Divyannsh Pincha
Kabir Bhalla
Sambhav Banthia
Shikhraj Singh

PREDICTING SLEEP ONSET LATENCY USING WEARABLES

Before the Dream Begins

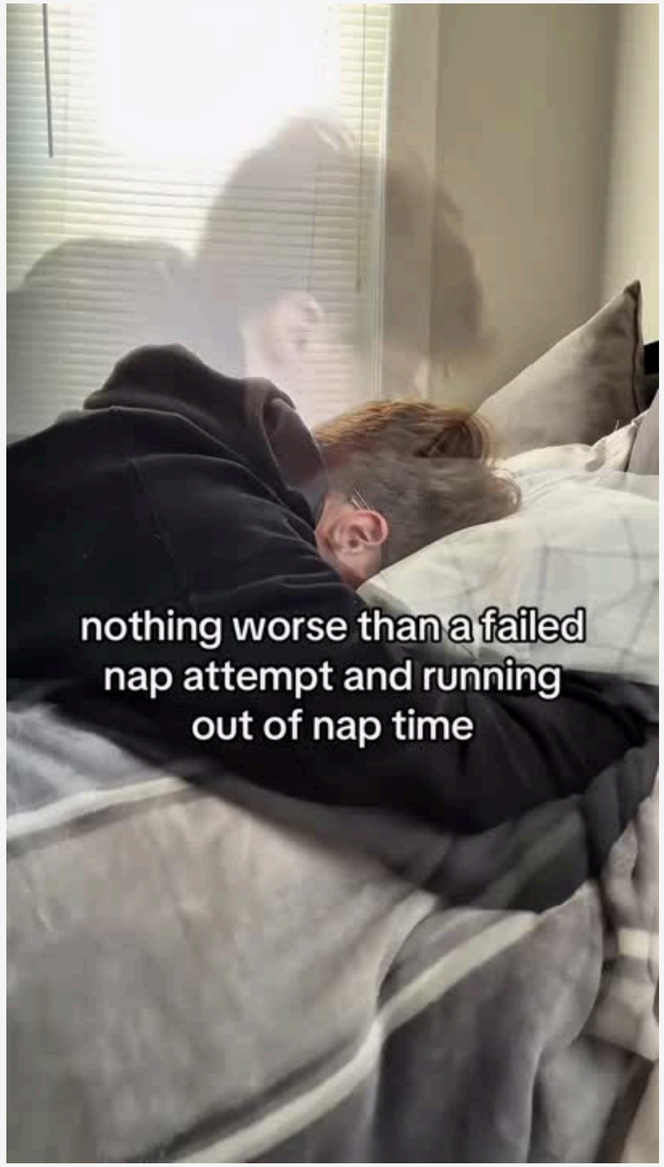
Take Leo DiCaprio in Inception:

He has to infiltrate the dreams of a billionaire and plant an idea in his head.

However, he first has to wait for him to fall asleep.

How will he know **how long** the billionaire will take to fall asleep?

Problem Statement



The Core Issue

People frequently spend extended periods lying awake in bed trying to sleep, which increases frustration, anxiety but most importantly, it **wastes time.**

Without an **objective signal**, people rely heavily on subjective feelings.

Context

Shifts from retrospective scoring to actionable pre-sleep guidance using wearables.

“If I go to bed now, what is the chance I will fall asleep quickly, if it’s low I would rather do something else. "

Slow sleep onset affects:
next-day focus
mood and productivity
long-term sleep quality*

*https://www.mayoclinic.org/diseases-conditions/insomnia/symptoms-causes/syc-20355167

Research Question

Given pre-bedtime wearable features and recent sleep history, can a machine learning model estimate:

P(SOL < 15 min | features, participant)

Core Definitions

SOL

Sleep Onset Latency — time from lights-off to first sustained sleep.

Success

SOL < 15 minutes. A quick onset bedtime attempt.

Slow Onset

SOL ≥ 15 minutes. Delayed or failed sleep onset attempt.

Sleepability

P(SOL < 15 min) — a probabilistic, personalized score per bedtime.

Target

Binary label: 1 if quick onset, 0 if slow onset.

Input

Wearable features from bedtime window

Output

Sleepability Score ∈ [0, 1]

Evaluation

PR-AUC, F1 (slow onset), Brier, ECE

PR- AUC: How well does the model rank high-risk slow-onset nights above normal nights?
F1: How good is the model at actually classifying slow-onset nights?
Brier Score: How close predicted probabilities are to reality.
Expected Calibration Error: Measures whether predicted probabilities match real-world frequencies.

Our model converts wearable data into a personalized sleepability score:

P(SOL<15 minutes)

This means the system can estimate **before bedtime** whether a person is likely to fall asleep quickly.

Potential Applications

1. Wearable-Based Sleep Coaching

"Your sleepability score tonight is low. Try starting wind-down 30 minutes earlier."

2. Personalized Pre-Bedtime Nudges
The system can recommend actions before sleep becomes a problem.
Possible nudges: e.g reduce screen exposure, avoid late caffeine.

3. Student Wellness Dashboard at Plaksha

Students can track how bedtime timing and recent sleep patterns affect their ability to fall asleep.

4. Habit Tracking and Behavior Feedback
The model gives feedback before sleep.
Shift:
From post-sleep tracking → pre-sleep prediction

5. Early Warning for Sleep Routine Disruption
If a student's recent sleep history shows increasing sleep onset latency, the system can flag a possible worsening sleep pattern early.

Impact

- Helps users understand when they are likely to struggle with sleep and take preventive action.
- Encourages better sleep hygiene through personalized, data-driven feedback.
- Could support student wellness programs by identifying broad sleep-risk trends in an anonymized, privacy-preserving way.

LITERATURE REVIEW

NOVELTY & EXISTING WORK

Has anyone worked on predicting SOL in **REAL TIME**?

To our knowledge, limited prior work directly predicts imminent sleep-onset latency from free-living consumer wearable signals as a calibrated pre-bedtime probability.

Resting-State Subjective Experience and EEG Biomarkers Are Associated with Sleep-Onset Latency


B. Alexander Diaz [1,2]

Richard Hardstone [1,2]


Huibert D. Mansvelder [1,2]

Eus J. W. Van Someren [1,2,3,4]


Klaus Linkenkaer-Hansen [1,2*]

What did they do?

- **13** healthy male participants
- **223 resting-state EEG trials**
- **Mixed models predicting SOL** from theta/alpha-band activity
- Requires **clinical EEG** equipment

Gaps we fill?

1. Requires EEG electrodes
2. Lab setting
3. **N=13**
4. Not scalable to app

Models for predicting sleep latency and sleep duration FREE

Francisco G Vital-Lopez, Thomas J Balkin, Jaques Reifman

Sleep, Volume 44, Issue 5, May 2021, zsaa263,

https://doi.org/10.1093/sleep/zsaa263

Published: 29 November 2020 **Article history**

What did they do?

A **biomathematical** model based on the **(homeostatic sleep pressure + circadian rhythm)**, to predict sleep-onset latency and sleep duration. Validated across 278 subjects from 18 **controlled laboratory studies**, producing 147 time-point measurements.

Gaps we fill?

1. **Doesn't** have **raw consumer wearable signals** like HR, HRV, actigraphy, temperature.
2. Predicts continuous **SOL (minutes)**, not a calibrated **probability** that SOL ≤ 15 min.
3. SOL in lab cohorts, **not from consumer wearables**;

Our work: first to use consumer wearable pre-sleep signals (HR, HRV, TEMP, ACC) to output a calibrated real-time P(SOL ≤ 15min) in free-living conditions

LITERATURE REVIEW

PROACTIVE SLEEP PREDICTION

Towards proactively improving sleep: machine learning and wearable device data forecast sleep efficiency 4–8 hours before sleep onset

Collin Sakal [1], Tong Chen [1], Wenxin Xu [1], Wei Zhang [1], Yu Yang [1], Xinyue Li [1]

Affiliations + expand

PMID: 40293116 DOI: 10.1093/sleep/zsaf113

What did they do?

