

Evaluating Symptom and Wearable Features for Machine-Learning Based Menstrual Bleeding Day Classification for Variable Cycles

Advita Goyal
advitagoyal01@gmail.com

ABSTRACT

Period-tracking applications are becoming increasingly common among younger generations. In period-tracking applications and models, accurately identifying menstrual bleeding days for highly variable cycles remains a weak spot. This can potentially disrupt lives for menstruators. Many studies have evaluated the use of different types of data to improve menstrual date classification, but there is a lack of research on types of data that specifically affect variable-cycle individuals. This study investigates the potential of symptom data (headaches, cramps, and more) and wearable device data (breathing rate, temperature, and more) to improve same-day menstrual bleeding classification compared to cycle history alone for variable cycles. In this study, three machine-learning models: logistic regression, gradient boosting, and random forest, were used to evaluate how same-day menstrual bleeding classification differed across four feature sets: cycle history only, cycle history + symptoms, cycle history + wearables, and all features. Among individuals with higher cycle variability, the highest AUC was achieved by the gradient boosting model using the cycle history + symptoms feature set (AUC: 0.831), but the cycle history feature set performed similarly (AUC: 0.827). In all three models, adding symptom and wearable features did not consistently improve performance compared to cycle history alone. This study served as a proof-of-concept exploratory study to potentially guide future research; future research should use larger and more diverse datasets to develop software to improve the classification of menstrual bleeding days among individuals with variable cycles to develop more accurate and clinically useful same-day menstrual bleeding classification tools.

INTRODUCTION & LITERATURE REVIEW

Period Tracking

Menstruation affects various aspects of daily life for menstruators. This can include physical comfort, academic/work responsibilities, exercise, social planning, emotional well-being, and more. Many menstruating individuals track their menstrual cycles to be physically and mentally prepared for

menstruation, family planning, and to be able to understand their bodies better in general (Epstein et al., 2017).

Early menstrual prediction was often done using calendar-based methods. These methods would often involve tracking one's menstruation dates over multiple cycles, averaging the cycle length, and using the average cycle length to predict future menstruation dates. However, calendar-based menstrual prediction methods often perform poorly at predicting future menstruation dates and accurately identifying menstrual phases (Wideman et al., 2012). This is especially relevant for people with variable cycles whose future menstruation dates may be even more difficult to estimate by cycle length alone.

The use of period-tracking applications is becoming more common among younger generations. One study found that 62.3% of millennial and Gen Z menstruating participants used period-tracking applications (Hong et al., 2024). Period-tracking applications help menstruators by arming them with the knowledge they need to be prepared for menstruation, empowering them to make informed decisions about their bodies and lives. The usefulness of period-tracking applications depends greatly on the quality and accuracy of the information they provide to the user. Studies evaluating period-tracking applications have found that the accuracy of apps can be limited, with one study evaluating 10 period-tracking apps finding that correct exact period-date prediction was rare and predictions were as much as 8 days before or after the actual date (Worsfold et al., 2021).

Unlike calendar-based methods, machine-learning models learn patterns directly from historical data rather than relying on fixed rules such as average cycle lengths. This allows machine-learning models to capture more complex, non-linear predictions, which may better represent women's cycles. Their predictions can also be updated as a user inputs more cycle data, allowing them to adapt to one's cycle history. Recent research involves using machine-learning models trained with individuals' menstruation dates to predict their future menstruation dates. For example, researchers have used self-tracked cycle histories to build models that predict future menstruation dates and update predictions as more data gets logged, even being able to account for skipped period tracking (Li et al., 2021).

Additional Data Types

Recent research has begun to incorporate different metrics to train machine-learning models for menstrual-cycle tracking, phase identification, bleeding classification, and menstruation prediction rather than relying only on cycle history information. Menstruation is also associated with changes in symptoms and physiological signals. Symptom-based data such as cramps and headaches can help identify bleeding days or nearby menstrual phases (Li et al., 2020). Wearable devices can also report useful physical information such as body temperature, resting heart rate, and respiratory patterns, which can be used to identify and predict menstrual phases (Yu et al., 2022; Lyzwinski et al., 2024).

