Machine Learning for Menstrual Recovery Prediction Using Synthetic Wearable Data

Lillian Shen[1,2], Pouria Mortezaagha[1,2] and Arya Rahgozar[1,2]

[1]Ottawa Hospital Research Institute, Ottawa, Canada
[2]University of Ottawa, Ottawa, Canada

Abstract

Secondary amenorrhea is a heterogeneous condition with implications for reproductive, cardiovascular, and bone health. Existing machine learning approaches in menstrual health focus on cycle prediction rather than recovery modeling in pathological conditions. We present a proof-of-concept framework to model menstrual recovery within three months from non-invasive wearable-derived physiological features and self-reported inputs, including heart rate variability, resting heart rate, sleep, physical activity, skin temperature, perceived stress, age, and duration of amenorrhea. Using a synthetically generated dataset of 5,000 individuals encoding physiologically informed feature-outcome relationships, eight baseline models were evaluated. The best-performing model (XGBoost) achieved AUC of 0.914. Permutation-based null models confirmed non-trivial predictive structure (AUC = 0.503), and XGBoost outperformed rule-based baselines (ΔAUC = 0.044). SHapley Additive exPlanations (SHAP) analysis identified perceived stress and heart rate variability as dominant predictors, consistent with the data-generating structure. As the wearable-derived features are routinely captured by consumer devices and the self-reported inputs require brief periodic assessment, this framework establishes a foundation for wearable-based modeling of menstrual recovery, with future work required for real-world clinical validation and integration into wearable-based health monitoring systems.

Keywords

secondary amenorrhea, menstrual recovery, machine learning, non-invasive monitoring, wearable devices, synthetic data.

1. Introduction

The convergence of artificial intelligence (AI) and wearable health technologies is expanding the scope of reproductive health medicine, as non-invasive, continuous monitoring now offers a window into biological processes that remain largely opaque in traditional clinical settings [1]. Secondary amenorrhea—the absence of menstruation for at least three months in previously menstruating individuals—arises from disruption of the hypothalamic–pituitary–ovarian axis [2]. These disturbances impair hormonal signaling and feedback mechanisms governing the menstrual cycle, potentially leading to long-term complications including osteopenia and cardiovascular dysfunction if not detected early [3].

Consumer wearable devices now enable continuous, non-invasive measurement of physiological signals that reflect these same regulatory systems. Metrics such as heart rate variability (HRV), sleep patterns, and skin temperature are closely linked to autonomic and thermoregulatory processes [4]. While machine learning approaches using wearable-derived physiological signals have shown strong performance in menstrual cycle prediction, existing work has primarily focused on regular cycle modeling rather than pathological menstrual dysfunction [5]. Existing amenorrhea-focused prediction studies are limited to a specific treatment-induced population and relied on invasive clinical and hormonal biomarkers [6].

To address this gap, we developed a machine learning framework that leverages non-invasive features from consumer wearable devices and brief self-report inputs to model recovery from secondary amenorrhea. Using a synthetic dataset constructed to reflect physiologically plausible relationships

Joint Proceedings of the AIME 2026 Workshops: 1st International Workshop on Multicentric and Privacy-preserving Learning in Healthcare, Foundation Models for Public Health and Epidemiology: From Promise to Practice, and First International Workshop on Knowledge Graphs for Health, July 10, 2026, Ottawa, Canada
lillianshen88@gmail.com (L. Shen)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-08-30

between behavioral, autonomic, and endocrine signals, we evaluated whether machine learning models can recover meaningful patterns related to menstrual recovery under controlled conditions.

2. Results

Table 1
Performance comparison of machine learning models for modeling menstrual recovery.