Developed **CatBoost** (gradient boosting) and **CNN-LSTM** models trained **exclusively on accelerometer** data from **80,811 adults** in the UK Biobank to forecast low sleep efficiency
Higher activity 4–6 hours before bed had moderate beneficial associations with sleep efficiency

Gaps we fill?

1. Predicts sleep **efficiency**, not sleep onset, **different outcome**
2. **Uses accelerometer only** - no HR, HRV, or skin temperature
3. **4–8 hour prediction horizon is far longer** than our real-time bedtime decision

Mapping the road to better sleep: forecasting sleep quality using actigraphy-based machine learning hours before bedtime

Ankit Parekh

What did they do?

- Peer-reviewed commentary in Sleep journal
- Calls for **calibrated probability outputs** ; not just **classification**
- Explicitly **identifies SOL prediction** as the next critical frontier

Gaps we fill?

we answer Parekh's call directly: SOL-specific, real-time, calibrated probability from consumer wearable signals

LITERATURE REVIEW

PRE-SLEEP HRV PREDICTS SLEEP ONSET

Pre-sleep heart rate variability predicts chronic insomnia and measures of sleep continuity in national-level athletes

Qinlong Li [1], Xiaochen Lei [2], Wenlang Yu [1], Charles J Steward [3], Yue Zhou [1]

What did they do?

Measured **pre-sleep HRV for 5 minutes** using Polar H10 chest strap before a single PSG night. Used **binary logistic regression** and **MLP neural network** to predict **insomnia** and sleep quality outcomes.

Gaps we fill?

1. Elite athletes only
2. Single night PSG
3. Chest strap

Heart rate variability at bedtime predicts subsequent sleep features

M P Tramonti Fantozzi, F Artoni, U Faraguna

What did they do?

- Pre-sleep LF/HF ratio correlated with **SOL and sleep architecture**
- Evening chronotypes showed **different autonomic profiles at the same clock time**
- **Pre-sleep window** (immediately before lights-off) is the **correct measurement epoch**

Gaps we fill?

1. They developed a correlational model
2. Lab experiment

Our work: extracts HRVand LF/HF from consumer wrist wearable in the timeframe before bedtime, operationalising this validated mechanism at scale

LITERATURE REVIEW

SKIN TEMPERATURE AS SOL PREDICTOR

Subjective sleep onset latency is influenced by sleep structure and body heat loss in human subjects

Ryusei Iijima, Akari Kadooka, Kairi Sugawara, Momo Fushimi, Mizuki Hosoe, Sayaka Aritake-Okada

What did they do?

Examined relationship between **subjective SOL, sleep structure, skin/body temperature changes**, and sleep evaluation in healthy adults. Used **stepwise regression** to identify the strongest predictor of subjective SOL

Gaps we fill?

1. Explains subjective vs objective SOL discrepancy - important for why diary labels are biased
2. Lab study

Functional link between distal vasodilation and sleep-onset latency?

K. Kräuchi, C. Cajochen, +1 author A. Wirz-Justice · Published in American Journal of... 1 March 2000 · Medicine, Environmental Science

What did they do?

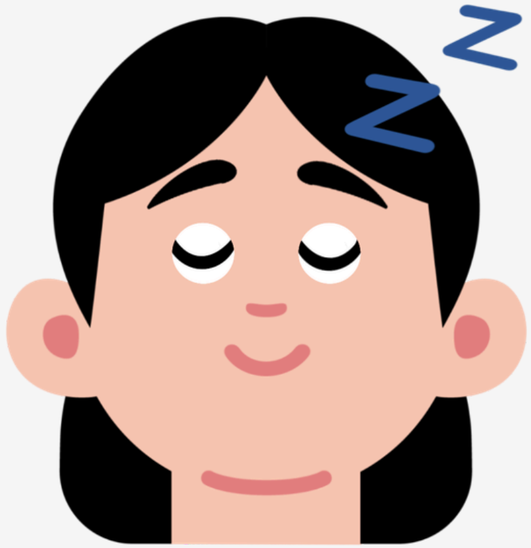
18 subjects in a **controlled constant-routine protocol**. Measured distal and **proximal skin temperatures, core body temperature, HR, melatonin**, and subjective sleepiness simultaneously. Used **stepwise regression to identify the best predictor - distal-to-proximal skin temperature gradient**

Gaps we fill?

1. Controlled laboratory constant-routine — not free-living
2. Manual measurement, not wearable sensor
3. DPG requires both distal (wrist) AND proximal (abdomen) temperature — wrist-only wearables give only one measurement

Our work: Empatica E4 wrist sensor captures distal skin temperature at 4Hz

LITERATURE REVIEW



There is prior work that:

1. Wearables that estimate SOL retrospectively
   - Fitbit,Apple watches etc.
   - They give you last night's SOL not the prediction in real time
2. Predictive work on other sleep outcomes (but not imminent SOL)
   - Machine-learning forecasting of low sleep efficiency 4–8 hours before sleep onset from accelerometer data
   - These target night-level efficiency or alertness, not SOL at the moment

Identified Gap: No work has

- Used consumer-grade wearables (HR, HRV, actigraphy, temp) in free-living conditions to build a supervised model that outputs a calibrated probability that SOL ≤ 15 min if the user attempts sleep now.

- Delivered that probability as a continuous, real-time "sleepability score" across the evening, so the user can choose when to initiate a sleep attempt.

LITERATURE REVIEW

FEASIBILITY

Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography

Daniel M Roberts [1], Margeaux M Schade [2], Gina M Mathew [2], Daniel Gartenberg [1], Orfeu M Buxton [2]

What did they do?

Validated **multi-sensor consumer wearables** (Apple Watch, Oura, Fitbit, Basis) for **sleep detection compared to PSG and research-grade actigraphy received >85%+ accuracy.** Epoch-level comparison of HR and motion signals.

How it helps?

Validates that consumer-grade wearable HR and motion are sufficiently accurate for sleep-related modelling. Justifies using Empatica E4 (clinical-grade consumer wearable) as a data source.

Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device

Olivia Walch [1,✉], Yitong Huang [2], Daniel Forger [3], Cathy Goldstein [1]

What did they do?

Multi-sensor wearable model using actigraphy + HR + circadian phase features. Achieved **~79% accuracy for 3-class sleep staging in large cohorts.**

How it helps?

Directly validates that wearable HR/HRV + motion signals are sufficiently informative to model transitions from wake to sleep

FEASIBILITY??

Can we even predict SOL using wearables and contexual data **accurately**?

The overwhelming majority of research validates consumer-grade wearables for estimating, not predicting SOL after the fact by analyzing physiological signals collected during attempted sleep.

These estimates are compared against PSG or actigraphy as ground truth and are useful for tracking trends over time or evaluating interventions like CBT-I.

