Self-Reported Symptoms Enable Four-Phase Menstrual Cycle Classification with Hormonally Validated Labels

Bernhard Specht*[1,2], Mohammad EL-Khozondar[1], Samaher Garbaya[1], Reinhard Schneider[3], Djamel Khadraoui[4], and Zied Tayeb[1]

[1]MyelinZ, Manchester, UK / San Francisco, USA / Leuven, Belgium / Luxembourg
[2]University of Luxembourg, Esch-sur-Alzette, Luxembourg
[4]Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg
[3]Luxembourg Centre for Systems Biomedicine (LCSB), Esch-sur-Alzette, Luxembourg

Abstract

Accurate inference of physiological state across the menstrual cycle has important applications in reproductive health and in understanding symptom dynamics, yet most non-hormonal approaches rely on wearable sensors or calendar-based tracking. Whether self-reported symptoms alone can support prospective, cross-subject phase classification remains unresolved. Here, we introduce a hybrid modelling framework that combines a gradient-boosted classifier with a Hidden Semi-Markov Model to infer four menstrual cycle phases (menstrual, follicular, fertile, and luteal) from self-reported data. The classifier captures non-linear symptom patterns, while the temporal model imposes biologically grounded constraints, including cyclic ordering and realistic phase durations. In a leave-one-subject-out evaluation using hormonally annotated data from 41 participants, the model achieved 67.6% accuracy and a macro F1 score of 0.662. Features reflecting short-term symptom variability were more informative than absolute symptom levels, indicating that within-person fluctuation provides a more generalisable signal of cycle phase than symptom intensity alone. These findings demonstrate the feasibility of low-burden, device-free menstrual health monitoring, establish symptom dynamics as a basis for scalable digital biomarkers, and expand access to tracking in resource-constrained settings and populations underserved by wearable-based approaches.

1 Introduction

Self-reported symptoms are the most widely collected form of menstrual cycle data, recorded daily by millions of people through period-tracking apps. Yet it remains unknown whether these subjective reports contain sufficient physiological signal to distinguish cycle phases, or whether inter-individual variability and reporting noise render them uninformative beyond menstrual bleeding itself. Answering this question matters for two reasons: it clarifies what symptom reports can and cannot tell us about underlying cycle biology, and it informs whether future lightweight collection methods (fewer questions, less frequent check-ins) are worth pursuing.

The menstrual cycle is governed by the hypothalamic-pituitary-ovarian axis, producing hormonal fluctuations that drive a recurring sequence of four clinically distinct phases: Menstrual, Follicular, Fertility (Ovulatory), and Luteal (Figure 1) [1, 2]. The cycle lasts approximately 28 days, though normal lengths range from 25 to 30 days. Ground-truth phase boundaries are established through daily hormonal assays (LH, estradiol, progesterone), making accurate phase tracking clinically relevant for fertility planning, contraception, and management of cycle-related conditions.

Current non-hormonal approaches to phase tracking vary widely in accuracy depending on the data modality, task definition, and evaluation protocol. Calendar-based methods, which rely solely on cycle

length history, predict ovulation day with at best 21% accuracy [3], and fertility apps perform similarly poorly, with only 4 of 53 evaluated web sites and apps correctly predicting the fertile window [4].

Wearable sensors that passively record continuous physiological signals (skin temperature, heart rate, electrodermal activity) achieve substantially higher accuracy. A systematic review of 27 studies reported pooled accuracy of 88% for *binary* fertile window detection (fertile vs. not fertile) from wearable data [5], and a study using the Ava wristband reported 90% for the same binary task [6]. However, these numbers reflect a simpler two-class problem with objective, high-frequency sensor data. When the task is extended to multi-phase classification under strict leave-one-subject-out (LOSO) evaluation, wearables achieve 87% on three phases (Period, Ovulation, Luteal) using fixed-window features, but only 63–65% on four phases under LOSO evaluation [7].

Self-reported symptom data from mobile health apps has also been explored. Large-scale analysis of 4.9 million self-tracked cycles demonstrated statistically significant associations between self-reported symptom patterns and cycle timing [8], and a generative HSMM applied to self-tracked records achieved 93% accuracy for retrospective cycle labeling [9]. However, this reflects per-user model fitting on each individual's full tracking history, not cross-subject generalization. To date, no study has attempted prospective, cross-subject four-phase classification from self-reported symptoms alone with hormonally validated ground truth.

The fundamental challenge of prospective, cross-subject four-phase classification from self-reported symptoms lies in the subjective, noisy nature of the data. Symptoms such as fatigue, mood swings, and headaches exhibit high inter-individual variability, recall bias, and non-random missingness. The Fertility phase is particularly difficult: lasting only a few days and driven by a transient estradiol peak rather than the sustained progesterone plateau that characterizes the Luteal phase, it produces subtle symptom shifts that are hard to distinguish from adjacent phases.

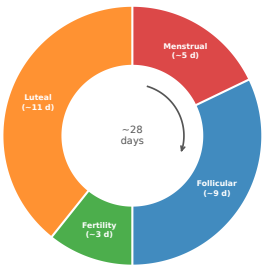

Figure 1: The four phases of the menstrual cycle with approximate durations. Ground-truth boundaries are determined by hormonal assays.

We address this challenge with a hybrid discriminative-generative architecture. A CatBoost gradient-boosted classifier produces per-day phase probabilities from 83 engineered symptom features but is agnostic to temporal structure. These probabilities are fed as discriminative emission scores into an HSMM that enforces the biological constraints of the menstrual cycle: cyclic phase ordering (Menstrual → Follicular → Fertility → Luteal → Menstrual) and Gaussian duration priors that prevent physiologically implausible phase lengths. The classifier handles complex, non-linear feature interactions; the HSMM provides temporal coherence that the classifier alone cannot enforce.

Our contributions are:

1. A hybrid gradient-boosted classifier and HSMM architecture for four-phase menstrual cycle classification from self-reported menstrual and symptom data, achieving 67.6% LOSO accuracy and 0.662 macro F1 on hormonally annotated data.
2. The finding that symptom *variability*, captured by rolling standard deviations, is the dominant predictive signal, not absolute symptom levels, suggesting that within-person fluctuation may be a more portable signal than symptom intensity.
3. To our knowledge, the first cross-subject machine learning benchmark on the mcPHASES dataset using exclusively self-reported data, quantifying the phase information content of subjective symptoms.

More broadly, the dominance of variability features over absolute symptom levels suggests that phase-related signal in self-reports is carried by within-person fluctuation rather than symptom intensity. This distinction may inform feature selection in future symptom-based systems and the design of shorter, more targeted questionnaires.

