	0.35	0.717	0.048	59	0.712	1.458
CLIDEM031	-60	0.863	1.504	-60	0.938	0.245	-0.35	0.788	0.055	0	0.389	1.523
CLIDEM05	-254	0.167	0.230	-237	0.916	0.201	-0.35	0.718	0.060	-250	0.575	1.112
CLIDEM06	-88	0.875	0.092	-88	0.988	0.091	-1.00	0.581	0.070	0	0.491	0.653
CLIDEM07	-184	0.890	2.513	300	-0.426	0.598	0.35	0.684	0.062	0	0.644	1.202
CLIDEM08	-73	0.784	0.067	-65	0.437	0.376	-0.85	0.686	0.042	-197	0.697	2.056
CLIDEM09	211	-0.064	4.704	300	-0.154	0.851	1.00	0.702	0.090	0	-0.324	1.254
CLIDEM11	-39	0.938	0.208	-43	0.922	0.387	-0.05	0.688	0.066	0	0.406	0.754
CLIDEM12	1	0.756	7.526	-48	0.781	0.280	0.55	0.802	0.041	0	0.830	1.138
CLIDEM13	-257	0.747	0.028	300	-0.415	1.139	0.00	0.911	0.032	0	0.968	0.275
CLIDEM14	-89	0.759	2.672	-202	0.691	0.379	0.00	0.740	0.081	0	0.384	1.046
CLIDEM15	-63	0.856	2.752	-58	0.979	0.184	0.90	0.576	0.072	0	0.634	1.709
CLIDEM16	-83	0.745	13.175	-67	0.787	0.419	0.80	0.920	0.044	-82	0.305	1.174
CLIDEM17	301	0.965	7.543	300	-0.474	0.543	0.95	0.839	0.040	0	0.767	0.704
CLIDEM18	-299	0.758	0.046	-222	0.961	0.384	0.15	0.579	0.070	0	0.788	0.396
CLIDEM19	-83	0.807	0.769	-205	0.933	0.181	-0.75	0.676	0.066	0	0.638	1.091
CLIDEM20	-172	0.234	0.374	301	0.211	0.265	0.00	NaN	0.044	0	0.382	8.759
CLIDEM21	-230	0.905	0.678	-191	0.959	0.193	-0.90	0.632	0.074	0	0.545	1.154
CLIDEM22	-89	0.950	1.384	-102	0.810	0.442	0.60	0.717	0.052	0	0.821	0.873
CLIDEM23	300	-0.260	1.528	-299	0.834	0.277	0.40	0.682	0.068	-290	0.856	0.521
CLIDEM24	1	0.899	7.879	-65	0.924	0.264	0.40	0.776	0.069	0	0.629	2.206
CLIDEM25	-74	0.967	0.570	-71	0.959	0.185	-0.95	0.759	0.064	0	0.708	1.100
CLIDEM26	-183	0.782	3.992	-185	0.749	0.187	-1.00	0.366	0.044	0	0.923	0.564
CLIDEM27	301	0.788	3.714	-73	0.958	0.223	-1.00	0.806	0.042	0	0.474	1.765
CLIDEM29	-299	0.220	8.052	-188	0.746	0.290	-0.15	0.670	0.056	0	0.681	1.346

B.2 Predictions Visualizations

Figure B.1 contains the prediction plots from the Mania cohort model.

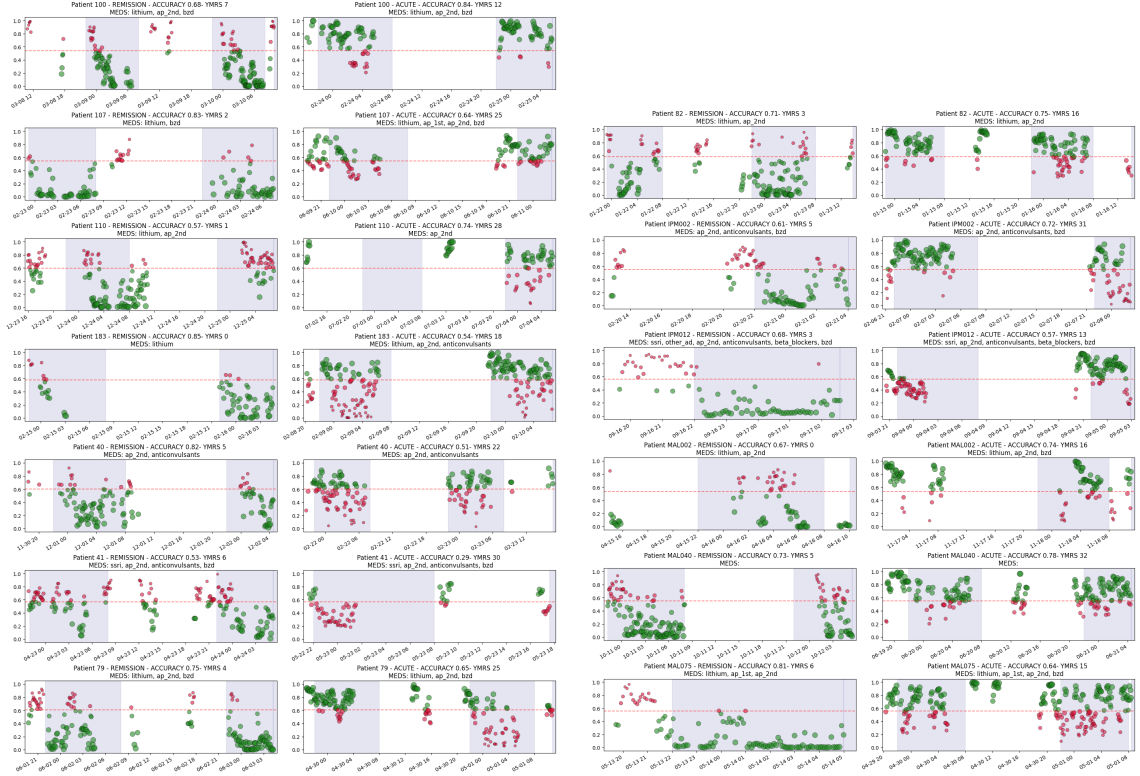


Figure B.1: Plots of sample predictions over time for the Mania cohort model.