Wearables Can Reliably Capture the Needed Physiology and Behavior

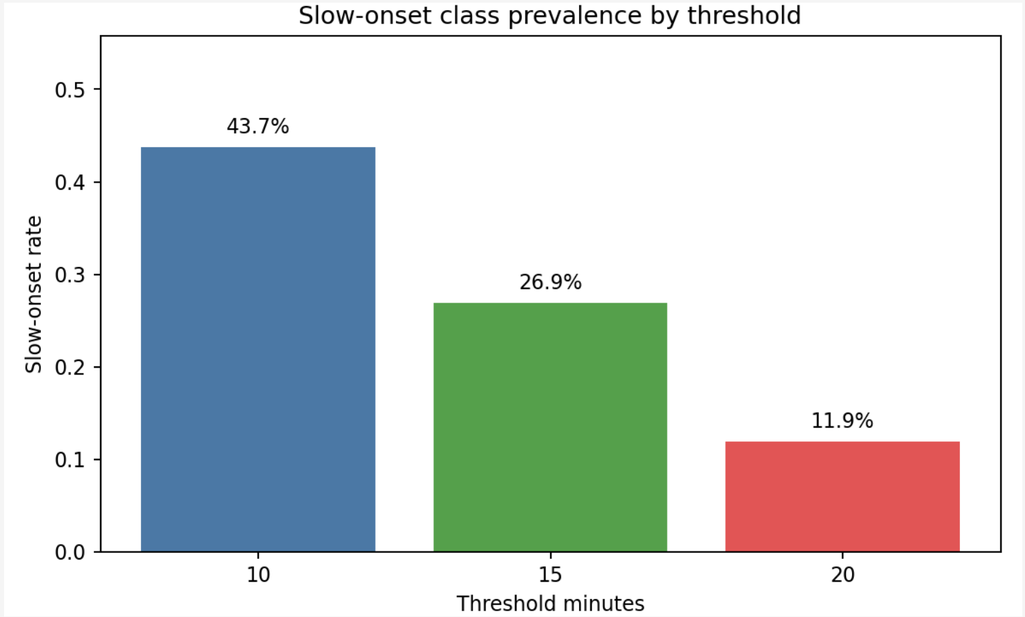
- *Multi-sensor devices (Apple Watch, Oura, Fitbit, Basis, etc.) provide heart rate, motion, and sometimes temperature that are strongly correlated with PSG and research-grade actigraphy at the epoch level*
- *Similar AI approaches (SLAMSS) using actigraphy + HR achieve ~79% accuracy for 3-class staging and can accurately estimate stage durations in large cohorts*

Implication: Wearable HR/HRV + motion signals are sufficiently informative to model transitions from wake to sleep.

WHY WE CHOSE 15 MINUTES

15 minutes was selected as the best balance between behavioral meaning and learnability.

Threshold	Slow-Onset Definition	Slow-Onset Count	Quick-Onset Count	Slow-Onset Rate	Interpretation
10 min	SOL ≥ 10 min	1748	2249	43.70%	Balanced but too mild/common
15 min	SOL ≥ 15 min	1076	2921	26.90%	Best compromise
20 min	SOL ≥ 20 min	475	3522	11.90%	Meaningful but too imbalanced



We chose 15 minutes because it preserved behavioral meaning without making the slow-onset class too rare.

DATASET SELECTION

Three non-negotiable requirements drove every selection decision

≥ 7 nights / person | **Pre-sleep signal window** | **Objective SOL label**

REJECTED

KAIST

Galaxy Watch · n = 49 · 28 nights

- No skin temperature sensor
- SOL = next-morning diary (~80% recall bias)
- Samsung PPG HRV inaccurate during waking
- RMSSD error too high in pre-sleep window

DREAMT v2.1

Empatica E4 + PSG ·n = 100· 1 night

- 1 night only — sleep debt impossible
- Clinical OSA patients, not free-living
- PhysioNet credentialing too slow
- No bedtime regularity computable

MESA

Actigraphy · n =2,237 · 7-day

- Zero pre-sleep physiology signals
- PSG SOL missing in 55% of rows
- Mean age 69.6 — wrong population
- No HR, HRV, or temperature

SELECTED

UCI COVID

Oura + Samsung · n = 21 · avg 7.8 m

Sleep debt

Rolling 7-day TST via Oura · 8 months depth

Pre-sleep HRV

Samsung 5-min HRV windows all evening

Validated SOL

Oura onset_latency · 79% PSG agreement

Bedtime regularity

Months of data → SD of bedtimes computable

UCI COVID is the only dataset meeting all three requirements — selected as the sleep debt validation set

DATASET CONSTRUCTION
DRYAD: MULTIMODAL LONGITUDINAL WEARABLE DATASET

Content:

- 23 valid participants
- 3997 valid sleep episodes
- 2921 success episodes
- 1076 slow-onset episodes
- Main anchor: Oura sleep.csv

Raw participant folders
↓
Oura sleep episode extraction
↓
15-min success target
↓
Leakage-safe feature engineering

Data Modalities

Oura Sleep	Anchor + SOL target
Oura Activity / Readiness	Daily readiness scores
Samsung HRV	Pre-bedtime HR & HRV
Samsung Pedometer	Steps, movement
EMA Mood	Affect, anxiety self-report
Personicle	Additional context

The dataset is multimodal, but sleep episodes are anchored using Oura onset latency

CLASS IMBALANCE & PARTICIPANT-AWARE VALIDATION

Imbalance Handling

Problem:

27% slow onset vs 73% quick onset — default classifiers prefer the majority class

Strategy:

Random oversampling of minority (slow onset) class to ~40% in TRAIN fold only

Critical rule:

Oversampling is NEVER applied to test fold — prevents artificially inflated performance

Result:

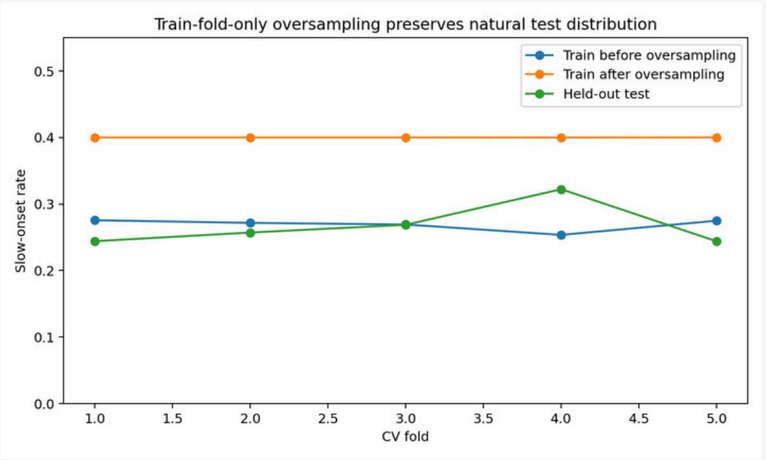
After oversampling: ~60% quick onset / ~40% slow onset in each training fold

Why not SMOTE?

Tested; random oversampling gave comparable or better generalization on this dataset

Cross-Validation: StratifiedGroupKFold (5-fold)

Fold	Test N	Test Success%	Test SlowOnset%
1 (4 ptps)	408	83.1%	16.9%
2 (6 ptps)	1187	62.8%	37.2%
3 (5 ptps)	903	82.5%	17.5%
4 (4 ptps)	703	75.1%	24.9%
5 (4 ptps)	796	70.9%	29.1%



Validation:

- StratifiedGroupKFold by participant
- held-out participants in each fold
- oversampling only inside training folds
- test folds left untouched

Methodology : From Raw Wearable Data to Sleepability Score

1

Raw Participant Folders

Oura, Samsung, EMA, Personicle

2

Sleep Episode Extraction

Oura anchor → SOL → valid episodes

3

Leakage Audit

Remove same-night features

4

Feature Engineering

7 feature groups, rolling histories

5

Ablation + Feature Selection

Standalone + incremental testing

6

Cross-Validation

StratifiedGroupKFold, train-fold oversampling

7

Model Training

LR → RF → CatBoost ladder

8

Sleepability Score

P(SOL < 15 min) output

FEATURE PRE PROCESSING

1. Raw wearable records were aligned to each sleep episode using bedtime_start_timestamp.

Purpose: only use data available before sleep attempt.

```python
def _window(df: pd.DataFrame, bedtime, hours: int) -> pd.DataFrame:
    return df[
        (df["dt"] < bedtime)
        & (df["dt"] >= bedtime - pd.Timedelta(hours=hours))
    ]
```