2 Methods

2.1 Dataset

We used the mcPHASES dataset [10], a publicly available longitudinal dataset hosted on PhysioNet comprising data from 42 menstruating participants recruited in the Greater Toronto Area, Canada. Participants were aged 18–29 (mean 20.6, median 20) and were required to have had no hormonal therapy or contraception for at least three months prior to enrolment; 14 of the 50 originally enrolled participants self-identified as having irregular cycles [11]. No body mass index data were collected. Participants were tracked over an initial three-month observation period (2022), with a subset of 20 returning for an ad-

ditional three months (2024). The dataset includes daily self-reported symptoms, continuous wearable physiology (Fitbit Sense), continuous glucose monitoring (Dexcom G6), and ground-truth hormonal assays (LH, Estradiol, Progesterone via Mira Plus urinalysis). Ground-truth phase labels (Menstrual, Follicular, Fertility, Luteal) were assigned by the dataset authors based on these hormonal assays.

We restricted our analysis to the 2022 observation period, as 99.4% of daily self-report entries in the 2024 period had no symptom data recorded. From this period, we used exclusively the self-reported symptom channels: 14 daily variables comprising flow volume (8-point scale from "Not at all" to "Very Heavy"), flow color (9 categories), and 12 symptom scales rated on a 6-point Likert scale from "Not at all" to "Very High" (appetite, exercise level, headaches, cramps, sore breasts, fatigue, sleep issues, mood swings, stress, food cravings, indigestion, bloating).

We excluded cycles containing ≥5 consecutive days with no self-report data. Because the feature set relies on rolling statistics computed over 5-, 7-, and 14-day windows, a gap of 5 or more consecutive days renders these statistics unreliable for the surrounding period, propagating missing-data artifacts into the classifier input. This filter removed 140 days (3.8% of the 2022 data) and one participant whose data consisted entirely of gappy cycles, reducing the cohort from 42 to 41 participants. After filtering, 3,557 participant-days with valid phase labels remained (2,983 after the 14-day warmup required for rolling features). Figure 2 summarizes the participant flow.

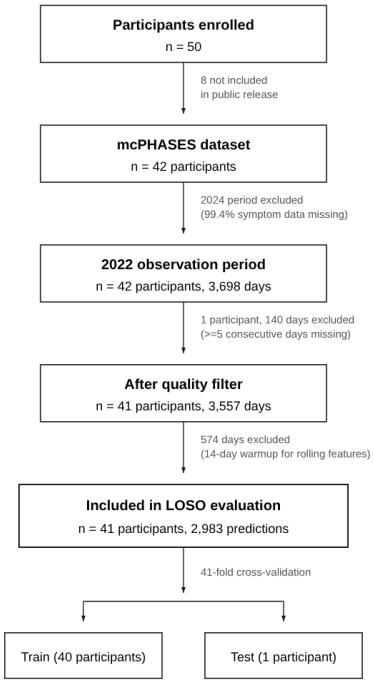

Figure 2: Participant and data flow from enrolment to the final evaluation set. Numbers at each stage reflect the cumulative effect of exclusion criteria.

2.2 Feature Engineering

We constructed 97 features from the 14 raw self-report variables, organized into four categories (see Appendix A for the complete glossary).

Cycle position features (13 features). We computed days-since-bleed counters using a grid of bleeding thresholds $t \in \{1, 2, 4\}$ and minimum consecutive-day streak lengths $c \in \{1, 2, 3\}$, yielding 9 variants. Multiple thresholds are necessary because menstrual bleeding patterns vary: some subjects exhibit inter-menstrual spotting (low flow between periods), and requiring a higher flow threshold or a longer streak prevents these episodes from resetting the counter prematurely. From the primary counter (days since ≥1 consecutive day with flow ≥2), we derived four additional features: a linear cycle progress (clipped to [0, 1] assuming a 28-day cycle), a modular cycle day, and a cyclic encoding pair:

$$\sin_{\text{cyc}} = \sin\Big(\frac{2\pi\, d}{T}\Big), \quad \cos_{\text{cyc}} = \cos\Big(\frac{2\pi\, d}{T}\Big), \qquad (1)$$

where d is the days-since-bleed count and $T = 28$ is the assumed cycle length. This encoding ensures that cycle positions near the boundary (e.g., day 27 and day 1) remain numerically close, which the linear counter alone does not guarantee.

Rolling standard deviations (42 features). Standard deviations computed over 5-, 7-, and 14-day sliding windows for all 14 self-report variables. These capture symptom *variability*, the rate at which the physiological state fluctuates, rather than absolute symptom levels.

Rolling means (28 features). Means computed over 7- and 14-day windows for all 14 variables, representing medium- to long-term symptom trends.

Raw values (14 features). The original 14 self-report variables were retained as features to preserve same-day symptom levels.

The first 14 days of each participant's data were excluded as a warmup period to ensure sufficient history for rolling statistics. We then removed features with pairwise Pearson $|r| > 0.9$ to reduce multicollinearity. This step removed approximately 14 features per fold (predominantly rolling means correlated with their counterpart at the other window length, and redun-

dant cycle position variants), yielding a consistent set of approximately 83 features per LOSO fold.

2.3 Classification Model

We used CatBoost [12], a gradient-boosted decision tree algorithm, for per-day phase classification. CatBoost was selected for its native handling of missing values [13] (common in self-reported data) and ordered boosting to reduce prediction shift on small sample sizes.

Hyperparameters were tuned via Optuna [14] Bayesian optimization on an 80/20 participant-level split, optimizing macro F1 over 3-fold cross-validation. The final configuration used depth = 7, learning rate = 0.107, L2 regularization = 5.92, and minimum 7 samples per leaf.

2.4 Temporal Smoothing via HSMM

Although CatBoost incorporates temporal context through rolling and lag features, it still scores each day independently and cannot enforce biologically valid phase ordering. This can produce impossible sequences (e.g., Luteal → Menstrual → Luteal within a single cycle). We address this with a Hidden Semi-Markov Model (HSMM) that overlays temporal structure on the classifier's output (Figure 3).

Let $S_t \in \{0, 1, 2, 3\}$ denote the phase state at day t and $\mathbf{x}_t$ the corresponding feature vector. The HSMM combines three components: discriminative emission scores derived from CatBoost, explicit duration priors for each phase, and left-right cyclic transition constraints. The Viterbi algorithm jointly optimizes the state sequence to maximize the total score across all three components.