Several studies have explored the use of wearable data collection devices for menstruation date prediction. One study found that wearable-based physiological measures enabled machine-learning

models to be able to identify phase patterns and classify menstrual phases (Kilungeja et al., 2025). Another study used temperature and heart-rate data from wearable devices to predict menstruation dates and fertile windows (Yu et al., 2022). Another study found that heart-rate patterns could be used to classify menstrual phases (Masuda et al., 2025). Similarly, one study found that skin-temperature data from a wearable device revealed menstrual cycle patterns (Gombert-Labedens et al., 2024). Together, these studies suggest wearable-device data contain meaningful menstrual-cycle information that may be useful for both menstrual-phase classification and prediction.

Other research has examined whether self-reported symptoms provide additional information about menstrual-cycle patterns. In a study of data from the period and cycle tracking app Clue, researchers found statistically significant relationships between cycle variability and self-reported qualitative symptoms based on data from over 378,000 users and 4.9 million cycles (Li et al., 2020), suggesting that symptom-based data also can contain meaningful insight for classifying and predicting menstrual phases.

Variable Cycles

Around 14% to 25% of women experience irregular cycles, which causes individuals to not be prepared for bleeding (Eunice Kennedy Shriver National Institute of Child Health and Human Development, n.d.). Existing studies suggest that prediction tools are less reliable for individuals with variable cycles. In one review of 10 period-tracking apps, researchers found that all apps predicted cycle length correctly for a simulated user with a constant 28-day cycle. However, the app predicted dates 0 to 8 days off for simulated users with variable and irregular cycles (Worsfold et al., 2021). Another group of researchers developed machine-learning algorithms to predict fertile windows and menstruation using body temperature and heart rate. Ultimately, they found that while performance was good for regular menstruators, prediction for variable cycles was not ideal and remained a weak spot (Yu et al., 2022).

This Study

Although prior work has shown that menstrual-cycle prediction can be improved through machine learning, self-reported symptom data, and wearable data, several gaps remain. First, many studies focus on one type of information at a time, such as one particular type of wearable data, potentially decreasing the overall effectiveness of the period tracking application. Second, variable cycles are often more difficult to track and predict, and there is limited research examining whether symptom and wearable data improve same-day bleeding classification, particularly among individuals with greater cycle variability.

Unlike prior studies that primarily focused on predicting future menstruation dates, this study examines a related but different task: classifying whether menstrual bleeding is occurring on the same day using symptom and wearable information collected that day.

This study addresses this gap by using the mcPHASES dataset to use both symptom-based and wearable data to answer the question: How do cycle-history-based, symptom-based, wearable-based, and combined

machine-learning models differ in their ability to classify same-day menstrual bleeding among individuals with more variable cycles?

This study will conduct a small-scale empirical analysis of which data types are most useful for same-day menstrual bleeding classification among individuals with more variable cycles. The aim of the study is not to develop a fully functional clinically useful period-tracking application. While this study is exploratory, identifying which feature sets can improve bleeding-day classification for individuals with variable cycles may guide future research using larger and more diverse datasets and support the eventual development of more accurate clinically useful prediction tools for variable cycle individuals.

METHODOLOGY

Dataset

This study used the publicly available menstrual cycle Physiological, Hormonal, and Self-Reported Events and Symptoms (mcPHASES) dataset. The dataset contains self-reported and wearable-device data from 42 menstruators collected through daily surveys and wearable devices across two three-month periods (Lin et al., 2025).