	SVM	XGBoost	Random Forest	TabPFN	PyTorch DeepNN	Logistic Regression	kNN	Decision Tree
AUC	0.9009	0.9137	0.9068	0.9133	0.8924	0.8983	0.8772	0.8748
Accuracy	0.8430	0.8420	0.8410	0.8370	0.8330	0.8320	0.8140	0.8140
Precision	0.8654	0.8416	0.8788	0.8692	0.8522	0.8393	0.8584	0.8054
F1	0.8069	0.8115	0.8005	0.7965	0.7946	0.7966	0.7615	0.7786
Sensitivity	0.7558	0.7834	0.7350	0.7350	0.7442	0.7581	0.6843	0.7535
Specificity	0.9099	0.8869	0.9223	0.9152	0.9011	0.8887	0.9134	0.8604
CV Mean AUC	0.9089	0.9162	0.9070	0.9195	0.9080	0.9108	0.8997	0.8762
CV SD AUC	0.0077	0.0066	0.0057	0.0061	0.0101	0.0064	0.0112	0.0088
CV Mean Accuracy	0.8398	0.8438	0.8415	0.8500	0.8380	0.8475	0.8345	0.8205
CV SD Accuracy	0.0041	0.0034	0.0056	0.0045	0.0069	0.0063	0.0108	0.0083
CV Mean F1	0.8090	0.8182	0.8082	0.8178	0.8094	0.8207	0.7935	0.7902
CV SD F1	0.0055	0.0044	0.0068	0.0057	0.0106	0.0095	0.0127	0.0160

Model performance is summarized in Table 1. Across all evaluated models (n = 8), performance clustered within a relatively narrow range, with AUC values spanning 0.862–0.914, indicating that most models achieved moderate-to-high discrimination within the synthetic dataset regardless of architecture. Ensemble-based models consistently occupied the top performance tier, with XGBoost, TabPFN, and Random Forest achieving AUC values ≥ 0.906. In contrast, simpler models (e.g., decision tree, kNN) formed a lower-performance cluster (AUC ≤ 0.877), suggesting that model complexity confers improved discrimination within this framework, albeit with diminishing returns.