Discriminative emissions. Rather than learning emission distributions from scratch, we convert the CatBoost predicted class probabilities $P(S_t \mid \mathbf{x}_t)$ into log-emission scores: $e_{s,t} = \alpha \cdot \log\big(P(S_t{=}s \mid \mathbf{x}_t) + \epsilon\big)$, where $\epsilon = 10^{-10}$ prevents numerical issues when a class probability is near zero. The scaling parameter α controls the relative influence of the classifier's confidence versus the duration priors: larger α trusts the classifier more, while smaller α lets the duration model dominate. We select α per LOSO fold via an inner validation split from the candidates $\{0.5, 1.0, 2.0, 5.0, 10.0\}$.

Duration priors. Standard HMMs assume geometric (memoryless) duration distributions, which poorly match the structured timing of menstrual cycle phases. Our HSMM instead models phase durations explicitly. For each phase s and each training fold, we compute the empirical mean μ_s and standard deviation σ_s of observed phase durations, then construct a discrete probability distribution $D_s(d) \propto \mathcal{N}(d; \mu_s, \sigma_s)$ for durations $d \in [d^{(s)}_{\min}, d^{(s)}_{\max}]$, where the bounds are the minimum and maximum durations observed in the training data. Durations outside these bounds receive zero probability. This ensures physiologically plausible phase lengths (e.g., preventing a 1-day Luteal phase) while allowing natural variation within observed ranges.

Transition constraints. The transition matrix enforces strict left-right cyclic order: Menstrual → Follicular → Fertility → Luteal → Menstrual. Because the start phase of a test sequence is unknown, we run the Viterbi algorithm from all four possible start states and select the path with the highest total score.

2.5 Evaluation

We evaluated using Leave-One-Subject-Out (LOSO) cross-validation: for each of the 41 participants, the model was trained on the remaining 40 and tested on the held-out participant. The held-out participant was excluded from all stages of each fold: CatBoost training, feature correlation filtering, and HSMM emission-scaling (α) tuning, which used a separate inner validation participant drawn from the training set. LOSO is the strictest form of participant-level evaluation, ensuring that no data from the test participant influences training. This is critical for self-reported symptom data, where individual reporting styles and baseline symptom levels vary substantially; a random train/test split would allow the model to memorize subject-specific patterns rather than learning generalizable phase signatures. With only 41 subjects, LOSO also maximizes the training set size in each fold (40 of 41 subjects), which is preferable to a held-out test split that would reduce an already small sample. The feature engineering pipeline was fit independently per fold to prevent data leakage. We report accuracy, macro F1 score, per-phase precision/recall/F1, and per-participant performance distributions.

Baselines. We compared against five baselines evaluated under the same LOSO protocol: (1) a majority-class predictor assigning all days to the most frequent phase (Luteal); (2) a bleed-only logistic regression using only flow volume and flow colour; (3) a calendar-only logistic regression using only cycle position features (days-since-bleed counters, cyclic encoding, cycle progress); (4) a logistic regression using all 83 features; and (5) a random forest (500 trees) using all features. All baselines were trained per fold on the same participant-level split as the primary model.

3 Results

3.1 Overall Performance

Table 1 summarizes the overall LOSO performance across all models. The CatBoost+HSMM pipeline achieves 67.6% accuracy and 0.662 macro F1, out-

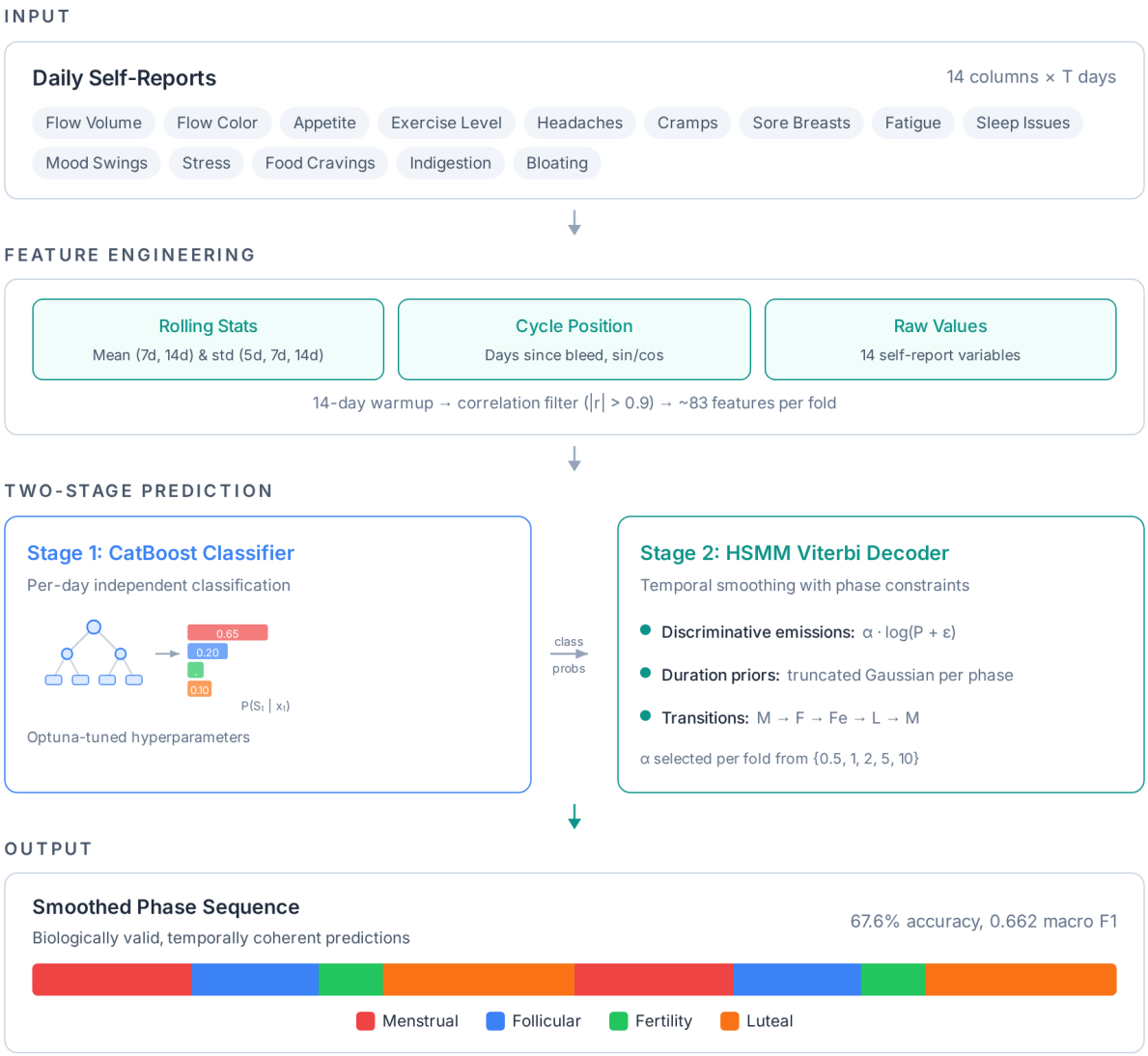

Figure 3: Overview of the two-stage prediction pipeline. Daily self-reported symptoms are transformed into rolling, positional, and interaction features, then classified by CatBoost. The HSMM Viterbi decoder smooths the resulting class probabilities using discriminative emissions, duration priors, and left-right cyclic transition constraints to produce a biologically valid phase sequence.