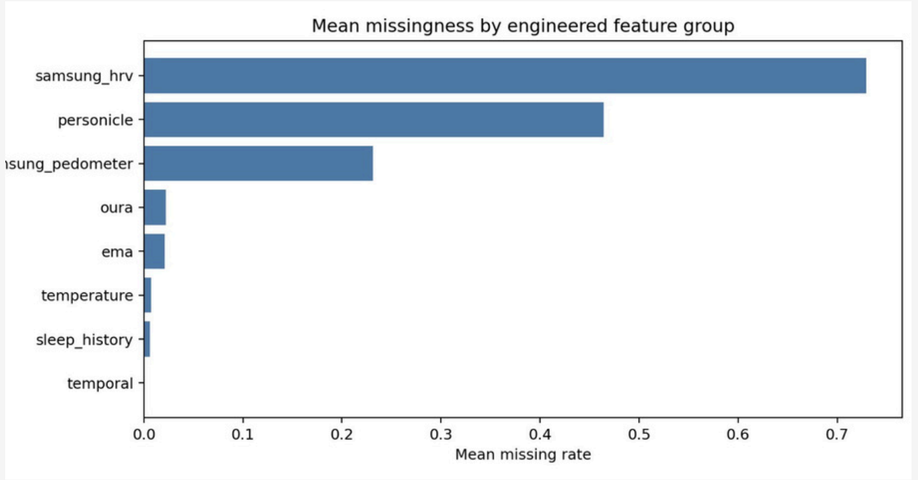
2. Missing value handling

For HRV, because missingness was high, we used median based imputation Missing values were not imputed globally. Imputation was fitted only on training folds and then applied to test folds to avoid leakage.

```python
keep["target_sleep_success_15"] = (
    keep["onset_latency_seconds"] < cfg["success_threshold_seconds"]
).astype(int)
keep["target_slow_onset_15"] = 1 - keep["target_sleep_success_15"]
```

3. Scaling

Scaling was used only for models which needed it : Logistic Regression only



4. Cyclic Encoding of time

Bedtime hour was transformed into sine/cosine features for circadian phase:
$\sin(2\pi h/24), \cos(2\pi h/24)$

5. **Lagged Oura summaries are attached only from before bedtime**

5. Leakage Prevention: The Model Cannot See the Future

✗ FORBIDDEN — Post-bedtime / Same Night	✓ ALLOWED — Pre-bedtime Only
✗ Same-night SOL / onset latency	✓ Bedtime hour, day-of-week, weekend flag
✗ Same-night sleep duration	✓ Previous night SOL (lag-1)
✗ Same-night sleep efficiency	✓ Rolling 3-night / 7-night SOL mean
✗ Same-night sleep score	✓ Previous sleep duration & efficiency
✗ Same-night sleep stages (REM, deep...)	✓ Bedtime hour variability (rolling 7n)
✗ Same-night sleeping HR / HRV	✓ Pre-bedtime HRV / HR windows
✗ Raw timestamps	✓ Steps / movement before bedtime
✗ Participant ID as a feature	✓ EMA mood/affect (same-day, pre-sleep)

```
grouped = df.groupby("participant_id", group_keys=False)
df["prev_onset_latency_minutes"] = grouped["onset_latency_minutes"].shift(1)
df["prev_sleep_duration_hours"] = grouped[
    "sleep_duration_hours_same_night_for_lag_only"
].shift(1)

shifted = df.groupby("participant_id")[[
    "onset_latency_minutes",
    "sleep_duration_hours_same_night_for_lag_only",
    "sleep_efficiency_same_night_for_lag_only",
    "bedtime_hour",
]].shift(1)

df["rolling_7d_sol_mean"] = (
    shifted["onset_latency_minutes"]
    .groupby(df["participant_id"])
    .rolling(7, min_periods=1)
    .mean()
    .reset_index(level=0, drop=True)
)
```

We removed the post/during sleep data and any sort of variable which gave away the answer

FEATURE ENGINEERING

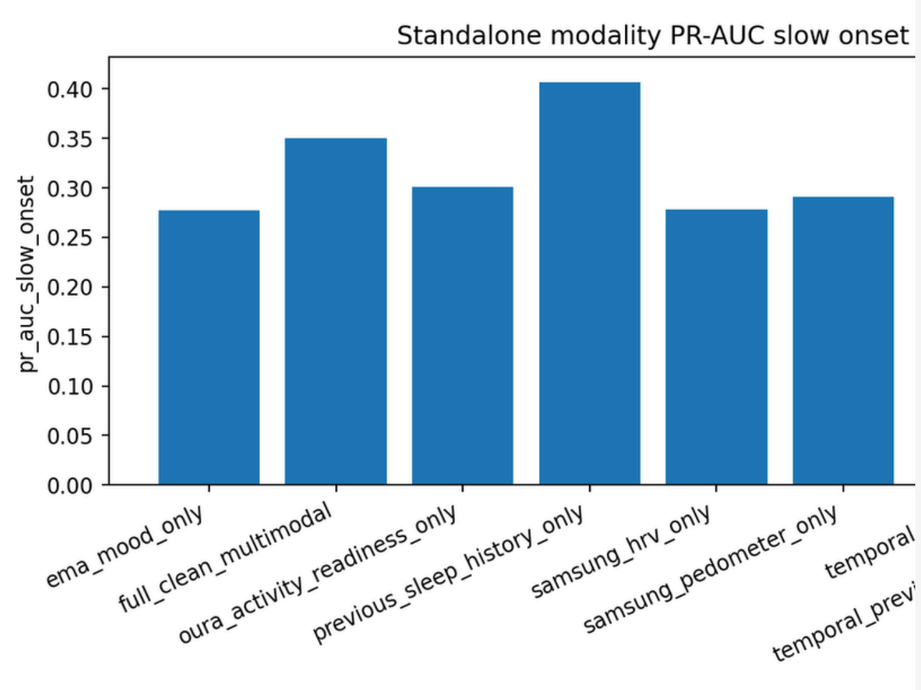
Features were grouped by sleep mechanisms, not randomly selected.

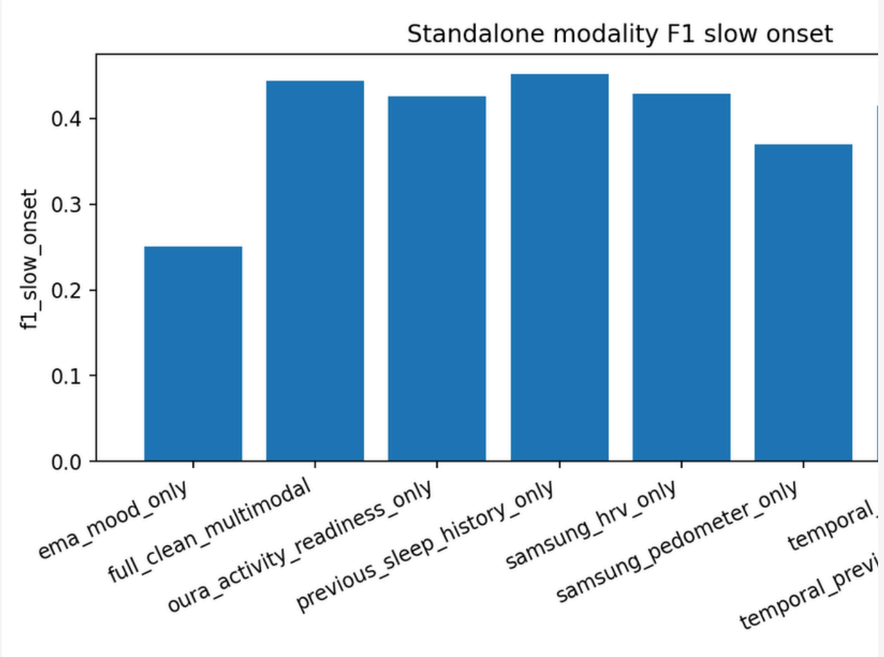
Feature Group	Examples	Why It Matters
Temporal features	bedtime hour, day index, cyclical time	Captures circadian timing
Previous sleep history	previous SOL, rolling SOL, previous bedtime	Captures sleep habit and sleep debt
Oura activity/readiness	activity score, readiness, recovery	Captures recovery state
Samsung HRV	HR mean, RMSSD, SDNN, HRV count	Captures autonomic arousal
Pedometer/activity	steps, movement, calories	Captures daytime load
EMA mood	nervousness, worry, affect	Captures emotional arousal

We tested whether each modality added useful predictive signal.

STANDALONE MODALITY ABLATION

Which Modality Works Best by Itself?



