Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders

Will Ke Wang KE.WANG064@DUKE.EDU
Jiamu Yang JERRY.YANG@DUKE.EDU
Leeor Hershkovich LEEOR.HERSHKOVICH@DUKE.EDU
Hayoung Jeong HAYOUNG.JEONG@DUKE.EDU
Bill Chen BILL.CHEN@DUKE.EDU
Karnika Singh KARNIKA.SINGH@DUKE.EDU
Ali R Roghanizad ALI.ROGANIZAD@DUKE.EDU
Md Mobashir Hasan Shandhi MOBASHIR.SHANDHI@DUKE.EDU
Andrew R Spector ANDREW.SPECTOR@DUKE.EDU
Jessilyn Dunn JESSILYN.DUNN@DUKE.EDU
Duke University, United States of America

Abstract

Sleep is crucial for health, and recent advances in wearable technology and machine learning offer promising methods for monitoring sleep outside the clinical setting. However, sleep tracking using wearables is challenging, particularly for those with irregular sleep patterns or sleep disorders.

In this study, we introduce a dataset collected from 100 patients from the Duke Sleep Disorders Center who wore an Empatica E4 smartwatch during an overnight sleep study with concurrent clinical-grade polysomnography (PSG) recording. This dataset encompasses diverse demographics and medical conditions. We further introduce a new methodology that addresses the limitations of existing modeling methods when applied on patients with sleep disorders. Namely, we address the inability of existing models to account for 1) temporal relationships while leveraging relatively small data, by introducing a LSTM post-processing method, and 2) group-wise characteristics that impact classification task performance (i.e., random effects), by ensembling mixed-effects boosted tree models. This approach was highly successful for sleep onset and wakefulness detection in this sleep disordered population, achieving an F1 score of 0.823 ± 0.019, an AUROC of 0.926 ± 0.016, and a 0.695 ± 0.025 Cohen's Kappa. Overall, we demonstrate the utility of both the data that we collected, as well as our unique approach to address the existing gap in wearable-based sleep tracking in sleep disordered populations.

Data and Code Availability This paper uses the DREAMT (**D**ataset for **R**eal-time sleep stage **E**stim**A**tion using **M**ultisensor wearable **T**echnology) dataset collected at the Duke Sleep Disorders Center and is made publicly available on PhysioNet. The code repository is available at github/DREAMT and the updated link to the dataset is available at PhysioNet/DREAMT.

Institutional Review Board (IRB) This study was IRB approved with IRB number: #Pro00108961

1. Introduction

1.1. Wearable technology for sleep tracking

Polysomnography (PSG) is the gold-standard for sleep assessment but is costly, labor-intensive, and unrepresentative of natural sleep settings (Jafari and Mohsenin (2010); Perez-Pozuelo et al. (2020)). Recently, alternative technologies such as wearable sensors have emerged for sleep tracking, offering reliability, user-friendliness, and accuracy (Perez-Pozuelo et al. (2020); Shaffer and Ginsberg (2017)). Wearables measure cardiorespiratory signals, movements, and skin temperature, offering a convenient, low-cost, objective sleep assessment in natural environments (Roebuck et al. (2014); Garbarino et al. (2014); de Zambotti et al. (2019)).

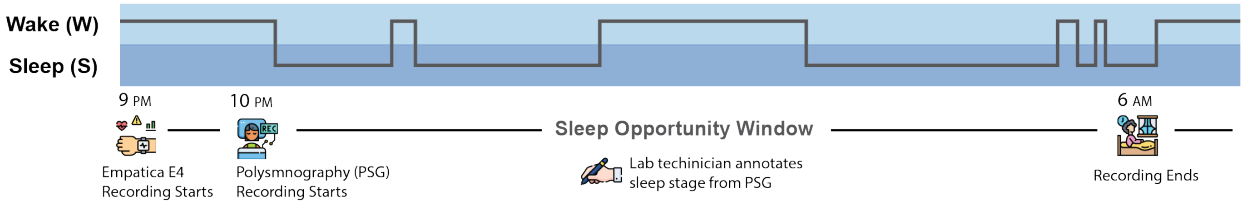


Figure 1: Our data collection process for the DREAMT dataset. Participants arrive around 9 PM, after which the researchers equipped them with the Empatica E4 on their wrists with their consent. Both PSG and E4 continued to collect data concurrently until the overnight study was completed and both PSG and E4 were taken off at the same time.

However, existing wearable-based sleep tracking algorithms often lack reliability in populations with sleep disorders, since most algorithms are developed using data from healthy adults with typical sleep patterns, excluding those with sleep disorders, leading to oversimplified models and inaccuracies in diverse populations (Imtiaz (2021); de Zambotti et al. (2019); Walch et al. (2019); Irish et al. (2015); Stenholm et al. (2019); Dorn et al. (2020); Dominguez et al. (2018); Erickson et al. (2022)). Recently, Pedro Fonseca et al., demonstrated the efficacy of neural network algorithms in sleep tracking with a dataset that consists of more than 1000 participants who have various sleep disorders (Fonseca et al. (2023)). However, the device, the dataset, and the algorithm remain proprietary to Philips, Inc, limiting broader accessibility and external validation of their proposed method. The scarcity of open-source datasets and algorithms for estimating sleep tracking from wearables exacerbates the challenge of developing universally effective models. Building models that cater to a broad population with varied sleep characteristics requires publicly accessible datasets that encompass various sleep profiles.

