https://doi.org/10.1038/s44294-025-00082-y

Validation of sleep-based actigraphy machine learning models for prediction of preterm birth

Benjamin C. Warner[1,2] ✉, Peinan Zhao[3] ✉, Erik D. Herzog[4] ✉, Antonina I. Frolova[3] ✉, Sarah K. England[1,3] ✉ & Chenyang Lu[1,2] ✉

Disruptive sleep is a well-established predictor of preterm birth. However, the exact relationship between sleep behavior and preterm birth outcomes remains unknown, in part because prior work has relied on self-reported sleep data. With the advent of smartwatches, it is possible to obtain more reliable and accurate sleep data, which can be utilized to evaluate the impact of specific sleep behaviors in concert with machine learning. We evaluate motion actigraphy data collected from a cohort of participants undergoing pregnancy, and train several machine learning models based on aggregate features engineered from this data. We then evaluate the relative impact from each of these actigraphy features, as well as features derived from questionnaires collected from participants. Our findings suggest that actigraphy data can predict preterm birth outcomes with a degree of effectiveness, and that variability in sleep patterns is a relatively fair predictor of preterm birth.

Preterm birth (PTB), which is generally defined as delivery before 37 weeks of gestation, is the single largest cause of death in children under the age of 5[1] with ~1 million deaths occuring per year[2]. While some etiologies of PTB have been identified, many remain unknown. Previous literature has shown that disruptive maternal sleep patterns have been associated with PTB outcomes[3–6].

One major limitation with previous studies is the reliance on self-reported sleep patterns, which is limited by a patient's ability to recall their sleep patterns accurately and consistently[7]. Wearable devices can alleviate this problem as they provide a more reliable and detailed stream of data[8,9]. Previous literature has found that wearable sensor data can be used to make predictions regarding both physical and mental health issues, ranging from pancreatic complications[10] to depression[11].

Using data collected from wearables, we evaluate predictions of binary PTB outcomes with patients from a cohort study conducted at Washington University in St. Louis/BJC HealthCare[12]. Participants from this cohort study were given actigraphy watches to wear for 2 weeks over the course of each trimester, capturing high-resolution sleep data. The collected actigraphy data are then transformed into interpretable quantitative features and used as input for several shallow machine learning (ML) models. These models are then evaluated to assess the relative impact of these features, offering several clinical insights into the relative importance of individual sleep and non-sleep behaviors, as well as insights for more complex ML models.

Previous work with this dataset has attempted to evaluate regression models of unengineered time-series data to predict the entire spectrum of gestational age (GA) directly from individual actigraphy samples[13], which is intrinsically different in both objective and approach from predicting binary-outcome PTB from statistics across a pregnancy. The authors noted that measured mean absolute error between actual and predicted GA was higher overall in PTB patients, but did not evaluate any classifier performance with respect to binary-outcome PTB. Moreover, the models presented in[13] are limited in their explanability as a result of both learning non-linear representations and at attempting to predict GA at a sample level. In addition, previous work has also examined direct correlations between engineered actigraphy features and PTB, evaluating the risk associated with each individual feature[5,6].

This paper evaluates the performance of binary-outcome classification of PTB from engineered actigraphy features and selected patient history features. The models presented here are computationally simpler and interpretable, which offer engineering and clinical insights about potential approaches for more complicated models. Overall, we validate the usage of sleep measures derived from actigraphy data in ML models for the prediction of binary-outcome PTB. From these models, relative comparisons of

[1]AI for Health Institute, Washington University in St. Louis, St. Louis, MO, USA. [2] Department of Computer Science & Engineering, Washington University in St. Louis, MO St. Louis, USA. [3]Center for Reproductive Health Sciences, Department of Obstetrics & Gynecology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA. [4]Department of Biology, Washington University in St. Louis, St. Louis, MO, USA. ✉e-mail: b.c.warner@wustl.edu; zhao.p@wustl.edu; herzog@wustl.edu; frolovaa@wustl.edu; englands@wustl.edu; lu@wustl.edu

the impact of actigraphy and patient history features on predictions are examined. We finally offer interpretations of each of the tested models, and guidance for future works.

Results

Among the 1523 patients who participated in the cohort study, we analyze the 665 patients who had actigraphy data in at least the first or second trimester of their pregnancy and had a recorded delivery date. The average patient had 39.1 (±32.2) day-level samples throughout the duration of their pregnancy, with the first trimester having 15.7 (±10.4) samples on average, the second trimester having an average of 24.0 (±18.9) samples, and the third trimester having an average of 17.0 (±10.3) samples. The overall distribution of samples collected from all patients can be seen in Fig. 1. Of these patients, the mean age was 29.2 (±5.29) years, and the majority (55.34%) of the patients were multiparous. A minority of patients (14.18%) experienced a PTB outcome. Full details about the demographics of the patients used in this dataset can be found in Section 1 in the Supplementary Materials, and details about the actigraphy features and numerical case report form features can be found in Table 1.

We compare the performance of models trained on the two primary sources of data, the engineered actigraphy features and case report form responses collected at each visit, in Table 2 and Fig. 2. Performance curves of the models trained only on the actigraphy or case report form data can be found in Section 3 of the Supplementary Material. Confusion matrices comparing the best model by area under the receiver-operator curve (AUROC) are provided in Table 3, and Tukey's honest significant difference test (HSD) results comparing each are provided in Tables 4, 5, and 6.

We find that, using actigraphy features and case report form survey data, it is possible to make reasonable predictions about binary-outcome PTB. As seen in Table 2, actigraphy features appear to underperform features from case report forms at predicting PTB when comparing the best models for each configuration. The combined performance is better than either source of data individually.

Gestational age and model performance

Figure 3 shows the performance of each model as samples up to a specified GA are included. As seen, the performance of the models does not change consistently as the GA upper-bound is increased, although it does increase noticeably in performance as the full GA spectrum is enabled.

This lack of consistent performance change likely occurs for several reasons. First, the distribution of study participants who have data up to a given GA is variable, and for those that do have data up to a specified GA, the duration and lengths are also variable. In addition, the aggregation used for all actigraphy features, mean and standard deviation, does not change linearly as the amount of data increases. This variability in AUROC and area under the precision-recall curve (AUPRC) appears to weakly correspond to the sample trends seen in Fig. 1, which is roughly centered around the boundaries in each trimester.

Feature explanations

To assess the importance of each feature in each model, we evaluate the features with SHapley Additive exPlanations (SHAP) scores[14], which provide relative estimates of how the output of a model will change as the input features change. Figure 4 shows the feature explanations for the best performing model with all features.

When all features are used, we find that the features that affect the output of the model the most are related to the number of complications that occurred during previous births. This is consistent with the literature, which finds that past PTB is a strong predictor of future PTB outcomes[15,16]. Features relating to socioeconomic status, highlighted in green, also rank highly, which is consistent with prior literature as race, ethnicity, and employment status are associated with preterm birth[17,18].

Actigraphy features were impactful to a lesser degree, with the highest ranked feature being the average day-to-day variability between sleep start. Other similarly ranked features following this included sleep start time, the variance of the start of the sleep cycle, and day-to-day variability in the duration of the sleep cycle, *etc*. Overall, actigraphy features relating to variability in sleep patterns appeared to rank higher than those derived from averages across a patient's pregnancy. Prior studies have shown that shift work is associated with higher incidence of PTB[19]. The variability captured by our actigraphy features may help explain this association, as shift work often disrupts regular sleep patterns and leads to increased day-to-day variability.

When we evaluate the best performing actigraphy-only model, shown in Fig. 5, we find a similar ordering of relevant features, with features reflecting variance between daily actigraphy measurements appearing towards the top. Some of this difference in ordering can be attributed to the issue of dimensionality, as the number of examples is smaller, although the limited sample size prevents us from making any conclusive orderings.

