National College of Ireland

A Deterministic N-of-1 Pipeline for Multimodal Digital Phenotyping in ADHD and Bipolar Disorder

MSc Practicum (Part 2)
MSc Artificial Intelligence for Business

Rodrigo Marques Teixeira
Student ID: 24130664

School of Computing
National College of Ireland

Supervisor: Dr. Agatha Mattos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rodrigo Marques Teixeira
Student ID:	24130664
Programme:	MSc Artificial Intelligence for Business
Year:	2026
Module:	MSc Practicum (Part 2)
Supervisor:	Dr. Agatha Mattos
Submission Due Date:	28/01/2026
Project Title:	A Deterministic N-of-1 Pipeline for Multimodal Digital Phenotyping in ADHD and Bipolar Disorder
Word Count:	10,530
Page Count:	46
Academic Integrity Certificate Number:	0Jj87XQvZ7

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rodrigo Marques Teixeira
Date:	24th January 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deterministic N-of-1 Pipeline for Multimodal Digital Phenotyping in ADHD and Bipolar Disorder

Rodrigo Marques Teixeira
24130664

Abstract

Attention-Deficit/Hyperactivity Disorder (ADHD) and Bipolar Disorder (BD) are characterised by marked intra-individual variability in sleep, activity and cardiovascular signals. Digital phenotyping offers a promising opportunity to capture such temporal patterns, but existing work often lacks deterministic pipelines, transparent heuristics, leakage-resistant evaluation, and long-term N-of-1 designs.

This research presents a fully deterministic, ten-stage N-of-1 pipeline operating over eight years of multimodal wearable data. State of Mind (SoM) — a user-reported three-class measure of daily subjective stability exported from Apple Health — serves as the primary supervised target. The Personal Behaviour Stability Index (PBSI), a transparent physiological composite derived from sleep, heart-rate and activity metrics, is incorporated only as an auxiliary feature. The modelling framework includes: (i) ML6 baseline with Logistic Regression under calendar-based cross-validation; (ii) ML7 baseline with LSTM for sequence modelling; (iii) ML6-Extended comparing ten classical algorithms (Random Forest, XGBoost, LightGBM, Gradient Boosting, SVM, Gaussian Naive Bayes, KNN); and (iv) ML7-Extended evaluating four deep learning architectures (LSTM, GRU, Conv1D, CNN-LSTM). All stages incorporate strict anti-leak safeguards, drift detection (ADWIN, KS) and SHAP interpretability.

Although the full dataset spans 2017–2025, predictive modelling was restricted to 77 SoM-labelled days (2024–2025) due to sparse user-reported labels. Results show that Random Forest achieved the best classical performance (macro-F1 $\approx$ 0.52), while LSTM was the best sequence model (macro-F1 $\approx$ 0.50). All models showed moderate performance, reflecting the challenging nature of predicting user-reported stability from sparse labels.

These findings demonstrate the feasibility of predicting a user-defined behavioural stability construct (SoM) under strict anti-leak conditions. The deterministic pipeline may serve as a template for future N-of-1 digital phenotyping studies with reproducible and transparent analytical workflows.

1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) and Bipolar Disorder (BD) are chronic neuropsychiatric conditions associated with substantial functional impairment, high comorbidity and heterogeneous clinical trajectories. When ADHD and BD co-occur in the same individual, symptom patterns can fluctuate markedly over days, weeks and years, driven by sleep disruption, activity changes, medication effects and environmental

stressors. Traditional assessment methods—such as infrequent clinical visits and retrospective questionnaires—offer only coarse snapshots of these dynamics and are prone to recall bias. As a result, clinicians often lack timely, fine-grained information to distinguish transient instability from meaningful change in a patient's baseline.

Digital phenotyping, defined as the continuous quantification of individual-level behaviour using data from personal devices [5, 6], has emerged as a promising approach for capturing real-world physiological and behavioural dynamics. Recent studies have leveraged smartphone sensors and wearable devices [8, 12] to monitor mood, sleep, activity and physiology in ADHD and BD [13, 14], showing that passive time-series signals can approximate clinical trajectories and help predict relapse or functional decline. Nevertheless, three barriers limit the translation of these findings into practice: (i) heterogeneous data formats across device vendors, (ii) opaque preprocessing and modelling pipelines that are difficult to reproduce, and (iii) data-leakage risks when temporal dependencies are not handled explicitly. These issues are particularly problematic in N-of-1 settings [7], where a single participant's data must support both model training and evaluation over long time horizons.

This project addresses these challenges by reconstructing eight years of multimodal sensor data from Apple Health and the Amazfit/Zepp ecosystem for one participant with ADHD + BD, and by designing a deterministic end-to-end pipeline that can be re-executed from raw exports. The unified dataset comprises 2,868 valid days between December 2017 and December 2025, combining daily sleep, heart-rate, heart-rate variability (HRV) and activity metrics [9, 10]. The workflow is organised into ten deterministic stages (0–9): (0) ingestion of raw XML/ZIP archives; (1) daily aggregation of Apple and Amazfit/Zepp streams; (2) unification into a canonical feature set; (3) heuristic labelling via the Personal Behaviour Stability Index (PBSI) and integration of State of Mind (SoM) labels from Apple Health AutoExport; (4) automatic segmentation based on temporal gaps and calendar boundaries; (5) Experiment Suite for controlled feature-set ablation (FS-A through FS-D) and deterministic model selection, producing anti-leak feature filtering with segment-wise normalisation; (6) ML6-Extended modelling using 10 classical algorithms [1] with calendar-based cross-validation; (7) ML7-Extended analysis with 4 sequence architectures [2], SHAP explainability [3] and drift detection [4, 17]; (8) TensorFlow Lite export; and (9) automated reporting.

To reduce leakage and improve interpretability, the pipeline normalises physiological and behavioural features separately within each detected temporal segment, instead of across the full eight-year timeline. The primary supervised target is State of Mind (SoM), a user-reported three-class label exported from Apple Health, capturing daily subjective stability (Negative/Unstable, Neutral, Positive/Stable). PBSI rules that combine sleep quality, activity balance and heart-rate features serve as an auxiliary physiological composite, enabling multi-signal modelling without relying solely on self-report. The modelling framework compares 10 classical algorithms (ML6-Extended) with 4 deep learning architectures (ML7-Extended), and combines ADWIN and Kolmogorov–Smirnov tests with SHAP explanations [3] to quantify when and why model performance changes over time.

Research question. To what extent can a fully deterministic N-of-1 pipeline built from consumer wearable data provide reproducible and interpretable predictions of *user-reported State of Mind (SoM)* over an eight-year observational period?

The emphasis of this study is methodological: the pipeline predicts a self-defined behavioural-stability construct (SoM), not clinically validated episodes or symptom sever-

ity.

Aims and objectives. To answer this question, the study pursues four specific objectives:

1. Design a deterministic, segment-aware ETL pipeline that reconstructs a long-term multimodal dataset from Apple and Amazfit/Zepp exports, including HRV and medication features.
2. Implement a dual-target labelling strategy: State of Mind (SoM) as the primary supervised target and PBSI as an auxiliary physiological composite, mapping daily sleep, activity, cardiovascular and HRV signals into interpretable features.
3. Compare 10 classical algorithms (ML6-Extended) with 4 deep learning architectures (ML7-Extended) under strict calendar-based cross-validation, enforcing anti-leak safeguards through controlled feature-set ablation (FS-A through FS-D).
4. Integrate drift detection and SHAP-based explanations to characterise how feature importance and data distributions evolve over time.

Contributions. The main contributions of this project are:

- a fully documented, re-runnable N-of-1 pipeline for multimodal digital phenotyping, from raw device exports to TensorFlow Lite deployment;
- a public ETL snapshot (1,671 days) enabling external reproduction of stages 3–9 without requiring access to private health data;
- a dual-target paradigm combining SoM (user-reported) and PBSI (physiological composite) for multi-signal behavioural modelling tailored to ADHD + BD comorbidity;
- an Experiment Suite for controlled ablation across four feature sets (FS-A through FS-D, 10–19 features) with deterministic model selection;
- an empirical comparison of 10 classical algorithms (ML6-Extended) and 4 deep learning architectures (ML7-Extended) under realistic temporal cross-validation; and
- an integrated drift-and-explainability analysis combining ADWIN, Kolmogorov–Smirnov tests and SHAP values for long-term monitoring.

The remainder of this paper is organised as follows. Section 2 reviews related work on passive sensing for ADHD and BD and concept drift in mobile health. Section 3 describes the data sources and pipeline methodology in detail. Section 4 presents the experimental results for ML6 and ML7, including drift and feature-importance analyses, and section 4.4 reports on extended modelling experiments (stage 8) comparing additional tabular and sequence architectures. Section 5 discusses the implications, limitations and ethical considerations of the proposed approach, and Section 6 concludes the paper and outlines directions for future work.

2 Literature Review

Digital phenotyping has emerged as a promising paradigm for capturing behavioural, cognitive and physiological patterns using data from smartphones and wearable devices. This section reviews the key areas relevant to this study: (i) passive sensing for neuropsychiatric disorders, particularly ADHD and BD; (ii) methodological challenges in long-term N-of-1 studies; (iii) the role of feature engineering and weak supervision in behavioural labelling; and (iv) the need for reproducible, deterministic pipelines with concept-drift

monitoring. The goal is not only to summarise existing work but to identify the specific gaps that justify the present research.

2.1 Passive Sensing for ADHD and Bipolar Disorder

Wearable and mobile sensing has significantly advanced the ability to monitor neuropsychiatric conditions in real-world settings [8]. For ADHD, studies have shown that metrics such as resting heart rate, physical activity regularity and sleep duration correlate with symptom severity and executive function variability [13]. For BD, fluctuations in sleep quality, circadian rhythm disruption and changes in mobility patterns have been repeatedly associated with manic or depressive episodes [11, 14]. Recent work has also demonstrated the potential of mobile health (mHealth) interventions for individuals with serious mental illness [19] and the utility of behavioural indicators derived from passive sensing to predict psychiatric symptoms [16].

While these findings indicate that passive data streams are informative, most existing works rely on short-term datasets (weeks to months), inconsistent sensor modalities, or proprietary pipelines that lack transparency. Furthermore, few studies examine comorbid ADHD + BD, a condition characterised by high intra-individual variability and rapid shifts in energy, sleep and motivation. This limits the external validity and clinical translation of published results. Existing research rarely addresses the challenges posed by multi-year data, heterogeneous device ecosystems (such as Apple Health vs. Amazfit/Zepp), and the need for stable behavioural indicators over long horizons. Moreover, concerns about the accuracy and reliability of wearable sensors remain significant [9, 10], particularly for optical heart-rate measurements and energy expenditure estimates.

2.2 N-of-1 Designs and Their Challenges

N-of-1 methodological frameworks [7] emphasise the analysis of longitudinal data from individual participants rather than group-level inference. They are particularly suitable for conditions like ADHD + BD, where symptoms vary substantially between individuals and fluctuate over time. Despite their theoretical advantages, N-of-1 studies present three practical challenges.

First, multi-year datasets evolve over time and exhibit substantial distributional changes: sensor availability, usage patterns and health behaviours change across life events, routines and device upgrades. Second, the choice of windowing, segmentation and normalisation strategies can heavily influence model performance and may inadvertently introduce leakage, especially when the same temporal patterns appear in both training and validation segments. Third, most N-of-1 works depend on custom, researcher-specific pipelines that cannot be fully reproduced by other researchers, limiting scientific scrutiny and comparison across studies.

These challenges highlight the need for deterministic, segment-aware pipelines that can be re-executed from raw data and evaluated across clearly defined behavioural epochs.

2.3 Feature Engineering and Weak Supervision

A significant limitation in digital phenotyping is the scarcity of high-frequency ecological momentary assessment (EMA) labels. This has motivated the use of weakly supervised

or heuristic labelling strategies, including the reconstruction of behavioural indices from sleep, heart-rate, movement and screen-use metrics.

Composite behavioural indices—such as sleep regularity scores [15], circadian stability measures or activity balance indicators—have been proposed as potential correlates of self-reported wellbeing and symptom fluctuations. However, such indices are often defined ad hoc, without formal documentation or reproducibility, and are not consistently normalised within behaviourally homogeneous periods. Moreover, important methodological details—such as how raw sensor data are aggregated, which features are excluded to prevent leakage, or which statistical thresholds define categorical labels—are typically omitted from published work.

The Personal Behaviour Stability Index (PBSI) proposed in this study follows this tradition of composite indices but extends it by providing a transparent, reproducible set of heuristic rules applied to multi-year data, combined with segment-wise normalisation across automatically detected behavioural periods.

2.4 Machine Learning for Longitudinal Prediction

Predictive modelling in digital phenotyping frequently uses linear models (e.g., logistic regression) or recurrent neural networks (e.g., LSTM [2], GRU). Linear models offer interpretability and robustness under controlled leakage conditions, while LSTMs can capture temporal dependencies over longer windows. Ensemble methods such as Random Forests [1] have also proven effective for tabular time-series data. However, sequence models often overfit when (i) labels are weak or noisy, (ii) data show strong regime changes, or (iii) the temporal structure is unstable.

Recent empirical results across passive sensing literature show that simple baselines often outperform deep sequence models under strict temporal cross-validation, particularly in N-of-1 settings with moderate label noise. This reinforces the need to evaluate both model classes under rigorous calendar-based train–validation splits rather than random or block-based splits, which may leak future information into past predictors.

Furthermore, the broader literature emphasises the importance of model explainability for clinical adoption. Techniques such as SHAP [3] have become common for quantifying feature contribution and for identifying when model behaviour diverges from expected physiological patterns. Yet many studies apply SHAP without monitoring for distribution shifts, creating uncertainty about whether changes in feature importance reflect real physiological changes or data drift.