1.2. Existing Datasets

Typical open-source benchmark datasets for sleep tracking mainly consist of raw PSG signals sometimes combined with actigraphy (Supplementary Table 1). Yet, the environments for PSG collection poorly represent natural sleep settings. Moreover, actigraphy lacks crucial sleep-related physiological data like heart rate and skin temperature. This gap highlights the need for new, multi-modal wearable sensor datasets for developing accurate, real-world applicable sleep tracking models.

Our literature review reveals a scarcity of publicly available datasets incorporating wearable sensor data (Supplementary Table 2). These existing datasets often fail to represent diverse demographics, as they predominantly feature young, female, sleep-typical individuals without sleep disorders (Gao et al. (2021); Walch et al. (2019)). Even though the Multi-Ethnic Study of Atherosclerosis (MESA) dataset includes a broader age range and gender balance, it focuses on cardiovascular conditions, not sleep-specific disorders (Chen et al. (2015); Zhang et al. (2018)). Furthermore, many studies, such as the one by Borazio et al. Borazio et al. (2014), rely solely on a single type of sensor (i.e., accelerometer), omitting key biometrics such as heart rate and heart rate variability. These datasets' limitations extend to the lack of gold-standard sleep stage annotations, as seen in the ECSMP dataset, which, despite its comprehensive physiological signal collection via the Empatica E4 and EEG, does not include annotated sleep stages (Gao et al. (2021)).

2. Data Collection and Dataset Description

2.1. Research grade wearable device Empatica E4

The Empatica E4 wristband has been used for various clinical applications (Campanella et al. (2023); Kyriakou et al. (2019); Sevil et al. (2021)).

The E4 collects several signals such as blood volume pulse (BVP), accelerometry (ACC), electrodermal activity (EDA), and skin temperature (TEMP).

Table 1: Demographic and medical information from DREAMT
AHI = Apnea/Hypopnea Index,
OAHI: Obstructive Apnea/Hypopnea INdex,
SaO2: Arterial Oxygen Saturation.

		Age Group			Totall
		Young Adults [18-39]	Mid Adults [40-64]	Older Adults [≥65]	
Count, n		20	47	33	100
Gender, n	Male	5	21	19	45
	Female	15	26	14	55
Age (years)		32 ± 6	54 ± 8	74 ± 1	56 ± 17
BMI (kg/m^2)		33.9 ± 13.2	33.7 ± 7.3	30.1 ± 4.0	33.7 ± 8.6
Obesity, n (%)		16 (80%)	34 (72%)	18 (55%)	68 (68%)
Mean SaO_2 (%)		94.8 ± 4.7	94.0 ± 2.2	93.7 ± 3.1	94.1 ± 3.1
Arousal Index (count per hour)		33.6 ± 23.0	35.7 ± 26.5	43.2 ± 26.0	37.7 ± 25.7
AHI (count per hour)		26.6 ± 45.0	22.5 ± 24.5	18.9 ± 21.3	22.1 ± 28.7
OAHI (count per hour)		26.0 ± 44.4	20.1 ± 22.6	14.3 ± 19.4	19.4 ± 27.5

E4, using their proprietary algorithms, processes the light received from green and red light sensors to obtain the BVP signal, which is then further processed to obtain the interbeat interval (IBI) signal (emp (a)). Instantaneous HR (beats per minute) is then derived from the IBI signal. When IBI is unavailable, the HR is interpolated using the value from the latest available data point(emp (b)). BVP, IBI, and HR are all synchronized.

The derived measurements from BVP signals of E4 (i.e., the interbeat interval (IBI) and heart rate (HR)) have been compared against Holter monitors and ECG, which are the gold standards to evaluate heart rate and heart rate variability (Van Voorhees et al. (2022); Stuyck et al. (2022)). Specifically, high inter-device correlations and intraclass correlations (ICCs) were observed between E4 and Holter monitors for IBIs at 1-second and 50-minute intervals (Van Voorhees et al. (2022)). Compared to an electrocardiogram (ECG), the E4 validly estimated HR with intervals as short as 10s (Stuyck et al. (2022)). The accuracy and reliability of BVP, IBI, and HR estimations from E4 are affected by motion artifacts, proper wear, and the environmental condition of where the data is collected (e4s (2020)). However, these factors are less of a concern during sleep as participants would have less movement during sleep compared to daytime and there is less risk for the device to be disconnected or improperly worn (Böttcher et al. (2022)).

2.2. Data collection process

A total of 100 unique participants were recruited from the Duke Sleep Disorder Lab to participate in the study between May, 2022 and September, 2022 (Additional Study Details 7.5). Upon arrival at the sleep lab, each participant received comprehensive information about the study according to IRB approved procedures. Written informed consent was obtained for all study participants.