Table 1. Self-Reported Symptom and Wearable Device Features Used in Model Training

Self-Reported Symptoms	Wearable Device Data
Headaches	Sleep Score
Cramps	Composition Score
Sore breasts	Revitalization Score
Fatigue	Duration Score
Sleep Issues	Deep Sleep Duration (Minutes)
Mood Swings	Sleep Resting Heart Rate
Stress	Restlessness
Food Cravings	Wrist Temperature Difference
Indigestion	Stress Score
Bloating	Sleep Points
	Responsiveness Points
	Exertion Points

	Full Sleep Breathing Rate
	Deep Sleep Breathing Rate
	Light Sleep Breathing Rate
	REM Sleep Breathing Rate
	Temperature Samples
	Nightly Temperature
	Baseline Relative Nightly Temperature Standard Deviation

Table 2. Sample Characteristics of the Eligible Analysis Sample. Participant counts, participant-days, bleeding outcomes, subgroup sizes, and missingness for symptom and wearable features.

Characteristic	Value
Eligible participants	30
Participant-days	2385
Bleeding episodes	117
Bleeding days	519
Bleeding-day prevalence	21.80%
Lower variability participants	15
Higher variability participants	15
Symptom missingness	5.50%
Wearable missingness	19.10%

The characteristics of the final analysis sample are summarized in Table 2.

Methodology Overview

This analysis consisted of four main steps. First, self-reported and wearable device data were merged into one dataset. Second, bleeding-day labels and predictor variables, including cycle history features, were

constructed. Participants were grouped into “lower variability” and “higher variability” groups. Third, three machine-learning models were trained using four different feature sets. Finally, model performance was evaluated according to multiple metrics.

Outcome and Feature Construction

In the main analysis dataset, each row represented one participant on one day of the study, called a participant-day. The outcome variable was a binary indicator that represented whether the participant reported menstrual bleeding on that day. A participant-day was labeled as a bleeding day if the flow_volume column in hormones_and_selfreport.csv from mcPHASES was present and not equal to “Not at all”. Otherwise, it was labeled as a non-bleeding day.

Two cycle history features were constructed using information from bleeding episodes that occurred before the current participant-day. Days since previous bleeding onset represented the number of days since the most recent previous bleeding-episode start. Previous cycle length represented the number of days between the two most recent bleeding-episode starts. Bleeding episodes beginning on the current day were not used to construct either feature. This ensured that that current-day bleeding status was not included in the predictor variables.

Table 3. Features Included in Each Feature Set. Cycle history, symptom, and wearable variables used to train the models.

Feature Set	Included Feature Groups
Cycle history only	Cycle history features
Cycle history + Symptoms	Cycle history and symptom features
Cycle history + Wearables	Cycle history and wearable features
All features	Cycle history, symptom, and wearable features

Table 3 shows the four feature sets created to compare how wearable and symptom-based features affect model performance. The cycle-history-only feature set included day in study and weekend status. The symptom feature set added self-reported ratings. The wearable feature set added physiological data derived from wearable devices. The combined feature set included cycle history, symptom, and wearable features.

Data Preprocessing

Symptom ratings were converted from categories into increasing numeric values from least to most severe. If any feature values were missing, median imputation was used to fill those gaps. For logistic regression, numeric features were standardized after imputation. For the random forest and the gradient

boosting models, median imputation was used but standardization was not required. Preprocessing was performed within machine-learning pipelines. This way, imputation values and scaling parameters were learned using only the training participants for each repeated train/test split before being applied to the corresponding test participants. This prevented information from the test data from influencing model training. Wearable-device data were merged with the main analysis dataset using participant ID, study interval, and day in study.

Model Training and Evaluation

Three supervised machine-learning models were evaluated: logistic regression, random forest, and gradient boosting. These simpler models were chosen because of the relatively small dataset. While the original dataset included 42 participants, only 30 had enough bleeding episode information to calculate cycle length variability. This resulted in an eligible analysis sample of 30 participants.

Model evaluation was performed using 30 repeated participant-level train/test splits. For each repetition, participants were randomly divided into 75% training participants and 25% testing participants, while preserving the proportion of lower- and higher-variability participants. All participant-days from a participant were assigned entirely to either the training or testing set to prevent data leakage between sets.