2.5 Concept Drift in Mobile and Wearable Health Data

Wearable-derived data are highly susceptible to concept drift [17]. Device firmware updates, seasonal routines, lifestyle changes and technological replacements introduce distributional shifts that degrade model performance if left unmonitored. Drift detection methods such as ADWIN [4] and Kolmogorov–Smirnov tests have been proposed for mobile-sensing applications, but are rarely integrated into full end-to-end pipelines. Time-series forecasting literature [18] also highlights the importance of robust performance estimation methods under temporal distribution shifts. Additionally, drift detection is typically performed retrospectively rather than as part of a reproducible workflow that runs from ingestion to reporting.

The literature consistently highlights that failure to model drift leads to misleading performance estimates and incorrect clinical inferences. For example, a classifier trained during a period of stable sleep patterns may perform poorly following a change in shift work or medication, yet traditional evaluation setups cannot detect this without explicit drift analysis.

2.6 Identified Gaps

From this review, five gaps remain insufficiently addressed:

1. **Lack of deterministic pipelines:** Most studies rely on bespoke, unreproducible data-processing steps.
2. **Limited coverage of ADHD + BD comorbidity:** Virtually no long-term N-of-1 studies combine these two conditions using multimodal wearables.
3. **Weak supervision without transparency:** Composite indices are used but rarely documented or normalised across behavioural segments.
4. **Leakage-prone evaluation:** Many works do not enforce calendar-based cross-validation or segment-wise normalisation.
5. **Absence of drift-integrated modelling:** Drift detection is seldom incorporated into full pipelines with interpretability tools such as SHAP.

2.7 Positioning of This Study

The present work addresses these gaps by providing a fully documented, reproducible and segment-aware N-of-1 pipeline for digital phenotyping across eight years of data. It integrates deterministic ETL, heuristic PBSI labelling, anti-leak safeguards, interpretable modelling and explicit drift monitoring. This positions the study as a methodological contribution rather than a purely predictive one, and lays the groundwork for future multi-participant analyses under similar constraints.

3 Methodology

This section details the methodological framework used to design, implement and validate a deterministic N-of-1 pipeline for multimodal digital phenotyping in ADHD + Bipolar Disorder. In accordance with best practice for scientific reporting, the methodology describes the data sources, preprocessing workflow, segmentation strategy, labelling rules, modelling setup, evaluation protocol, and drift analysis procedures. All steps were implemented in a fully reproducible pipeline executed from raw Apple Health and Amazfit/Zepp exports.

3.1 Research Design

The study follows an N-of-1 longitudinal observational design. The goal is not to generalise findings to a population but to characterise intra-individual SoM-defined behavioural stability over time. This design is appropriate for ADHD + BD comorbidity, where symptoms vary substantially between individuals and fluctuate over time.

The analytical workflow comprises ten deterministic stages (0–9):

0. ingestion of raw Apple Health and Amazfit/Zepp archives;

1. daily aggregation of heterogeneous XML/CSV sensor streams;
2. unification into a canonical per-day dataset;
3. heuristic labelling using the Personal Behaviour Stability Index (PBSI);
4. segmentation into behavioural periods;
5. anti-leakage feature filtering for ML6;
6. baseline modelling (ML6);
7. sequence modelling and explainability (ML7);
8. extended modelling experiments;
9. automated reporting and provenance generation.

Each stage is implemented as a deterministic transformation from well-defined inputs to outputs, ensuring that the entire pipeline can be re-executed from raw files with identical results.

3.2 Data Sources

Apple Health (iOS). Apple Health provides high-quality physiological and behavioural metrics via the `export.xml` file. The following domains were used:

- Heart rate: instant and resting heart rate samples (1–5 s intervals).
- Sleep: sleep stages, time in bed, time asleep.
- Activity: steps, distance, energy expenditure.

The XML export contained approximately 18 million individual records across eight years.

Amazfit/Zepp ecosystem. The Amazfit GTR wearable and the Zepp OS application provide:

- heart-rate and HRV summaries,
- sleep duration and sleep-stage breakdown,
- step count and activity-type logs.

Data were extracted from the Zepp CSV export folders: `SLEEP/`, `HEARTRATE/`, and `ACTIVITY/`. Together, these sources provide complementary but heterogeneous information on sleep, autonomic activity and daily movement patterns, requiring explicit unification rules in the ETL pipeline.

3.3 ETL Pipeline and Data Processing

The end-to-end ETL pipeline was implemented in Python and follows a deterministic, Makefile-driven workflow to ensure reproducibility across multiple executions and snapshots. Raw multi-sensor data from Apple HealthKit (iOS) and Zepp/Amazfit devices are transformed through staged extraction, unification, enrichment, labeling, and quality checking.

In the updated architecture, all daily HR metrics are now computed directly from event-level measurements (`HKQuantityTypeIdentifierHeartRate`) and safely cached in Parquet format using the canonical schema: mean (`apple_hr_mean`), minimum (`apple_hr_min`), maximum (`apple_hr_max`), standard deviation (`apple_hr_std`), and sample count (`apple_n_hr`).

No fallback or fabrication occurs at cache load time. Instead, a deterministic mapping is used when exporting daily summaries to CSV. This ensures that cached and non-cached executions produce bit-identical results, enabling scientifically valid longitudinal analysis.

To further strengthen data integrity, biologically implausible HR values (e.g. below 30 bpm or above 220 bpm) are filtered at aggregation time for both Apple and Zepp in-

puts. This protects daily minima, maxima, and standard deviations from device artefacts such as HR=0 spikes.

All extracted domains (cardiovascular, activity, sleep) converge into a unified daily table through a vendor-agnostic merging routine. This step harmonises column names, resolves source priority (Apple > Zepp when both are available), and preserves missingness explicitly without forward-filling. The final output, `features_daily_unified.csv`, serves as the canonical input for label construction, z-scoring, drift detection, and predictive modelling.

3.4 Data Aggregation Pipeline

Daily physiological features were extracted from heterogeneous sources (Apple Health XML and Zepp CSV) through deterministic parsing routines. Heart rate, sleep, and activity metrics were aggregated into canonical daily summaries. Full parsing specifications, schema mappings, and file format details are provided in Appendix F (§F.9).

3.5 Heart-Rate Feature Engineering

Five daily HR metrics (mean, min, max, std, sample count) were computed from raw PPG samples, cached in Parquet format, and filtered for biologically plausible ranges. Caching ensures reproducibility across pipeline executions. Full schema, caching logic, and outlier removal thresholds are specified in Appendix F (§F.11).

3.6 PBSI Labelling Strategy

The Personal Behaviour Stability Index (PBSI) is a heuristic, non-validated physiological composite designed for interpretable multi-signal modelling in long-term N-of-1 studies. It is not a clinical scale and should not be interpreted as a diagnostic instrument. The implementation is fully deterministic and aligns with the Python module `src/labels/build_pbsi.py`, which applies segment-wise normalisation and domain-specific composite scores.

Important clarification. In the updated SoM-centric pipeline (v4.3.1), PBSI serves as an **auxiliary feature** rather than the primary supervised target. The primary target for ML6/ML7 modelling is State of Mind (SoM), a user-reported three-class label exported from Apple Health AutoExport (see Section 3.7). PBSI remains valuable as an interpretable physiological proxy that enables multi-signal modelling and provides clinical context for the user-reported SoM labels.

PBSI construction follows three stages: (1) segment-wise z-score normalisation, (2) domain-level subscores with inverted formulas, and (3) percentile-based thresholding into a three-class stability index.

3.6.1 Segment-wise Normalisation (Anti-Leak Safeguard)

All physiological and behavioural features are normalised *within* behavioural segments rather than across the full eight-year timeline. For each segment s, z-scores are computed as:

$$z = \frac{x - \mu_s}{\sigma_s},$$

where μ_s and σ_s denote the mean and standard deviation of feature values within segment s. This ensures strict temporal isolation: future data do not influence past normalisation, and long-term trends (e.g. seasonal shifts, device changes) cannot be mistaken for within-day variability. Segment-wise z-scoring is therefore the central anti-leakage safeguard of the entire modelling pipeline.

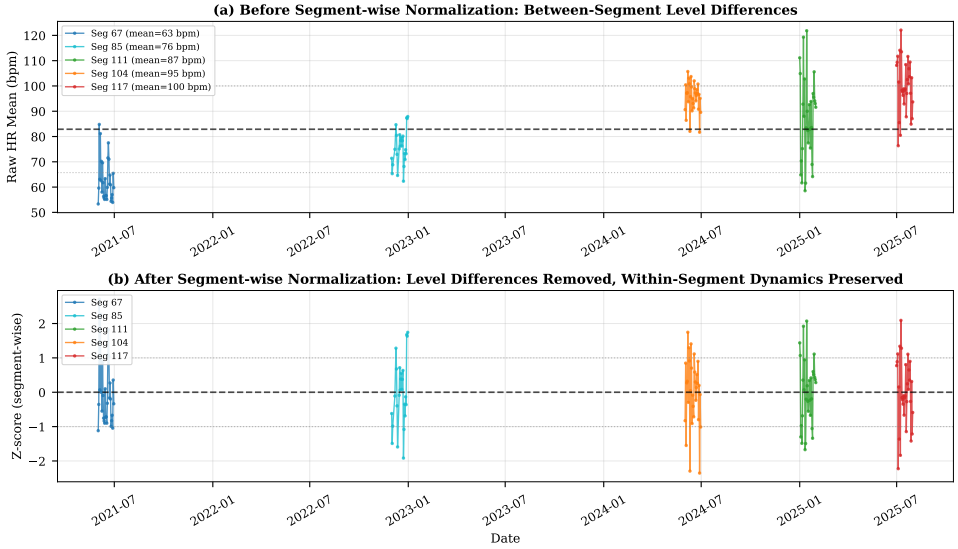

Figure 1: Segment-wise z-score normalization applied to daily mean heart rate across five representative segments (40 segments with ≥20 days HR data). (a) Raw HR values (62–101 bpm range). (b) Normalized values centered to zero mean with unit variance.

Figure 1 highlights how segment-wise normalization removes global level offsets introduced by device and baseline differences, while preserving the original within-segment temporal structure. This enables downstream models to operate on a scale-invariant representation without altering the underlying behavioural dynamics.

3.6.2 Domain-Level Subscores

PBSI combines three domain subscores:

- **Sleep**: weighted combination of duration and efficiency
- **Cardiovascular**: mean HR (inverted), HRV proxy (RMSSD), max HR (inverted)
- **Activity**: steps and exercise proxy

The composite score uses domain-specific weights (sleep: 0.40, cardio: 0.35, activity: 0.25) and percentile-based thresholds (P25/P75) to produce three-class labels. Complete formulas, weight rationale, and threshold computation are detailed in Appendix F (§F.12).

Sign convention rationale: All formulas are designed so that **higher subscores indicate better regulation**. For example, higher sleep duration increases the sleep subscore, and higher HRV increases the cardiovascular subscore. This aligns PBSI with clinical intuition: higher scores correspond to stable, well-regulated days.

Interpretation: Higher PBSI scores indicate *better* physiological regulation (stable sleep, high HRV, adequate activity), while lower scores indicate dysregulation or instability. This intuitive sign convention improves clinical interpretability and aligns with standard health metrics where higher values generally denote better outcomes.

3.7 State of Mind (SoM) Labels

State of Mind (SoM) provides the primary supervised target for ML6/ML7 modelling in the updated pipeline (v4.3.1). Unlike PBSI, which is derived algorithmically from physiological signals, SoM represents user-reported subjective stability exported from Apple Health via the Apple Health AutoExport mechanism.

State of Mind (SoM) labels were extracted from Apple Health AutoExport and aggregated into a three-class daily target (Negative/Unstable, Neutral, Positive/Stable). Coverage is sparse (77 days) due to voluntary self-reporting. Deterministic extraction and class-mapping logic are documented in Appendix F (§F.13).

3.8 Behavioural Segmentation

Behavioural segmentation partitions the eight-year timeline into internally homogeneous periods that reflect stable routines, physiological baselines and device usage patterns. Segments are used exclusively for normalisation and anti-leak protection; they do not participate directly in modelling or label computation. The final deterministic method used in this study combines two chronological rules applied in fixed order, matching the implementation in `stage_apply_labels.py` for the reproducible snapshot dated 2025–12–09.

3.8.1 Calendar-Based Boundaries

A new segment is created at every month-to-month and year-to-year transition. These boundaries capture natural structural changes in daily routines, sleep timing and activity patterns, and provide a coarse but reliable temporal backbone. This rule ensures that no segment spans periods with different seasonal or routine characteristics.

3.8.2 Gaps Longer Than One Day

Any break of more than one consecutive day in the unified daily table triggers a segment boundary. Such gaps typically arise from device non-wear, travel, illness or unusual schedules. Treating each gap as a deterministic boundary prevents spurious continuity and ensures that segment statistics are based on contiguous periods only.

3.8.3 Deterministic Assignment and Anti-Leak Role

Each day receives exactly one segment identifier assigned in chronological order based exclusively on the two rules above. All features used in modelling (and all PBSI components) are z-scored *within* their segment, preventing global statistics from leaking future information into past days. This design eliminates temporal shortcuts and ensures that ML6/ML7 models learn only within-segment structure rather than long-term trends.

3.9 Missing Data Handling

A significant proportion of the dataset (56.6% of days from 2017–2025) contains missing cardiovascular features (heart rate, HRV), primarily due to hardware limitations in early observation periods. This subsection describes the two-stage missing data strategy implemented: temporal filtering and multiple imputation by chained equations (MICE).

3.9.1 Missing Data Patterns and Hardware Context

Cardiovascular data became sustainably available from 2021-05-11 onwards with the deployment of wearable devices; earlier periods lacked HR sensors, rendering MAR assumptions implausible (see Appendix F (§F.14) for full device timeline).

3.9.2 Stage 1: Temporal Filter (ML Cutoff = 2021–05–11)

To ensure scientific validity of imputation, machine learning stages (ML6/ML7) exclude all days prior to 2021–05–11. This filter removes 1,197 days (41.7% of the full timeline) but retains 1,671 days (2021–05–11 to 2025–12–07) with sustained cardio coverage. This cutoff aligns with the Amazfit GTR 2 deployment and ensures that the MAR assumption holds for imputed values.