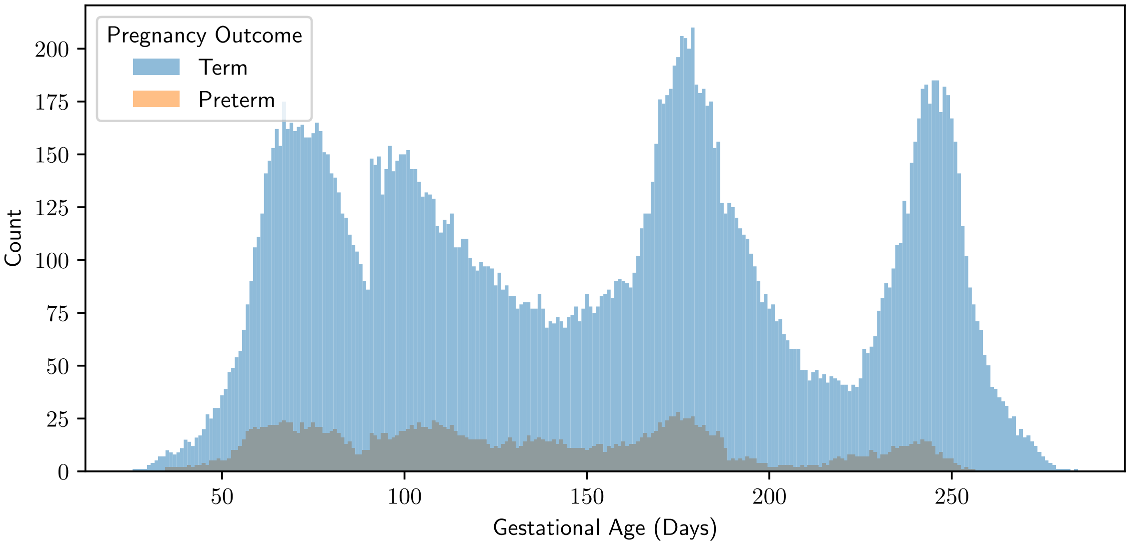


Fig. 1 | Histogram of the collected actigraphy samples. Results are stratified by whether the patient experienced a positive or negative preterm birth outcome. In the first trimester there are 263 negative patients and 39 positive patients who had a sample, 262 negative patients and 47 positive patients who have a sample, and in the third trimester there are 46 positive patients and 8 negative patients who have a sample.

Table 1 | Actigraphy and numerical case report form features, stratified by trimester where applicable

Feature	Trimester	Preterm birth	Non-preterm birth	Corrected *p*
Number of days sampled	1	23.41 (22.54, 24.28)	21.88 (21.65, 22.11)	$1.78 \times 10^{-04*}$
	2	35.50 (34.45, 36.54)	40.97 (40.48, 41.46)	$1.19 \times 10^{-13*}$
	3	15.11 (14.74, 15.48)	26.59 (26.17, 27.02)	$3.20 \times 10^{-36*}$
	All	29.20 (28.48, 29.91)	32.52 (32.23, 32.82)	$5.45 \times 10^{-12*}$
Day-to-day variability between sleep duration	1	−1:59 (−1:48, 0:11)	−1:59 (−1:55, 0:03)	1.00
	2	−1:60 (−1:53, 0:07)	−1:60 (−1:57, 0:02)	1.00
	3	−1:58 (−1:42, 0:14)	−1:60 (−1:57, 0:03)	1.00
	All	−1:59 (−1:54, 0:05)	−1:60 (−1:58, 0:01)	1.00
Day-to-day variability between sleep middle	1	−1:59 (−1:53, 0:06)	−1:60 (−1:58, 0:02)	1.00
	2	0:01 (−1:56, 0:05)	1:00 (−1:59, 0:01)	1.00
	3	0:02 (−1:55, 0:10)	−1:60 (−1:58, 0:02)	1.00
	All	0:01 (−1:57, 0:04)	−1:60 (−1:59, 0:01)	1.00
Day-to-day variability between sleep end	1	−1:59 (−1:52, 0:07)	−1:59 (−1:57, 0:02)	1.00
	2	0:01 (−1:55, 0:06)	−1:60 (−1:58, 0:02)	1.00
	3	0:01 (−1:52, 0:11)	−1:60 (−1:58, 0:02)	1.00
	All	1:00 (−1:56, 0:04)	−1:60 (−1:59, 0:01)	1.00
Day-to-day variability between sleep start	1	−1:60 (−1:50, 0:09)	1:00 (−1:57, 0:03)	1.00
	2	0:01 (−1:55, 0:07)	1:00 (−1:58, 0:02)	1.00
	3	0:03 (−1:51, 0:16)	−1:60 (−1:57, 0:02)	1.00
	All	0:01 (−1:56, 0:06)	1:00 (−1:59, 0:01)	1.00
Total motion count during sleep time	1	13590.90 (12743.07, 14438.73)	12150.57 (11887.88, 12413.25)	0.00^{*}
	2	12633.96 (12098.49, 13169.43)	11333.77 (11167.64, 11499.91)	$2.54 \times 10^{-06*}$
	3	14090.49 (13072.20, 15108.79)	12560.09 (12321.40, 12798.78)	0.03^{*}
	All	13108.81 (12694.23, 13523.40)	11875.15 (11753.36, 11996.94)	$5.48 \times 10^{-09*}$
Sleep length between start and end	1	9:20 (9:10, 9:29)	9:22 (9:19, 9:25)	1.00
	2	8:53 (8:47, 8:58)	9:07 (9:05, 9:09)	$4.97 \times 10^{-06*}$
	3	9:07 (8:55, 9:20)	8:60 (8:57, 9:02)	1.00
	All	9:02 (8:57, 9:07)	9:08 (9:07, 9:10)	0.04^{*}
End of the sleep cycle	1	7:38 AM (7:30 AM, 7:45 AM)	7:42 AM (7:39 AM, 7:44 AM)	1.00
	2	7:49 AM (7:44 AM, 7:55 AM)	7:34 AM (7:32 AM, 7:35 AM)	$1.68 \times 10^{-07*}$
	3	7:58 AM (7:48 AM, 8:08 AM)	7:35 AM (7:33 AM, 7:38 AM)	$4.08 \times 10^{-05*}$
	All	7:48 AM (7:44 AM, 7:52 AM)	7:36 AM (7:35 AM, 7:37 AM)	$1.68 \times 10^{-07*}$
Frequency of motion counts during sleep time	1	0.19 (0.18, 0.19)	0.18 (0.18, 0.18)	0.57
	2	0.19 (0.19, 0.19)	0.18 (0.17, 0.18)	$9.71 \times 10^{-13*}$
	3	0.21 (0.21, 0.22)	0.19 (0.18, 0.19)	$2.47 \times 10^{-13*}$
	All	0.19 (0.19, 0.20)	0.18 (0.18, 0.18)	$9.05 \times 10^{-17*}$
Halfway between sleep start and sleep end	1	2:57 AM (2:51 AM, 3:04 AM)	3:01 AM (2:58 AM, 3:03 AM)	1.00
	2	3:23 AM (3:18 AM, 3:28 AM)	2:60 AM (2:58 AM, 3:01 AM)	$3.74 \times 10^{-19*}$
	3	3:24 AM (3:16 AM, 3:33 AM)	3:05 AM (3:03 AM, 3:07 AM)	$1.72 \times 10^{-04*}$
	All	3:16 AM (3:13 AM, 3:20 AM)	3:01 AM (3:00 AM, 3:03 AM)	$2.44 \times 10^{-14*}$
Start of the sleep cycle	1	10:18 PM (10:09 PM, 10:26 PM)	10:20 PM (10:17 PM, 10:23 PM)	1.00
	2	10:57 PM (10:50 PM, 11:03 PM)	10:26 PM (10:24 PM, 10:28 PM)	$1.18 \times 10^{-22*}$
	3	10:51 PM (10:40 PM, 11:02 PM)	10:35 PM (10:33 PM, 10:38 PM)	0.04^{*}
	All	10:46 PM (10:41 PM, 10:50 PM)	10:28 PM (10:26 PM, 10:29 PM)	$3.93 \times 10^{-14*}$
Total daily motion count	1	235784.82 (227642.00, 243927.63)	226901.10 (224501.82, 229300.39)	0.13
	2	254095.03 (248684.76, 259505.31)	243520.87 (241601.36, 245440.38)	0.00^{*}
	3	231434.40 (219811.23, 243057.56)	240551.62 (238066.34, 243036.90)	0.71
	All	245803.74 (241582.54, 250024.93)	238978.93 (237679.47, 240278.38)	0.01^{*}

Table 1 (continued) | Actigraphy and numerical case report form features, stratified by trimester where applicable