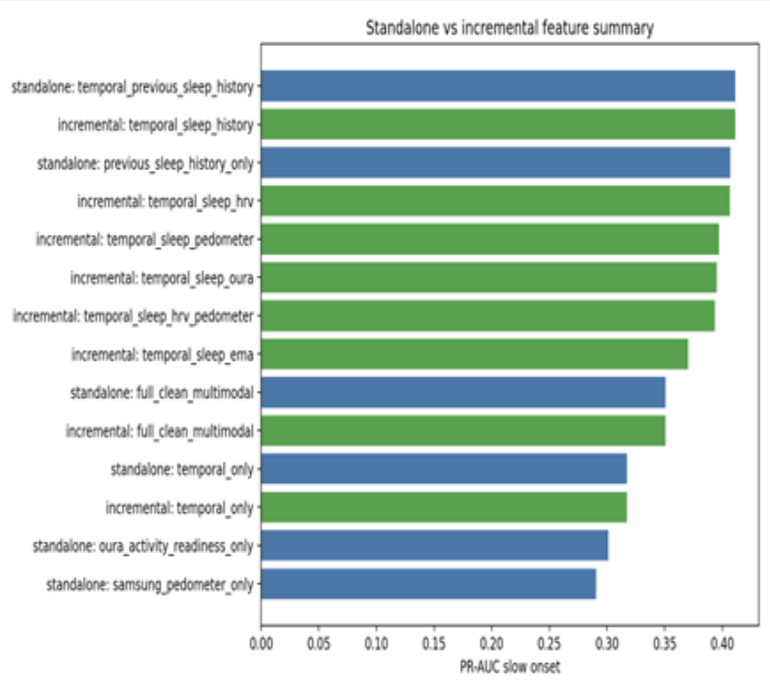


Feature Set	Features	Missingness	PR-AUC Slow	F1 Slow Onset
Temporal +	18	0.40%	0.411	0.434
Previous sleep	12	0.60%	0.407	0.452
Full	172	24.50%	0.35	0.444
Temporal only	6	0.00%	0.317	0.415
Oura only	16	2.40%	0.301	0.426
Samsung	42	23.10%	0.291	0.37
Samsung HRV	36	72.90%	0.278	0.429
EMA mood	50	2.10%	0.278	0.251

For the standalone modality ablation, we trained a separate Logistic Regression model for each feature group.

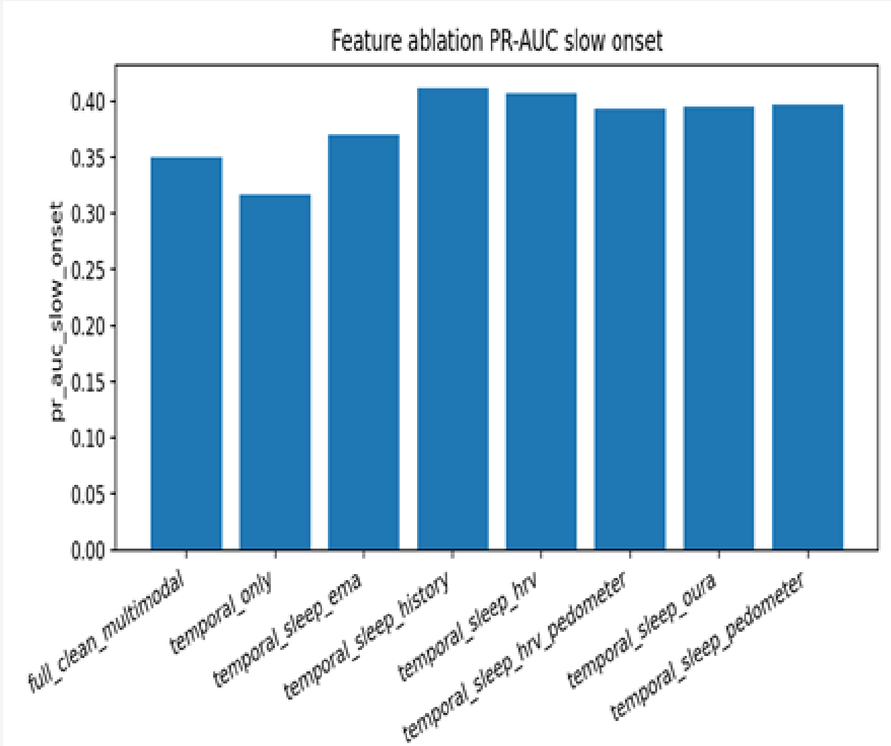
Sleepability was more history-dependent than sensor-dependent in this dataset.
hrv was a good predictor with a high f1 score but had higher missing values

INCREMENTAL FEATURE ABLATION



Do Extra Modalities Improve the Sleep-History Baseline?

If we already know timing and sleep history, does adding HRV/Oura/EMA/pedometer help?



Feature Set	# Features	Missingness	PR-AUC (slow onset)
Temporal + History	18	0.4%	0.411
+ Oura	34	1.3%	0.395
+ Sleep History Only	12	0.6%	0.400
+ HRV	**54**	**48.8%**	**0.457**
Full Multimodal	172	24.5%	0.350

Temporal + Sleep History + HRV showed the strongest signal but had high missingness, we imputed

MODEL PIPELINE

We moved from interpretability to nonlinear modeling to boosted tabular learning.

Logistic Regression
↓
Random Forest
↓
CatBoost

Model	Question
Logistic Regression	Can sleepability be explained by interpretable linear effects?
Random Forest	Are there nonlinear interactions that logit misses?
CatBoost	Can stronger boosting improve tabular performance?
Extended benchmark	Do additional models improve over the final model?

We did not start with the most complex model. We first built an interpretable baseline, then tested whether nonlinear models improved prediction.

LOGISTIC REGRESSION I: STATISTICAL RESULTS

Diagnostic	Result	
Observations	3,997	Large episode-level dataset
Features	18	Compact final feature set
LR test p-value	0.001	Features jointly improve prediction
Hosmer–Lemeshow p-value	0.127	No strong evidence of poor fit
Converged	Yes	Model fit was stable

Which features are associated with higher or lower odds of falling asleep within 15 minutes?

Before using black-box models, we checked whether the feature set has meaningful interpretable signal.

LOGISTIC REGRESSION II: PREDICTIVE BASELINE

After using Logistic Regression statistically, we also used it as a predictive ML baseline
We ask:

How well can a simple linear model predict sleep success on unseen participants?

Predictive Setup
Target: P(SOL<15 min)
Feature set: Temporal + sleep history + HRV
Validation: participant-aware cross-validation
Imbalance: training-fold-only oversampling
Evaluation: held-out participants

HRV-Enhanced Logistic Regression Results

Diagnostic	Result
PR-AUC slow onset	0.41
F1 slow onset	0.443
Brier score	0.203
ECE	0.126
Top-20% risk lift	1.681

Logistic Regression gives a strong transparent baseline and performs well for ranking high-risk slow-onset nights.

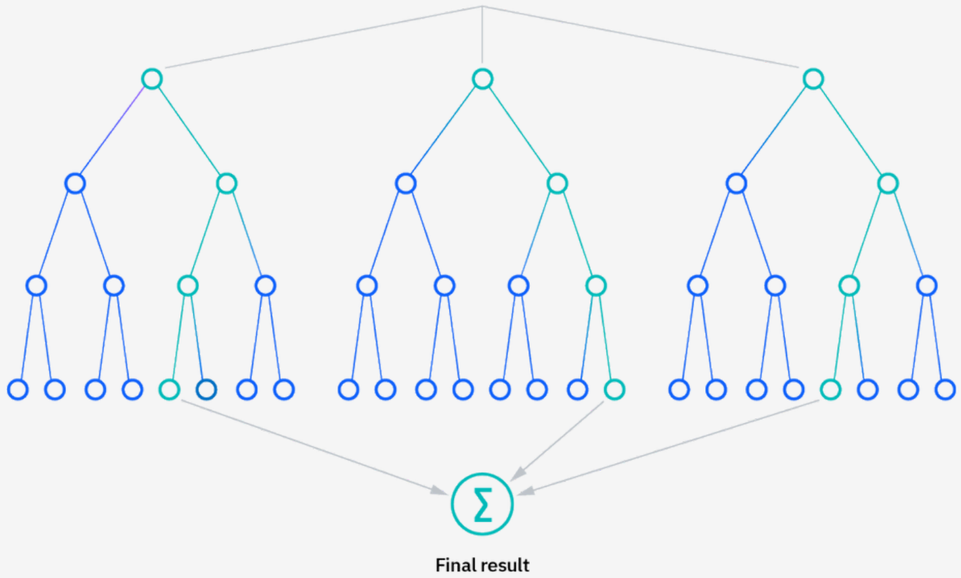
RANDOM FOREST

Why Random Forest?