Table 1: Overall LOSO classification performance with 95% bootstrap CIs (10,000 subject-level resamples).

Model	Accuracy [95% CI]	Macro F1 [95% CI]	MCC	Cohen's κ
Majority class	.352 [.319, .383]	.130 [.121, .139]	.000	.000
Bleed-only (LR)	.467 [.428, .505]	.354 [.329, .378]	.291	.215
Logistic regression	.609 [.566, .650]	.590 [.547, .629]	.465	.463
Calendar-only (LR)	.630 [.584, .672]	.606 [.560, .650]	.493	.490
Random forest	.653 [.605, .700]	.633 [.582, .681]	.525	.522
CatBoost	.651 [.608, .692]	.629 [.586, .671]	.523	.519
CatBoost+HSMM	**.676** [.633, .717]	**.662** [.618, .702]	**.556**	**.553**

performing all baselines including CatBoost alone (65.1%, 0.629), random forest (65.3%, 0.633), and a calendar-only model using cycle position features (63.0%, 0.606). A majority-class baseline achieves 35.2% accuracy. All 95% confidence intervals are computed via subject-level bootstrap (10,000 resamples).

Table 2: Per-phase classification report (CatBoost+HSMM, LOSO).

Phase	Prec.	Rec.	F1	Support
Menstrual	0.812	0.739	0.774	556
Follicular	0.669	0.652	0.661	733
Fertility	0.512	0.420	0.462	645
Luteal	0.694	0.816	0.750	1,049
Macro avg	0.672	0.657	0.662	2,983

3.2 Ablation: CatBoost vs CatBoost+HSMM

Figure 4 presents the confusion matrices for both models alongside per-phase F1 scores. The HSMM temporal smoothing provides a consistent improvement over CatBoost alone, increasing accuracy from 65.1% to 67.6% (+2.4 pp) and macro F1 from 0.629 to 0.662 (+3.2 pp). The improvement was statistically significant (paired Wilcoxon signed-rank test on per-participant F1, $W = 658$, $p = 0.001$), with the HSMM improving 29 of 41 participants. The improvement is observed across all four phases, with the largest gains for Fertility, the most challenging phase due to its short duration and subtle symptom signature (F1 0.398 $\rightarrow$ 0.462, +6.4 pp).

The dominant error pattern is Fertility misclassified as Luteal (33%) or Follicular (25%), consistent with the physiological difficulty of detecting the transient ovulatory window from secondary symptoms.

Table 3 shows the effect of removing each feature group on CatBoost classification performance (without HSMM smoothing). Cycle position features contribute the largest individual drop in macro F1 (-1.4 pp), consistent with their role as a structural backbone for phase prediction. Rolling standard deviations and rolling means show equal ablation impact (-0.8 pp each), while raw values have the smallest impact (-0.2 pp), suggesting that temporal aggregations capture most of the information present in same-day symptom levels. All four groups contribute to performance, with no single group accounting for the majority of the signal.

Table 3: Feature group ablation (CatBoost only).

Group removed	F1	ΔF1
None (full model)	.629	–
Cycle position (10)	.615	−.014
Rolling mean (18)	.621	−.008
Rolling std (41)	.622	−.008
Raw values (14)	.627	−.002

3.3 Feature Importance

Figure 5 shows SHAP-based feature importance aggregated across all LOSO folds. Rolling standard deviations dominate the top features, confirming that the model detects phase transitions through *changes in symptom stability* rather than absolute symptom levels. Cycle position features (days-since-bleed) also rank highly, providing a strong structural backbone for phase prediction.

3.4 Per-Phase SHAP Analysis

Figure 6 presents per-phase SHAP beeswarm plots revealing distinct feature signatures for each phase. Menstrual classification relies primarily on flow-related features and cycle position. Follicular and Luteal phases are distinguished by different patterns of symptom variability. Fertility classification, the hardest phase, depends on a combination of cycle position and subtle variability signals, explaining its lower F1 score.

3.5 Per-Subject Analysis

Per-participant F1 varied substantially (mean 0.648 $\pm$ 0.145), reflecting the inherent inter-individual variability in symptom expression. Figure 7 shows per-participant F1 for CatBoost (x-axis) vs CatBoost+HSMM (y-axis); the HSMM improved classification for 29 of 41 participants (71%).