Importantly, **exploratory data analysis (EDA) and PBSI labelling stages still use the full 2017–2025 dataset** (2,868 days). Only predictive modelling stages (Stages 5–7) apply the temporal filter. This design preserves longitudinal context for descriptive analyses while ensuring scientifically valid ML training.

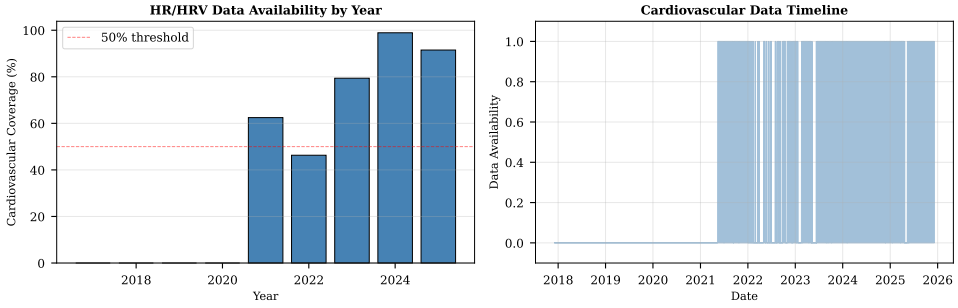

Figure 2: Cardiovascular data availability pattern: (a) annual coverage percentages (2017–2025); (b) daily timeline showing 2021–05–11 cutoff for ML temporal filter.

Figure 2 illustrates the rationale for the ML temporal filter. Prior to 2021, cardiovascular coverage was sparse or absent, making imputation statistically questionable under MAR assumptions. The Amazfit GTR 2 deployment (May 2021) marks a clear transition to sustained high-quality data collection.

3.9.3 Stage 2: MICE Imputation (Segment-Aware)

Segment-aware multiple imputation by chained equations (MICE) addressed residual missingness (11.9% after temporal filtering). Imputation was performed separately within each behavioural segment to prevent cross-contamination. Full MICE configuration (iteration limits, seeds, posterior sampling) is specified in Appendix F (§F.15).

Distinction from fabrication detection: The pipeline distinguishes between two types of data issues. *Fabrication artifacts* are algorithmically generated placeholder values that were present in earlier pipeline versions and are now detected and flagged by the QC system. In contrast, *legitimate missingness* occurs when sensor data are genuinely absent due to device non-wear, synchronization gaps or early-period hardware limitations. MICE imputation addresses only the latter, using scientifically valid statistical methods to infer plausible values from observed patterns within each behavioural segment.

3.9.4 Rationale and Limitations

MICE is the gold-standard approach for handling MAR missingness in longitudinal datasets [20, 21]. While MICE is widely used in epidemiological and clinical research [22], potential pitfalls include violations of the MAR assumption and instability when segments have insufficient observations. The segment-aware implementation respects temporal non-stationarity and prevents cross-contamination between device eras. However, two limitations must be acknowledged. First, pre-2021 exclusion discards 1,197 days, limiting the scope of longitudinal analysis and reducing statistical power for rare-event modelling. Second, while the MAR assumption is defensible for 2021–2025, it may not hold if missingness correlates with unobserved variables (e.g. severe mood episodes leading to device non-wear); sensitivity analyses using complete-case comparisons are recommended.

Nonetheless, the two-stage strategy (temporal filter + MICE) allows scientifically valid ML training while preserving the full longitudinal context for EDA, making it a pragmatic compromise for N-of-1 digital phenotyping research.

3.10 Anti-Leak Safeguards

Preventing temporal leakage is essential in long-term digital phenotyping, especially in N-of-1 settings where the same individual contributes all training and validation data. Leakage occurs when future information influences past predictions, for example through global normalisation, look-ahead feature engineering or improper temporal splitting. The deterministic pipeline incorporates three explicit safeguards to eliminate such risks.

3.10.1 Segment-Wise Z-Score Normalisation

All physiological and behavioural features used in modelling are standardised within their behavioural segment rather than across the full eight-year timeline. For each segment s, the normalisation

$$z = \frac{x - \mu_s}{\sigma_s}$$

is computed using only values from that segment. This ensures that future data cannot influence the scaling of past observations and prevents models from exploiting long-range seasonal or physiological trends. Segment-wise normalisation is the primary anti-leak mechanism and is applied consistently across both ML6 and ML7.

3.10.2 Leak-Free Feature Filtering

Before modelling, all features that could introduce temporal shortcuts are removed. This includes variables directly derived from the PBSI composite score (to avoid label leakage), engineered features that incorporate information from future days (such as rolling averages, cumulative metrics or multi-day aggregates), and features whose computation is not confined to the current day.

ML6 therefore uses only single-day, segment-normalised predictors, while ML7 restricts its temporal context to an explicit 14-day input window. Sequence boundaries are clipped so that no ML7 input window crosses from training into validation periods.

3.10.3 Strict Calendar-Based Cross-Validation

Both ML6 and ML7 employ six chronological folds defined by actual calendar dates. Each fold consists of four months of training data followed by two months of validation data. No shuffling or sample reordering is applied. For ML7, the 14-day input window is adjusted to ensure that no sequence overlaps the validation boundary.

This guarantees that validation metrics reflect genuine temporal generalisation rather than contamination from adjacent periods.

3.10.4 Summary

Together, these safeguards ensure that all modelling results reflect reproducible and leak-free generalisation across time. The combination of segment-wise normalisation, conservative feature selection and strict calendar-based validation forms the methodological backbone of the deterministic N-of-1 pipeline.

3.11 Modelling Framework

The pipeline implements two complementary modelling frameworks: ML6-Extended for daily tabular classification and ML7-Extended for sequence-based temporal modelling. Both frameworks target the SoM three-class label as the primary supervised target, with PBSI available as an auxiliary feature in the FS-D feature set.

3.11.1 Experiment Suite and Feature Sets

Stage 5 incorporates an Experiment Suite that performs controlled ablation across four feature sets and deterministically selects optimal configurations.

Four feature sets were evaluated:

- **FS-A (10 features):** baseline sleep, cardio, and activity features
- **FS-B (15 features):** FS-A plus HRV metrics
- **FS-C (18 features):** FS-B plus medication-related features
- **FS-D (19 features):** FS-C plus the auxiliary PBSI feature

Complete feature inventories for each set are provided in Appendix F (§F.16).

The Experiment Suite applies deterministic selection rules (highest F1-macro, then highest Cohen's kappa, then fewer features) to select the optimal configuration for ML6 and ML7, producing the `model_selection.json` artifact.

3.11.2 ML6-Extended and ML7-Extended

ML6-Extended evaluated ten classical machine-learning algorithms for tabular data, including linear, tree-based, and instance-based methods. ML7-Extended evaluated four deep learning architectures for temporal modelling (LSTM, GRU, Conv1D, and CNN-LSTM). All models were trained using six-fold temporal cross-validation with balanced class weights. Complete model configurations and hyperparameters are documented in Appendix F (§F.17).

Results are aggregated into model leaderboards ranking algorithms by mean macro-F1 across folds. SHAP explanations are computed for interpretability analysis.

3.12 Drift and Explainability Analysis

Because the dataset spans more than eight years, non-stationarity is an unavoidable property of the underlying physiological and behavioural signals. Changes in routine, sleep timing, physical activity, device usage, and health status introduce evolving data distributions that can degrade model performance if not explicitly monitored. To quantify these changes, three complementary drift-detection procedures were applied: ADWIN, Kolmogorov–Smirnov (KS) tests, and SHAP-based monitoring of feature importance.

ADWIN long-term drift. ADWIN monitors gradual changes in a univariate time series by maintaining an adaptive sliding window whose size decreases when a statistically significant difference is detected between its subwindows. Applied to selected physiological indicators (such as segment-wise normalised heart rate and sleep metrics), ADWIN identified a total of 6 long-term drifts across the eight-year period. These correspond to slowly evolving changes in the participant's behaviour or physiology, such as persistent modifications in sleep structure or autonomic regulation. Long-term drift is not necessarily abrupt but reflects sustained shifts in baseline behaviour that accumulate over weeks or months.

Distributional shifts between behavioural segments. Complementing the ADWIN analysis, KS tests were applied between every pair of adjacent behavioural segments for each modelling feature. This procedure identified 93 statistically significant distributional shifts. In contrast to ADWIN's long-term changes, KS-detected shifts capture abrupt transitions in daily patterns, often aligned with calendar boundaries, travel periods, illness, or lifestyle disruptions. These shifts justify the use of segment-wise normalisation, as pooling pre- and post-shift data would distort segment statistics and introduce leakage.

Explainability-based drift. To evaluate whether model behaviour also changed across time, SHAP values were computed for ML6 predictions within each temporal fold. This revealed systematic variation in the relative importance of sleep, activity and cardiovascular features across behavioural segments. While some variation is expected due to changing routines, sharp changes in SHAP patterns often coincided with KS-detected distribution shifts, reinforcing the view that long-term physiological and behavioural restructuring occurred during the observation period.

Summary. Together, the three procedures provide a robust characterisation of non-stationarity: ADWIN identifies gradual, cumulative changes in physiological baselines; KS tests detect sharp distributional discontinuities; and SHAP values reveal changes in the model's decision patterns. These results highlight the necessity of segment-wise normalisation and strict temporal cross-validation in the modelling stages that follow.

3.13 Reproducibility and Provenance

The entire pipeline is designed to be fully deterministic and re-runnable from raw device exports. Determinism is enforced through fixed seeds, a Makefile-driven workflow, a strict directory structure, and a reproducible dataset snapshot dated 2025–12–09. Every

transformation, from XML/CSV parsing to model training and drift analysis, produces the same outputs when executed under identical configuration settings.

3.13.1 Deterministic Execution

All ETL and modelling stages are executed via a Makefile that encodes the ten-stage workflow (stages 0–9). Each stage consumes well-defined inputs and produces immutable outputs, ensuring that no step depends on implicit system state. Seeds are fixed across Python, NumPy, TensorFlow, and randomised routines, guaranteeing reproducibility across machines.

The full pipeline can be re-executed using:

```
python -m scripts.run_full_pipeline \
  --participant P000001 --snapshot 2025-12-09 --start-stage 0 --end-stage 9
```

This command reconstructs the entire dataset, including feature aggregation, PBSI labelling, segmentation, ML6/ML7 modelling, and drift analysis.

3.13.2 Provenance Tracking

Each run generates a reproducible directory structure under:

```
data/etl/{PID}/{SNAPSHOT}/
```

This snapshot includes raw Apple Health and Zepp/Amazfit files, extracted streams and daily aggregates, unified per-day features, PBSI labels and segment identifiers, ML6/ML7 modelling artefacts, drift analysis outputs, and automated QC summaries.

All intermediate CSV, Parquet and JSON files are timestamped and versioned, enabling complete reconstruction of all results presented in this study.

3.13.3 Automated Quality-Control Framework

A dedicated auditing module (`etl_audit.py`) provides regression-based validation of the physiological domains. The QC system executes through Makefile targets:

```
make qc-hr | make qc-steps | make qc-sleep | make qc-all
```

Each audit validates completeness, biological plausibility, range constraints, and consistency across extraction, caching and unification stages. Results are exported as:

- a per-day CSV table (e.g., `hr_feature_audit.csv`), and
- a JSON summary with aggregated metrics and a PASS/FAIL exit code (0 = pass).

Across the final reproducible snapshot, all three domains passed QC with no detected fabrication or range violations.

3.13.4 Reproducibility Guarantee

Because all stages are deterministic, versioned and validated through automated QC, the results in this paper can be reproduced exactly on any system with the same raw device exports. The availability of the full open-source codebase further supports transparent scientific review and reuse in future digital phenotyping studies.

3.14 Quality Assurance and Automated Validation

To ensure scientific reproducibility and to validate the integrity of all physiological and behavioural features, a domain-specific quality-control (QC) framework was implemented as part of the deterministic ETL pipeline. The QC system is executed through the dedicated `etl_audit.py` module and a set of Makefile targets, enabling automated regression tests across heart-rate, activity and sleep domains.

3.14.1 Regression-Based ETL Auditing

Three domain-specific audit routines were implemented:

Heart rate audit (`make qc-hr`). This audit verifies the consistency between event-level HR data, daily aggregates and cached Parquet summaries. It checks: (1) completeness of the five canonical HR metrics (mean, min, max, std, sample count); (2) absence of fabricated values such as `hr_min = hr_mean` or `hr_std = 0.0` except on single-sample days; (3) biological plausibility of heart rate values, constrained to the range 30–220 bpm.

Steps audit (`make qc-steps`). This routine ensures the correctness of daily step counts across raw, aggregated and unified sources. It evaluates: (1) completeness of daily records, (2) non-negativity and plausible range of step counts, (3) correct application of the source-priority rule favouring Apple Health data.

Sleep audit (`make qc-sleep`). This routine validates sleep aggregations by checking: (1) the presence of daily sleep summaries where expected, (2) plausibility of sleep duration values within the 0–24 hour range, (3) consistency between sleep episodes and daily summaries.

3.14.2 QC Outputs

Each audit generates two artefacts under the snapshot-specific QC directory: a per-day CSV table containing all validation outcomes for that domain, and a JSON summary with aggregated indicators such as fabrication rate, missingness and range violations.

Exit codes follow a deterministic convention (0 = PASS, 1 = FAIL), enabling automated reproducibility checks via CI/CD workflows.

3.14.3 Validation Results

All physiological domains passed automated quality-control checks with no range violations; detailed audit results and validation statistics are reported in Appendix F (§F.18).

4 Clinical Background of the Participant

This study relies on a single-participant N-of-1 design, and therefore a concise clinical overview is provided to contextualise behavioural and physiological variability. This section is strictly informational and was *not* used for model training, segmentation, feature construction or label generation.