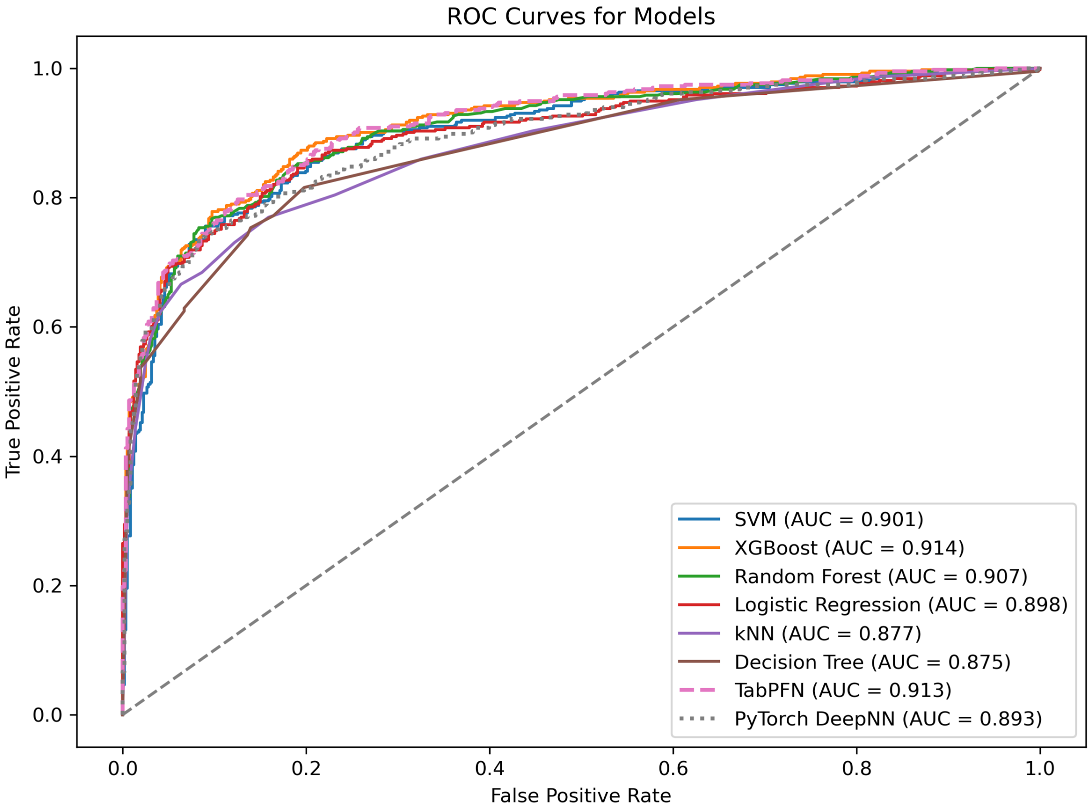

Figure 1: ROC curves for evaluated models.

ROC curves shown in Fig. 1 further highlight the limited separability among top-performing models. Curves for XGBoost (AUC = 0.914), TabPFN (0.913), and Random Forest (0.907) closely overlap across the full range of false positive rates, indicating near-equivalent ranking performance and reinforcing that model choice has minimal impact on discrimination under threshold-independent evaluation.

Based on overall discrimination, calibration, and stability metrics, XGBoost, Random Forest, and SVM were selected as top-performing models.

Table 2
Comprehensive evaluation of top-performing models.

Metric	XGBoost	Random Forest	Support Vector Machine
AUC (95%)	0.914 (0.895–0.931)	0.907 (0.887–0.925)	0.901 (0.881–0.920)
Brier Score	0.113	0.117	0.120
F1 Score at Optimal Threshold	0.817	0.809	0.807
Optimal Threshold	0.526	0.471	0.479
Feature Stability	0.872	0.827	0.871
False Negatives	99	107	113
False Positives	58	53	55
Total Errors	157	160	168

Further evaluation of the top three models revealed differences not captured by discrimination alone in Table 2. Calibration analysis showed that all models were reasonably well-calibrated, with predicted probabilities broadly aligning with the diagonal reference. XGBoost exhibited closer alignment with perfect calibration across mid-to-high probability ranges ($\approx$0.4–0.9), whereas SVM showed mild overestimation at lower probabilities and Random Forest demonstrated greater variability in intermediate ranges. These findings are reflected in Brier scores, with XGBoost achieving the lowest error (0.113), followed by Random Forest (0.117) and SVM (0.120), indicating improved agreement between predicted probabilities and observed outcomes. Threshold optimization yielded similar performance across models, with only marginal differences in F1 score (XGBoost: 0.817 at threshold 0.526; Random Forest: 0.809 at 0.471; SVM: 0.807 at 0.479), suggesting that model ranking was not sensitive to threshold selection. Confidence intervals for AUC overlapped substantially (XGBoost: 0.895–0.931; Random Forest: 0.887–0.925; SVM: 0.881–0.920), further supporting comparable discrimination performance. Feature stability analysis demonstrated higher consistency for XGBoost (Spearman ρ = 0.872) and SVM (ρ = 0.871) compared with Random Forest (ρ = 0.827), indicating more robust and reproducible feature attribution across resampled datasets for these models.

Taken together, although discrimination and threshold-dependent metrics were similar across models, XGBoost provided the most favorable balance between calibration accuracy, feature stability, and error distribution.

SHAP analysis identified perceived stress, HRV, sleep duration, and resting heart rate as dominant predictors, consistent with the imposed synthetic data structure. Lower perceived stress and higher HRV were associated with increased predicted probability of recovery, aligning with established links between autonomic regulation, stress physiology, and restoration of menstrual function. These findings confirm that the model recovered feature importance patterns consistent with the imposed synthetic data structure.