For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.

Figure B.2 contains all the prediction plots from the Depression Increaser model.

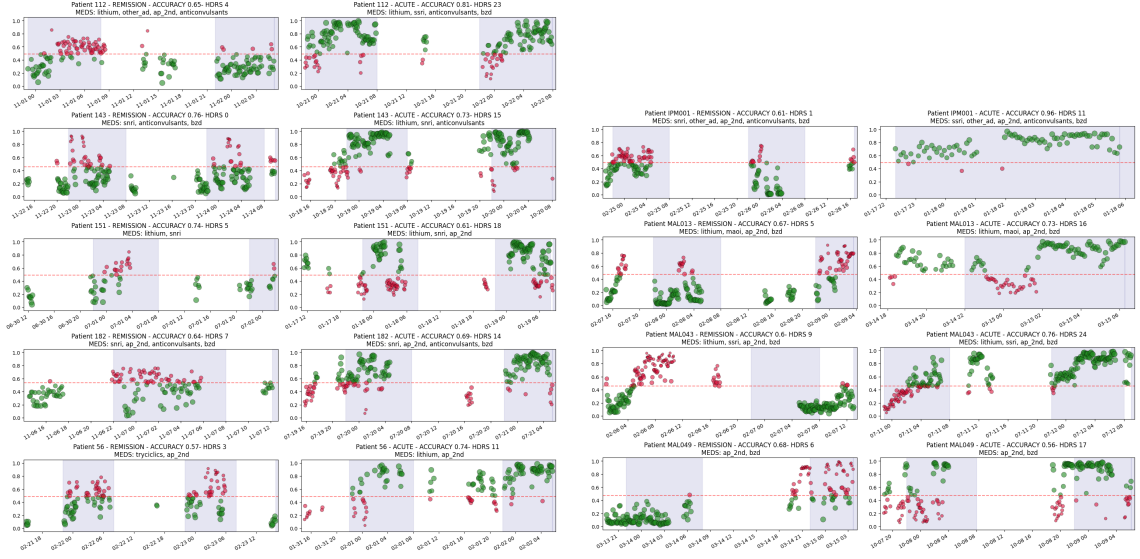


Figure B.2: Plots of sample predictions over time for the Depression Increaser sub-cohort model.

For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.

Figure B.3 contains all the prediction plots from the Depression Dropper model.

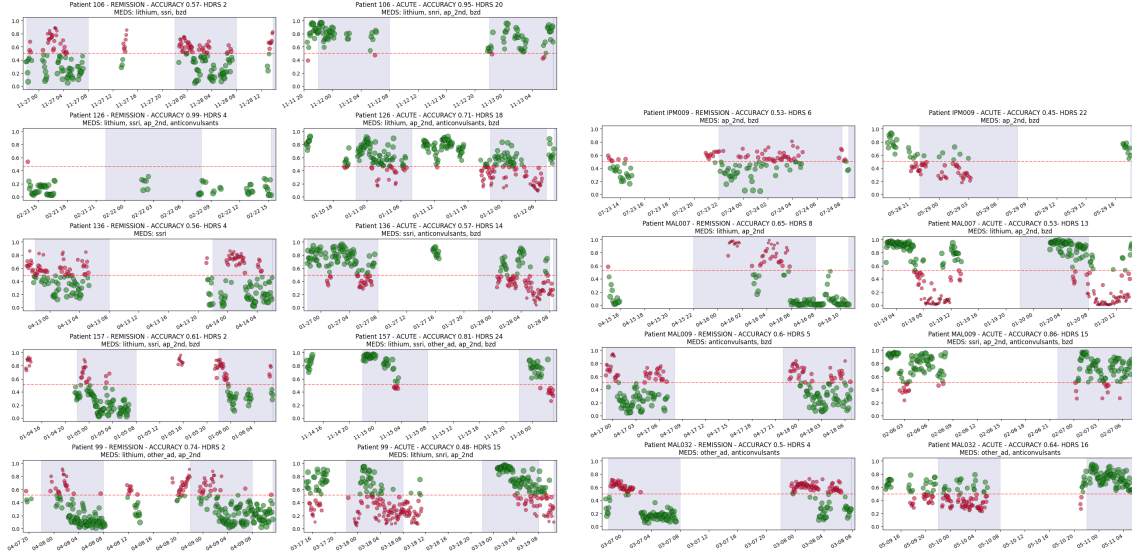


Figure B.3: Plots of sample predictions over time for the Depression Dropper sub-cohort model.

For each subplot the predictions for a single session are displayed. Each dot represents the prediction of a single sample, plotted at the time of sampling on the x axis, with the confidence of acute prediction on the y axis. The dots are colored green if accurately classified, red otherwise, while the dotted red line marks the threshold used for that fold. The shadowed part of the graph corresponds with night time, to evaluate the impact of circadian cycles on the predictions.