For each outer training set, the classification threshold was selected using only the training participants. Training participants were further divided into four participant-level stratified folds, and out-of-fold predicted probabilities were used to identify the threshold that maximized the F1 score. A final model was then fit using all outer-training participants and evaluated once on the held-out test participants using the selected threshold.

Each model was evaluated using the four feature sets (cycle history only, cycle history + symptoms, cycle history + wearables, and all features). Performance was summarized across the 30 repeated participant splits using accuracy, precision, recall, F1 score, receiver operating characteristic area under the curve (ROC-AUC), and precision-recall area under the curve (PR-AUC).

Hyperparameters were selected manually in order to provide reasonable model complexity for the relatively small dataset. These hyperparameters were held constant across all feature sets to ensure fair comparisons. No additional hyperparameter tuning was performed because instead of trying to maximize performance, the primary objective of this study was to compare feature sets under a consistent modeling setup.

Logistic regression used L2 regularization with $C = 1.0$, the lbfgs solver, a maximum of 1,000 iterations, and balanced class weights. Gradient boosting used 300 estimators, a learning rate of 0.03, a maximum tree depth of 2, and a minimum of 10 samples per leaf. Random forest used 300 estimators, no maximum tree depth, a minimum of 5 samples per leaf, and balanced class weights. A base random seed of 42 was used for all repeated train/test splits.

Cycle Variability Analysis

Cycle length was estimated from each participant based on the gaps between bleeding episode start days. The start of a bleeding episode was defined as a day of a participant's bleeding with the preceding day not classified as bleeding. Participants having at least three bleeding episode start dates were included so that at least two cycle-length gaps could be calculated.

Variability for cycles was measured by the standard deviation at the participant level. A median split was used to create similarly sized higher and lower variability groups for this exploratory analysis. For the eligible participants, the median standard deviation of cycle length was 4.45 days. Participants with cycle-length standard deviations at or below the median of 4.45 days were classified as lower variability, while those above the median were classified as higher variability.

Model performance was evaluated separately for each model-feature combination. Then, model performance was summarized based on lower and higher variability groups across the repeated participant-level test splits.

Software and Implementation

All analyses were conducted in Python. Pandas and NumPy were used for data cleaning and preprocessing. Scikit-learn was used for machine-learning models, preprocessing pipelines, repeated participant-level train/test splitting, participant-level cross-validation, and evaluation. Results were organized into comparison tables by model type, feature set, and cycle-variability group.

RESULTS

Overall Model Performance

Table 4. Overall Model Performance Across Feature Sets. Performance metrics for logistic regression, gradient boosting, and random forest models using the four feature sets.

Model	Feature Set	Threshold	Accuracy	Precision	Recall	F1	ROC AUC	PR AUC
Logistic Regression	Cycle history only	0.651	0.883	0.748	0.691	0.717	0.822 ± 0.011	0.732 ± 0.029
	Cycle history + Symptoms	0.611	0.821	0.584	0.670	0.617	0.809 ± 0.023	0.637 ± 0.066

	Cycle history + Wearables	0.643	0.833	0.627	0.599	0.605	0.796 ± 0.017	0.635 ± 0.057
	All features	0.622	0.804	0.553	0.631	0.581	0.794 ± 0.021	0.597 ± 0.062
Gradient Boosting	Cycle history only	0.410	0.883	0.752	0.684	0.715	0.895 ± 0.022	0.745 ± 0.039
	Cycle history + Symptoms	0.403	0.879	0.733	0.698	0.713	0.894 ± 0.026	0.773 ± 0.037
	Cycle history + Wearables	0.410	0.883	0.755	0.678	0.713	0.879 ± 0.020	0.742 ± 0.037
	All features	0.409	0.876	0.731	0.686	0.704	0.884 ± 0.025	0.756 ± 0.037
Random Forest	Cycle history only	0.701	0.881	0.756	0.664	0.705	0.892 ± 0.022	0.742 ± 0.041
	Cycle history + Symptoms	0.534	0.873	0.701	0.724	0.711	0.885 ± 0.025	0.747 ± 0.037
	Cycle history + Wearables	0.506	0.878	0.729	0.693	0.709	0.875 ± 0.017	0.737 ± 0.037
	All features	0.461	0.874	0.705	0.717	0.708	0.881 ± 0.022	0.741 ± 0.036