Each participant checked in at approximately 9 PM on the scheduled day of their appointment (Figure 1). The E4 wristband was placed on the participant's left wrist immediately after the consent form was signed at participant's arrival. The PSG data collection was started soon after. The recording continued throughout the night to monitor the participant's sleep condition and ended when the participant awakened around 6 AM the next morning naturally. The E4 device was also deactivated at the time of awakening.

From the participant's schedule and the PSG recordings, we recorded the corresponding sleep stage labels as "Wake" (W), "REM" (R), "Stage 1 Non-REM" (N1), "Stage 2 Non-REM" (N2), "Stage 3 Non-REM" (N3), and "Missing" (No stage labeled) annotated by the Sleep Disorder Lab's technicians (Figure 1). These sleep stage labels were time-aligned with the timestamps from the E4 data. In the published dataset, all timestamps are time-shifted.

We retrieved 100 recordings from 100 unique participants (45 male and 55 female) with ages ranging from 21 to 87 years. Clinical measurements relevant to understanding sleep behavior and/or disorders were taken for all participants as part of the protocol during the sleep study. These measurements include Body Mass Index (BMI), Obstructive Apnea-Hypopnea Index (OAHI), and Apnea-Hypopnea Index (AHI). In addition, we retrospectively recorded labels for sleep apnea events (including central apnea, hypopnea, and obstructive apnea). The average age was 56.2 ± 16.6 years, BMI: 33.7 ± 8.6 kg/m^2, OAHI: 19.4 ± 27.5 /h, AHI: 22.1 ± 28.7 /h (AHI < 5/h is healthy). Among all participants, 68 were obese or severely obese (BMI ≥ 30 kg/m^2). Among the 23 participants who had severe OSA (AHI > 30 /h), 17 were obese or severely obese. The numbers of participants with different apnea severity levels and different obesity categories are included in supplementary Figure 1.

In terms of medical history and sleep disorders, common sleep disorders or symptoms reported by the participants include snoring (n=40), excessive daytime sleepiness (n=34), sleep apnea (n=56), obstructive sleep apnea (n=33), restless sleep (e.g., restless leg syndrome) (n=23), difficulty breathing (e.g., gasping during sleep) (n=22). In the DREAMT dataset, for each participant, we include the most commonly observed comorbidities and sleep disorder diagnoses among this cohort.

2.3. Dataset Description

We coined our dataset as DREAMT (**D**ataset for **R**eal-time sleep stage **E**stim**A**tion using **M**ultisensor wearable **T**echnology).

Each sensor in the Empatica E4 employs a different sampling frequency (Empatica (2020)). Triaxial accelerometry (ACC) is sampled at 32 Hz, with each axis named ACC_X, ACC_Y, and ACC_Z. Blood volume pulse (BVP) derived from the photoplethysmography (PPG) sensor is sampled at 64 Hz. Both electrodermal activity (EDA) from the galvanic skin response sensor and skin temperature (TEMP) from the infrared thermopile sensor are sampled at 4 Hz. Heart rate (HR), estimated from the BVP signal, is reported every 1 second (i.e., 1 Hz). The technician-annotated sleep labels derived from PSG are recorded every 30 seconds. Overall, the time-aligned dataset consists of six raw E4 signals (BVP, ACC_X, ACC_Y, ACC_Z, EDA, TEMP), two derived signals (HR and IBI), the sleep-stage label (REM Sleep, Non-REM Sleep, Wake), and the true timestamp of every epoch (Figure 2). The PSG dataset will be available upon reasonable request, upon which the details of the PSG dataset can be provided. Further details of the E4 wearable dataset can be found in Appendix B.

Counting the 30-second epochs recorded, there are in total 8,636 REM epochs, 52,915 NREM epochs (stages 1-3 combined), and 20,334 wake epochs. Each individual, on average had (label percentage per participant) 12±6% REM, 64±14% NREM, and 25±17% Awake instances. (See Supplementary Figure 1)

2.4. Ethics statement

Study information was provided by the Duke Sleep Disorders Lab care team following the ethics protocols established by the DUHS IRB, including written informed consent (IRB #Pro00108961) with explicit permission to share de-identified data. In the publicly-available data, all direct identifiers are removed and all timestamps are time shifted to protect participant identities.

Our dataset aims to aid the efforts in developing, testing, and evaluating machine learning algorithms for real-time sleep tracking and sleep apnea event detection. We anticipate development of novel algorithms and validation of existing methods to follow with the release of DREAMT, especially for continuous sleep tracking and sleep disorder monitoring to improve sleep health.

Privacy is the major ethical concern of our data collection studies. We strictly follow the IRB rules to anonymize and protect participants' data. Anyone outside our core data collection group cannot access direct individually-identifiable information.

We also eliminated the data for users who stopped their participation at any time during the study. Since some sensitive sensor data (continuous biometric) can disclose identities, we do not provide the exact starting timestamp of each recording. We also do not provide participants' full medical history for privacy concerns.

2.5. Preprocessing

To synchronize all E4 data, we upsampled E4 signal channels that have less than 64 Hz frequency to match the highest sampling frequency (64 Hz from BVP) by repeating the value at each time point (last observation carried forward). The preprocessing and

feature extraction steps are explained in detail below, and can be referenced in our published code repo.