3. Discussion and Limitations

This study demonstrates that wearable-derived physiological and behavioral features contain sufficient structure to model menstrual recovery in secondary amenorrhea within a synthetic benchmarking framework. Among evaluated models, XGBoost demonstrated the strongest overall performance. SHAP analysis identified perceived stress, HRV, sleep-related metrics, and activity patterns as dominant predictors, consistent with the imposed physiological relationships encoded during synthetic data generation.

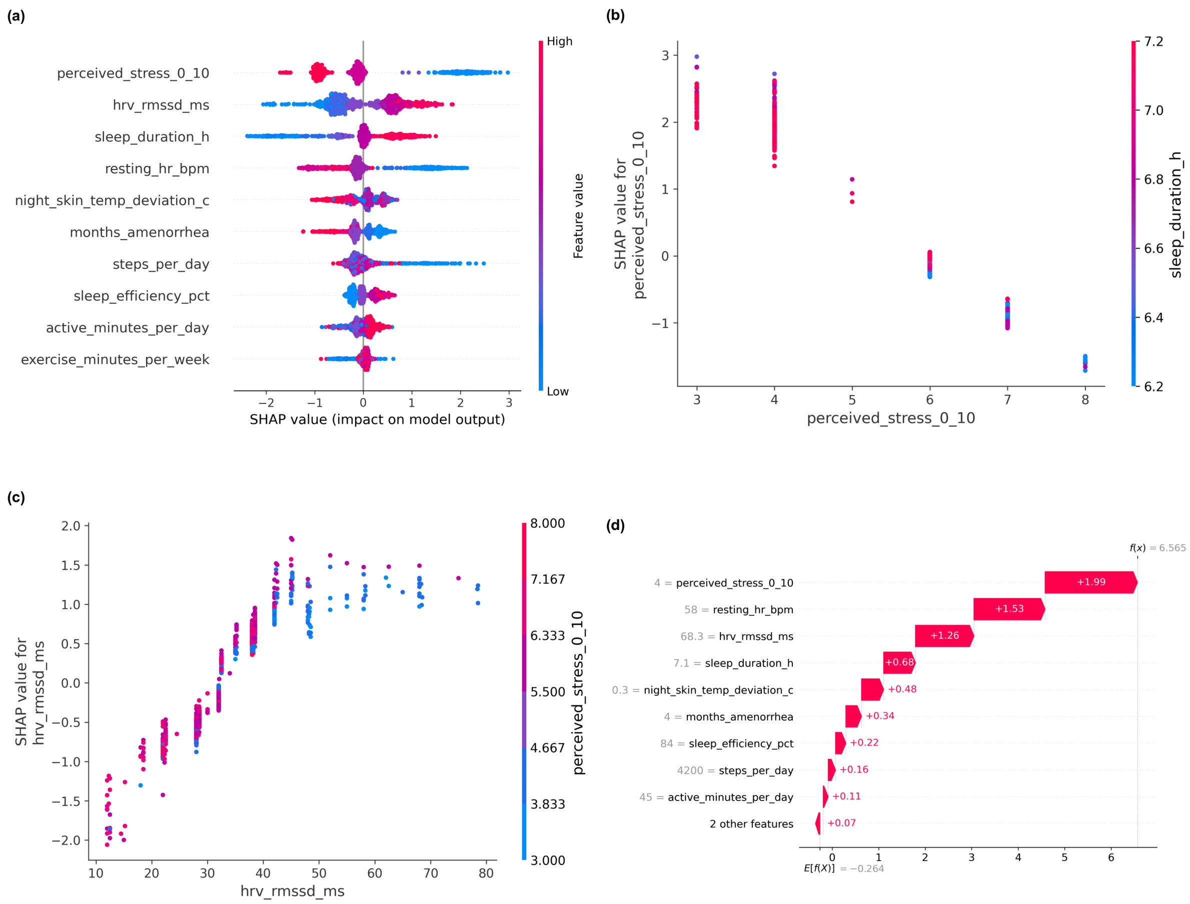

Figure 2: SHAP analysis of the XGBoost model.