Feature	Trimester	Preterm birth	Non-preterm birth	Corrected *p*
Longest contiguous subsequence of zero actigraphy activity	1	49.40 (48.37, 50.44)	47.28 (46.93, 47.64)	7.09×10^{-04}*
	2	49.09 (48.37, 49.81)	49.19 (48.94, 49.44)	1.00
	3	49.41 (47.83, 50.98)	50.84 (50.49, 51.18)	0.44
	All	49.22 (48.67, 49.78)	49.25 (49.08, 49.43)	1.00
Maternal Age at Enrollment/Consent	All	29.46 (28.32, 30.60)	29.16 (28.73, 29.59)	1.00
BMI at 1st Prenatal Visit (calculated)	All	31.04 (29.27, 32.80)	27.65 (26.98, 28.32)	0.00*
Day-to-day variability between sleep duration (Mean)	All	1:55 (1:47, 2:03)	1:41 (1:38, 1:44)	0.01*
Day-to-day variability between sleep duration (Std. Dev.)	All	1:34 (1:28, 1:40)	1:24 (1:21, 1:26)	0.02*
Day-to-day variability between sleep middle (Mean)	All	1:06 (1:02, 1:11)	0:58 (0:56, 2:00)	0.01*
Day-to-day variability between sleep middle (Std. Dev.)	All	0:54 (0:50, 0:58)	0:49 (0:48, 0:51)	0.24
Day-to-day variability between sleep end (Mean)	All	1:16 (1:09, 1:23)	1:09 (1:07, 1:11)	0.22
Day-to-day variability between sleep end (Std. Dev.)	All	1:08 (1:02, 1:14)	1:04 (1:02, 1:06)	1.00
Day-to-day variability between sleep start (Mean)	All	1:34 (1:28, 1:41)	1:20 (1:17, 1:22)	0.00*
Day-to-day variability between sleep start (Std. Dev.)	All	1:16 (1:11, 1:22)	1:08 (1:05, 1:10)	0.05*
Total motion count during sleep time (Mean)	All	13647.79 (12323.81, 14971.77)	12707.36 (12244.70, 13170.01)	1.00
Total motion count during sleep time (Std. Dev.)	All	8463.87 (7533.06, 9394.68)	7747.39 (7382.87, 8111.92)	1.00
Sleep length between start and end (Mean)	All	9:04 (8:51, 9:17)	9:11 (9:07, 9:15)	1.00
Sleep length between start and end (Std. Dev.)	All	1:48 (1:41, 1:54)	1:38 (1:35, 1:40)	0.04*
End of the sleep cycle (Mean)	All	7:50 AM (7:36 AM, 8:03 AM)	7:44 AM (7:39 AM, 7:50 AM)	1.00
End of the sleep cycle (Std. Dev.)	All	1:19 (1:13, 1:26)	1:13 (1:11, 1:15)	0.37
Frequency of motion counts during sleep time (Mean)	All	0.20 (0.18, 0.21)	0.19 (0.18, 0.19)	0.55
Frequency of motion counts during sleep time (Std. Dev.)	All	0.06 (0.05, 0.06)	0.05 (0.05, 0.05)	0.05
Halfway between sleep start and sleep end (Mean)	All	3:17 AM (3:05 AM, 3:30 AM)	3:08 AM (3:03 AM, 3:14 AM)	1.00
Halfway between sleep start and sleep end (Std. Dev.)	All	1:08 (1:03, 1:14)	1:01 (0:59, 1:03)	0.05*
Start of the sleep cycle (Mean)	All	10:46 PM (10:30 PM, 11:01 PM)	10:33 PM (10:27 PM, 10:39 PM)	1.00
Start of the sleep cycle (Std. Dev.)	All	1:33 (1:26, 1:40)	1:21 (1:19, 1:24)	0.02*
Total daily motion count (Mean)	All	248621.24 (227847.62, 269394.86)	238254.11 (231486.01, 245022.21)	1.00
Total daily motion count (Std. Dev.)	All	61098.30 (54031.17, 68165.42)	60132.28 (57293.21, 62971.36)	1.00
Longest contiguous subsequence of zero actigraphy activity (Mean)	All	49.05 (47.20, 50.90)	49.04 (48.45, 49.63)	1.00
Longest contiguous subsequence of zero actigraphy activity (Std. Dev.)	All	11.68 (11.15, 12.22)	11.96 (11.76, 12.17)	1.00

p values corrected with the Benjamini-Yekutieli[34] method are shown, significant differences (α = 0.05) are highlighted with an *.

First-time/nulliparous pregnancies

Prior PTB complications are a strong predictor of future PTB complications, but such foresight does not exist in the case of nulliparious pregnancies. To evaluate these pregnancies, we train separate models on nulliparous patients. For training, we replace case report form features relating to delivery history with empty values. Results from training on nulliparous patients only are reported in Table 2.

As seen in Table 2, we find that the performance of actigraphy features is distinctive when we use Gaussian Naïve Bayes as the classifier. For all remaining model types, the performance is comparable both in contrast or together with case report form data, differing by relatively small amount for both AUROC and specificity at 90% sensitivity. This indicates that actigraphy data may provide performance comparable to or better than what can be assessed in a clinical survey specifically with regards to nulliparous patients.

In addition, the actigraphy features become a larger component of the most impactful features, as seen in Fig. 6, although part of this can be attributed to the reduced dimensionality. Of the case report form features included, those relating to socioeconomic status appear to be the most impactful. When examining the actigraphy features only, as seen in Fig. 7, we find that features evaluating variability are still among the most impactful features, whether they are averages of the day-to-day variability features or variances of the daily features.

Discussion

Overall, we find that actigraphy data compiled into simple measures of sleep can aid in the prediction of PTB, and that simpler ML architectures appear to perform better at this. For all ablations tested, we find that Gaussian Naïve Bayes (Gaussian NB) has the highest average AUROC. This is remarkable since it is architecturally simpler than other models, and suggests that the underlying features exhibit some independence from each other. This independence argument is furthered by the lower performance from our XGBoost models, as they learn decision trees where learned relationships may have dependencies. We do note that the small sample size and reduced dimensionality may enable this difference.

We also find that for the actigraphy-only models, there is a noticeable split in the explanability between aggregating variability and averages of actigraphy features. Among the highest performing features, we find that those capturing variability in sleep patterns—either at the day-to-day or whole-sample level—were the most explanatory features. Conversely, features examining a patient's average behavior generally ranked lower, which suggests that consistent sleep patterns are more important than any specific sleep metric. This insight could inform the development of intervention strategies focused on sleep hygiene, emphasizing the importance of reducing variability in sleep patterns rather than targeting sleep duration or timing alone.

Table 2 | Comparison of models for all patients and nulliparous patients

All Patients		**AUROC**		**AUPRC**		
Model	**Ablation**	**Pooled**	**95% CI**	**Pooled**	**95% CI**	**Specificity at 90% Sensitivity**
Gaussian naive Bayes	All features	0.697	**0.724** (0.692–0.755)	**0.383**	**0.382** (0.348–0.416)	**0.335** (0.215–0.456)
Linear support vector classifier	All features	0.669	0.672 (0.641–0.702)	0.311	0.337 (0.290–0.384)	0.156 (0.101–0.211)
Logistic regression	All features	**0.709**	0.722 (0.692–0.753)	0.329	0.373 (0.325–0.420)	0.349 (0.243–0.454)
Nonlinear support vector classifier	All features	0.689	0.720 (0.678–0.762)	0.307	0.345 (0.306–0.385)	0.269 (0.143–0.395)
XGBoost	All features	0.662	0.668 (0.637–0.698)	0.228	0.271 (0.240–0.303)	0.113 (-0.018–0.245)
Average	All features	0.685	0.701 (0.686–0.716)	0.312	0.342 (0.323–0.361)	0.244 (0.194–0.295)
Gaussian naive Bayes	Actigraphy only	**0.657**	**0.666** (0.623–0.709)	**0.222**	**0.254** (0.225–0.284)	**0.255** (0.157–0.353)
Linear support vector classifier	Actigraphy only	0.498	0.514 (0.447–0.581)	0.149	0.202 (0.167–0.236)	0.104 (0.026–0.183)
Logistic regression	Actigraphy only	0.601	0.619 (0.561–0.678)	0.175	0.217 (0.189–0.245)	0.185 (0.100–0.270)
Nonlinear support vector classifier	Actigraphy only	0.608	0.612 (0.575–0.649)	0.189	0.223 (0.201–0.246)	0.155 (0.100–0.209)
XGBoost	Actigraphy only	0.587	0.582 (0.537–0.626)	0.177	0.237 (0.182–0.291)	0.083 (-0.014–0.180)
Average	Actigraphy only	0.590	0.599 (0.574–0.623)	0.182	0.227 (0.212–0.241)	0.156 (0.120–0.193)
Gaussian naive Bayes	Case report forms only	0.672	**0.712** (0.683–0.742)	**0.379**	**0.390** (0.346–0.435)	0.218 (0.147–0.289)
Linear support vector classifier	Case report forms only	0.668	0.671 (0.635–0.706)	0.308	0.323 (0.287–0.359)	0.184 (0.134–0.234)
Logistic regression	Case report forms only	**0.703**	0.711 (0.690–0.732)	0.329	0.350 (0.306–0.394)	**0.287** (0.231–0.343)
Nonlinear support vector classifier	Case report forms only	0.623	0.645 (0.581–0.710)	0.258	0.292 (0.228–0.356)	0.171 (0.075–0.267)
XGBoost	Case report forms only	0.597	0.606 (0.568–0.644)	0.218	0.298 (0.246–0.349)	0.000 (-)
Average	Case report forms only	0.653	0.669 (0.650–0.688)	0.298	0.331 (0.309–0.352)	0.172 (0.136–0.208)