The gradient boosting and random forest models generally had higher AUC values than the logistic regression models. The gradient boosting model using the “cycle history only” feature set had the highest AUC (0.895 ± 0.022), followed closely by the gradient boosting “cycle history + symptoms” model (0.894 ± 0.026) and the random forest “cycle history only” model (0.892 ± 0.022). The gradient boosting “cycle history + symptoms” model had the highest PR-AUC (0.773 ± 0.037).

Across all three models, adding symptom and wearable features did not consistently improve performance compared to the “cycle history only” feature set. For logistic regression, the “cycle history only” model had the highest AUC (0.822 ± 0.011), F1 score (0.717), precision (0.748), and accuracy (0.883). For gradient boosting and random forest, the “cycle history only” feature set also had the highest AUC.

However, the small differences between some feature sets were within the variability observed across repeated splits, suggesting that cycle history features provided the strongest performance overall, while symptom and wearable features did not consistently improve same-day bleeding classification.

Cycle Variability Subgroup Analysis

Tables 5-7 show each model's performance by cycle variability group. Results should be interpreted cautiously due to the small subgroup sizes. Results for the lower-variability group are included in the tables below for comparison but were not the primary focus of this study.

Table 5. Logistic Regression Model Performance by Cycle Variability. Performance metrics for participants with higher and lower cycle variability.

Variability Group	Feature Set	Threshold	Accuracy	Precision	Recall	F1	AUC	PR-AUC
Lower variability	Cycle history only	0.651	0.920	0.858	0.766	0.808	0.848 ± 0.011	0.825 ± 0.032
	Cycle history + Symptoms	0.611	0.838	0.637	0.749	0.675	0.842 ± 0.027	0.706 ± 0.083
	Cycle history + Wearables	0.643	0.858	0.699	0.668	0.674	0.830 ± 0.020	0.725 ± 0.061
	All features	0.622	0.821	0.601	0.717	0.641	0.836 ± 0.029	0.661 ± 0.082
Higher variability	Cycle history only	0.651	0.844	0.638	0.612	0.620	0.788 ± 0.023	0.635 ± 0.052
	Cycle history + Symptoms	0.611	0.802	0.532	0.588	0.550	0.775 ± 0.042	0.586 ± 0.082
	Cycle history + Wearables	0.643	0.806	0.557	0.525	0.529	0.757 ± 0.032	0.556 ± 0.086
	All features	0.622	0.787	0.510	0.540	0.510	0.751 ± 0.040	0.558 ± 0.081

The logistic regression model performed better for the lower variability group than for the higher variability group across all four feature sets. For the lower variability group, the “cycle history only” feature set achieved the highest AUC (0.848 ± 0.011), followed by “cycle history + symptoms” (0.842 ± 0.027). For the higher variability group, the “cycle history only” feature set also achieved the highest AUC (0.788 ± 0.023). Adding symptom or wearable features did not improve AUC for either variability group. The higher variability group also showed greater variability in AUC across repeated splits for most feature sets.

Table 6. Gradient Boosting Model Performance by Cycle Variability. Performance metrics for participants with higher and lower cycle variability.