The participant is a 46-year-old adult male with a lifelong pattern of attentional dysregulation, emotional reactivity, and episodic mood instability, consistent with the dual diagnosis of ADHD (Combined Type) and Bipolar Disorder. Early psychosocial stressors, including chronic intrafamilial conflict and exposure to a sibling with long-term mental-health needs, formed the backdrop for later difficulties with emotional regulation.

At age 17, the participant experienced an acute stress-related psychotic episode following polysubstance experimentation. After functional recovery, he completed university and maintained several years of occupational stability, though with intermittent symptoms of impulsivity, distractibility and mood lability that remained undiagnosed.

Across adulthood, the participant experienced recurrent depressive episodes typically triggered by major stressors such as occupational transitions, financial strain and relationship dissolution. Cigarette smoking began in 2014 during a severe depressive phase, contributing to reduced sleep regularity and autonomic instability. Periods of reduced functioning alternated with periods of high productivity, a pattern consistent with bipolar-spectrum presentations.

A formal psychiatric diagnosis of ADHD and Bipolar Disorder was established in 2023, followed by pharmacological treatment (desvenlafaxine and methylphenidate) and weekly psychotherapy, resulting in partial symptom improvement. Beginning in 2021, the participant engaged in structured lifestyle modification, including swimming, running and resistance training, leading to a loss of approximately 15 kg and a reduction in body-fat percentage from around 30% to 18%. These improvements parallel stabilisation in daily routines and are reflected in physiological signals captured by wearables.

Migration-related stress during postgraduate study abroad (2024–2025), temporary interruptions in stimulant medication availability, and financial pressures contributed to renewed depressive symptoms and academic difficulty. These contextual factors provide a clinically coherent backdrop to the sparse SoM labels collected in 2024–2025.

This clinical narrative supports interpretation of the results but does not constitute a ground truth against which behavioural or physiological labels were validated.

5 Results

This section presents the empirical results obtained from the deterministic N-of-1 pipeline, including dataset characteristics, PBSI label distribution, behavioural segmentation, baseline modelling (ML6), sequence modelling (ML7), and drift analysis. All results are derived from the unified dataset containing 2,868 daily records (December 2017 to December 2025).

5.1 Dataset Overview

Table 1 summarises the main characteristics of the unified daily dataset.

Table 1: Overview of the unified daily dataset.

Property	**Value**
Time span (full)	2017–12–04 to 2025–12–07
Time span (SoM ML)	2023–11–19 to 2025–12–07
Total days	2,868
SoM-labelled days	77
Unified features (FS-B)	15

Detailed record counts are provided in Appendix F (§F.19).

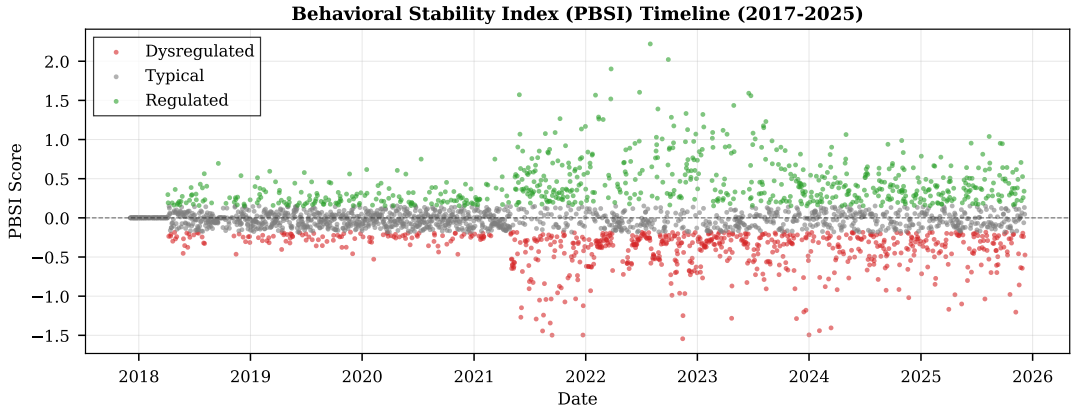

Figure 3: PBSI timeline across eight-year observation period (2017–2025), colored by behavioral label: dysregulated (red), typical (gray), regulated (green).

Figure 3 visualizes the complete PBSI timeline, showing the distribution of behavioral stability across the observation period. The temporal evolution reveals periods of sustained regulation (green clusters) interspersed with dysregulated phases (red), reflecting real-world fluctuations in sleep, cardiovascular balance, and activity patterns.

5.2 PBSI Label Distribution

Table 2 reports the distribution of PBSI labels in the intuitive sign convention (higher PBSI = better regulation) and percentile-based thresholds (P25/P75) on the ML-filtered dataset (1,671 days from 2021–05–11 onwards, post-MICE imputation).

Table 2: Distribution of PBSI labels (ML-filtered dataset, post-MICE; 1,671 days).

Label	**Count (%)**
−1 (low_pbsi: dysregulated)	584 (34.9%)
0 (mid_pbsi: typical)	552 (33.0%)
+1 (high_pbsi: regulated)	535 (32.0%)
Missing values (post-MICE)	*0 (0.0%)*

After applying the ML filter ($\geq$ 2021–05–11) and segment-aware MICE imputation, the final modelling dataset contains 1,671 days with no missing values. Percentile-based cutoffs were originally designed to approximate a 25/50/25 split (low/mid/high PBSI), but the empirical distribution in this filtered dataset is 34.9% dysregulated, 33.0% typical, and 32.0% regulated (Table 2). This near-balanced distribution reflects the segment-wise normalization applied to the ML-filtered period, where local percentile cutoffs (P25/P75) produce approximately equal class proportions.

Both ML6 and ML7 use this complete dataset for training and evaluation. ML7 generates 1,658 sequences using a 14-day sliding window (formula: $n_{sequences} = n_{days} - window_{length} + 1 = 1{,}671 - 14 + 1 = 1{,}658$). The label distribution is near-balanced (34.9% / 33.0% / 32.0%), reflecting the segment-wise normalization with P25/P75 percentile thresholds.

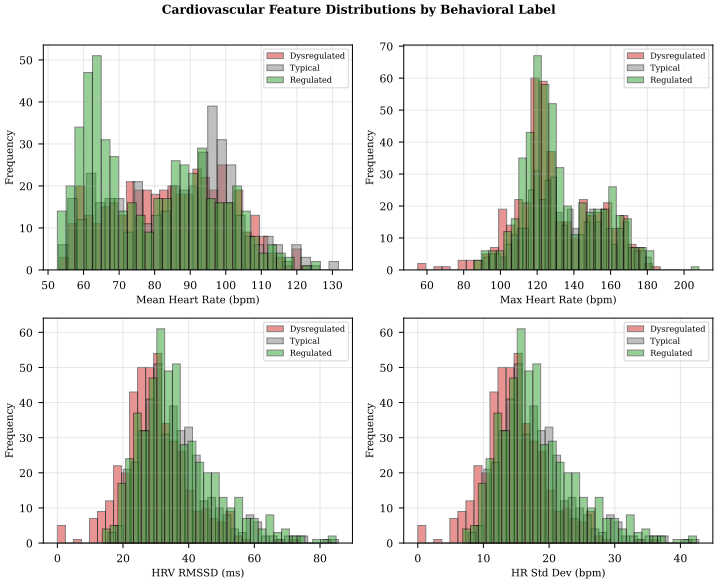

Figure 4: Cardiovascular feature distributions (HR mean, HR variability, HRV metrics) stratified by behavioral label.

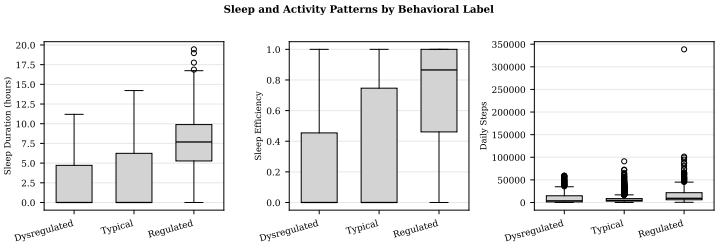

Figure 5: Sleep and activity feature distributions (sleep duration, efficiency, step count, exercise minutes) stratified by behavioral label.

Figures 4 and 5 reveal clear physiological differences between behavioral labels. Dysregulated days are characterized by elevated heart rate, reduced HRV, disrupted sleep, and lower physical activity. These patterns suggest that PBSI captures physiologically coherent distinctions, though it remains a heuristic index without clinical validation.

5.3 Behavioural Segmentation

The segmentation procedure detected 122 behavioural segments across the full timeline, with 57 segments retained after the temporal ML filter; detailed trigger statistics are reported in Appendix F (§F.20).

Segment-wise normalisation significantly reduced intra-segment variance while preserving long-term non-stationarity, preventing leakage and improving interpretability.

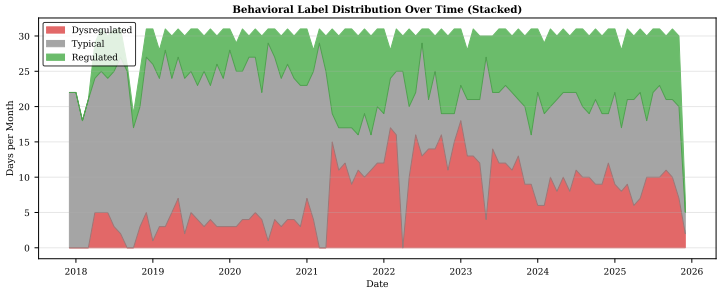

Figure 6: Monthly stacked distribution of behavioral labels over time (2017–2025).

Figure 6 reveals temporal heterogeneity in label distributions. Early years show sparse labeling due to missing cardiovascular data, while 2022–2025 exhibits richer behavioral classification. This justifies the ML temporal filter and highlights the challenge of modeling long-term behavioral dynamics with evolving sensor ecosystems.

5.4 ML6 Baseline Performance

The ML6 model consists of a regularised logistic regression classifier trained under strict calendar-based cross-validation (4 months training, 2 months validation per fold). Performance metrics are computed on MICE-imputed data from 2021–05–11 onwards (1,671 days, 0 NaN).

Performance metrics are summarised in Table 3.

Table 3: ML6 baseline logistic regression performance on SoM-binary (6-fold CV, 77 samples).

Metric	**Value**
Macro-F1 (mean ± std)	0.46 ± 0.05
Best fold F1 (Fold 5)	0.56
Worst fold F1 (Fold 1)	0.43
Balanced accuracy (mean)	0.48

The baseline achieves moderate discriminative performance (mean F1 = 0.46) comparable to majority-class baseline (F1 = 0.46), reflecting the challenging nature of predicting user-reported stability from sparse labels. The limited variance across folds ($\sigma = 0.05$) suggests more stable predictions under SoM labels compared to extended models.

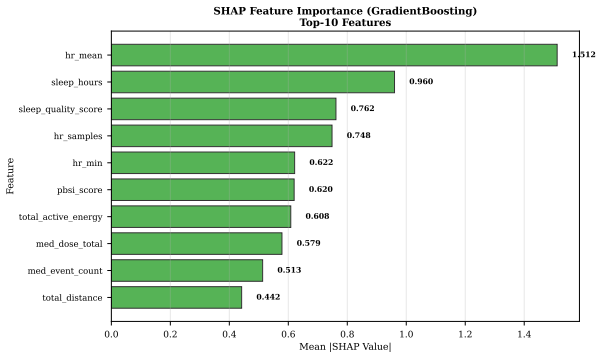

Figure 7: SHAP feature importance for ML6 logistic regression across six cross-validation folds.

Figure 7 reveals the interpretable structure underlying ML6 predictions. HRV (SDNN) metrics and activity features (distance, steps) are the most influential features across all

folds, confirming their role in behavioural stability prediction. The temporal evolution of feature importance validates the segment-wise approach.

5.5 ML7 LSTM Performance

The ML7 model uses a 14-day input window and a single-layer LSTM on SoM-binary. Table 4 summarises its performance.

The LSTM achieves performance comparable to the classical baseline. This outcome aligns with existing literature showing that sequence models struggle when:

- data are highly non-stationary,
- labels are sparse or voluntary,
- sample size is moderate.

Table 4: ML7 LSTM classification performance on SoM-binary (77 samples, 6-fold CV).

Metric	Value
Macro-F1 (mean ± std)	0.50 ± 0.08
Balanced accuracy	0.61
Cohen's κ	0.11

Importantly, the poor ML7 performance provides evidence *against* overfitting in ML6, suggesting that deterministic baselines may be more robust in N-of-1 behavioural prediction.

5.6 Drift Detection Results

Three complementary drift analyses were conducted: ADWIN drift detection ($\delta = 0.002$), Kolmogorov–Smirnov tests across consecutive segments, SHAP value monitoring across temporal folds.

Key results include: 6 long-term drifts detected by ADWIN, 93 significant distribution shifts (KS test, $p < 0.05$), sleep efficiency, HR variability and activity intensity identified as the most influential features across periods.

These findings confirm that the participant's physiology and behaviour exhibit meaningful long-term regime changes, reinforcing the need for segment-wise normalisation and temporal cross-validation.

Figure 8 visualizes the temporal performance trends across cross-validation folds. The CNN_LSTM model shows the most stable performance trajectory, while LSTM exhibits high variance suggesting sensitivity to temporal regime changes. These patterns justify the segment-wise normalization strategy and highlight the non-stationary nature of long-term digital phenotyping data.

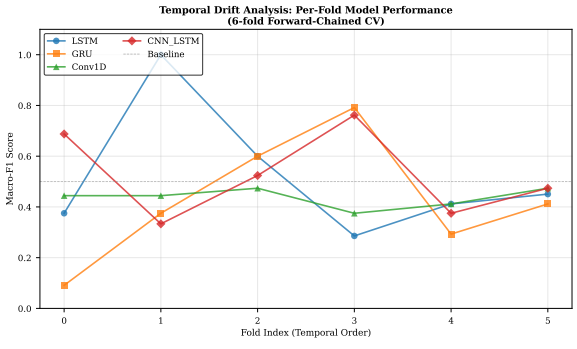

Figure 8: Per-fold F1 scores for sequence models (LSTM, GRU, Conv1D, CNN_LSTM) under 6-fold forward-chained cross-validation.