Nulliparous Patients		**AUROC**		**AUPRC**		
Model	**Ablation**	**Pooled**	**95% CI**	**Pooled**	**95% CI**	**Spec. at 90% Sens.**
Gaussian naive Bayes	All features	0.639	0.641 (0.571–0.711)	**0.451**	**0.437** (0.381–0.492)	0.187 (0.025–0.348)
Linear support vector classifier	All features	0.539	0.525 (0.444–0.606)	0.131	0.207 (0.174–0.240)	0.188 (0.088–0.288)
Logistic regression	All features	**0.670**	**0.671** (0.593–0.749)	0.156	0.247 (0.200–0.295)	**0.315** (0.138–0.493)
Nonlinear support vector classifier	All features	0.534	0.537 (0.498–0.577)	0.119	0.209 (0.188–0.230)	0.142 (0.056–0.229)
XGBoost	All features	0.622	0.622 (0.534–0.709)	0.152	0.274 (0.213–0.335)	0.188 (0.005–0.372)
Average	All features	0.601	0.599 (0.567–0.631)	0.202	0.275 (0.245–0.305)	0.204 (0.146–0.263)
Gaussian naive Bayes	Actigraphy only	0.677	**0.677** (0.595–0.759)	**0.169**	0.262 (0.222–0.302)	**0.325** (0.136–0.514)
Linear support vector classifier	Actigraphy only	0.524	0.496 (0.381–0.610)	0.128	0.218 (0.174–0.263)	0.131 (0.033–0.229)
Logistic regression	Actigraphy only	**0.679**	**0.677** (0.605–0.749)	0.160	0.257 (0.219–0.296)	0.298 (0.131–0.466)
Nonlinear support vector classifier	Actigraphy only	0.615	0.629 (0.543–0.714)	0.169	0.284 (0.207–0.362)	0.187 (0.081–0.292)
XGBoost	Actigraphy only	0.602	0.601 (0.515–0.687)	0.135	**0.347** (0.251–0.443)	0.138 (-0.033–0.310)
Average	Actigraphy only	0.619	0.616 (0.577–0.655)	0.152	0.274 (0.247–0.301)	0.216 (0.154–0.278)
Gaussian naive Bayes	Case report forms only	0.537	0.532 (0.468–0.596)	**0.415**	**0.408** (0.331–0.485)	0.129 (0.045–0.212)
Linear support vector classifier	Case report forms only	0.486	0.454 (0.362–0.546)	0.104	0.189 (0.153–0.226)	0.063 (0.014–0.112)
Logistic regression	Case report forms only	**0.632**	**0.662** (0.537–0.788)	0.162	0.331 (0.197–0.465)	**0.271** (0.052–0.491)
Nonlinear support vector classifier	Case report forms only	0.446	0.468 (0.400–0.536)	0.108	0.221 (0.176–0.266)	0.046 (0.001–0.092)
XGBoost	Case report forms only	0.612	0.618 (0.566–0.671)	0.140	0.288 (0.236–0.340)	0.127 (-0.065–0.318)
Average	Case report forms only	0.543	0.547 (0.507–0.587)	0.186	0.287 (0.251–0.324)	0.127 (0.070–0.185)

Area under the receiver-operator curve (AUROC), area under the precision-recall curve (AUPRC), and specificity at 90% sensitivity highlighted for each of the trained models, grouped by which sources of data were included. Model averages for area under the receiver-operator curve (AUROC) and area under the precision-recall curve (AUPRC) are obtained by pooling classifier results together, and 95% confidence intervals are obtained by averaging all folds.
Best performance metrics over each ablation are bolded.

Fig. 2 | Reciever-operator and precision-recall curves for models using all features. Pooled **a** reciever-operator curves and **b** precision-recall curves for all models using all data sources.

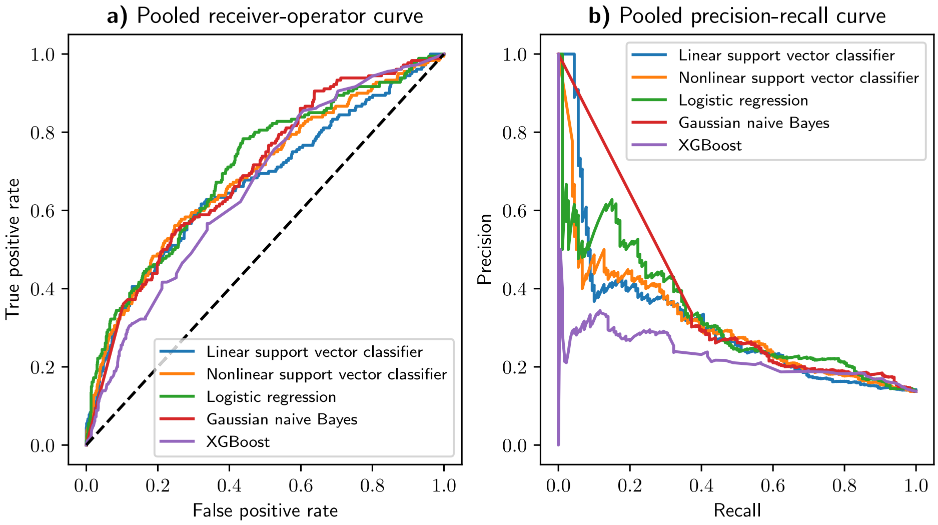


Table 3 | Confusion matrices for classifiers with the best area under the receiver-operator curve (AUROC) among all patients, with the threshold set to match a 50% true positive rate

(a) Actigraphy/Gaussian naive Bayes		
	PN	**PP**
TN	0.605	0.258
TP	0.076	0.061
(b) Case Reports/Gaussian naive Bayes		
	PN	**PP**
TN	0.701	0.162
TP	0.076	0.061
(c) All/Gaussian naive Bayes		
	PN	**PP**
TN	0.734	0.129
TP	0.095	0.043

PN is predicted negative, PP is predicted positive, TN is true negative, and TP is true positive.

Table 4 | Confusion matrices for classifiers with the best area under the receiver-operator curve (AUROC) among nulliparous patients, with the threshold set to match a 50% true positive rate

(a) Actigraphy/Logistic Regresssion		
	PN	**PP**
TN	0.674	0.222
TP	0.069	0.034
(b) Case Reports/Logistic Regression		
	PN	**PP**
TN	0.697	0.200
TP	0.069	0.034
(c) All/Logistic Regression		
	PN	**PP**
TN	0.683	0.214
TP	0.069	0.034

PN is predicted negative, PP is predicted positive, TN is true negative, and TP is true positive.

Table 5 | Tukey's honest significant difference test for area under the receiver-operator curve (AUROC) for across all models staratifed by feature set

	Actigraphy (1)	All (2)	Case Reports (3)	Nulliparous Actigraphy (4)	Nulliparous All (5)	Nulliparous Case Reports (6)
(1)	–	$2.62 \times 10^{-05*}$	0.01^{*}	0.96	1.00	0.14
(2)	$2.62 \times 10^{-05*}$	–	0.65	$9.08 \times 10^{-04*}$	$2.87 \times 10^{-05*}$	$3.12 \times 10^{-11*}$
(3)	0.01^{*}	0.65	–	0.12	0.01^{*}	$2.31 \times 10^{-07*}$
(4)	0.96	$9.08 \times 10^{-04*}$	0.12	–	0.97	0.01^{*}
(5)	1.00	$2.87 \times 10^{-05*}$	0.01^{*}	0.97	–	0.13
(6)	0.14	$3.12 \times 10^{-11*}$	$2.31 \times 10^{-07*}$	0.01^{*}	0.13	–