Variability Group	Feature Set	Threshold	Accuracy	Precision	Recall	F1	AUC	PR-AUC
Lower variability	Cycle history only	0.410	0.924	0.899	0.741	0.811	0.958 ± 0.012	0.884 ± 0.038
	Cycle history + Symptoms	0.403	0.923	0.879	0.763	0.814	0.956 ± 0.009	0.899 ± 0.027
	Cycle history + Wearables	0.410	0.921	0.900	0.723	0.800	0.942 ± 0.016	0.876 ± 0.042
	All features	0.409	0.913	0.864	0.736	0.789	0.946 ± 0.013	0.880 ± 0.035
Higher variability	Cycle history only	0.410	0.839	0.621	0.625	0.618	0.827 ± 0.039	0.619 ± 0.053
	Cycle history + Symptoms	0.403	0.833	0.602	0.632	0.612	0.831 ± 0.043	0.656 ± 0.059
	Cycle history + Wearables	0.410	0.843	0.629	0.630	0.626	0.815 ± 0.034	0.624 ± 0.061
	All features	0.409	0.837	0.615	0.634	0.619	0.822 ± 0.041	0.643 ± 0.060

The gradient boosting model also performed better for the lower variability group than for the higher variability group across all feature sets. For the lower variability group, the highest AUC was achieved by the “cycle history only” feature set (0.958 ± 0.012), while the highest PR-AUC was achieved by “cycle history + symptoms” (0.899 ± 0.027). For the higher variability group, “cycle history + symptoms” produced the highest AUC (0.831 ± 0.043) and PR-AUC (0.656 ± 0.059), but these differences were small relative to the variability across repeated splits.

Table 7. Random Forest Model Performance by Cycle Variability. Performance metrics for participants with higher and lower cycle variability.

Variability Group	Feature Set	Threshold	Accuracy	Precision	Recall	F1	AUC	PR-AUC
Lower variability	Cycle history only	0.701	0.921	0.888	0.736	0.803	0.952 ± 0.011	0.878 ± 0.033
	Cycle history + Symptoms	0.534	0.920	0.857	0.774	0.810	0.945 ± 0.011	0.876 ± 0.033
	Cycle history + Wearables	0.506	0.920	0.894	0.724	0.798	0.933 ± 0.012	0.855 ± 0.037
	All features	0.461	0.919	0.862	0.759	0.805	0.942 ± 0.014	0.863 ± 0.044
Higher variability	Cycle history only	0.701	0.838	0.629	0.589	0.603	0.821 ± 0.048	0.610 ± 0.078
	Cycle history + Symptoms	0.534	0.823	0.570	0.670	0.614	0.819 ± 0.049	0.612 ± 0.065
	Cycle history + Wearables	0.506	0.833	0.596	0.660	0.623	0.809 ± 0.033	0.619 ± 0.057
	All features	0.461	0.826	0.576	0.672	0.616	0.818 ± 0.040	0.623 ± 0.058

The random forest model showed higher performance for the lower variability group than for the higher variability group across all feature sets. For the lower variability group, the “cycle history only” feature set achieved the highest AUC (0.952 ± 0.011). For the higher variability group, the “cycle history only” feature set also produced the highest AUC (0.821 ± 0.048), while the all-features feature set achieved the highest PR-AUC (0.623 ± 0.058). However, the differences among feature sets in the higher variability group were small relative to the variability across repeated splits.

DISCUSSION

The modest performance of the machine-learning models for classifying menstrual bleeding days for variable cycles is consistent with previous research showing that period-tracking applications and menstrual prediction models perform less reliably for variable and irregular cycles (Yu et al., 2022; Worsfold et al., 2021). Although wearable and symptom features have been shown to contain useful menstrual cycle information (Kilungeja et al., 2025; Li et al., 2020; Masuda et al., 2025; Gombert-Labeledens et al., 2024; Lyzwinski et al., 2024), adding these features did not consistently improve performance compared to cycle history alone in this study. This study may guide future research

using larger and more diverse datasets and support the eventual development of more accurate clinically useful same-day menstrual bleeding classification tools.