3-axis Accelerometer (ACC) We applied a fifth-order Butterworth band-pass filter (3-11 Hz), following the same parameters reported by Oura (Altini and Kinnunen (2021)). We then extracted statistical features such as the trimmed mean (10% on each side removed), max, and interquartile range (IQR) of each axis in successive 30-second windows from the absolute value of the filtered signal. From the raw ACC signal, we calculated the Mean Absolute Deviation (MAD) of each axis, which was based on the deviation from the vector magnitude of the current epoch. The trimmed mean, max, and IQR were also aggregated every 30 seconds.

Skin Temperature (TEMP) We first applied winsorizing to the skin temperature values, clipping temperature values to within 31-40 °C. The mean, min, max, and standard deviation were then extracted as statistical features. The winsorization only corrected 10% of the recorded temperature values.

Blood Volume Pressure (BVP) We applied Oura's PPG preprocessing methods to BVP, including a Chebyshev type II bandpass filter (0.5-20 Hz) for noise reduction (Altini and Kinnunen (2021)). Then, using the NeuroKit2 Python package (Makowski et al. (2021)), we extracted HRV metrics such as the root mean square of successive RR interval differences (rMSSD), standard deviation of NN intervals (SDNN), percentage of successive RR intervals that differ by more than 50 ms (pNN50), power in the low frequency bands (LF: 0.05 $\sim$ 0.15 Hz) and high frequency bands (HF: 0.15 $\sim$ 0.4 Hz) bands, the main frequency peak in the LF and HF bands, total power, normalized power, and breathing rate. These particular spectral divisions (LF and HF bands) were chosen because physiological mechanisms related to HRV manifest themselves within these bands (Shaffer and Ginsberg (2017)).

Electrodermal Activity (EDA) We implemented the preprocessing techniques introduced in Anusha et al. (2022) to preprocess the raw EDA signal. We detrended the signal by fitting a least-squares regression line on each 5-second segment and subtracting the fitted regression line from each segment. A Butterworth low-pass filter was applied to the detrended EDA to remove any high-frequency noise. We decomposed the preprocessed signal into the tonic skin conductance level (SCL) and the phasic skin conductance response (SCR) (Makowski et al. (2021)) using the Python functions provided by NeuroKit2 (Makowski et al. (2021)). After decomposition, we extracted the relevant SCR features, including the parameters of the peak (height and amplitude) and the temporal characteristics of the peak (time to reach the peak amplitude and time to recover from the peak amplitude). We then calculated the mean and maximum of these features in each epoch.

Signal Quality Based on Moscato et al. (Moscato et al. (2022))), we defined our signal quality assessment based on the following criteria. A 30-second segment was deemed as an artifact if an extreme activity is detected (Activity Index larger than 0.4125, Moscato et al. (2022)), if the BVP raw signal we obtained is outside of the normal range (-500 to 500), or if the signl-to-noise (SNR) ratio is smaller than 10 dB. The frequency range of interest of BVP is defined to be from 0.5 Hz to 20 Hz.

Feature Engineering Initially, feature engineering was conducted for each 30-second epoch, the details of which can be found in our github repo. The exact features used can be found in our code repo. In addition to the original extracted features, we also add further processed features. For every feature derived from an epoch, we apply Gaussian filters temporally for each participant's entire night of data for each extracted feature. An example of how this Gaussian filter is applied is demonstrated in Figure 3. Additionally, the temporal derivative of the Gaussian smoothed feature was computed. Lastly, we incorporated the variance of this feature, calculated over a moving window, to enhance the robustness of our feature set.

2.6. Mixed Effects Boosted Trees

Currently existing sleep tracking models do not account for intra-individual characteristics of sleep, which reduces their performance in real world deployment. To address this gap, we introduce mixed-effects modeling with a Gaussian process in a boosted tree algorithm. (Sigrist (2020) The boosted tree algorithm we chose is Light Gradient Boosting Machine (LightGBM) (Altini and Kinnunen (2021)) due to its efficiency with large datasets and its high performance in terms of training speed and accuracy.

Gaussian Process (GP) Boosting (Sigrist (2020)), extends the gradient boosting framework by incorporating Gaussian processes, allowing the model to

capture complex, nonlinear relationships in the data, while relaxing on the independence assumption. This approach is useful when the dataset naturally falls into certain groups that have differing characteristics.

By including obesity or apnea severity as random effects, our model can account for individual variations that are not explained by the fixed effects. This is particularly important in sleep studies where individual physiological characteristics can significantly impact sleep patterns. Random effects allow for the modeling of these individual differences, enhancing the performance of the sleep tracking algorithm in a diverse population.