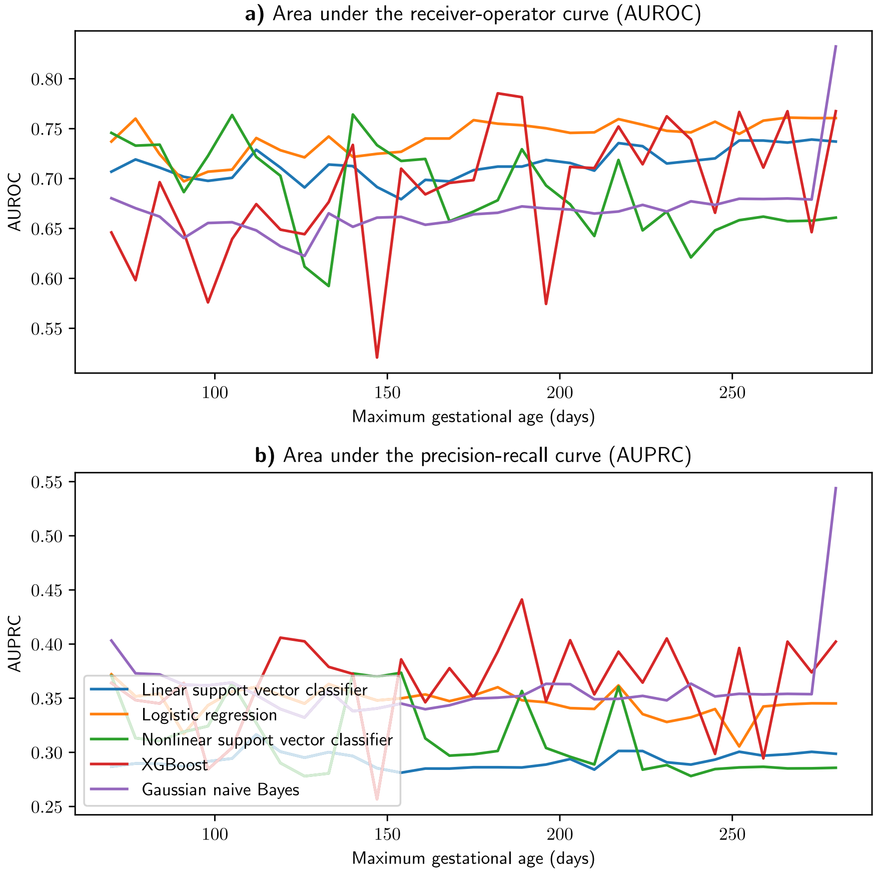
Significant differences (α = 0.05) are highlighted with an *.

Table 6 | Tukey's honest significant difference test for area under the precision-recall curve (AUPRC) across all models stratified by feature set

	Actigraphy (1)	All (2)	Case Reports (3)	Nulliparous Actigraphy (4)	Nulliparous All (5)	Nulliparous Case Reports (6)
(1)	–	$1.48 \times 10^{-08*}$	$4.12 \times 10^{-07*}$	0.10	0.09	0.01^{*}
(2)	$1.48 \times 10^{-08*}$	–	0.99	0.00^{*}	0.00^{*}	0.04^{*}
(3)	$4.12 \times 10^{-07*}$	0.99	–	0.02^{*}	0.03^{*}	0.17
(4)	0.10	0.00^{*}	0.02^{*}	–	1.00	0.98
(5)	0.09	0.00^{*}	0.03^{*}	1.00	–	0.98
(6)	0.01^{*}	0.04^{*}	0.17	0.98	0.98	–

Significant differences (α = 0.05) are highlighted with an *.

Fig. 3 | Reciever-operator and precision-recall curves for data up to a given gestational age (GA). Selected **a** receiver-operator curves and **b** precision-recall curves using with all features calculated with features up to a maximum GA using one random seed.



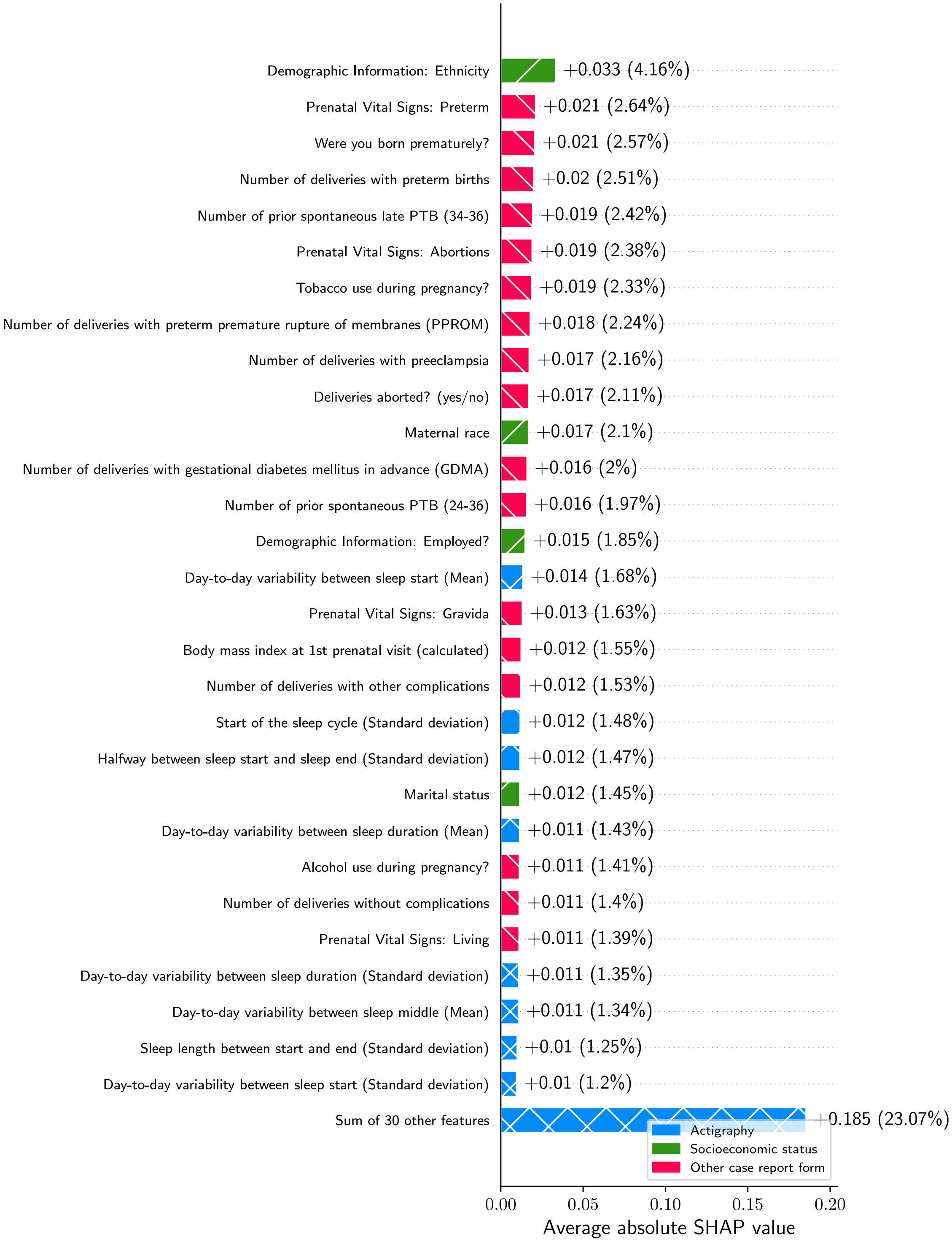


Fig. 4 | SHapley Additive exPlanations (SHAP) analysis of Gaussian NB with all features. Features indicative of socioeconomic status are highlighted green, and other patient history variables are highlighted red.

For the case report form features included, we find that some of the most explanatory features are past pregnancy complications, which is consistent with previous literature[18,20]. We also find that various socioeconomic features are predictive of PTB, as they may be proxy measures of maternal sleep. Race and ethnicity have been linked to increased sleep disturbances and poorer sleep quality in tandem with more frequent PTB outcomes[21,22]. Similarly, employment status and income have also been associated with differences in PTB outcomes[23], as the effects of employment range from physical overexertion[24] to direct conflicts with sleep[25].