Performance was worse for participants with higher variability cycles compared to lower variability cycles across all three machine-learning models. Adding symptom and wearable features did not consistently improve performance for the higher variability group. Cycle history alone generally achieved the highest or similar AUC compared to the other feature sets, suggesting that additional symptom and wearable information did not consistently provide more useful information for same-day bleeding classification in this dataset.

One goal of this study was to examine whether symptom and wearable features could improve classification specifically for participants with higher cycle variability. The results suggest that higher cycle variability remained a challenge across all three models and that adding these features was not enough to consistently improve classification. This supports the need for future research focused specifically on individuals with more variable cycles and on identifying other types of information or modeling approaches that may improve performance.

LIMITATIONS & FUTURE WORK

This study had several major limitations.

Dataset Limitations

First, the dataset was relatively small, especially for the cycle variability subgroup analysis. The dataset came from an extremely limited number of participants. Additionally, the study that collected the dataset took place over two 3-month intervals, meaning that the number of cycles per participant was limited, which is especially important in a study that evaluated the regularity of participants' cycles. Additionally, the dataset did not focus specifically on individuals with high-variability cycles, so models were trained on data from both high and low variability cycles.

Second, the dataset included missing values in symptoms and wearable data. Symptoms and wearable data were not available for each participant-day. Imputation was used to train the models, which may have reduced the strength of symptoms and wearable signals.

Cycle Variability Limitations

Third, the distinction between “high variability” and “low variability” cycles is subjective. There are various ways to define a variable cycle. For this study, the standard deviation of cycle lengths was used to separate participants into lower and higher variability groups. With a different cutoff, definition, or

participant data, the participants assigned to each group could have changed significantly, potentially significantly changing the results. Additionally, the small number of cycles per participant in the dataset makes it difficult to analyze the cycle variability of each participant.

Model Limitations

Fourth, this study used relatively simple machine-learning models. These models were appropriate for a small dataset and allowed comparison across feature sets, but they may not capture complex menstrual patterns.

Finally, performance remained lower for participants with higher cycle variability than for participants with lower cycle variability. Even the best-performing models for the higher variability group had lower AUCs, suggesting that classifying menstruation days remained more difficult for individuals with more variable cycles. This was likely due to all the limitations listed above. Because of these limitations, results should be viewed as exploratory and not definitive.

Future Work

Future work should focus on addressing these limitations by using a larger dataset with more participants and more cycles per participant to train models. Data should be collected specifically on individuals with more variable cycles in order for models to be designed around higher variability participants. Additionally, different definitions of cycle variability should be tested (such as clinical cutoffs) to see if results change. More advanced models should also be used to improve performance. Future work that extends this approach to advance prediction should emphasize reducing false negatives, since missed bleeding days could be harmful to users who rely on predictions to prepare for menstruation.

CONCLUSION

The challenge of accurately classifying same-day menstrual bleeding remains difficult for variable cycle individuals, for whom methods that rely solely on cycle history are often unreliable. In this study, it was found that combining cycle history, wearable, and symptom data did not consistently improve the machine-learning models' ability to classify same-day bleeding and non-bleeding days compared to cycle history alone. However, performance was consistently lower for participants with higher cycle variability than for participants with lower cycle variability, and the study was held back by various limitations. These included limitations with the dataset size and values, defining cycle variability, and model limitations. This study serves as a proof-of-concept showing that classifying menstrual bleeding days remains more challenging for individuals with more variable cycles and that further work must be done to identify features and modeling approaches that can improve performance in larger-scale and more diverse datasets.