2.7. Time series classification post-processing

Our post-processing method aimed to learn the behaviors of sleep state transitions through time, using the probability outputs and most important features identified from the epoch-by-epoch classifier. Previous efforts surrounding post-processing sleep epochs for sleep staging algorithm development have used Hidden Markov Models (HMM) or rule-based methods (Fedorin et al. (2019); Trinh and Ha (2022)). HMM has difficulty capturing long-term dependencies which are likely to be important in sleep tracking, given that past states affect current and future states. We therefore implemented a shallow Long Short-Term Memory (LSTM) as our post-processing method. LSTM is a type of Recurrent Neural Network (RNN) designed to avoid the long-term dependency problem in RNNs, making them more effective at learning from sequences of data where the context spans over long intervals (Hochreiter and Schmidhuber (1997)). Traditional RNNs struggle with long-range dependencies due to the vanishing gradient problem, where gradients become exceedingly small during back-propagation, resulting in ineffective learning. LSTMs mitigate this issue with their gated structure, making them adept at capturing relationships in data over longer periods. We propose to use a single-layer LSTM to perform sequence-sequence classification as a post-processing method. This shallow LSTM module prevents overfitting to a dataset small in participant number while still learning temporal information, where each participant is considered to have one time series data overnight. Our method offers a viable solution for enhancing predictive performance from time series data for small datasets.

3. Methods

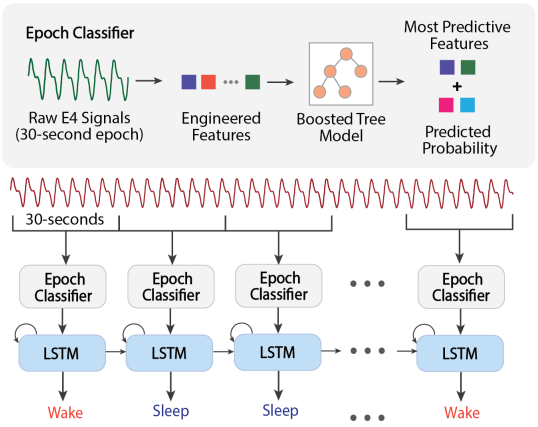


Figure 2: Representation of our model pipeline. *Epoch Classifier:* The raw wearable signals from Empatica E4 are preprocessed and the engineered features are fed into boosted tree models. *Post-processing module:* The negative (sleep) and positive (wake) output probabilities and the most predictive features are collected as inputs to a shallow LSTM model to predict sleep vs wake labels through time.

3.1. Experiments

We next performed a highly conservative quality control step, removing any participants with artifacts in >20% of their entire night's epochs to ensure that all datasets used in the subsequent analysis were of optimal quality. 80 participants remained.

We treat wake epochs as positive labels because 1) the wake epochs during the sleep opportunity window are notably less numerous than sleep epochs and 2) detecting wake epochs during sleep opportunity window has important clinical implications. For all experiments listed below, hyperparameter tuning for all models was performed on the training set using HyperOpt (Bergstra et al. (2013)) python package, with the validation set used for early stopping. We used SMOTE from Python's Imbalance Learn package (Lemaître et al. (2017)) to balance the dataset for training, but we did not use SMOTE to balance the validation set nor the testing set during model

Table 2: Experiment Results

Random Effects	LSTM post-processing	F1 Score	AUROC	AUPRC	Accuracy	Cohen's Kappa
None	No	0.777 ± 0.009	0.895 ± 0.007	0.885 ± 0.008	0.816 ± 0.008	0.605 ± 0.024
Obesity	No	0.785 ± 0.020	0.902 ± 0.015	0.891 ± 0.015	0.825 ± 0.013	0.622 ± 0.023
Apnea Severity	No	0.782 ± 0.015	0.898 ± 0.016	0.886 ± 0.017	0.826 ± 0.010	0.623 ± 0.026
None	Yes	0.805 ± 0.025	0.915 ± 0.019	0.906 ± 0.018	0.836 ± 0.018	0.649 ± 0.023
Obesity	Yes	0.822 ± 0.019	0.926 ± 0.011	0.914 ± 0.008	0.853 ± 0.012	0.683 ± 0.028
Apnea Severity	Yes	**0.823 ± 0.019**	**0.926 ± 0.016**	**0.915 ± 0.020**	**0.857 ± 0.016**	**0.694 ± 0.025**

tuning and evaluation. We report on the model performances in terms of accuracy, F1-score, Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Cohen's Kappa, all respective to the positive class (wake).

We include the boosted tree's predicted probabilities for the negative and positive labels (sleep and wake, respectively), as well as the two strongest predictors as the inputs into the LSTM post-processing step. The two most predictive features from GPBoost were the accelerometry index (ACC_INDEX, Moscato et al. (2022)) and Higuchi Fractal Dimension (HRV_HFD, Makowski et al. (2021)). The rationale for including two of the strongest predictors is to ensure that the post-processing LSTM module is aware of the context on which the predicted probabilities are based while building a parsimonious model. We selected two strongest predictors, one from BVP signals and one from ACC signals since these two signals are deemed important for sleep tracking. To select the two features, we conducted a separate train-test experiment using the baseline LightGBM model prior to the cross-validation step. In this experiment, we randomly selected 56 participants (70%) to be in the training set, 8 participants (10%) in the validation set, and 16 participants (20%) in the hold-out testing set. After analyzing the Shapley values on the training data using this LightGBM model, we found the most predictive features to be ACC_INDEX and HRV_HFD.