This study is limited by the use of a synthetic dataset generated under imposed physiological assumptions. Consequently, model performance reflects learnability of the data-generating structure and may overestimate real-world predictive accuracy. While supplementary analyses demonstrated non-trivial predictive structure, external validation using clinical and real-world wearable datasets is required.

Methods

Due to the absence of publicly available, comprehensive datasets integrating non-invasive wearable-derived physiological signals with menstrual recovery outcomes in secondary amenorrhea, a synthetic dataset was generated to enable controlled model development. Synthetic data were generated using the NVIDIA NeMo DataDesigner framework with physiologically constrained feature relationships and probabilistic recovery outcomes. A total of 5,000 samples were generated, each representing an individual with secondary amenorrhea, defined as $\geq$3 months of absent menses at baseline.

Features included age, duration of amenorrhea, daily steps, active minutes, exercise duration, resting heart rate, heart rate variability (HRV), night skin temperature deviation, sleep duration, sleep efficiency, sleep variability, and perceived stress.

Eight machine learning models were evaluated using an 80:20 stratified train-test split with five-fold cross-validation. All preprocessing steps were implemented within pipeline structures to ensure proper separation between training and test data and prevent data leakage. Missing values were handled using k-nearest neighbors imputation (`KNNImputer`, `n_neighbors = 10`), fitted on training folds and applied to validation and test data. Feature scaling (`StandardScaler`) was applied to models sensitive to feature magnitude, while tree-based and ensemble models were trained directly on imputed, unscaled data. Default or literature-informed hyperparameters were used for all models, with class balancing applied where supported.

Discrimination was assessed using AUC, accuracy, precision, recall, sensitivity, specificity, and F1 score. Model stability was assessed using five-fold stratified cross-validation. Top three architectures were selected based on both cross-validation and test set performance and further evaluated across calibration curves, threshold optimization, confidence intervals, and error analysis. The final model was selected based on overall balance and performance, robustness, and clinical applicability. Model interpretability was assessed using SHAP. Global feature importance, local explanations, and dependence plots were generated using a subsample of 200 instances and a background sample of 100 instances from the test set to characterize feature effects and ensure alignment with known physiological mechanisms.

All code used for synthetic data generation and model development is publicly available at https://github.com/lillianshn/secondary-amenorrhea-/. The data supporting the findings of this study are openly available at https://www.kaggle.com/datasets/lillianshen88/secondary-amenorrhea-recovery.

Acknowledgments

This study received no funding.

Disclosure of Interests

The authors declare no competing interests.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. R. Bull, et al., Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles, npj Digital Medicine 2 (2019) 83. doi:10.1038/s41746-019-0152-7.

[2] M. Lord, S. M. Jenkins, M. Sahni, Secondary amenorrhea, StatPearls, 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK431055/.

[3] B. Meczekalski, O. Niwczyk, G. Bala, A. Szeliga, Stress, kisspeptin, and functional hypothalamic amenorrhea, Current Opinion in Pharmacology 67 (2022) 102288. doi:10.1016/j.coph.2022.102288.

[4] A.-M. Schyvens, et al., A performance validation of six commercial wrist-worn wearable sleep-tracking devices for sleep stage scoring compared to polysomnography, SLEEP Advances 6 (2025) zpaf021. doi:10.1093/sleepadvances/zpaf021.

[5] K. Li, I. Urteaga, A. Shea, V. J. Vitzthum, C. H. Wiggins, N. Elhadad, A predictive model for next cycle start date that accounts for adherence in menstrual self-tracking, Journal of the American Medical Informatics Association 29 (2021) 3–11. doi:10.1093/jamia/ocab182.

[6] L. Song, et al., Development of a machine learning model for predicting treatment-related amenorrhea in young women with breast cancer, Bioengineering 12 (2025) 1171. doi:10.3390/bioengineering12111171.