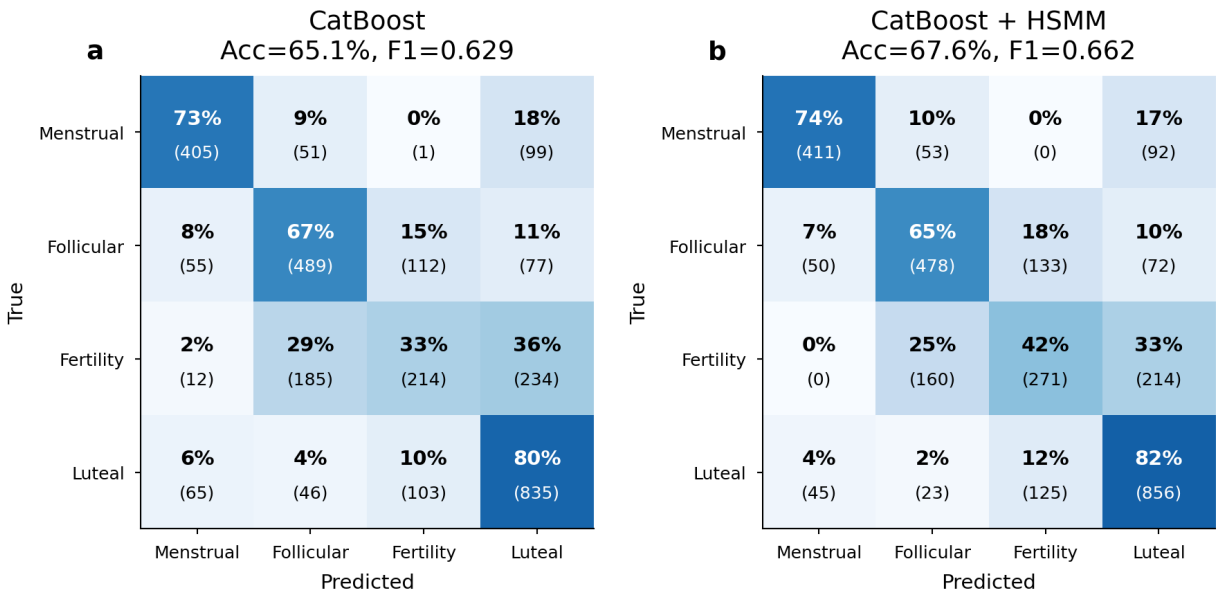

Figure 4: Normalized confusion matrices under LOSO cross-validation for CatBoost alone (a) and CatBoost+HSMM (b). Cells show percentage (bold) and absolute count.

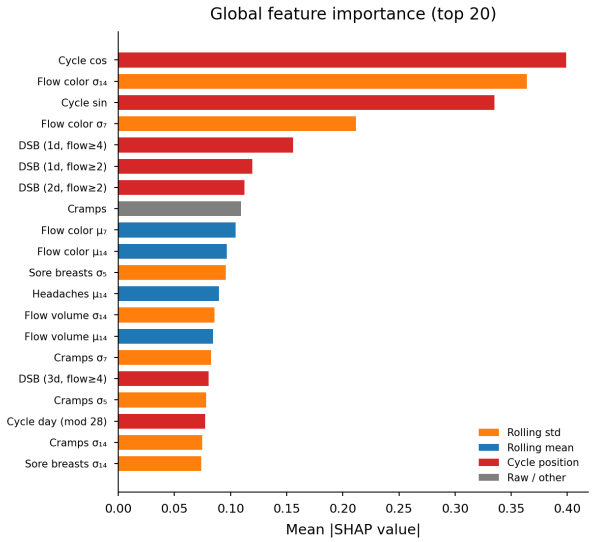

Figure 5: Global SHAP feature importance (mean |SHAP| across all classes and LOSO folds). Colors indicate feature category.

3.6 Phase Timeline Visualization

Figure 8 presents phase timelines for the three best- and three worst-performing participants alongside daily flow intensity. Best participants exhibit clear cyclic flow patterns and near-perfect alignment between true and predicted phases. Worst participants show sparse or absent flow data, demonstrating the model's dependence on flow-derived cycle position features.

4 Discussion

4.1 Symptom Signal Is Stronger Than Expected

The central finding is that self-reported symptoms alone carry enough phase-discriminative information to reach 67.6% LOSO accuracy on four-phase classification. This is notable given the inherent subjectivity, recall bias, and missingness of the data. For context, Kilungeja et al. [7] reported 63–65% on the same four-phase task using wearable sensors (skin temperature, EDA, heart rate), though direct comparison is limited by differences in dataset, cohort size, and preprocessing.

It is worth emphasizing what this result does *not* claim. Daily completion of 14 Likert scales is not a practical alternative to passive wearable monitoring; the reporting burden alone makes it unsuitable for routine use. Rather, the value of this baseline is scientific: it quantifies the phase information content of subjective symptoms and establishes a floor that lighter-touch collection methods (fewer questions, weekly check-ins, free-text parsing) could aim to approach. It also suggests that future multimodal systems combining even a small number of symptom questions with passive sensor data may outperform either modality alone.