For nulliparous patients, we find that the overall performance of the actigraphy data is more comparable in performance to models trained on the case report form data only. When compared to whole-cohort models, the performance is similar for the actigraphy-based models, while the performance of the models trained on the case report form data drops noticeably. In addition to past PTB being a

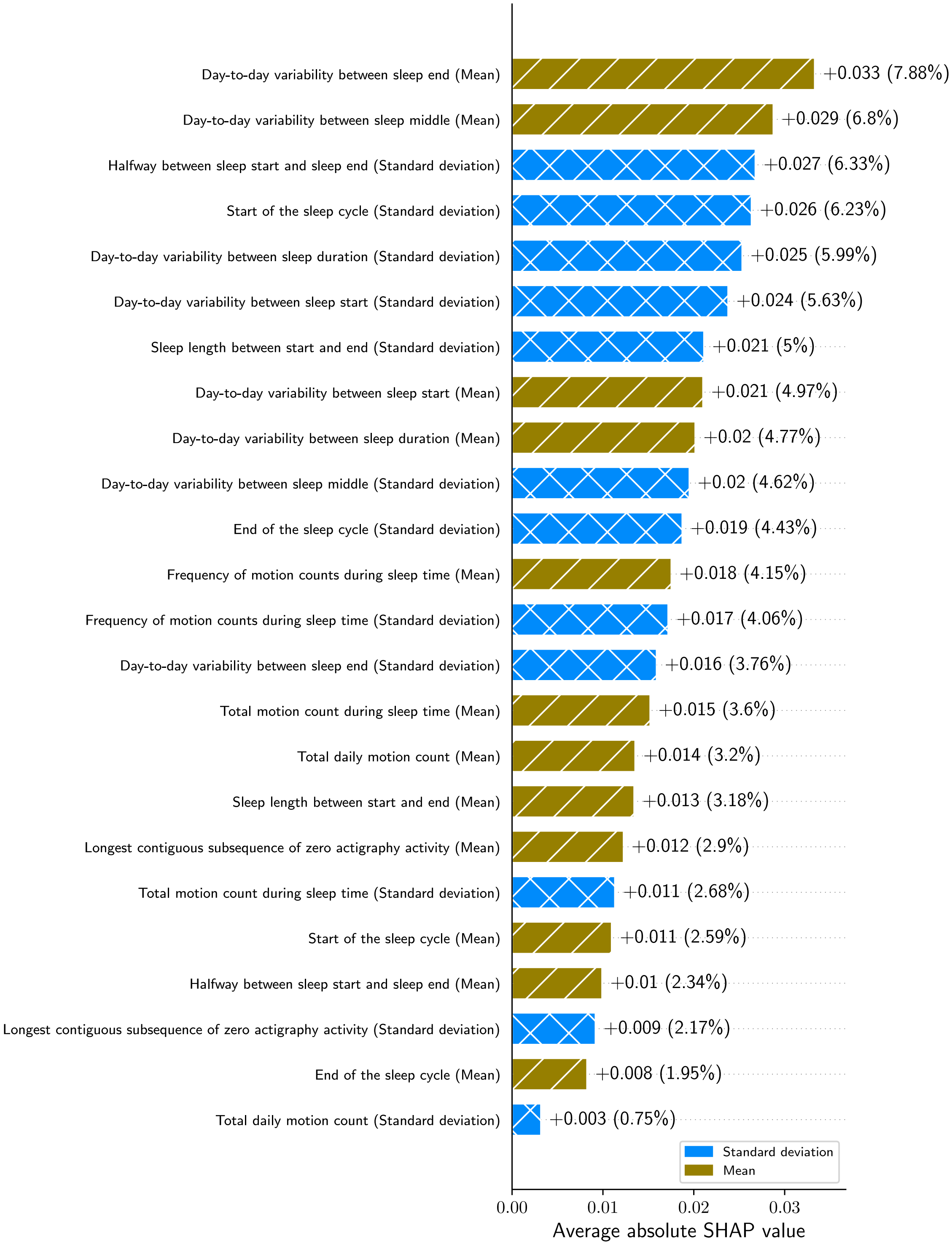


Fig. 5 | SHapley Additive exPlanations (SHAP) analysis of Gaussian Naïve Bayes (Gaussian NB) with actigraphy features only. Features aggregating average patient behavior are highlighted in gold, and features aggregating the standard deviation of patient behaviors are highlighted in blue.

strong predictor of future PTB, this may suggest that monitoring sleep patterns is more necessary for nulliparous patients.

One limitation of this approach is that we do not evaluate categorical features as one-hot values, as the sample size would not be able to counterbalance the large number of features generated by one-hot categorical features. As a result, it is more difficult to interpret the impact of some categorical variables that do not actually have an ordinality to them (e.g., race, marital status). Similarly, we discard non-numerical features from the case report form features, as incorporating them with vision/language models would significantly increase the overall dimensionality; future models may incorporate these for improved performance.

Sample size, particularly with regards to the nulliparous pregnancies, is another limiting issue, as it makes noise more prominent when training and evaluating these models. To mitigate this issue, we employed multiple random shufflings of the data for training and evaluation. However, we note that this is limited given the notable discrepancy between the AUROC/AUPRC metric poolings and their corresponding confidence intervals,

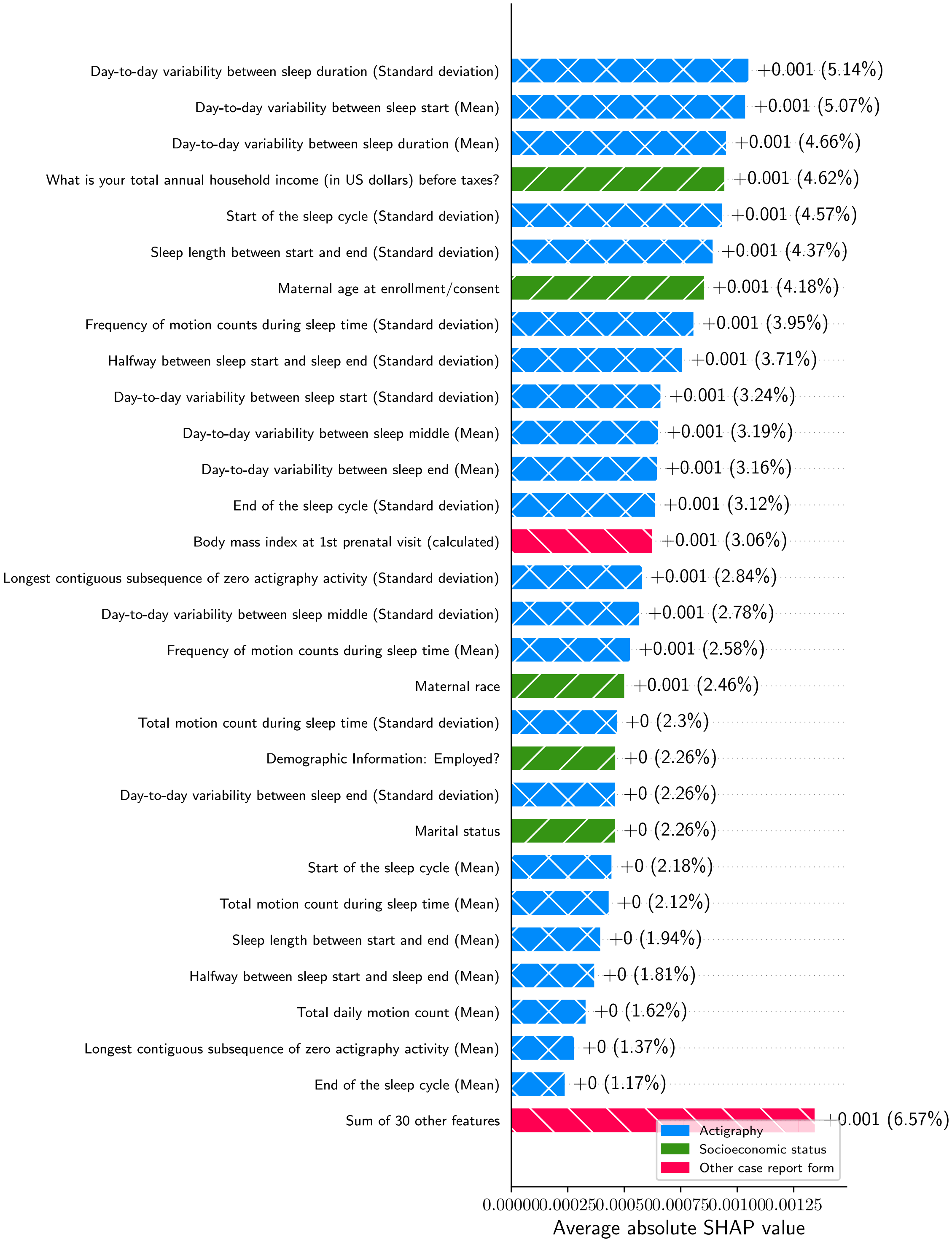


Fig. 6 | SHapley Additive exPlanations (SHAP) analysis of logistic regression with all features for nulliparous patients. Features indicative of socioeconomic status are highlighted green, and other patient history variables are highlighted red.

which may result from the wide performance differences between each shuffling and how they interact when averaging together. Sample size is a limitation not only in cohort size, but also the amount of actigraphy data, as the duration and frequency at which study participants wore their actigraphy watches was not consistent. Further studies should evaluate larger cohorts of patients to ensure accurate performance measurements, as well as cohorts from other locations to validate the performance with respect to different demographics. Moreover, longer and more consistent usage of actigraphy watches may also reveal more reliable patterns of motion behavior that predict PTB.