We performed 5-fold cross-validation at participant level, where every fold has data from 16 participants, and the rest of the 64 participants were divided into 56 training participants and 8 validation participants.

1. LightGBM
2. GPBoost + Obesity random effect
3. GPBoost + Apnea Severity random effect
4. LightGBM + LSTM post-processing
5. GPBoost + Obesity random effect + LSTM post-processing
6. GPBoost + Apnea Severity random effect + LSTM post-processing

A note about missing sleep stage annotations Sleep stage is labeled as missing when the gold standard PSG was missing or sleep annotations were not made. The epochs labeled as missing were omitted from the mixed effects modeling directly. Significant missingness has been found in only two participants, who had their PSG re-setup during the overnight study, which resulted in 15 minutes of consecutive missing labels each. We also found one epoch with missing label each in four other participants. Due to the extremely low occurrence of the "missing" label, we treated the epochs before and after as adjacent to each other, while omitting the epoch(s) with "missing" label, when performing LSTM post-processing.

4. Results and Discussion

In the models we tested, the best performing algorithm was the mixed-effects GPBoost using apnea severity as the random effect, and combined with LSTM for signal post-processing. This model was able to improve upon the baseline LightGBM model and achieve F1-scores of 0.823 ± 0.019, accuracy scores of 0.857 ± 0.016, AUROC scores of 0.926 ± 0.016, AUPRC scores of 0.915 ± 0.020, and Cohen's Kappa scores of 0.694 ± 0.025, beating out a model with Obesity as the random effect slightly.

As seen in Table 2, the combination of LightGBM and mixed-effects modeling using Gaussian processes can potentially lead to higher accuracy and better performance overall. LightGBM efficiently handles various types of data, while Gaussian processes capture complex patterns and relationships. Using this approach of boosted tree with mixed-effects modeling for sleep tracking allows researchers to collect datasets coming from a heterogeneous population and develop algorithms while taking this heterogeneity into account for the model building. Incorporating obesity or apnea severity as random effects enables the model to make semi-personalized predictions, leading to a fairer approach that accounts for individual physiological differences, making our algorithm robust to diverse sleep behaviors and physiological characteristics. A notable observation is that using either Apnea Severity or Obesity as the random effect results in similar performance improvements, which is unsurprisingly, given the potentially strong correlation between participants with moderate or severe obesity developing sleep apnea.

LSTM post-processing enables sequence-to-sequence classification, accounting for the temporal patterns in sleep states even when using a small dataset, which is common in clinical research due to the cost and effort associated with data collection. LSTM post-processing adds the use of temporal information inspired by deep neural networks without having to resort to large databases.

We also conducted the same set of experiments on only the participants without apnea (n=22), and the results are summarized in Supplementary Table 3. We found that the LSTM post-processing step still improved on the algorithm performance but adding random effects in the modeling process did not, which corroborated our expectation.

In Figure 3, we plot the example raw signals against true sleep and wake labels as well as the predicted labels from GPBoost (adding obesity as the random effects) with and without LSTM post-processing. We chose ACC_INDEX and HRV_MinNN because these features are easy to interpret physiologically. ACC_INDEX is a directly measurement of how much wrist activity there is in a 30-second epoch, and HRV_MinNN means the mininum NN intervals, the intervals between heart beats as defined by adjacent QRS complexes (usually, R-R peaks) in the electrocardiogram, corresponding inversely to the quickest instantenous heart beat. As can be seen comparing the third and fourth panels in Figure 3, the LSTM post-processing method improves the GPBoost predictions by taking into account longer-range temporal dynamics. Specifically, the LSTM post-processing method is able to correct for false positives and false negatives by learning that wake epochs are more likely to appear at the start of and near the end of the sleep opportunity window, while wake episodes are expected to be more sporadic in the middle of the sleep opportunity window.

Our work is unique in its introduction of a novel dataset containing high resolution, research-grade wearable data from patients with varying degrees of sleep disorders as well as sleep stages and sleep apnea annotations based on gold-standard PSG, including detailed sleep stage annotations and apnea event labels by expert sleep technicians. Our approach is also unique and useful in that we are the first to use Gaussian Process Boosted Tree Models to model random effects in the application of wearable sleep tracking, and that we introduce a new post-processing method based on LSTM to learn temporal information from the outputs of the trained Boosted Tree Models.

4.1. Limitations

In this study, we focused on two random effects, obesity and apnea severity, to demonstrate the utility of our mixed-effects modeling approach. However, this dataset contains numerous variables of interest for further exploration, such as mental health conditions, cardiovascular illness, or diagnosis of a sleep disorder other than sleep apnea. Our study here is not a comprehensive analysis of the entire dataset. Further research should consider exploring using different mixed-effects to potentially achieve better sleep tracking performance.