Logistic Regression assumes mostly additive linear effects.

Random Forest was used to capture:

- nonlinear thresholds
- feature interactions
- irregular sleep-history patterns
- HRV effects that may matter only under certain bedtime-history conditions



Why It Fits Our Problem

Sleepability may depend on combinations such as:

recent high SOL + late bedtime + low HRV

These interactions are difficult for Logistic Regression but natural for Random Forest.

RANDOM FOREST: HOW WE USED IT

Setup:

Feature set:

Temporal + Previous Sleep History + HRV

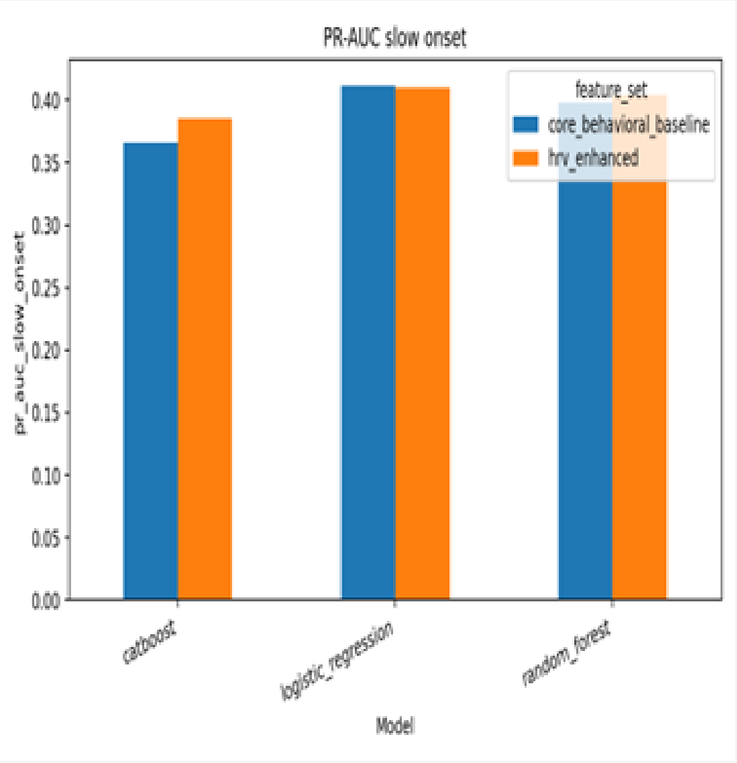
- StratifiedGroupKFold by participant
- oversampling only inside training folds
- test folds untouched

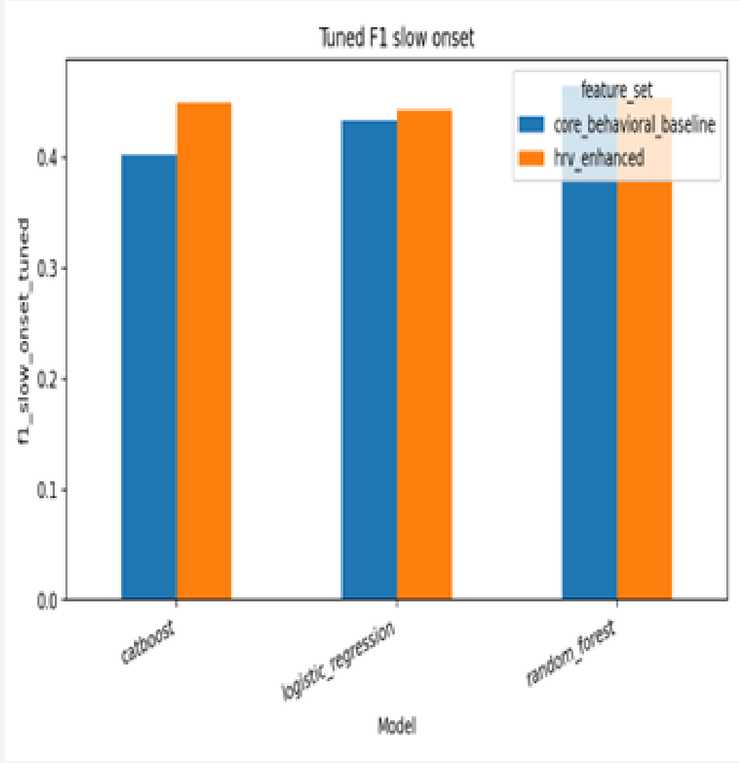
Why These Parameters?
We used a regularized forest to reduce overfitting because the dataset has many sleep episodes but only 23 participants.

The final HRV-enhanced Random Forest used a tuned success cutoff of approximately 0.60, selected within training folds. This means the model only predicts sleep success when $P(SOL<15) \geq 0.60P$

Parameter	Value
n_estimators	300
max_depth	5
min_samples_leaf	5
max_features	sqrt
class_weight	None
imbalance handling	train-fold-only oversampling

RANDOM FOREST: RESULTS





Random Forest improves practical slow-onset classification compared with Logistic Regression on F1-style decision performance.

It also maintains better probability error than Logistic Regression:
Brier: 0.197 vs 0.203

Random Forest is better suited for nonlinear sleepability interactions.

Parameter	Value
PR-AUC slow onset	0.404
F1 slow onset	0.454
Brier score	0.197
ECE	0.109
Top-20% risk lift	1.607

CATBOOST

CatBoost is a strong gradient-boosted tree model for tabular data.

We tested it because it can:

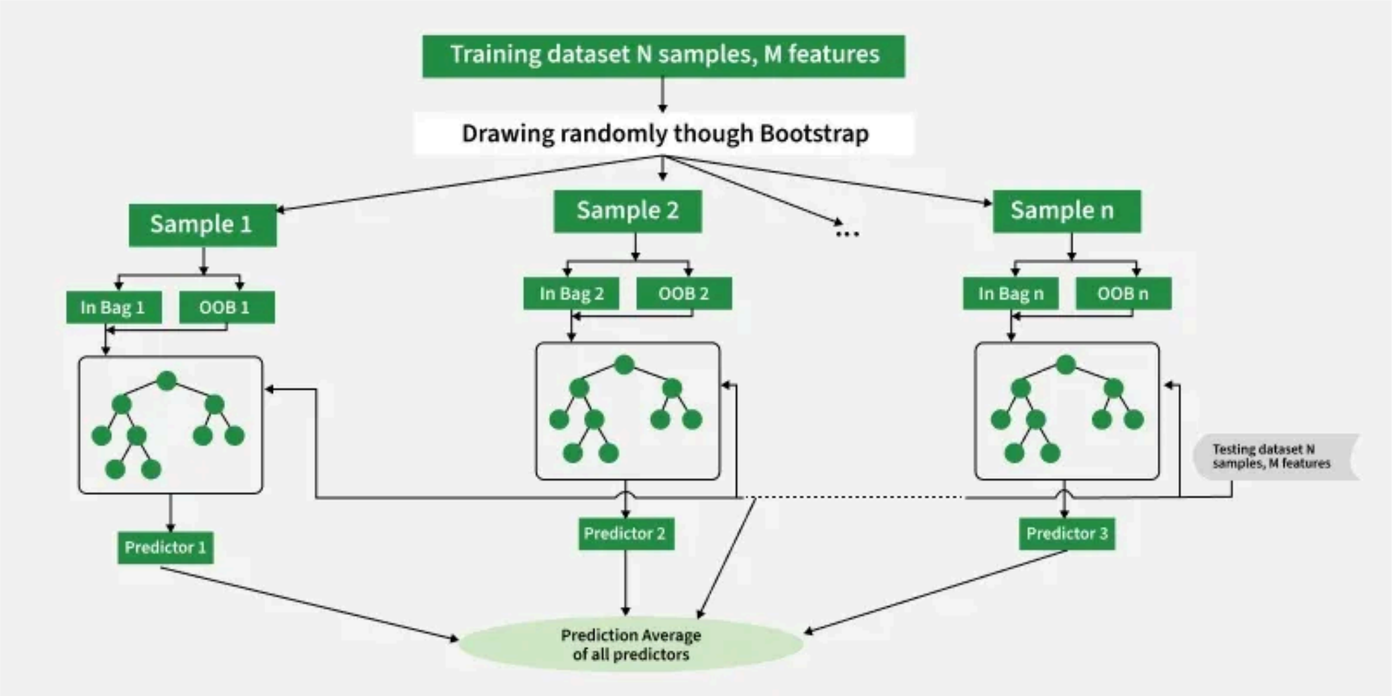
- capture nonlinear relationships
- handle feature interactions
- work well with mixed and missing tabular features
- serve as a stronger boosting benchmark against Random Forest