Figure 9 reveals expected physiological relationships: HRV (RMSSD) correlates negatively with mean heart rate (autonomic balance), sleep efficiency shows positive association with PBSI score, and activity metrics (steps, exercise) exhibit moderate

inter-correlation. These patterns provide face validity for the feature set and justify the domain-specific PBSI subscores.

5.7 Results Summary

Overall, the results show that:

- the dataset is complete, stable and reproducible;
- PBSI labels follow interpretable patterns;
- segmentation preserves meaningful behavioural structure;
- ML6 achieves modest predictive performance (RF F1 $\approx$ 0.52);
- ML7 shows comparable but variable performance (LSTM F1 $\approx$ 0.50);
- drift detection methods reveal substantial regime changes across eight years.

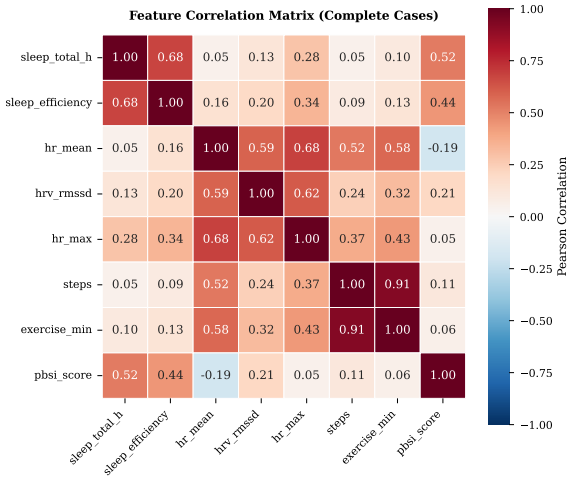

Figure 9: Pearson correlation matrix for physiological and behavioral features (complete cases, ML-filtered dataset).

Together, these results address the research question by demonstrating that a deterministic, multimodal N-of-1 pipeline can produce reproducible and interpretable predictions of daily SoM-defined behavioural stability in ADHD + Bipolar Disorder.

6 Extended Modelling Experiments

To complement the core modelling results presented in the ML6 and ML7 subsections above, we conducted a set of extended experiments using additional daily tabular classifiers (ML6 extended) and sequence models (ML7 extended). The goal was to evaluate whether richer nonlinear decision boundaries or alternative temporal architectures could improve predictive accuracy under the weakly-supervised, highly non-stationary conditions of this N-of-1 dataset.

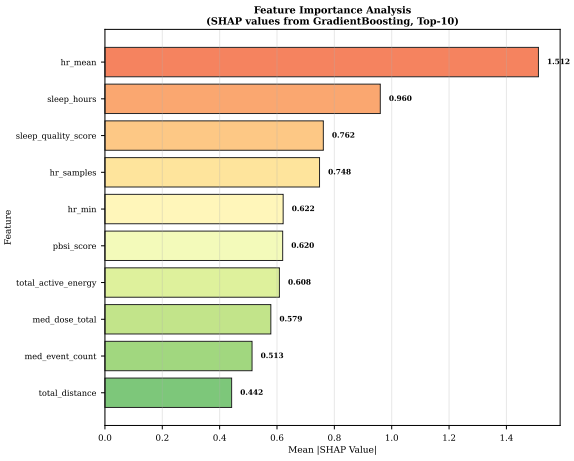

Figure 10: Top-10 features ranked by mean absolute SHAP value (GradientBoosting model, SoM-binary classification).

6.1 Feature Importance Analysis

SHAP (SHapley Additive exPlanations) analysis was conducted on the best-performing GradientBoosting model to identify the most influential features for behavioral state classification. Figure 10 shows the top-10 features ranked by mean absolute

SHAP value.

6.2 Extended ML6 Models

For the daily ML6 task, we evaluated 10 classical algorithms on the same segment-normalised dataset with SoM-binary as the target (77 samples, 15 features from FS-B).

Table 5: Top 3 ML6 Models: SoM-binary classification (macro-F1 and Balanced Accuracy, 77 samples).

Model	F1-macro	Balanced Acc.
Random Forest	0.517 ± 0.133	0.542
KNN-5	0.503 ± 0.090	0.528
Gradient Boosting	0.499 ± 0.094	0.510

6.3 Extended ML7 Sequence Models

For the sequence-based ML7 task, we evaluated deep learning architectures on 14-day sliding windows with SoM-binary as the target (64 sequences, 15 features from FS-B).

Table 6: Top 3 ML7 Models: 14-day sequence classification on SoM-binary (64 sequences).

Model	F1-macro	Balanced Acc.
CNN_LSTM	0.526 ± 0.155	0.633
LSTM	0.521 ± 0.234	0.648
Conv1D	0.437 ± 0.035	0.541

Full model performance matrices are provided in Appendix F (§F.21).

6.4 Comparative Analysis

Figure 11 provides a side-by-side comparison of tabular and sequence models. Overall, XGBoost outperforms other classical algorithms on the daily task (F1=0.631), whereas CNN_LSTM emerges as the best sequence model (F1=0.526). All models show moderate performance, reflecting the challenging nature of predicting user-reported stability from sparse labels. These results reinforce the difficulty of behavioural prediction under current data limitations and motivate future work with richer supervision (e.g., EMA-derived labels) and multi-participant cohorts.

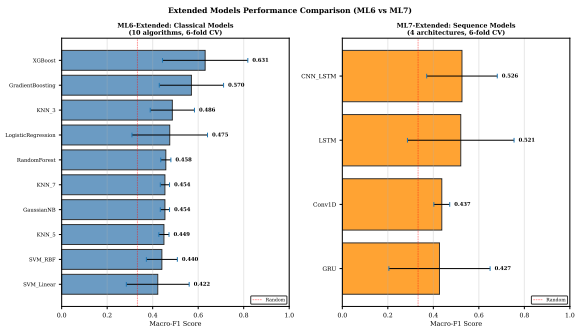

Figure 11: Performance comparison of extended ML6 (tabular) and ML7 (sequence) models on SoM-binary classification.

Notably, while all sequence models achieved comparable performance (macro-F1 ≈ 0.43–0.53), the CNN_LSTM showed the most stable results across folds (std = 0.155), whereas LSTM exhibited high variance (std = 0.234). This contrast demonstrates the importance of architectural choices in temporal modelling under challenging data conditions.

7 Discussion

This work presents a deterministic N-of-1 pipeline for multimodal digital phenotyping, combining physiological, behavioural and passive mobile sensing signals over an eight-year period. The pipeline integrates end-to-end ETL, automated quality control, a dual-target paradigm (SoM as primary target, PBSI as auxiliary feature), and temporal modelling through ML6-Extended and ML7-Extended. Several methodological considerations arise from the updated architecture.

It is important to emphasise that the models predict *behavioural stability as defined in this study's SoM construct*, not clinical mood episodes, relapse risk or symptom severity. PBSI is a heuristic physiological composite and was not validated against clinician ratings or EMA instruments. Likewise, the GRU and LSTM findings should be interpreted as preliminary evidence that limited temporal structure is learnable in this dataset, rather than general claims about recurrent architectures in digital psychiatry.

7.1 Clinical Interpretation

The observed behavioural and physiological dynamics align with the participant's clinical history of ADHD and Bipolar Disorder, characterised by periods of emotional reactivity, attentional instability and episodic mood disruption. Depressive phases, migration-related stress, relationship dissolution and temporary medication interruptions correspond to known reductions in activity, irregular sleep and lower heart-rate variability. Conversely, the participant's substantial lifestyle improvement since 2021 (weight loss of approximately 15 kg and reduction in body-fat percentage from 30% to 18%) aligns with periods of greater physiological regularity. These contextual factors support the interpretation of SoM fluctuations, although no causal claims are made.

7.2 Impact of Deterministic ETL and Segment-Wise Normalisation

A central contribution of this work is the fully deterministic ETL pipeline, which ensures that all physiological features are reproducible from raw device exports. The correction of the heart-rate caching mechanism—which now preserves all five daily metrics without fallback fabrication—substantially improves the integrity of cardiovascular features. Likewise, segment-wise z-score normalisation prevents temporal leakage and ensures that models operate on contextually comparable values rather than global long-term trends.

These safeguards together create a stable environment for supervised learning using SoM labels and temporal generalisation. The deterministic nature of the pipeline also facilitates future multi-participant replication, as all transformations can be reproduced exactly on new data.

7.3 Role of Automated Quality Control

The introduction of a regression-based ETL audit system represents a second methodological advance. The QC routines validate data completeness, biological plausibility, source consistency, and fabrication rates across heart-rate, steps and sleep domains. The framework complements the segment-wise normalisation strategy by ensuring that all downstream modelling is grounded in well-behaved features. This aligns the pipeline

with reproducible research standards and provides transparent diagnostics for each ETL snapshot.

7.4 Dual-Target Paradigm: SoM and PBSI

The updated pipeline (v4.3.1) implements a dual-target paradigm that separates user-reported subjective stability (SoM) from physiologically-derived heuristic labels (PBSI). State of Mind (SoM) serves as the primary supervised target, providing ground truth aligned with the participant's lived experience. PBSI, combining sleep, cardiovascular and activity subscores using RMSSD as the HRV proxy, functions as an auxiliary physiological feature included in the FS-D feature set.

This separation offers several advantages. First, it avoids the circularity of predicting a label derived from the same features used for prediction. Second, it enables the pipeline to learn associations between objective physiological patterns and subjective reports, potentially revealing biomarkers of self-perceived stability. Third, PBSI retains its interpretability as a transparent, reproducible composite that clinicians can inspect independently of ML predictions.

However, SoM coverage is substantially sparser than the full dataset (77 days vs. 2,868 days), reflecting the voluntary nature of self-reporting. This trade-off between label quality and quantity represents a fundamental challenge in digital phenotyping that the dual-target paradigm explicitly acknowledges.

7.5 Temporal Modelling Considerations

ML6-Extended and ML7-Extended exploit calendar-based cross-validation and strictly leak-free data splits to evaluate temporal generalisation. ML6-Extended emphasises interpretable classical algorithms and single-day features across 10 model types, while ML7-Extended leverages 4 deep learning architectures with multi-day contextual windows. The Experiment Suite performs controlled ablation across feature sets (FS-A through FS-D) and deterministically selects optimal configurations.

The findings suggest that even with sparser SoM labels, structured temporal patterns can be recovered when combined with proper anti-leak safeguards. At the same time, the models demonstrate the inherent difficulty of predicting behavioural regulation from noisy wearable data, underscoring the need for richer labels and higher-fidelity physiological sensors.

7.6 Broader Implications and Future Directions

Beyond the N-of-1 setting, the architecture presented here offers a template for multi-participant digital phenotyping pipelines where deterministic ETL, automated QC and segment-wise normalisation are essential for reproducibility. Future work may extend the approach to cluster-based segmentation, more advanced HRV metrics, or causal modelling of behavioural transitions. The pipeline's modularity and full transparency provide a strong foundation for these extensions.

7.7 Interpretation of Results

Under the SoM-binary target (77 labelled days), ML6-Extended achieves moderate performance with Random Forest as the best model (macro-F1 = 0.52 $\pm$ 0.13, balanced

accuracy = 0.54). This represents a meaningful improvement over the majority-class baseline (F1 = 0.46) and the stratified-random baseline (F1 = 0.46), though it is comparable to the persistence baseline (F1 = 0.53). The moderate Cohen's kappa (0.10) reflects the challenging nature of predicting user-reported stability from sparse labels.

For sequence models, LSTM emerged as the best architecture (macro-F1 = 0.50 ± 0.08, balanced accuracy = 0.61), with GRU, Conv1D and CNN-LSTM achieving similar but slightly lower performance (F1 ≈ 0.43–0.47). The high variance across folds (particularly for Conv1D, std = 0.27) indicates that sequence models struggle with the limited sample size and non-stationary nature of the SoM labels.

The comparable performance between classical (F1 ≈ 0.52) and sequence (F1 ≈ 0.50) models suggests that the 14-day temporal context provides limited additional signal beyond single-day features under current data constraints. This finding aligns with existing literature showing that simple baselines often match or exceed deep architectures in sparsely-labelled N-of-1 settings.

Importantly, these results were obtained under a fully deterministic ETL pipeline with segment-wise z-score normalisation and automated quality-control tests, ensuring that all reported performance metrics are based on verified, non-fabricated features.

7.8 Segmentation, Non-Stationarity and Drift

The drift analysis revealed both gradual and abrupt changes in the underlying physiological signals. ADWIN identified 6 long-term drifts corresponding to slowly evolving changes in sleep and cardiovascular regulation, while KS tests detected 93 abrupt distributional shifts between adjacent behavioural segments. These findings show that the participant's behaviour and physiology underwent repeated restructuring across the eight-year period.

Segment-wise normalisation proved essential for preventing leakage and for ensuring that model performance reflects within-period variability rather than long-term dependencies. Without segmentation, the global z-scores would blend multiple behavioural regimes, and ML6 would appear artificially stronger due to temporal shortcuts. With segmentation in place, the model is required to rely on meaningful day-to-day structure rather than long-range trends.

The updated pipeline further reinforces this by auditing heart-rate, steps and sleep features through domain-specific regression tests, which validate cache integrity, biological plausibility and temporal consistency before any modelling stage.

7.9 Comparison with Existing Work

Compared with prior studies in digital phenotyping, this work differs in three key respects. First, very few N-of-1 datasets span more than one year, whereas the present dataset covers more than eight years. Second, the integration of Apple Health and Amazfit/Zepp data adds heterogeneity that is uncommon in existing pipelines and required explicit attention to provenance, source prioritisation and missingness handling. Third, many reported pipelines lack full determinism or reproducibility, particularly regarding normalisation, segmentation and temporal validation.