Another limitation of our study is that our feature extraction pipeline is not data-driven, but rather it is domain-driven. We engineered the features to be

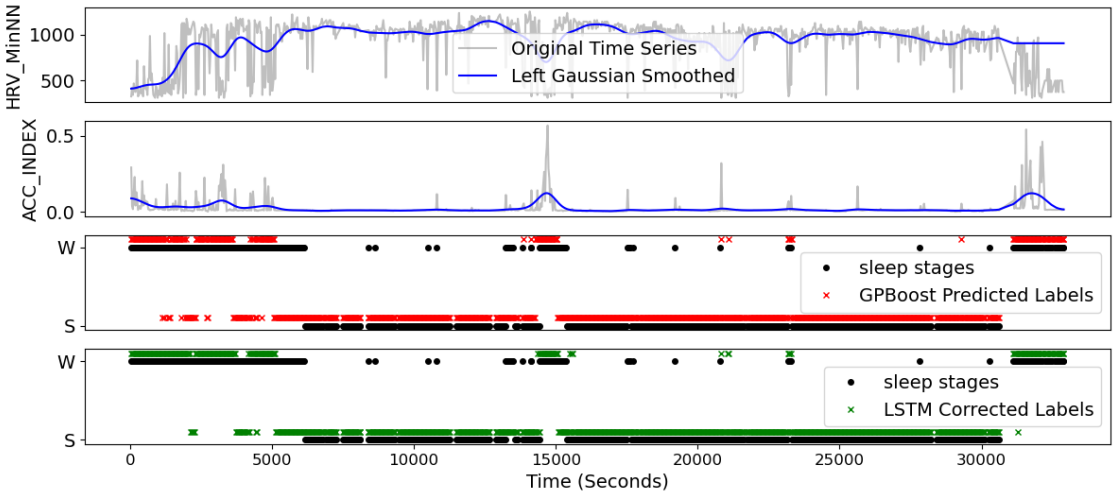


Figure 3: Example output from a validation participant after GPBoost modeling and LSTM post-processing. The top two panels plot the two example features, HRV_MinNN, and ACC_INDEX, with grey-colored original values and blue-colored gaussian smoothed values. The third panel shows the true sleep vs wake labels in black and GPBoost predicted labels in red, while the fourth panel shows the true sleep vs wake labels in black and the corrected labels using LSTM post-processing in green.

used in the modeling tasks based on domain knowledge and exploratory data visualization. However, due to the large number of epochs we have in the dataset, it is possible to use deep learning models or other complex time series models to automate the process of feature extraction that can be used for further prediction.

Contrary to other existing open source datasets, (see Supplementary Table 2), we do not have a large number of participants who present normal sleep behaviors (i.e., without any sleep disorder diagnosis). Because these data were collected on real patients who had some type of sleep abnormality that led to their clinical assessment in the sleep lab as a part of their usual care, all 100 participants in this dataset are expected to have some type of sleep disorder. There are 26 participants who do not have sleep apnea, but they do have other conditions including depression, anxiety, hypertension, or asthma that are expected to affect sleep. In future research, we plan to focus on augmenting this dataset by including healthy adults.

Lastly, every participant started their sleep in the clinic around 10 PM, woke up around 6 AM, and slept in the conditions of the sleep clinic; these conditions may not necessarily correspond with each participant's regular sleep schedule or habits. The differences between the conditions under which sleep was assessed in the clinic vs how it takes place in the real world might hinder our method's performance for sleep tracking were it to be deployed outside of the sleep clinic setting. However, as no gold-standard method yet exists to obtain ground truth sleep labels outside of a clinical setting, validated real-world sleep tracking remains out of reach.

5. Conclusion

In this paper we introduce a novel dataset consisting of high-resolution research-grade wearable device data from 100 participants who have varying degrees of sleep disorders, with sleep stages and sleep apnea annotations based on PSG. To our knowledge, our study is the first to release a complete dataset for sleep tracking and staging using a publicly available research-grade wearable device from such a large and diverse population. We are also making available our code available for loading, preprocessing and analyzing our dataset in full. This dataset can provide a benchmark for future sleep tracking and sleep staging algorithm development, especially for a population

with sleep disorders, promoting greater equity and generalizability in the potential application of sleep tracking.

We also introduce two novel methods of sleep tracking algorithm development that are specifically designed for this dataset: mixed-effects modeling in a boosted tree model (GPBoost) and LSTM-based post-processing. Both methods show improvement in sleep vs wake detection compared to the baseline LightGBM model. The GPBoost model adds generalizability by learning the global context while taking into consideration that there are different physiological subgroups within the population. The LSTM post-processing method uses a very shallow LSTM layer and adds robustness to the base model by learning temporal information, which is especially beneficial when training with relatively small datasets.

References

Utilizing the PPG/BVP signal, a. URL https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal.

E4 data - IBI expected signal – Empatica Support, b. URL https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal.

Decoding wearable sensor signals - what to expect from your E4 Data | Research | Blog | Empatica, 2020. URL https://www.empatica.com/blog/decoding-wearable-sensor/-signals-what-to-expect-from-your-e4-data.html.