Why It Fits Our Problem

Wearable sleep data is:

- tabular
- noisy
- nonlinear
- missingness-heavy
- interaction-heavy

CatBoost is designed for this kind of predictive setting.



CATBOOST

CatBoost tested whether stronger boosting could outperform Random Forest on the HRV-enhanced feature set.

CatBoost uses gradient boosting , that builds a sequence of decision trees, where each new tree tries to fix the errors of the previous trees. Start with a basic prediction (like the average label). Compute residuals (the mistakes). Train new trees to correct those mistakes. Repeat for many rounds, each tree gradually improves the model

Feature Set: Temporal + Previous Sleep + HRV

Parameter	**Strategy**
Loss function	Logloss
Evaluation	AUC / probability metrics
Tree depth	modest depth
Learning rate	regularized
Class weights	None
Missingness	handled through model +

Validation:

- participant-aware folds
- training-fold-only oversampling
- test folds untouched

Results

Metric	Value
PR-AUC slow onset	0.385
F1 slow onset	0.45
Brier score	0.199
ECE	0.08
Top-20% risk lift	1.611

CatBoost had the best calibration error ECE=0.080

But it did not beat Random Forest on slow-onset F1 or Brier score.

Conclusion

CatBoost was useful as a strong benchmark, but it did not replace Random Forest as the final predictive model.

MODEL COMPARISION

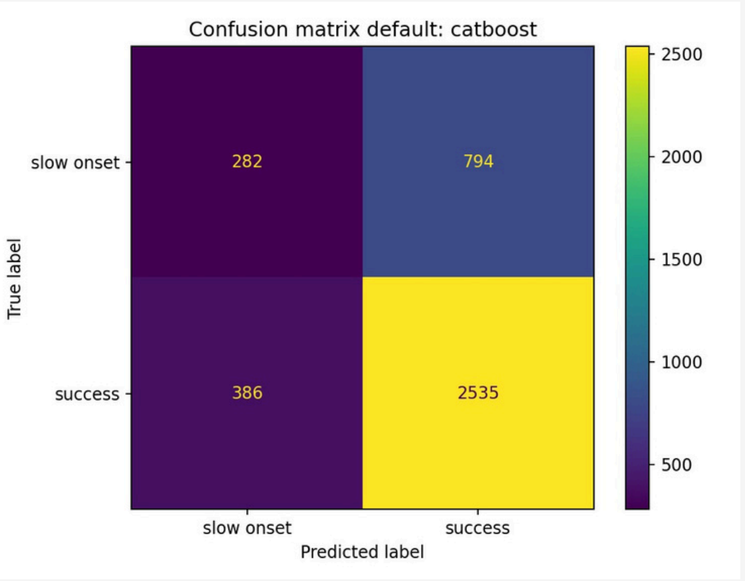
Model	PR-AUC Slow Onset	F1 Slow Onset	Brier	ECE	Top-20% Lift
Logistic Regression	0.41	0.443	0.203	0.126	**1.681**
Random Forest	0.404	**0.454**	**0.197**	0.109	1.607
CatBoost	0.385	0.45	0.199	**0.08**	1.611

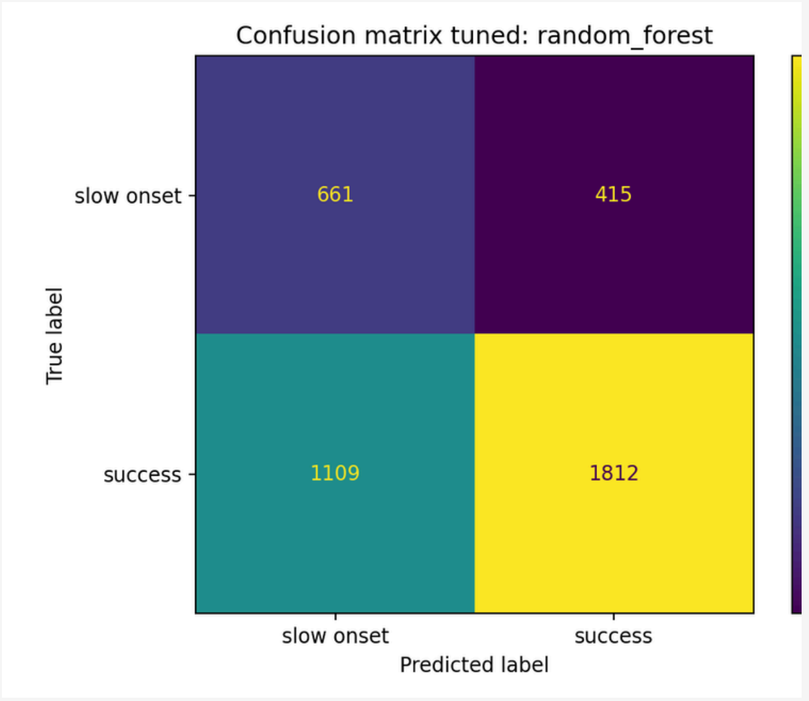
- Logistic Regression had the best top-risk lift.
- Random Forest had the best slow-onset F1 and Brier score.
- CatBoost had the best calibration error.
- Random Forest gave the best overall practical tradeoff.

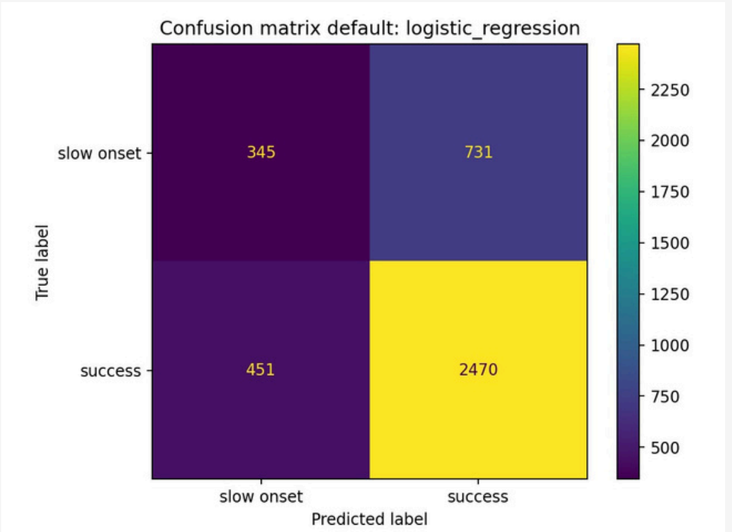
Feature Set: Temporal + Previous Sleep History + Pre-bedtime HRV