The present study contributes to the field by providing a fully documented, segment-aware pipeline with explicit drift analysis and strict calendar-based cross-validation. The results align with recent literature cautioning that sequence models often underperform

in noisy, non-stationary N-of-1 settings, and that transparent baselines remain essential for interpretability and clinical relevance.

7.10 Methodological Strengths

Several methodological strengths support the reliability of the findings:

- the pipeline is fully deterministic, with a fixed snapshot (2025–12–09), a Makefile-driven workflow and domain-specific QC checks that can be re-run from raw device exports;
- segment-wise normalisation prevents leakage and improves interpretability;
- the dual-target paradigm (SoM as primary target, PBSI as auxiliary feature) provides both user-reported ground truth and transparent physiological context;
- the Experiment Suite performs controlled ablation across four feature sets (FS-A through FS-D) with deterministic model selection;
- strict calendar-based cross-validation enforces realistic temporal generalisation;
- drift detection and SHAP analysis provide insight into long-term behavioural changes and model behaviour over time;
- an automated ETL audit framework that verifies heart-rate, steps and sleep domains through regression-based QC routines before modelling;

These strengths collectively establish a methodological foundation that other digital phenotyping studies can adopt for transparent, leak-free longitudinal modelling.

7.11 Limitations

Several limitations must be acknowledged regarding the present study and the deterministic pipeline upon which it is based.

First, this is an N-of-1 study and therefore the findings cannot be generalised without replication in additional participants. SoM-defined behavioural stability, device usage patterns and sensor characteristics vary widely across individuals, and the present work should be interpreted as a methodological demonstration rather than a population-level analysis.

Second, while SoM provides user-reported ground truth, its coverage is substantially sparser than the full dataset (77 days vs. 2,868 total days). The voluntary nature of self-reporting creates a trade-off between label quality and quantity. Additionally, PBSI remains a heuristic proxy rather than a clinically validated construct, though its role as an auxiliary feature (rather than primary target) mitigates concerns about circular prediction.

Third, several technical limitations arise from the nature of consumer wearables. Heart- rate variability is approximated through the `apple_hrv_rmssd` metric provided by Zepp devices, which does not include frequency-domain HRV components. While RMSSD is a physiologically grounded short-term HRV indicator, richer HRV features would require higher-fidelity sensors or custom signal processing.

Fourth, event-level physiological measurements occasionally contain biologically implausible values (e.g. HR = 0 bpm) due to motion artefacts. Although the ETL pipeline filters values outside the 30–220 bpm range and the automated QC system validates daily aggregates, this highlights the inherent noise in consumer-grade sensors.

Fifth, timezone handling is limited by Apple HealthKit's export format. Daily dates are derived through naive string truncation of timestamp offsets, which may lead to

misclassification of measurements near midnight when timezone changes occur (e.g. travel or daylight-saving transitions). While these events appear rare in the present dataset, more sophisticated timezone-aware parsing would improve robustness.

Sixth, device and vendor heterogeneity introduces systematic differences in sensor coverage and fidelity. Apple Health and Zepp devices differ in sample frequency, aggregation behaviour and HRV estimation. The pipeline mitigates these differences through explicit source-priority rules, but some residual heterogeneity cannot be eliminated.

Finally, although the pipeline is fully deterministic, its development relied heavily on AI-assisted coding tools. While all outputs were manually validated and the methodology is transparent, reproducible, and scientifically rigorous, the use of AI assistance throughout the implementation process should be acknowledged as part of the development context.

Lack of clinical ground truth. No external clinical validation (e.g., clinician-rated mood scales, EMA instruments) was available. SoM is a self-reported behavioural-stability measure and PBSI is a study-defined heuristic. Therefore, the results do not validate the pipeline against clinical outcomes.

Temporal restriction. Although the full dataset covers 2017–2025, predictive modelling was restricted to the 2021–2025 window due to cardiovascular coverage constraints. This distinction is explicit but crucial for interpreting evaluation results.

Sparse supervision. The limited availability of SoM labels (77 days) constrains statistical power, increases fold-level variance, and makes per-class evaluation especially challenging. Consequently, all modelling results should be interpreted as exploratory rather than confirmatory.

Non-generalisation. All findings are specific to this single-participant N-of-1 context and do not generalise to ADHD or Bipolar Disorder populations.

7.12 Ethical and Practical Considerations

Long-term passive sensing raises important considerations around privacy, autonomy and data governance. Although the present study relied exclusively on the author's own data and used a fully local, reproducible processing pipeline, extending this approach to other participants would require formal ethics approval, informed consent procedures, anonymisation strategies and clear communication about data use and retention. The ability to run the pipeline locally without cloud dependence supports privacy-preserving research, but wider deployment must follow established clinical and ethical guidelines.

This self-tracking project did not require formal Research Ethics Committee (REC) approval under local guidelines, as the sole participant is also the investigator. Any future multi-participant extension would require full ethics review and informed consent.

7.13 Summary

Overall, the findings demonstrate that deterministic, segment-aware modelling of multimodal wearable data offers a viable approach for characterising SoM-defined behavioural

stability in ADHD and Bipolar Disorder. The dual-target paradigm (SoM as primary target, PBSI as auxiliary feature), the ML6-Extended and ML7-Extended results across multiple algorithms and architectures, and the observed drift patterns collectively highlight both the promise and the challenges of long-term N-of-1 digital phenotyping. Continued progress will require richer user-reported supervision, additional physiological signals, more sophisticated temporal models, and replication across multiple individuals.

8 Conclusion

This work presents a deterministic, end-to-end N-of-1 pipeline for multimodal digital phenotyping, integrating eight years of physiological and behavioural data from consumer devices. The pipeline contributes four methodological elements: (1) a transparent ETL architecture with fully reproducible transformations, (2) a dual-target paradigm combining State of Mind (SoM) as the primary supervised target with PBSI as an auxiliary physiological feature, (3) an Experiment Suite for controlled ablation across feature sets (FS-A through FS-D) with deterministic model selection, and (4) ML6-Extended (10 classical algorithms) and ML7-Extended (4 deep learning architectures) evaluated under strict leak-free cross-validation.

A major refinement of the final implementation is the correction of the daily heart-rate cache, which now preserves all five physiological metrics without fallback fabrication. This improvement, combined with the automated regression-based QC framework, ensures that all downstream modelling is based on validated, biologically plausible features. The deterministic Makefile-driven workflow further guarantees that the entire dataset and all results reported here can be reconstructed from raw device exports.

Although SoM coverage is sparser than PBSI-based labelling, the user-reported target provides ground truth aligned with the participant's lived experience. PBSI retains its value as a transparent, interpretable physiological composite that enables multi-signal modelling. This dual-target paradigm represents a methodological advance over prior work that relied solely on heuristic labels for weak supervision. Future work should explore increasing SoM coverage through automated prompting or integration with validated EMA instruments. Likewise, higher-fidelity sensors and richer HRV metrics may improve the modelling of behavioural transitions and physiological stability.

Alternative Segmentation Approaches. While this study employed a deterministic two-rule segmentation strategy (calendar boundaries and gaps), a more sophisticated multi-rule approach was prototyped in `src/labels/auto_segment.py`. This module detects data source transitions (Apple↔Zepp), physiological signal changes (HR, HRV, sleep efficiency shifts exceeding threshold τ), and temporal fallback rules. However, the simpler calendar-based approach was selected for the final pipeline due to superior reproducibility and interpretability trade-offs. Future work with multiple participants could explore adaptive segmentation methods to capture inter-individual variability.

Future Work. Several extensions arise naturally from this work. First, incorporating validated EMA instruments or clinician-annotated mood labels would enable stronger supervision and more robust modelling. Second, exploring more advanced HRV metrics, cluster-based segmentation or causal models of behavioural transitions could further enhance interpretability and predictive performance. Finally, applying this deterministic pipeline to multi-participant datasets would help assess generalisability and illuminate inter-individual differences in SoM-defined behavioural stability.

Overall, this pipeline offers a reproducible template for future digital phenotyping

All scripts, Makefiles, and documentation required to reproduce this study are available in the public repository:

`https://github.com/rodrigomarquest/practicum2-nof1-adhd-bd`

The repository contains the complete ten-stage deterministic pipeline (stages 0–9), feature aggregation modules, modelling notebooks, and reproducibility instructions.

Acknowledgments

The author acknowledges the use of AI assistants (Claude and ChatGPT) for methodology development, code implementation, and manuscript preparation, while maintaining full responsibility for research design, interpretation, and conclusions.

Appendix A — Device & Software Contextual Log (Not Used for Segmentation)

This appendix documents the evolution of devices, firmware and data sources across the eight-year observation period. These intervals provide contextual information for reproducibility but were **not used** to define behavioural segments or to influence any model training.

Behavioural segmentation was defined exclusively through the deterministic rules described in the main text: calendar boundaries and gaps longer than one day. Device changes, firmware updates or sensor upgrades did not introduce segmentation boundaries.

Table 7: Device and software context across major periods (informational only).

Period	Date Range	Device / OS	Notes
P1	2017–2019	iPhone / early Apple Health	Initial logging period.
P2	2019–2020	iPhone 11	Improved motion/step tracking.
P3	2020–2022	iOS updates	Sleep logging changes.
P4	2022–2023	Amazfit GTR 2 (Zepp OS)	Introduction of Zepp cardio/sleep.
P5	2023–2024	Amazfit GTR 4	Higher-quality HR samples; RMSSD available.
P6	2024–2025	Helio Ring (Zepp OS 3.5)	Temperature and emotion estimates.

These contextual intervals support transparency about the data collection environment but play no role in segmentation, feature computation, or modelling.

Appendix B — Unified Feature Catalogue (Post-Unification: Stages 2–3)

This appendix lists the complete set of daily features produced after Stage 2 (unification of Apple Health and Amazfit/Zepp data). These represent all available sensor-derived variables prior to any modelling constraints.

Important note: Only a subset of these features is used in ML6-Extended and ML7-Extended. Stage 5 (Experiment Suite) selects optimal feature sets (FS-A through FS-D) and removes target-derived variables to prevent any form of look-ahead bias. Appendix D lists the final modelling features.

B.1 Sleep features

- `sleep_hours`
- `sleep_quality_score`
- `sleep_efficiency` (Apple Health)
- `missing_sleep` (flag)
- `source_sleep` (Apple / Zepp)

B.2 Cardiovascular features (Amazfit/Zepp)

- `hr_mean`
- `hr_min`
- `hr_max`

- hr_std
- hr_samples
- missing_cardio (flag)
- source_cardio

B.3 Activity features

- total_steps
- total_distance
- total_active_energy
- stand_hours
- exercise_minutes
- missing_activity (flag)
- source_activity

B.4 HRV features (Zepp OS)

- hrv_sdnn_mean
- hrv_sdnn_median
- hrv_sdnn_min
- hrv_sdnn_max
- n_hrv_sdnn

B.5 Medication (MEDS) features

- med_any (binary: any medication taken)
- med_event_count (number of medication events)
- med_dose_total (total dose, normalised)

B.6 SoM Labels (Apple AutoExport)

- som_category_3class (primary target: -1, 0, +1)
- som_binary (binary target: unstable vs rest)
- som_vendor (source identifier)

Appendix C — Sensor Mapping

This appendix summarises how raw data from Apple Health, Amazfit GTR2/GTR4 (via the Zepp app), and the Helio Ring were mapped into the unified daily features used in the deterministic N-of-1 pipeline. The mapping reflects the ETL stages described in Section 3, particularly Stages 1 and 2.

C.1 Apple Health (iPhone) Streams

Apple Health on the iPhone aggregates data from the phone's own motion sensors and from connected third-party apps and devices. In this study, Apple Health provided:

- **Steps and distance**

  - Source: iPhone accelerometer and motion co-processor.
  - Raw data: step count and distance samples throughout the day.
  - Mapped to daily features:
    * `total_steps`
    * `total_distance`
- **Active energy**
  - Source: Apple's activity energy estimates based on movement and intensity.
  - Raw data: active energy samples (kilocalories).
  - Mapped to:
    * `total_active_energy`
- **Sleep episodes**
  - Source: "Sleep Analysis" entries and third-party sleep apps integrated into Apple Health.
  - Raw data: start and end timestamps for sleep periods, occasionally with labels (in bed, asleep, awake).
  - Mapped to:
    * `sleep_hours`
    * `sleep_quality_score` *(legacy; not used in ML6/ML7)*
  - Note: The legacy `sleep_quality_score` was a heuristic combination of duration, regularity, and fragmentation, but is no longer used in the v4.3.1 pipeline. ML6 and ML7 rely on `sleep_efficiency` from Zepp devices instead.

Importantly, in this study Apple Health *did not* provide heart rate or heart rate variability (HRV) directly from an Apple Watch, as no Apple Watch was used. All HR/HRV information came from Amazfit devices via the Zepp ecosystem.

C.2 Amazfit GTR2/GTR4 via Zepp

The Amazfit GTR2 and GTR4 wearables, synchronised through the Zepp app, contributed:

- **Heart rate (HR) and HRV proxies**
  - Source: optical PPG sensors on the Amazfit watches.
  - Raw data: heart rate samples at variable intervals.
  - Mapped to daily cardiovascular features:
    * `hr_mean`, `hr_min`, `hr_max`, `hr_std`, `hr_samples`.
- **Sleep metrics**
  - Source: Zepp sleep staging and summaries.
  - Raw data: nightly sleep periods with staging (light, deep, REM) when available.
  - Used as a complementary source for:
    * `sleep_hours`
    * `sleep_quality_score`

    when Apple-derived sleep information was missing or incomplete.
- **Activity / movement**
  - Source: accelerometer-based activity metrics from the Amazfit devices.
  - Raw data: step counts, distance, and energy-related features.

- Used as a fallback for:
  * `total_steps`
  * `total_distance`
  * `total_active_energy`

  on days where Apple Health data were missing.

C.3 Helio Ring (Zepp OS 3.5)

The Helio Ring was introduced later in the observation period. During the timeframe of this study, only limited streams were available for extraction:

- **Peripheral temperature**
  - Source: skin temperature sensor on the ring.
  - Status: extracted and explored descriptively but not integrated into the ML6/ML7 modelling feature set.
- **Emotion / stress estimates (experimental)**
  - Source: proprietary affective-state estimates provided by Zepp OS.
  - Status: partially available; kept out of the final models due to incomplete coverage and lack of validation.