REFERENCES

- Epstein, D. A., Lee, N. B., Kang, J. H., Agapie, E., Schroeder, J., Pina, L. R., Fogarty, J., Kientz, J. A., & Munson, S. (2017, May 2). *Examining menstrual tracking to inform the design of personal informatics tools* [Paper presentation]. Proceedings of the 2017 CHI conference on human factors in computing systems. <http://dx.doi.org/10.1145/3025453.3025635>
- Eunice Kennedy Shriver National Institute of Child Health and Human Development. (n.d.). *What are menstrual irregularities?* U.S. Department of Health and Human Services. <https://www.nichd.nih.gov/health/topics/menstruation/conditioninfo/irregularities>
- Gombert-Labedens, M., Alzueta, E., Perez-Amparan, E., Yuksel, D., Kiss, O., de Zambotti, M., Simon, K., Zhang, J., Shuster, A., Morehouse, A., Alessandro Pena, A., Mednick, S., & Baker, F. C. (2024). Using wearable skin temperature data to advance tracking and characterization of the menstrual cycle in a real-world setting. *Journal of Biological Rhythms*, 39(4), 331-350. <https://doi.org/10.1177/07487304241247893>
- Hong, M., Rajaguru, V., Kim, K., Jang, S.-Y., & Lee, S. G. (2024). Menstrual cycle management and period tracker app use in millennial and generation Z individuals: Mixed methods study. *Journal of Medical Internet Research*, 26, e53146. <https://doi.org/10.2196/53146>
- Kilungeja, G., Graham, K., Liu, X., & Nasser, M. (2025). Machine learning-based menstrual phase identification using wearable device data. *Npj Women's Health*, 3(1). <https://doi.org/10.1038/s44294-025-00078-8>
- Li, K., Urteaga, I., Shea, A., Vitzthum, V. J., Wiggins, C. H., & Elhadad, N. (2021). *A generative, predictive model for menstrual cycle lengths that accounts for potential self-tracking artifacts in mobile health data* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2102.12439>
- Li, K., Urteaga, I., Wiggins, C. H., Druet, A., Shea, A., Vitzthum, V. J., & Elhadad, N. (2020). Characterizing physiological and symptomatic variation in menstrual cycles using self-tracked mobile-health data. *Npj Digital Medicine*, 3(1). <https://doi.org/10.1038/s41746-020-0269-8>
- Lin, B., Li, J. Y., Kalani, K., Truong, K., & Mariakakis, A. (2025). mcPHASES: A Dataset of Physiological, Hormonal, and Self-reported Events and Symptoms for Menstrual Health Tracking with Wearables (version 1.0.0). *PhysioNet*. RRID:SCR_007345. <https://doi.org/10.13026/zx6a-2c81>
- Lyzwinski, L., Elgendi, M., & Menon, C. (2024). Innovative approaches to menstruation and fertility tracking using wearable reproductive health technology: Systematic review. *Journal of Medical Internet Research*, 26, e45139. <https://doi.org/10.2196/45139>
- Masuda, H., Okada, S., Shiozawa, N., Sakaue, Y., Manno, M., Makikawa, M., & Isaka, T. (2025). Machine learning model for menstrual cycle phase classification and ovulation day detection based on sleeping heart rate under free-living conditions. *Computers in Biology and Medicine*, 187, 109705. <https://doi.org/10.1016/j.combiomed.2025.109705>
- Wideman, L., Montgomery, M. M., Levine, B. J., Beynon, B. D., & Shultz, S. J. (2012). Accuracy of calendar-based methods for assigning menstrual cycle phase in women. *Sports Health: A Multidisciplinary Approach*, 5(2), 143-149. <https://doi.org/10.1177/1941738112469930>

- Worsfold, L., Marriott, L., Johnson, S., & Harper, J. C. (2021). Period tracker applications: What menstrual cycle information are they giving women? *Women's Health*, 17. <https://doi.org/10.1177/17455065211049905>
- Yu, J.-L., Su, Y.-F., Zhang, C., Jin, L., Lin, X.-H., Chen, L.-T., Huang, H.-F., & Wu, Y.-T. (2022). Tracking of menstrual cycles and prediction of the fertile window via measurements of basal body temperature and heart rate as well as machine-learning algorithms. *Reproductive Biology and Endocrinology*, 20(1). <https://doi.org/10.1186/s12958-022-00993-4>