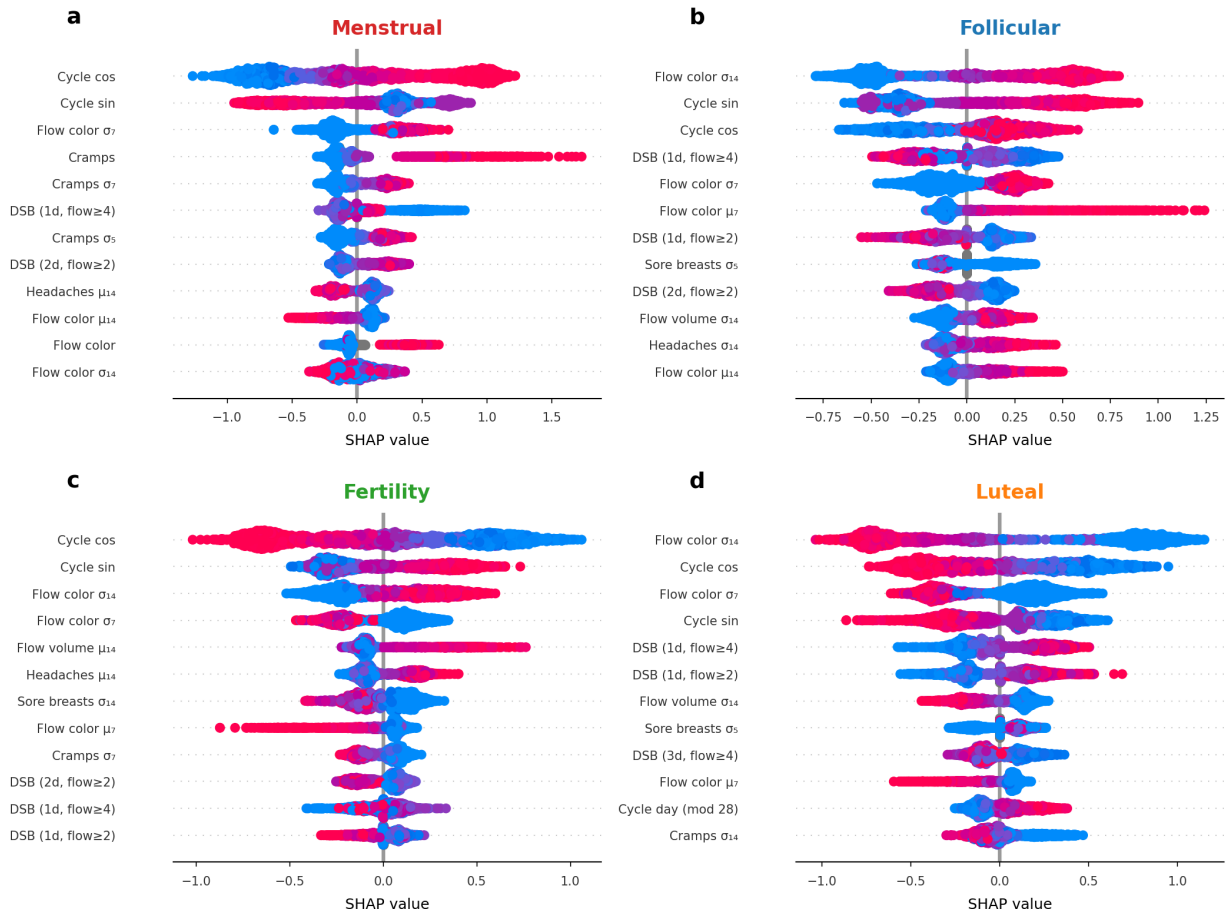

Figure 6: Per-phase SHAP beeswarm plots. Each panel shows the top 12 features driving classification of one phase. Dot color indicates feature value (blue = low, red = high); horizontal position indicates SHAP contribution to the phase prediction.

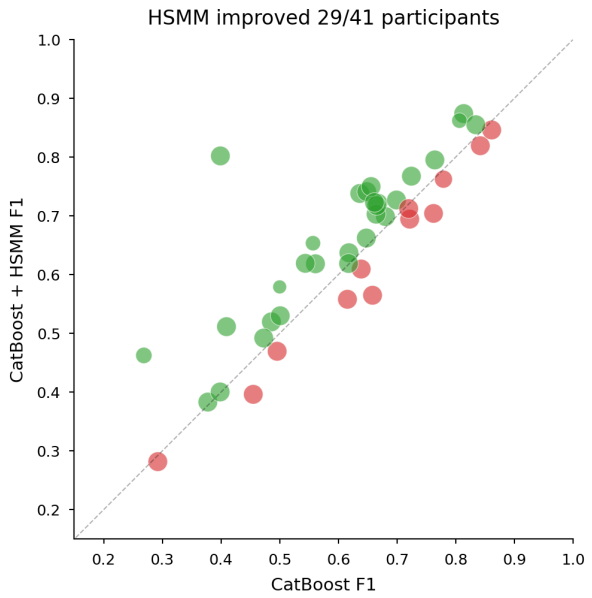

Figure 7: Per-participant F1 scores: CatBoost vs CatBoost+HSMM. Green points (above diagonal) indicate HSMM improvement. Point size reflects number of observation days.

Beyond benchmarking, these results have implications for menstrual health research more broadly. First, they demonstrate that self-reported symptom data, the most accessible form of cycle tracking, carries meaningful phase signal, supporting its use in low-cost reproductive health tools where wearable devices are unavailable or impractical. Second, the finding that variability features outperform absolute levels contributes to symptom science by suggesting that within-person dynamics, rather than cross-sectional symptom severity, are the more informative axis of variation. Third, because the approach requires no specialised hardware, it is applicable to digital phenotyping in populations historically underserved by device-first research, including people in resource-constrained settings and those who cannot or choose not to wear sensors.

4.2 Symptom Variability as a Physiological Proxy

Rolling standard deviations account for 45% of feature importance, far exceeding the contribution of absolute symptom levels. One plausible interpretation is that these features capture the downstream symptomatic effects of hormonal gradients at phase boundaries: the LH surge triggering ovulation and progesterone withdrawal initiating menstruation are rapid endocrine events that may transiently disrupt symptom homeostasis. A rolling standard deviation would register such disruptions as elevated dispersion, even when mean symptom levels remain moderate. We note, however, that this mechanistic link remains a hypothesis; the present study does not measure hormones directly.

This pattern also has a statistical explanation. Variance decomposition of daily mood reports across menstrual cycles shows that 79–98% of total variance arises from day-to-day fluctuations, with between-person differences accounting for up to 16% [15]. Because the menstrual cycle is fundamentally a within-person process [16], absolute symptom levels conflate true phase-related signal with idiosyncratic reporting baselines. A rolling standard deviation is invariant to each participant's intercept: whether a participant reports cramps as "3" or "6", the feature captures the *amplitude of change* rather than the level itself. This effectively achieves the same goal as formal person-centering [16], making variance features inherently more portable across participants than raw scores.

4.3 The Role of Temporal Constraints

The HSMM contributed a consistent +2.4 pp accuracy improvement by enforcing two biological constraints: cyclic phase ordering and realistic phase durations. The optimal emission scaling parameter ($\alpha = 0.5$, selected most frequently across folds) down-weights the CatBoost log-probabilities relative to the duration priors, indicating that the decoder benefits from relying more heavily on expected phase durations than on per-day classifier confidence.

In contrast to the generative HSMM of Symul and Holmes [9], which models emissions directly from symptom distributions, our approach uses the HSMM purely as a temporal regularizer over CatBoost class probabilities. This two-stage architecture delegates feature-level modeling to the gradient booster and sequential plausibility to the HSMM, avoiding the need to specify parametric emission distributions for high-dimensional engineered features.

4.4 Limitations

Fertility phase performance. The model achieves only 0.462 F1 on the Fertility phase, with 57% of Fertility days misclassified as Follicular or Luteal. Since fertility tracking is among the most clinically relevant applications of cycle phase prediction, this performance gap limits practical utility for the phase where accuracy matters most.

Dependence on flow reporting. Cycle position features (days-since-bleed variants) rank among the strongest predictors, and the worst-performing participants in our evaluation are those with sparse or absent flow data (Figure 8). This means the model partially operates as a calendar method informed by the last observed bleed. For users who do not re-