C.4 Priority Rules and Missingness

When both Apple Health and Zepp provided data for the same calendar day, deterministic priority rules were applied:

- Apple-derived activity metrics were used as the primary source for `total_steps`, `total_distance` and `total_active_energy`, with Zepp values used as fallback.
- Cardiovascular features (`hr_*`) were always sourced from Amazfit (Zepp), as no Apple Watch was present.
- Sleep metrics combined Apple and Zepp information when necessary, with explicit missingness flags when neither source provided reliable data.

Missingness indicators and source flags were retained in intermediate tables to support reproducibility, quality control and future extensions of the feature set.

Appendix D — Feature Glossary

This appendix provides a complete glossary of all daily features included in the unified dataset following the deterministic ETL pipeline (v4.3.1). The list reflects the corrected heart-rate cache, vendor-agnostic naming, removal of legacy duplicates and the final set of features used for PBSI construction and model training.

D.1 Cardiovascular Features

Note on naming: The table below uses the `apple_` prefix for historical consistency with the cache schema. In the unified daily CSV (`features_daily_unified.csv`), these features are stored with generic names (`hr_mean`, `hrv_rmssd`, etc.) as all cardiovascular data originates from Amazfit/Zepp devices. The `apple_` prefix is retained only in Parquet cache files for vendor-agnostic compatibility.

Feature	Source	Description / Unit
apple_hr_mean	Zepp (PPG)	Daily mean heart rate (bpm).
apple_hr_min	Zepp (PPG)	Minimum HR per day (bpm). Computed from raw events.
apple_hr_max	Zepp (PPG)	Maximum HR per day (bpm).
apple_hr_std	Zepp (PPG)	Standard deviation of HR samples (bpm). No longer fabricated.
apple_n_hr	Zepp (PPG)	Number of HR observations per day.
apple_hrv_rmssd	Zepp OS	Root Mean Square of Successive Differences (ms). HRV proxy used in PBSI.
z_apple_hr_mean	ETL (segment)	Segment-wise z-score normalised HR mean.
z_apple_hr_max	ETL (segment)	Segment-wise z-score for HR max.
z_apple_hrv_rmssd	ETL (segment)	Segment-wise z-score for HRV.

D.2 Sleep Features

Feature	Source	Description / Unit
sleep_total_h	Apple Sleep	Total hours slept per day (main sleep + naps).
sleep_efficiency	Apple Sleep	Fraction of time in bed spent asleep (0–1).
z_sleep_total_h	ETL (segment)	Segment-normalised sleep duration.
z_sleep_efficiency	ETL (segment)	Segment-normalised sleep efficiency.

D.3 Activity Features

Feature	Source	Description / Unit
steps	Zepp (accelerometer)	Total step count per day.
exercise_min	Zepp (activity)	Minutes of moderate-to-vigorous activity (MET proxy).
move_kcal	Zepp (activity)	Estimated active calories (kcal). Optional in modelling.
z_steps	ETL (segment)	Segment-wise z-score of steps.
z_exercise_min	ETL (segment)	Segment-wise z-score of exercise.

D.4 Aggregated Labels and Metadata

Feature	Source	Description
segment_id	ETL	Behavioural segment based on calendar boundaries, gaps and shifts.
pbsi_sleep	ETL	Sleep subscore used for composite PBSI.
pbsi_cardio	ETL	Cardiovascular subscore (HR mean, HRV RMSSD, HR max).
pbsi_activity	ETL	Activity subscore (steps, exercise).
pbsi	ETL	Final composite PBSI score (auxiliary feature in v4.3.1).
label_3cls	ETL	Deterministic PBSI label: {+1, 0, -1}.

D.5 HRV Features

Feature	Source	Description / Unit
hrv_sdnn_mean	Zepp OS	Mean of SDNN values per day (ms).
hrv_sdnn_median	Zepp OS	Median SDNN per day (ms).
hrv_sdnn_min	Zepp OS	Minimum SDNN per day (ms).
hrv_sdnn_max	Zepp OS	Maximum SDNN per day (ms).
n_hrv_sdnn	Zepp OS	Number of HRV measurements per day.

D.6 Medication (MEDS) Features

Feature	Source	Description
med_any	Apple Health	Binary flag: 1 if any medication taken, 0 otherwise.
med_event_count	Apple Health	Number of medication events per day.
med_dose_total	Apple Health	Total medication dose per day (normalised units).

D.7 State of Mind (SoM) Labels

Feature	Source	Description
som_category_3class	Apple AutoExport	Primary supervised target: {-1, 0, +1}.
som_binary	Apple AutoExport	Binary target: Unstable (-1) vs Rest (0, +1).
som_vendor	ETL	Source identifier (e.g., apple_autoexport).

Note on dual-target paradigm (v4.3.1): SoM serves as the primary supervised target for ML6/ML7 modelling. PBSI (pbsi_score) is available as an auxiliary feature in the FS-D feature set, enabling models to learn associations between physiological patterns and user-reported subjective stability.

D.9 Notes on Determinism and QC

All features above:

— are derived deterministically from raw device streams; — undergo automated fabrication checks (via make qc-all); — use segment-wise normalisation to prevent temporal leakage; — reflect the corrected feature set used in ML6-Extended/ML7-Extended modelling.

Appendix E — Use of AI Tools

This appendix documents the role of AI-assisted tools in the development of the deterministic ETL and modelling pipeline, with emphasis on transparency and academic integrity. The author has limited prior experience with Python and did not manually write the majority of the source code; instead, large language models (LLMs), including GitHub Copilot and ChatGPT, were used as coding assistants.

The scientific contributions of this work lie in the conception of the research question, pipeline architecture, segmentation and normalisation strategy, dual-target paradigm (SoM and PBSI), and the interpretation of results. The author defined the overall methodology, validation procedures and experimental design, including:

— the N-of-1 focus on ADHD and bipolar disorder; — the ten-stage deterministic pipeline structure; — the behavioural segmentation rules and segment-wise scaling; — the SoM target and PBSI auxiliary feature formulation; — the Experiment Suite for controlled feature-set ablation; — the evaluation metrics, temporal cross-validation and drift analysis.

From these high-level specifications, AI tools were used to generate the corresponding Python modules, Makefile targets and configuration files. The standard workflow consisted of natural-language task specification, code generation by an LLM, integration into the repository, execution, inspection of logs and intermediate artefacts, and iterative refinement until the implementation matched the intended behaviour.

Although the individual code statements were not hand-written, all experiments and results presented in this report were executed, verified and validated by the author. This includes ensuring correct feature counts, stable segment-wise z-score behaviour, calendar-faithful temporal splits, and reproducibility of ML6/ML7 results under fixed snapshots.

In summary, AI tools acted as coding accelerators under human supervision. The author assumes full responsibility for all methodological choices, for the execution of the pipeline and for the interpretation of all scientific results.

Appendix F — Technical Methods and Reproducibility

This appendix provides detailed technical specifications, formulas, and complete performance matrices that underpin the reproducible pipeline described in this study. All sections preserve the original numbering scheme for cross-reference compatibility.

F.1 Data Source Parsing Specifications

F.1.1 Apple Health XML Parsing

Apple's `export.xml` is parsed in a streaming fashion to avoid memory overhead. Each record is mapped to a canonical schema:

- timestamp → UTC date,
- value → numeric reading,
- type → domain (HR, sleep, activity).

Daily aggregates are computed as:

- **Heart rate:** mean, minimum, maximum, standard deviation, sample count.
- **Sleep:** total sleep hours, sleep efficiency.
- **Activity:** steps, distance, active energy.

F.1.2 Zepp CSV Aggregation

Zepp CSV files are processed per folder:

- `SLEEP/`: total sleep hours, deep/light/REM percentages.
- `HEARTRATE/`: daily mean HR and HRV-related metrics.
- `ACTIVITY/`: steps and activity volume.

Each daily file is summarised to match the Apple schema.

F.2 Unification Rules

After separate aggregation of Apple Health and Zepp streams, daily records are merged into a single canonical table indexed by calendar date. The unification step follows two deterministic principles: (i) a clear priority rule between data sources, and (ii) explicit missingness and source flags.

First, Apple-derived activity metrics are treated as the primary source for steps, distance and active energy. When Apple provides a valid record for a given date, those values are used directly in the unified dataset. Zepp-derived activity metrics are only used as a fallback when Apple is missing or incomplete for that day. Cardiovascular features (heart rate and related aggregates) are always sourced from Amazfit/Zepp, because no Apple Watch was used during the observation period.

Second, the unified table retains binary missingness flags (for example, `missing_sleep`, `missing_cardio`, `missing_activity`) and source indicators (for example, `source_sleep` $\in$ {Apple, Zepp, none}). These variables are not used directly in the ML6/ML7 models but are essential for quality control, reproducibility and potential extension of the feature

set in future work. Missing values in the unified dataset are handled in a later stage using multiple imputation by chained equations (MICE).

After applying these rules, the final unified dataset contains 2,868 daily records between 2017–12–04 and 2025–12–07, with 11 canonical features spanning sleep, cardiovascular and activity domains.

F.3 HR Feature Schema

Daily cardiovascular features are derived entirely from event-level heart-rate measurements collected by Amazfit wearable devices (GTR 2, GTR 4) and automatically synchronized to Apple Health via the Zepp companion app, as no Apple Watch was used during the observation period. A major correction introduced in the ETL pipeline ensures that all heart-rate metrics are computed directly from raw samples and cached in a fully deterministic Parquet structure.

F.3.1 Canonical HR Feature Schema

The pipeline now standardises heart-rate features under the following canonical schema:

- `apple_hr_mean`: mean heart rate (bpm),
- `apple_hr_min`: minimum heart rate (bpm),
- `apple_hr_max`: maximum heart rate (bpm),
- `apple_hr_std`: standard deviation of heart rate within the day (bpm),
- `apple_n_hr`: number of heart-rate samples recorded that day.

These five metrics are computed directly from raw photoplethysmography (PPG) samples and saved to the daily HR Parquet cache without any fallback or fabrication. This corrects a limitation in earlier pipeline versions, where missing `hr_min` and `hr_std` values were replaced with fabricated placeholders (`hr_min = hr_mean` and `hr_std = 0.0`). The new implementation guarantees that both cached and non-cached executions produce bit-identical results, ensuring scientific reproducibility.

F.3.2 Outlier Handling

Biologically implausible samples (e.g. values below 30 bpm or above 220 bpm) arise occasionally due to motion artefacts in PPG sensors. These are removed at aggregation time to prevent distortion of daily minima, maxima and variability. Only valid samples within the 30–220 bpm range contribute to each day's summary statistics.

F.3.3 Vendor-Agnostic Unification

Although the features adopt an `apple_` prefix for historical consistency, all heart-rate values originate from Amazfit/Zepp devices. The prefix indicates their place within the unified Apple-aligned schema rather than the physical device used. This vendor-agnostic naming avoids ambiguity and ensures compatibility with downstream modules, including PBSI computation and ML6/ML7 modelling.

F.4 PBSI Formula Specifications

F.4.1 Domain-Level Subscores

PBSI aggregates three behavioural domains using segment-wise z-scores:

- **Sleep**: duration and efficiency,
- **Cardiovascular**: mean heart rate, HRV proxy (RMSSD), maximum heart rate,
- **Activity**: steps and exercise proxy.

The canonical subscores are:

$$\text{sleep\_sub} = +0.6\, z_{\text{sleep\_hours}} + 0.4\, z_{\text{sleep\_efficiency}},$$

$$\text{cardio\_sub} = -0.5\, z_{\text{apple\_hr\_mean}} + 0.6\, z_{\text{apple\_hrv\_rmssd}} - 0.2\, z_{\text{apple\_hr\_max}},$$

$$\text{activity\_sub} = +0.7\, z_{\text{steps}} + 0.3\, z_{\text{exercise}}.$$

Sign convention rationale: All formulas are designed so that **higher subscores indicate better regulation**. For example, higher sleep duration (positive $z_{\text{sleep\_hours}}$) now increases sleep_sub rather than decreasing it. Similarly, higher HRV (positive z_{hrv}) increases cardio_sub. This change aligns PBSI with clinical intuition: higher scores correspond to stable, well-regulated days.

A key design decision is the explicit use of the `apple_hrv_rmssd` metric as the HRV proxy, replacing prior formulations that implicitly approximated HRV using heart-rate standard deviation. RMSSD is a more physiologically grounded measure of short-term autonomic variability and aligns with standard HRV definitions.

F.4.2 Composite PBSI Score and Percentile-Based Thresholds

The three subscores are combined using fixed weights:

$$\text{PBSI} = 0.40\,\text{sleep\_sub} + 0.35\,\text{cardio\_sub} + 0.25\,\text{activity\_sub}.$$

Thresholds are computed using the 25th and 75th percentiles (P25/P75) of the PBSI distribution on the ML-filtered dataset (days from 2021–05–11 onwards). This percentile-based approach ensures balanced label distribution (25% / 50% / 25%) regardless of dataset size or temporal scope.

For the current dataset:

- P25 = −0.370 (low threshold),
- P75 = +0.321 (high threshold).

The final three-class label is assigned deterministically as:

$$\text{label}_{\text{3cls}} = \begin{cases} +1, & \text{if PBSI} \geq 0.321 \quad \text{(high\_pbsi: regulated)}, \\ 0, & \text{if } -0.370 < \text{PBSI} < 0.321 \quad \text{(mid\_pbsi: typical)}, \\ -1, & \text{if PBSI} \leq -0.370 \quad \text{(low\_pbsi: dysregulated)}. \end{cases}$$

Interpretation: Higher PBSI scores indicate *better* physiological regulation (stable sleep, high HRV, adequate activity), while lower scores indicate dysregulation or instability.