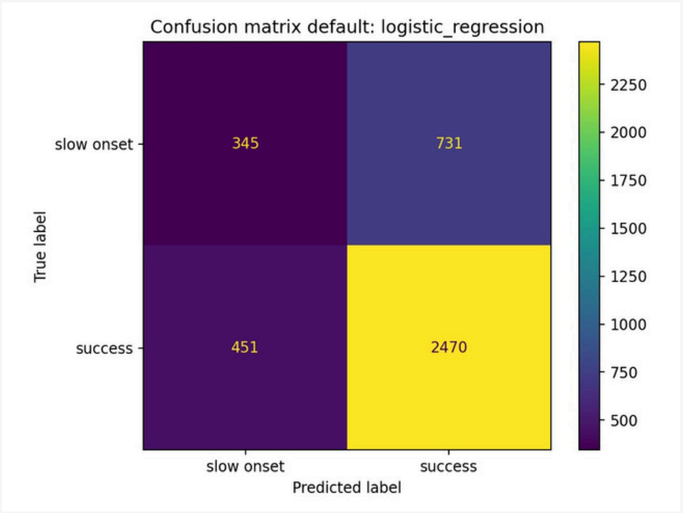
Reason	Explanation
Includes behavioral signal	previous sleep history and bedtime
Includes physiological signal	pre-bedtime HRV
Captures nonlinear effects	Random Forest handles interactions
Good decision performance	best slow-onset F1 among main HRV
Good probability quality	best Brier score among main HRV
Literature aligned	HRV reflects autonomic readiness

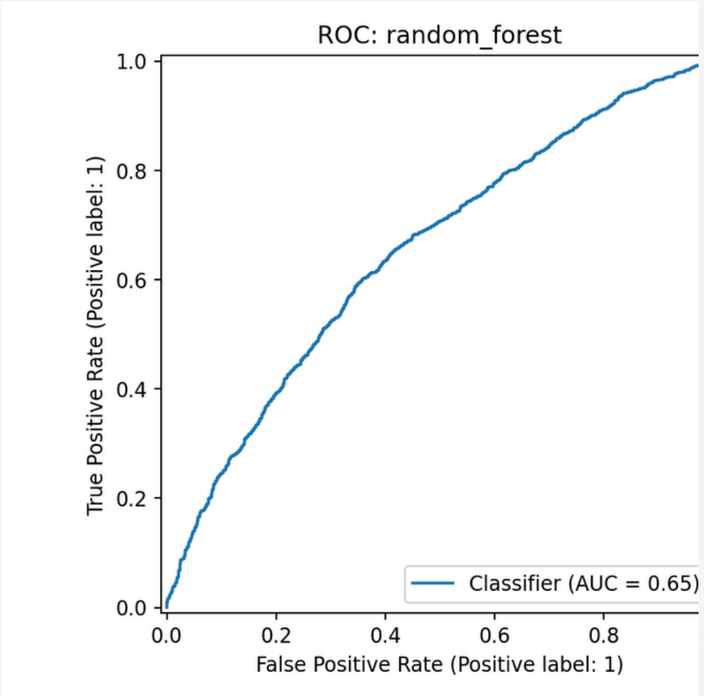
Random Forest was selected because it gave the best balance of prediction, probability quality, nonlinear modeling, and physiological interpretability.


Confusion matrix default: catboost
slow onset
success
282
794
386
2535
True label
slow onset
success
Predicted label
2500
2000
1500
1000
500


Confusion matrix tuned: random_forest
slow onset
success
661
415
1109
1812
True label
slow onset
success
Predicted label


Confusion matrix default: logistic_regression
slow onset
success
345
731
451
2470
True label
slow onset
success
Predicted label
2250
2000
1750
1500
1250
1000
750
500


Confusion matrix default: logistic_regression
slow onset
success
345
731
451
2470
True label
slow onset
success
Predicted label
2250
2000
1750
1500
1250
1000
750
500


ROC: random_forest
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Classifier (AUC = 0.65)
0.0
0.2
0.4
0.6
0.8
1.0

VALIDATION AGAINST BAD/RANDOM MODELS

- Because 73.08% of sleep episodes are quick-onset, a bad model can get high accuracy by always predicting “quick sleep.”
- But that model is useless for our actual goal because it detects 0% of slow-onset nights.

Model / Baseline	Accuracy	Balanced Accuracy	PR-AUC Slow Onset	Slow-Onset Recall	Slow-Onset F1	Interpretation
Always predict quick-onset	0.731	0.5	0.269	0	0	Looks accurate but detects no
Logistic Regression	0.704	0.583	0.41	0.321	0.443	Learns strongest slow-onset ranking
Random Forest	0.722	0.549	0.404	0.175	0.454	Strong probability quality and success prediction
CatBoost	0.705	0.565	0.385	0.262	0.45	Best calibrated probability model

A random model would have slow-onset PR-AUC around 0.269, equal to the slow-onset prevalence. Our models reach 0.385–0.41, giving a 39–45% relative improvement over random ranking.

THRESHOLD TUNING TURNS THE MODELS INTO SLOW-ONSET RISK DETECTORS

- The default 0.5 threshold is not ideal because the dataset is imbalanced.

Model	Threshold Type	Threshold	Balanced Accuracy	Slow-Onset Recall	Slow-Onset F1
Logistic Regression	Default	0.5	0.583	0.321	0.369
Logistic Regression	Tuned	0.599	0.594	0.6	0.441
Random Forest	Default	0.5	0.549	0.175	0.253
Random Forest	Tuned	0.639	0.593	0.651	0.447
CatBoost	Default	0.5	0.565	0.262	0.323
CatBoost	Tuned	0.632	0.582	0.54	0.421

The tuned thresholds ouput better metrics

BEST MODELS FOR EACH CASE

Logistic Regression is best for ranking:

- highest PR-AUC success: 0.8139
- highest PR-AUC slow onset: 0.41

Random Forest is best for deployment:

- highest tuned slow recall: 0.6506
- highest slow onset F1: 0.454
- lowest Brier score: 0.197

CatBoost is best for calibration:

- lowest ECE: 0.0818

Random Forest using Temporal + Previous Sleep History + HRV gives the best practical tradeoff for sleepability prediction on this wearable dataset.

FINAL TAKEAWAYS

- A 15-minute threshold provides a meaningful and learnable sleepability target.
- Temporal and previous sleep-history features provide the strongest behavioral foundation.
- HRV adds physiological readiness information and improves the model's alignment with wearable sleep literature.
- Logistic Regression served two roles: statistical interpretation and predictive baseline.
- Random Forest was selected as the final model because it captured nonlinear interactions and gave the best overall practical tradeoff.
- CatBoost and extended models did not improve enough to replace Random Forest.

predicted_probability_success
0.675413881
0.712282385
0.702795985
0.675529072
0.719083712
0.721152926
0.671368306
0.661708052
0.576526512
0.666206643
0.706370484
0.72289075
0.730275247
0.689129921
0.705593052
0.69292841
0.559900958
0.616899296
0.677214701
0.523477993
0.627678179
0.674099269
0.683880413
0.727838214
0.725961522

LIMITATIONS

Small Sample

Only 23 valid participants. External validity is limited; results may not generalise beyond this cohort.

Repeated Measures / Non-Independence

~174 episodes per participant. Also the reason why we rejected the idea of deep neural networks

Moderate Performance

ROC-AUC ~ 0.65. This is above chance but not clinical-grade accuracy.

Limited Diversity

The dataset was of students only in the year 2020. thus this is not enough to generalize.

This is a prototype model, not a clinically validated sleep recommendation system.

CONCLUSION

We built a leakage-controlled, participant-aware sleepability prediction pipeline.

The final model estimates:

- **P(SOL<15 minutes)**

using:

- **Temporal + Previous Sleep History + HRV**

FINAL STATEMENT

The model shows that sleepability is driven primarily by recent sleep behavior and bedtime timing, while HRV adds useful physiological context. The result is a wearable-aligned prototype for estimating bedtime success probability.

Thank You