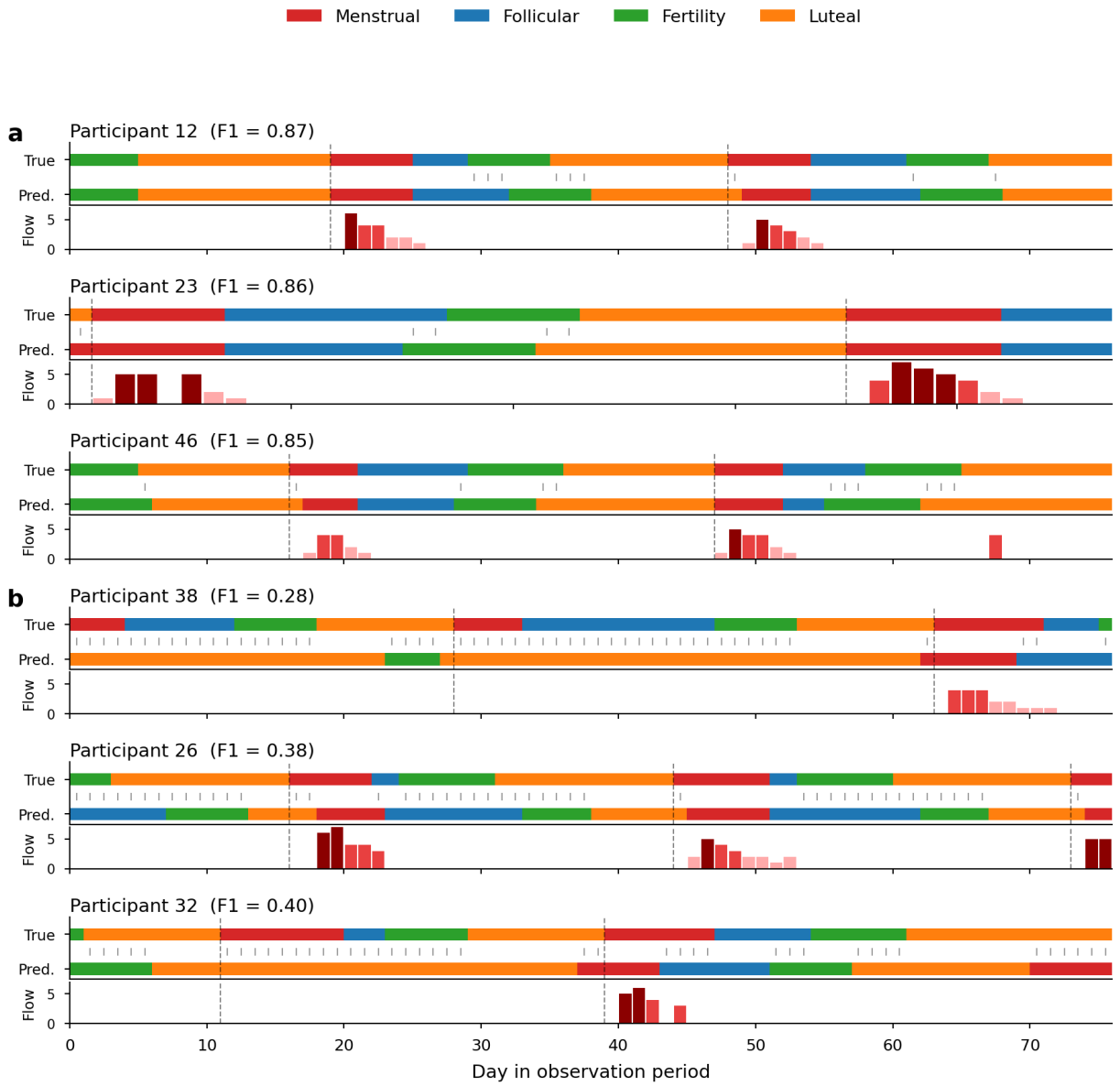

Figure 8: Phase timelines for (a) three best-performing and (b) three worst-performing participants. Each participant shows true phases (top strip), predicted phases (middle strip), and daily flow intensity (bars). Tick marks indicate misclassified days.

port flow consistently, performance is likely to degrade substantially.

Sample size and demographic homogeneity. The mcPHASES cohort comprises 41 young adult Canadian participants (aged 18–29), predominantly of East Asian and Caucasian background, with no BMI data collected. Hormonal contraception was an exclusion criterion, so generalisability to contraceptive users is unknown. Larger and more diverse cohorts are needed to assess external validity across age, ethnicity, BMI, contraceptive status, cycle regularity, and reproductive health conditions such as PCOS, endometriosis, and perimenopause.

Ground-truth ambiguity. Phase labels in mcPHASES are derived from hormonal assays, the current gold standard. However, phase boundaries are biologically continuous rather than discrete; labeling disagreements near transitions are inevitable and place a ceiling on achievable classification accuracy.

Self-report bias. Non-random missingness (users log more during symptomatic phases) and subjective rating scales remain inherent limitations of approaches based solely on self-reported data.

4.5 Future Directions

The present study deliberately restricted itself to self-reported symptoms in order to isolate their predictive value. A natural next step is multimodal fusion: the mcPHASES dataset includes continuous wearable channels (skin temperature, heart rate, electrodermal activity) that could compensate for the two weaknesses identified above. Skin temperature captures the post-ovulatory basal body temperature shift that self-reported symptoms largely miss, directly addressing the Fertility phase gap, while resting heart rate, which is elevated in the luteal phase, provides a passive cycle anchor that does not depend on user-reported flow [6].

Personalization is a second promising direction. Per-participant F1 varies substantially (0.648 ± 0.145), suggesting that a one-size-fits-all model underserves a substantial fraction of users. Given even a small amount of labeled data per user, the HSMM duration priors and emission scaling could be adapted to individual cycle characteristics, narrowing the gap between best- and worst-performing participants.

Finally, targeted improvements for the Fertility phase deserve attention, given its clinical importance and current 0.462 F1. Phase-aware loss weighting, oversampling of the short Fertility window, or auxiliary regression targets (e.g., days to ovulation) could help the classifier allocate more capacity to the phase that matters most for fertility-tracking applications.

5 Conclusion

We developed and evaluated a cross-subject four-phase menstrual cycle classifier using only daily self-reported menstrual and symptom data. In leave-one-subject-out evaluation on hormonally annotated data from 41 participants, the hybrid CatBoost-HSMM model achieved 67.6% accuracy and 0.662 macro F1, with symptom variability emerging as a stronger predictive signal than symptom intensity. Because variance features are invariant to individual reporting baselines, they may generalise across participants more readily than raw scores, with implications for questionnaire design and feature selection in future symptom-based systems.

The approach has clear limitations. The Fertility phase remains the weakest (F1 0.462), and the model depends heavily on self-reported flow for cycle anchoring. Practical use will require better fertility-phase detection, stronger validation in diverse cohorts, and direct comparison with multimodal approaches that combine targeted symptom questions with passive wearable signals.

Acknowledgements

The authors gratefully acknowledge the support of Innovate UK for co-funding this research. The authors also acknowledge the Luxembourg National Research Fund (FNR) for its partial support of this work.

Funding

This work was supported in part by an Innovate UK grant (BodyMirror AI+: 10173207) and by a PhD grant from the Luxembourg National Research Fund (FNR) under project reference 17223919/MMS/Industrial Fellowship.

Data Availability

This study used the publicly available mcPHASES dataset hosted on PhysioNet. Code and evaluation scripts are available at `https://github.com/mcphases-classification`.