In addition, future work with actigraphy data could incorporate luminosity sensor data, as it may provide additional signals and corroborate signals captured by an actigraphy sensor. Another area of future work are with models trained with self-supervised learning (SSL), which learn relationships between input features before being fine-tuned for a downstream task. SSL models are particularly effective as these learned relationships between features generalize well in supervised tasks[26].

In conclusion, our findings show that actigraphy data can help preterm birth (PTB) in both multiparous and nulliparous patients, with sleep variability emerging as a key predictive feature. These results highlight the

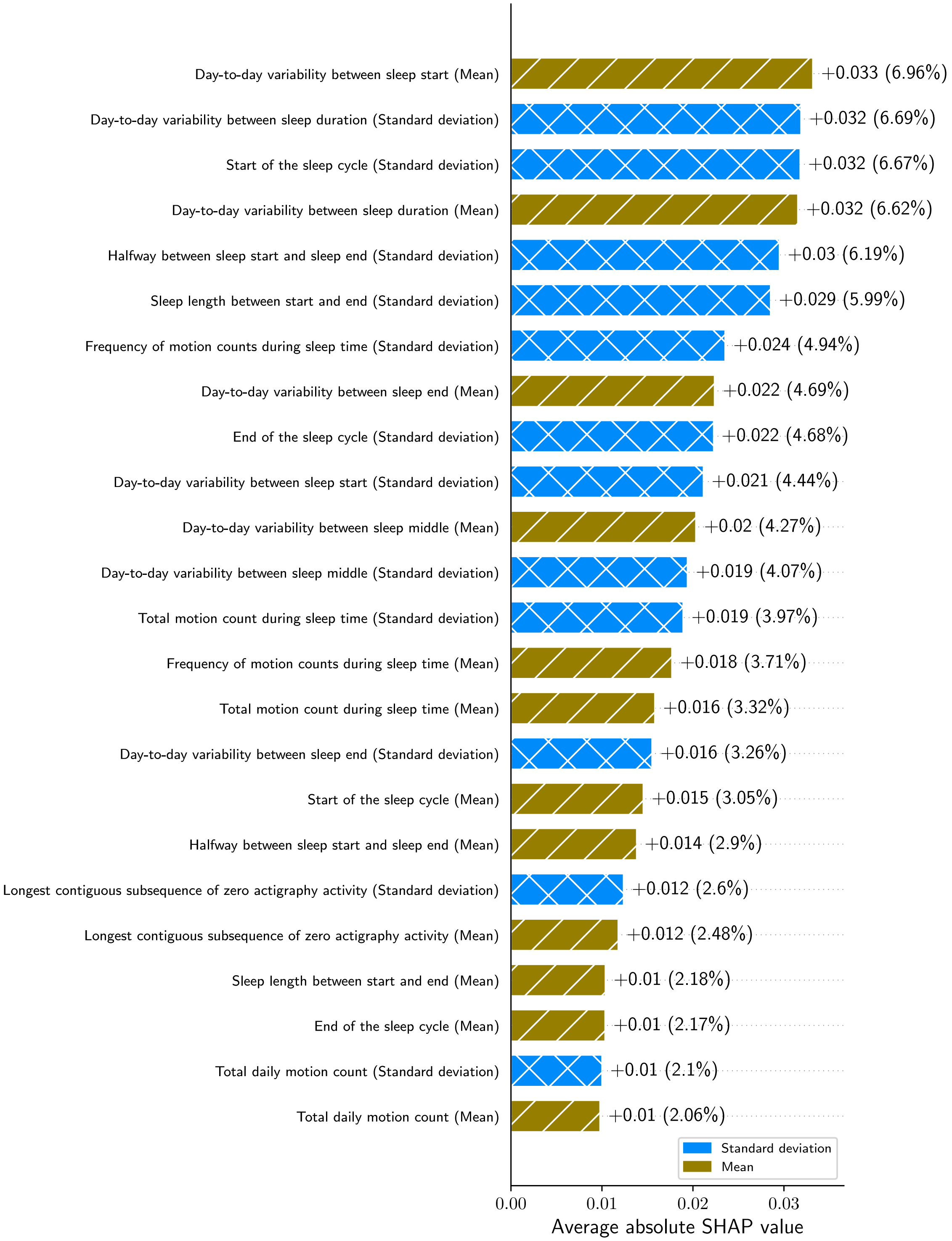


Fig. 7 | SHapley Additive exPlanations (SHAP) analysis of Gaussian Naïve Bayes (Gaussian NB) with actigraphy features only. Features aggregating average patient behavior are highlighted in gold, and features aggregating the standard deviation of patient behaviors are highlighted in blue.

potential of unobtrusive wearable measurements to enable early detection and intervention for PTB. Future work could explore larger or more diverse cohorts and develop targeted intervention strategies informed by these predictions to improve pregnancy outcomes.

Methods

Study characteristics

This study was completed as a part of of the March of Dimes Prematurity Research Center at Washington University in St. Louis/BJC HealthCare[12], which was approved by the Washington University IRB (reference #201612070) in accordance FDA Good Clinical Practices and the Declaration of Helsinki. Written informed consent was obtained from participants for the usage of their clinical, biospecimen, imaging, and questionnaire data. Patients were recruited at the Washington University Medical Campus if they had a singleton pregnancy with an estimated GA under 20 weeks, planned to deliver at Barnes-Jewish Hospital, and were age 18 or older.

Trained obstetric research staff used a series of case report forms to collect baseline maternal demographics, medical history, antepartum data and obstetric outcomes as previously described in ref. 12. Patient data were collected at scheduled study visits during each trimester and at delivery, where biological samples, imaging, actigraphy, and responses to standardized surveys were obtained from each patient.

Survey data included questions from eleven different validated surveys and standalone questions covering stress, schedule, sleep quality, physical activity, postnatal depression, diet, demographics, and overall lifestyle. We derive the label of PTB from the reported estimated date of confinement (EDC), labeling births that occur 3 full weeks before the listed EDC as PTB. EDC was derived from the patient's last menstrual period or first ultrasound[27].

Actigraphy feature design

Actigraphy measurements were collected over a 2-week period in each trimester (first trimester: 0–13 weeks and 6 days, second trimester: 14–27 weeks and 6 days, third trimester: ≥28 weeks) using the CamNtech MotionWatch 8. Measurements were collected at a minute-frequency over the duration a study participant wore their actigraphy watch. Patients were reminded through calls, emails, and texts to return their actigraphy watches after the capture period either at the next study visit or through a courier[12]. Patients who did not have actigraphy data in either their first or second trimester were filtered from the results for this analysis.

These features are very high-resolution, and to ensure the data is tractable for shallow ML model training, we engineer these raw time-series signals into aggregate features over day-level windows. On top of the day-level measurements, we also measure the absolute change between days where data is present. Section 2 of the Supplementary Materials contains a summary of these engineered features.

To generate these features, all actigraphy data is separated into days centered around midnight, from which we then attempt to estimate the sleep cycle that occurred for each given day. A summary of these calculated features that were used in the dataset for the ML models can be found in Section 2 of the Supplementary Materials.

Model design

For each study participant, we aggregate the day-level actigraphy features down to their mean and standard deviation across the entire duration of the pregnancy. When evaluating the window of GAs below a full-term pregnancy, we drop all actigraphy data with a GA below a set range (e.g., if we set the upper limit at 140 days, all data before 140 days are dropped, and the remainder is aggregated).

For the survey data, we select features with both domain knowledge and automatic techniques. We first select a predefined set of features based on pre-determined clinical knowledge, and sum values of questions regarding individual births together. After these features, we select an additional 10 features with the minimal-redundancy-maximal-relevance algorithm with semantic textual similarity scores generated with PubMedBERT[28] fine-tuned on several clinical and general datasets, as described in[29]. Features not represented numerically are dropped. The full list of features that we used can be found in Section 2 of the Supplementary Materials.

After this, we concatenate both sources of data, scale all numerical features to its normal distribution, and encode all categorical features as ordinal values. Missing values are imputed with either the mean, median, most common value, or the mean of the 5 nearest neighbors, which is learned during cross-validation (CV). The data is randomly split across a 80%/20% train/test split. For the whole cohort, 532 and 133 patients appear in in each split, with each split having 66 and 28 PTB patients, respectively. For the nulliparous cohort, this becomes 238 in the train set and 59 in the test set, with each of those splits having 24 patients and 9 patients respectively.