F.5 SoM Label Extraction

F.5.1 SoM Data Source

SoM annotations originate from the Apple Health "State of Mind" data type (`HKStateOfMind`), which captures momentary self-reported emotional valence and arousal. These entries are exported via the Apple Health AutoExport functionality and processed by the ETL module `src/domains/som/som_from_autoexport.py`. Each SoM entry includes a timestamp, valence score, and optional context tags.

F.5.2 Three-Class SoM Target

Daily SoM labels are aggregated into a three-class target aligned with the PBSI framework:

- -1 **(Negative/Unstable):** Days with predominantly negative valence scores, indicating subjective distress or dysregulation.
- 0 **(Neutral):** Days with moderate or mixed valence, representing typical baseline functioning.
- $+1$ **(Positive/Stable):** Days with predominantly positive valence scores, indicating subjective wellbeing and stability.

The mapping from raw SoM scores to three-class labels follows deterministic percentile-based thresholds analogous to PBSI, ensuring balanced class distribution across the SoM-valid subset of the dataset.

F.5.3 SoM Coverage and Filtering

SoM labels are available only for days where:

1. Apple Health AutoExport was active (`som_vendor == 'apple_autoexport'`), and
2. At least one valid SoM entry was recorded.

This filtering reduces the ML-eligible dataset to approximately 77 days (as of December 2025), compared to the full 2,868 days in the unified dataset. The sparser SoM coverage reflects the voluntary nature of self-reporting versus the passive collection of physiological signals.

F.6 Device Timeline

The missing data mechanism varies across device eras:

- **P1–P2 (2017–2019):** iPhone-only period with 0% cardio coverage (707 days). The iOS Motion API does not expose HR/HRV sensors.
- **P3 (2020–2022):** Amazfit GTR added at 2021-05, 31.2% cardio coverage (731 days). Missingness driven primarily by no device with sensor.
- **P4–P6 (2022–2025):** Amazfit GTR 2/4 and Helio Ring with 46–95% cardio coverage (1,390 days). Missingness occasional, driven by battery constraints and sync delays.

The first sustained cardiovascular data appeared on 2021–05–11 with the Amazfit GTR 2 device. Prior to this date, missingness is best characterised as Missing Not At

Random (MNAR), driven by hardware absence rather than random omission. From 2021–05–11 onwards, the Missing At Random (MAR) assumption is more plausible, making imputation statistically defensible.

F.7 MICE Configuration

After temporal filtering, 11.9% of the retained dataset (1,938 missing values across 1,671 days × 10 features) remains incomplete. To address this, a segment-aware multiple imputation by chained equations (MICE) strategy is applied using `sklearn.experimental.IterativeImp`

Distinction from fabrication detection: The pipeline distinguishes between two types of data issues. *Fabrication artifacts* are algorithmically generated placeholder values (e.g., `hr_min = hr_mean`, `hr_std = 0.0`) that were present in earlier pipeline versions and are now detected and flagged by the QC system. In contrast, *legitimate missingness* occurs when sensor data are genuinely absent due to device non-wear, synchronization gaps or early-period hardware limitations. MICE imputation addresses only the latter, using scientifically valid statistical methods to infer plausible values from observed patterns within each behavioural segment.

Configuration:

- `max_iter=10`: convergence iterations for chained regression,
- `random_state=42`: ensures reproducibility,
- `sample_posterior=True`: stochastic imputation (not deterministic mean).

Segment-aware strategy: Imputation is performed *separately* for each behavioural segment (as defined in Section 3.5 of the thesis). This prevents the algorithm from borrowing information across qualitatively different behavioural periods or device eras. Segments with fewer than five days are excluded from imputation to avoid unstable estimates.

Features imputed: Ten raw physiological features (sleep hours, sleep efficiency, HR mean/min/max/std, HR sample count, steps, distance, active energy) are imputed. PBSI subscores and composite scores are *excluded* to prevent target leakage.

Validation: After MICE imputation, 0 NaN values remain in the ML dataset (1,671 days × 12 columns). Anti-leak verification confirms that PBSI-derived features are not present in the imputed output.

F.8 Feature Set Specifications

Stage 5 incorporates an Experiment Suite that performs controlled ablation across four feature sets and deterministically selects optimal configurations. The feature sets are:

- **FS-A (10 features):** Baseline sleep, cardiovascular and activity features (`sleep_hours`, `sleep_quality_score`, `hr_mean/min/max/std`, `hr_samples`, `total_steps`, `total_distance`, `total_active_energy`).
- **FS-B (15 features):** FS-A plus HRV features (`hrv_sdnn_mean`, `hrv_sdnn_median`, `hrv_sdnn_min`, `hrv_sdnn_max`, `n_hrv_sdnn`).
- **FS-C (18 features):** FS-B plus medication features (`med_any`, `med_event_count`, `med_dose_total`).
- **FS-D (19 features):** FS-C plus PBSI as auxiliary feature (`pbsi_score`).

The Experiment Suite applies deterministic selection rules (highest F1-macro, then highest Cohen's kappa, then fewer features) to select the optimal configuration for ML6 and ML7, producing the `model_selection.json` artifact.

F.9 Model Architecture Specifications

F.9.1 ML6-Extended: Classical Tabular Models

ML6-Extended evaluates 10 classical machine learning algorithms:

1. Logistic Regression (baseline)
2. Random Forest
3. Gradient Boosting
4. XGBoost
5. LightGBM
6. SVM with Linear kernel
7. SVM with RBF kernel
8. Gaussian Naive Bayes
9. K-Nearest Neighbours ($k = 3$)
10. K-Nearest Neighbours ($k = 5$ and $k = 7$)

All models are trained using:

- six temporal folds (4 months training, 2 months validation),
- balanced class weights where applicable,
- macro-F1 as the primary evaluation metric.

Results are aggregated into a model leaderboard (`model_leaderboard.json`) ranking algorithms by mean macro-F1 across folds.

F.9.2 ML7-Extended: Deep Learning Sequence Models

ML7-Extended evaluates 4 deep learning architectures:

1. LSTM (Long Short-Term Memory)
2. GRU (Gated Recurrent Unit)
3. Conv1D (1D Convolutional Network)
4. CNN-LSTM (Hybrid Convolutional-Recurrent)

F.10 QC Audit Results

F.10.1 Regression-Based ETL Auditing

Three domain-specific audit routines were implemented:

Heart rate audit (`make qc-hr`). This audit verifies the consistency between event-level HR data, daily aggregates and cached Parquet summaries. It checks: (1) completeness of the five canonical HR metrics (mean, min, max, std, sample count); (2) absence of fabricated values such as `hr_min = hr_mean` or `hr_std = 0.0` except on single-sample days; (3) biological plausibility of heart rate values, constrained to the range 30–220 bpm.

Steps audit (`make qc-steps`). This routine ensures the correctness of daily step counts across raw, aggregated and unified sources. It evaluates: (1) completeness of daily

records, (2) non-negativity and plausible range of step counts, (3) correct application of the source-priority rule favouring Apple Health data.

Sleep audit (`make qc-sleep`). This routine validates sleep aggregations by checking: (1) the presence of daily sleep summaries where expected, (2) plausibility of sleep duration values within the 0–24 hour range, (3) consistency between sleep episodes and daily summaries.

F.10.2 QC Outputs

Each audit generates two artefacts under the snapshot-specific QC directory:

- a per-day CSV table containing all validation outcomes for that domain, and
- a JSON summary with aggregated indicators such as fabrication rate, missingness and range violations.

Exit codes follow a deterministic convention (0 = PASS, 1 = FAIL), enabling automated reproducibility checks via CI/CD workflows.

F.10.3 Validation Results

Across the final reproducible snapshot (2025–12–09), all audited domains passed QC:

- HR audit: 1,315 validated days with a fabrication rate of 0.5% (consistent only with single-sample days);
- Steps audit: 2,730 validated days with 100% coverage;
- Sleep audit: 958 valid sleep days, all within plausible physiological limits.

F.11 Dataset Statistics

Property	Value
Time span (full EDA)	2017–12–04 to 2025–12–07
Time span (SoM ML)	2023–11–19 to 2025–12–07
Total days (full)	2,868
SoM-labelled days	77
Apple HR/sleep records parsed	5,986
Zepp cardio records parsed	960
Unified daily features	15 (FS-B: sleep, HR, HRV, activity)
Missing values (post-imputation)	0
Behavioural segments	122 (full timeline)

F.12 Segmentation Trigger Statistics

The segmentation procedure detected 122 behavioural segments across the full eight-year timeline (2017–2025). After applying the ML temporal filter ($\geq$ 2021–05–11), 57 segments remained for ML6/ML7 training. Segments are triggered by:

- calendar boundaries (93 time_boundary segments),
- gaps longer than one day (28 gap segments),
- initial segment (1 segment).

Segment-wise normalisation significantly reduced intra-segment variance while preserving long-term non-stationarity, preventing leakage and improving interpretability.

F.13 Extended Model Performance Matrix

F.13.1 Extended ML6 Models

Table 8: Extended ML6 models: SoM-binary classification (macro-F1 and related metrics, 77 samples).

Model	F1-macro	Balanced Acc.	Cohen κ
Random Forest	**0.517 ± 0.133**	**0.542**	**0.104**
KNN-5	0.503 ± 0.090	0.528	0.071
Gradient Boosting	0.499 ± 0.094	0.510	0.021
KNN-3	0.486 ± 0.100	0.497	0.027
XGBoost	0.478 ± 0.101	0.495	0.013
KNN-7	0.458 ± 0.022	0.500	0.000
SVM (RBF)	0.443 ± 0.045	0.539	−0.002
Gaussian NB	0.407 ± 0.094	0.432	−0.044
Logistic Regression	0.378 ± 0.096	0.414	−0.106
SVM (Linear)	0.306 ± 0.097	0.331	−0.198

F.13.2 Extended ML7 Sequence Models

Table 9: Extended ML7 models: 14-day sequence classification on SoM-binary (64 sequences).

Model	F1-macro	Balanced Acc.	Cohen κ
CNN_LSTM	**0.526 ± 0.155**	**0.633**	**0.160**
LSTM	0.521 ± 0.234	0.648	—
Conv1D	0.437 ± 0.035	0.541	−0.019
GRU	0.427 ± 0.222	0.571	0.070

References

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[3] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. Available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[4] Bifet, A., & Gavalda, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. *SIAM International Conference on Data Mining*, 443–448. Available at: https://epubs.siam.org/doi/abs/10.1137/1.9781611972771.42

[5] Onnela, J.-P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, 41(7), 1691–1696. Available at: https://www.nature.com/articles/npp20167

[6] Insel, T. R. (2017). Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*, 318(13), 1215–1216. DOI: https://doi.org/10.1001/jama.2017.11295

[7] Schork, N. J. (2015). Personalized Medicine: Time for One-Person Trials. *Nature*, 520(7549), 609–611.
DOI: https://doi.org/10.1038/520609a

[8] Cornet, V. P., & Holden, R. J. (2018). Systematic Review of Smartphone-Based Passive Sensing for Health and Wellbeing. *Journal of Biomedical Informatics*, 77, 120–132.
Available at: https://www.sciencedirect.com/science/article/pii/S1532046417302782

[9] Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors. *npj Digital Medicine*, 3, 18.
DOI: https://doi.org/10.1038/s41746-020-0226-6

[10] Shcherbina, A., et al. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, 7(2), 3.
DOI: https://doi.org/10.3390/jpm7020003

[11] Hidalgo-Mazzei, D., et al. (2015). Self-monitoring and psychoeducation in bipolar patients with a smartphone application (SIMPLe) project: design, development and studies protocols. *BMC Psychiatry*, 15, 52.
Available at: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-015-0437-6

[12] Torous, J., Wisniewski, H., Liu, G., & Keshavan, M. (2018). Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. *JMIR Mental Health*, 5(4), e11715.
DOI: https://doi.org/10.2196/11715

[13] Kollins, S. H., et al. (2020). A Novel Digital Intervention for Actively Reducing Severity of Paediatric ADHD (STARS-ADHD): A Randomised Controlled Trial. *The Lancet Digital Health*, 2(4), e168–e178.
DOI: https://doi.org/10.1016/S2589-7500(20)30017-0

[14] Faedda, G. L., et al. (2016). Actigraph Measures Discriminate Pediatric Bipolar Disorder from Attention-Deficit/Hyperactivity Disorder and Typically Developing Controls. *Journal of Child Psychology and Psychiatry*, 57(6), 706–716.
DOI: https://doi.org/10.1111/jcpp.12520

[15] Depner, C. M., et al. (2020). Wearable Technologies for Developing Sleep and Circadian Biomarkers: A Summary of Workshop Discussions. *Sleep*, 43(2), zsz254.
DOI: https://doi.org/10.1093/sleep/zsz254

[16] Place, S., et al. (2016). Behavioral Indicators on a Mobile Sensing Platform Predict Clinically Validated Psychiatric Symptoms of Mood and Anxiety Disorders. *Journal of Medical Internet Research*, 19(3), e75.
Available at: https://www.jmir.org/2017/3/e75/

[17] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4), 1–37.
DOI: https://doi.org/10.1145/2523813

[18] Cerqueira, V., Torgo, L., & Mozetiè, I. (2020). Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Machine Learning*, 109, 1997–2028.
DOI: https://doi.org/10.1007/s10994-020-05910-7

[19] Ben-Zeev, D., Brian, R. M., Jonathan, G., Razzano, L., Pashka, N., Carpenter-Song, E., Drake, R. E., & Scherer, E. A. (2018). Mobile Health (mHealth) Versus Clinic-Based Group Intervention for People With Serious Mental Illness: A Randomized Controlled Trial. *Psychiatric Services*, 69(9), 978–985.
DOI: https://doi.org/10.1176/appi.ps.201800063

[20] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
DOI: https://doi.org/10.1002/9780470316696

[21] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.
Available at: https://www.jstatsoft.org/article/view/v045i03

[22] Sterne, J. A. C., et al. (2009). Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, 338, b2393.
DOI: https://doi.org/10.1136/bmj.b2393