References

[1] Beverly G. Reed and Bruce R. Carr. The normal menstrual cycle and the control of ovulation. In *Endotext*. MDText.com, Inc., 2000. URL `http://www.ncbi.nlm.nih.gov/books/NBK279054/`.

[2] Henry N. Jabbour, Rodney W. Kelly, Hamish M. Fraser, and Hilary O. D. Critchley. Endocrine regulation of menstruation. *Endocrine Reviews*, 27(1):17–46, 2006. doi: 10.1210/er.2004-0021.

[3] Sarah Johnson, Lorrae Marriott, and Michael Zinaman. Can apps and calendar methods predict ovulation with accuracy? *Current Medical Research and Opinion*, 34(9):1587–1594, 2018. doi: 10.1080/03007995.2018.1475348.

[4] Robert Setton, Christina Tierney, and Tony Tsai. The accuracy of web sites and cellular phone applications in predicting the fertile window. *Obstetrics and Gynecology*, 128(1):58–63, 2016. doi: 10.1097/AOG.0000000000001341.

[5] Yue Shi, Chi Chiu Wang, Yongkang Yang, Qin Li, Pui Wah Chung, and Yao Wang. The diagnostic accuracy of wearable digital technology in detecting fertility window and menstrual cycles: a systematic review and Bayesian network meta-analysis. *npj Digital Medicine*, 9:139, 2026. doi: 10.1038/s41746-025-02320-8.

[6] Brianna Mae Goodale, Mohaned Shilaih, Lisa Falco, Franziska Dammeier, Györgyi Hamvas, and Brigitte Leeners. Wearable sensors reveal menses-driven changes in physiology and enable prediction of the fertile window: Observational study. *Journal of Medical Internet Research*, 21 (4):e13404, 2019. doi: 10.2196/13404.

[7] Grentina Kilungeja, Krystal Graham, Xudong Liu, and Mona Nasseri. Machine learning-based menstrual phase identification using wearable device data. *npj Women's Health*, 3(1):29, 2025. doi: 10.1038/s44294-025-00078-8.

[8] Kathy Li, Iñigo Urteaga, Chris H. Wiggins, Anna Druet, Amanda Shea, Virginia J. Vitzthum, and Noémie Elhadad. Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data. *npj Digital Medicine*, 3:79, 2020. doi: 10.1038/s41746-020-0269-8.

[9] Laura Symul and Susan Holmes. Labeling self-tracked menstrual health records with hidden semi-markov models. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1297–1308, 2022. doi: 10.1109/JBHI.2021.3110716.

[10] Blue Lin, Jin Yi Li, Kaavya Kalani, Khai Truong, and Alex Mariakakis. mcphases: A dataset of physiological, hormonal, and self-reported events and symptoms for menstrual health tracking with wearables, 2025. URL `https://physionet.org/content/mcphases/1.0.0/`.

[11] Georgianna Lin, Brenna Li, Helen Li, Chloe Zhao, Khai N. Truong, and Alex Mariakakis. Users' perspectives on multimodal menstrual tracking using consumer health devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–24, 2024. doi: 10.1145/3678575.

[12] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. doi: 10.48550/arXiv.1706.09516.

[13] Yandex. Catboost documentation: Missing value support, 2024. URL `https://catboost.ai/en/docs/concepts/algorithm-missing-values-processing`.

[14] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019. doi: 10.1145/3292500.3330701.

[15] Tierney K. Lorenz, Amanda N. Gesselman, and Virginia J. Vitzthum. Variance in mood symptoms across menstrual cycles: Implications for premenstrual dysphoric disorder. *Women's Reproductive Health*, 4(2):77–88, 2017. doi: 10.1080/23293691.2017.1326248.

[16] Katja M. Schmalenberger, Hafsah A. Tauseef, Jordan C. Barone, Sarah A. Owens, Lynne Lieberman, Marc N. Jarczok, Susan S. Girdler, Jeff Kiesner, Beate Ditzen, and Tory A. Eisenlohr-Moul. How to study the menstrual cycle: Practical tools and recommendations. *Psychoneuroendocrinology*, 123:104895, 2021. doi: 10.1016/j.psyneuen.2020.104895.

A Feature Glossary

Table 4 lists all 97 features constructed before correlation filtering. After removing features with pairwise Pearson $|r| > 0.9$, approximately 83 remain per fold (see Section 2). Notation: σ_w = rolling standard deviation over a w-day window; μ_w = rolling mean over a w-day window; DSB = days since bleed.

Table 4: Complete feature glossary. Features are grouped by category. The "Source" column indicates the raw self-report variable from which the feature is derived.

Category	**Feature notation**	**Description**	**Source variable(s)**
Cycle position (13 features)			
	Cycle cos, Cycle sin	Cosine/sine encoding of cycle day (period = 28 d)	Flow volume
	Cycle day (mod 28)	Day within estimated cycle, modulo 28	Flow volume
	Cycle progress	Fraction of 28-day cycle elapsed (clipped 0–1)	Flow volume
	DSB (cd, flow$\geq t$)	Days since $\geq c$ consecutive days with flow $\geq t$ 9 variants: $c \in \{1, 2, 3\}$, $t \in \{1, 2, 4\}$	Flow volume
Rolling standard deviations (42 features: 14 variables × 3 windows)			
	$X\ \sigma_5$, $X\ \sigma_7$, $X\ \sigma_{14}$	Rolling std over 5-, 7-, 14-day windows	All 14 self-report variables
Rolling means (28 features: 14 variables × 2 windows)			
	$X\ \mu_7$, $X\ \mu_{14}$	Rolling mean over 7- and 14-day windows	All 14 self-report variables
Raw values (14 features)			
	Flow volume	Daily self-reported menstrual flow (0–7)	—
	Flow color	Menstrual flow color (0–7)	—
	Cramps	Menstrual cramp severity (0–5)	—
	Fatigue	Fatigue level (0–5)	—
	Bloating	Bloating severity (0–5)	—
	Mood swings	Mood swing severity (0–5)	—
	Sore breasts	Breast tenderness (0–5)	—
	Stress	Stress level (0–5)	—
	Headaches	Headache severity (0–5)	—
	Sleep issues	Sleep quality issues (0–5)	—
	Food cravings	Food craving intensity (0–5)	—
	Indigestion	Indigestion severity (0–5)	—
	Appetite	Appetite level (0–5)	—
	Exercise	Exercise level (0–5)	—

Note: After removing features with pairwise Pearson $|r| > 0.9$ per LOSO fold, approximately 83 of these 97 features remain. The exact set varies slightly across folds.