We train our models with several standard ML models, including logistic regression, linear support vector machine (SVM), kernelized/non-linear SVM[30], XGBoost[31], and Gaussian NB[32]. Logistic regression predicts the output class using the sigmoid of the linear combination of the input weights. Linear SVM predicts the class using a linearly-separated hyperplane, and kernelized SVM uses a kernel function to learn a non-linear separation of each class[30]. XGBoost is a gradient boosting method that builds an ensemble of decision trees to optimize predictive performance[31], and Gaussian NB models output class conditioned on normal distributions of each feature[32]. We evaluate the results across 10 random initializations for each model in Section "Results", and report the average AUROC and AUPRC through pooling[33], as well as the 95% confidence interval over all initializations. SHAP values are averaged across all random initializations. A graphical summary of this training pipeline can be seen in Section 2 of the Supplementary Materials.

To find the best hyperparameters for each of the tested models, we use 5-fold stratified CV, which preserves the class proportionality across each fold, using the training set. For XGBoost, the hyperparameter space ranges from 1 to 3 estimators, 1 to 3 maximum depth, a learning rate of 0.1, and a fitting objective of AUROC. For linear SVM, we test regularization parameters ranging logarithmically from 0.001 to 10, with 1000 iterations of training. For non-linear SVM, we evaluate polynomial and radial basis function kernels on top of the linear SVM parameters. For logistic regression, we evaluate regularization parameters from 0.001 to 10 with a L_2 penalty, and 1000 maximum iterations of training. For Gaussian NB, we use 10^{-9} as a fixed smoothing parameter.

Data availability

The data used in these findings can be obtained from the authors by request with permission from Washington University in St. Louis.

Code availability

We make the code used in this study available at https://github.com/bcwarner/mod-actigraphy-clf.

Received: 22 August 2024; Accepted: 26 May 2025;
Published online: 20 June 2025

References

1. Cao, G., Liu, J. & Liu, M. Global, regional, and national incidence and mortality of neonatal preterm birth, 1990-2019. *JAMA Pediatr.* **176**, 787–796 (2022).
2. Chawla, D. & Agarwal, R. Preterm births and deaths: from counting to classification. *Lancet Glob. Health* **10**, e1537–e1538 (2022).
3. Sutcliffe, S. et al. Risk of pre-term birth as a function of sleep quality and obesity: prospective analysis in a large Prematurity Research Cohort. *Sleep. Adv.* **4**, zpad043 (2023).
4. Wang, L. & Jin, F. Association between maternal sleep duration and quality, and the risk of preterm birth: a systematic review and meta-analysis of observational studies. *BMC Pregnancy Childbirth* **20**, 125 (2020).
5. Hoyniak, C. P. et al. The association between maternal sleep and circadian rhythms during pregnancy and infant sleep and socioemotional outcomes. https://www.researchsquare.com/article/rs-3937599/v1 (2024).
6. Hoyniak, C. P. et al. Sleep and circadian rhythms during pregnancy, social disadvantage, and alterations in brain development in neonates. *Dev. Sci.* **27**, e13456 (2024).
7. Li, R. et al. Sleep disturbances during pregnancy are associated with cesarean delivery and preterm birth. *J. Matern. Fetal Neonatal Med.* **30**, 733–738 (2017).
8. Cespedes, E. M. et al. Comparison of Self-Reported Sleep Duration With Actigraphy: Results From the Hispanic Community Health Study/Study of Latinos Sueño Ancillary Study. *Am. J. Epidemiol.* **183**, 561–573 (2016).
9. Nauha, L. et al. Comparison and agreement between device-estimated and self-reported sleep periods in adults. *Ann. Med.* **55**, 2191001 (2023).
10. Cos, H. et al. Predicting outcomes in patients undergoing pancreatectomy using wearable technology and machine learning: prospective cohort study. *J. Med. Internet Res.* **23**, e23595 (2021).
11. Dai, R. et al. Multi-task learning for randomized controlled trials: a case study on predicting depression with wearable data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **6**, 1–23 (2022).

12. Stout, M. J. et al. A multidisciplinary prematurity research cohort study. *PLoS ONE* **17**, e0272155 (2022).
13. Ravindra, N. G. et al. Deep representation learning identifies associations between physical activity and sleep patterns during pregnancy and prematurity. *npj Digit. Med.* **6**, 171 (2023).
14. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30**, https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (2017).
15. Varner, M. W. & Esplin, M. S. Current understanding of genetic factors in preterm birth. *BJOG* **112**, 28–31 (2005).
16. Hsieh, T.-T. et al. The impact of interpregnancy interval and previous preterm birth on the subsequent risk of preterm birth. *J. Soc. Gynecol. Investig.* **12**, 202–207 (2005).
17. Culhane, J. F. & Goldenberg, R. L. Racial disparities in preterm birth. *Semin. Perinatol.* **35**, 234–239 (2011).
18. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* **371**, 75–84 (2008).
19. Adane, H. A., Iles, R., Boyle, J. A., Gelaw, A. & Collie, A. Maternal occupational risk factors and preterm birth: a systematic review and meta-analysis. *Public Health Rev.* **44**, 1606085 (2023).
20. Kvalvik, L. G., Wilcox, A. J., Skjærven, R., Østbye, T. & Harmon, Q. E. Term complications and subsequent risk of preterm birth: registry based study. *BMJ* m1007. https://www.bmj.com/lookup/doi/10.1136/bmj.m1007 (2020).
21. Christian, L. M. et al. Pathways to maternal health inequities: Structural racism, sleep, and physiological stress. *Brain Behav. Immun.* **123**, 502–509 (2025).
22. Lucchini, M. et al. Racial/ethnic disparities in subjective sleep duration, sleep quality, and sleep disturbances during pregnancy: an ECHO study. *Sleep* **45**, zsac075 (2022).
23. Huang, L. et al. Association between sleep during pregnancy and birth outcomes: a prospective cohort study. *Reprod. Biol. Endocrinol.* **23**, 18 (2025).
24. Saurel-Cubizolles, M. J. Employment, working conditions, and preterm birth: results from the Europop case-control survey. *J. Epidemiol. Community Health* **58**, 395–401 (2004).
25. Wallace, D. A. et al. Associations between evening shift work, irregular sleep timing, and gestational diabetes in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). *Sleep* **46**, zsac297 (2023).
26. Erhan, D., Courville, A., Bengio, Y. & Vincent, P. Why does unsupervised pre-training help deep learning? *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 201–208. ISSN: 1938-7228. https://proceedings.mlr.press/v9/erhan10a.html (2010).
27. American College of Obstetrics and Gynecology Committee on Practice Bulletins-Obstetrics. ACOG Practice Bulletin Number 49, December 2003: Dystocia and augmentation of labor. *Obstet. Gynecol.* **102**, 1445–1454 (2003).
28. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
29. Warner, B. C., Xu, Z., Haroutounian, S., Kannampallil, T. & Lu, C. Utilizing Semantic Textual Similarity for Clinical Survey Data Feature Selection ArXiv:2308.09892 [cs]. *Find. Assoc. Comput. Linguist.: ACL 2025*. https://2025.aclweb.org/program/find_papers/ (2025).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
31. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge* 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785 (2016).
32. Chan, T. F., Golub, G. H. & LeVeque, R. J. Updating formulae and a pairwise algorithm for computing sample variances. In *COMPSTAT 1982 5th Symposium Held Toulouse 1982* (eds Caussinus, H., Ettinger, P. & Tomassone, R.) 30–41. http://link.springer.com/10.1007/978-3-642-51461-6_3 (Physica-Verlag HD, 1982).
33. Jack, H. & Niall, A. On averaging ROC curves. *Trans. Mach. Learn. Res*. https://openreview.net/forum?id=FByH3qL87G (2023).
34. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*. **29**, https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-4/The-control-of-the-false-discovery-rate-in-multiple-testing/10.1214/aos/1013699998.full (2001).

Acknowledgements

We thank the participants in the cohort study for their key contributions in advancing preterm birth science, as well as the staff in the Division of Clinical Research who facilitated the enrollment of participants and collection of data that made this study possible. This work was supported financially by a research grant from the March of Dimes Foundation and by the Fullgraf Foundation. This work was supported institutionally by St. Louis Children's Hospital, Barnes-Jewish Hospital, the Department of Obstetrics and Gynecology at Washington University School of Medicine in St. Louis, the Washington University AI for Health Institute, and the Washington University in St. Louis James McKelvey School of Engineering.

Author contributions

B.C.W. wrote the main manuscript text and contributed the code for the training, evaluation and interpretation of the models. P.Z. contributed the code for the features. E.D.H., S.K.E., A.I.F., and C.L. provided subject matter expertise. All authors were responsible for the interpretation of the models.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44294-025-00082-y.

Correspondence and requests for materials should be addressed to Benjamin C. Warner, Peinan Zhao, Erik D. Herzog, Antonina I. Frolova, Sarah K. England or Chenyang Lu.