Marco Altini and Hannu Kinnunen. The Promise of Sleep: A Multi-Sensor Approach for Accurate Sleep Stage Detection Using the Oura Ring. *Sensors*, 21(13):4302, January 2021. ISSN 1424-8220. doi: 10.3390/s21134302. URL https://www.mdpi.com/1424-8220/21/13/4302. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

A. S. Anusha, S. P. Preejith, Tony J. Akl, and Mohanasankar Sivaprakasam. Electrodermal activity based autonomic sleep staging using wrist wearable. *Biomedical Signal Processing and Control*, 75:103562, May 2022. ISSN 1746-8094. doi: 10.1016/j.bspc.2022.103562. URL https://www.sciencedirect.com/science/article/pii/S1746809422000842.

James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123. PMLR, February 2013. URL https://proceedings.mlr.press/v28/bergstra13.html. ISSN: 1938-7228.

Marko Borazio, Eugen Berlin, Nagihan Kucukyildiz, Philipp Scholl, and Kristof Van Laerhoven. Towards Benchmarked Sleep Detection with Wrist-Worn Sensing Units. In *2014 IEEE International Conference on Healthcare Informatics*, pages 125–134, September 2014. doi: 10.1109/ICHI.2014.24.

Sebastian Böttcher, Solveig Vieluf, Elisa Bruno, Boney Joseph, Nino Epitashvili, Andrea Biondi, Nicolas Zabler, Martin Glasstetter, Matthias Dümpelmann, Kristof Van Laerhoven, Mona Nasseri, Benjamin H. Brinkman, Mark P. Richardson, Andreas Schulze-Bonhage, and Tobias Loddenkemper. Data quality evaluation in wearable monitoring. *Scientific Reports*, 12(1):21412, December 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-25949-x. URL https://www.nature.com/articles/s41598-022-25949-x. Number: 1 Publisher: Nature Publishing Group.

Sara Campanella, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors (Basel, Switzerland)*, 23(7):3565, March 2023. ISSN 1424-8220. doi: 10.3390/s23073565. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10098696/.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, June 2015. ISSN 1550-9109. doi: 10.5665/sleep.4732.

Massimiliano de Zambotti, Nicola Cellini, Aimée Goldstone, Ian M. Colrain, and Fiona C. Baker. Wearable Sleep Technology in Clinical and Research Settings. *Medicine and Science in Sports and Exercise*, 51(7):1538–1557, July 2019. ISSN 1530-0315. doi: 10.1249/MSS.0000000000001947.

Jennifer E. Dominguez, Chad A. Grotegut, Mary Cooter, Andrew D. Krystal, and Ashraf S. Habib. Screening extremely obese pregnant women for obstructive sleep apnea. *American Journal of Obstetrics and Gynecology*, 219(6):613.e1–613.e10, December 2018. ISSN 1097-6868. doi: 10.1016/j.ajog.2018.09.001.

Aaron van Dorn, Rebecca E. Cooney, and Miriam L. Sabin. COVID-19 exacerbating inequalities in the US. *Lancet (London, England)*, 395(10232):1243–1244, April 2020. ISSN 1474-547X. doi: 10.1016/S0140-6736(20)30893-X.

Inc. Empatica. E4 data - BVP expected signal, January 2020. URL https://support.empatica.com/hc/en-us/articles/360029719792-E4-data-BVP-expected-signal.

Melissa L. Erickson, Will Wang, Julie Counts, Leanne M. Redman, Daniel Parker, Janet L. Huebner, Jessilyn Dunn, and William E. Kraus. Field-Based Assessments of Behavioral Patterns During Shiftwork in Police Academy Trainees Using Wearable Technology. *Journal of Biological Rhythms*, 37(3):260–271, June 2022. ISSN 0748-7304. doi: 10.1177/07487304221087068. URL https://doi.org/10.1177/07487304221087068. Publisher: SAGE Publications Inc.

Illia Fedorin, Kostyantyn Slyusarenko, Wonkyu Lee, and Nataliya Sakhnenko. Sleep Stages Classification in a Healthy People Based on Optical Plethysmography and Accelerometer Signals via Wearable Devices. In *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 1201–1204, Lviv, Ukraine, July 2019. IEEE. ISBN 978-1-72813-882-4. doi: 10.1109/UKRCON.2019.8879875. URL https://ieeexplore.ieee.org/document/8879875/.

Pedro Fonseca, Marco Ross, Andreas Cerny, Peter Anderer, Fokke van Meulen, Hennie Janssen, Angelique Pijpers, Sylvie Dujardin, Pauline van Hirtum, Merel van Gilst, and Sebastiaan Overeem. A computationally efficient algorithm for wearable sleep staging in clinical populations. *Scientific Reports*, 13:9182, June 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-36444-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10244431/.

Zhilin Gao, Xingran Cui, Wang Wan, Wenming Zheng, and Zhongze Gu. ECSMP: A dataset on emotion, cognition, sleep, and multi-model physiological signals. *Data in Brief*, 39:107660, December 2021. ISSN 2352-3409. doi: 10.1016/j.dib.2021.107660. URL https://www.sciencedirect.com/science/article/pii/S2352340921009355.

Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W. Picard, and Simone Tognetti. Empatica E3 — A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, pages 39–42, November 2014. doi: 10.1109/MOBIHEALTH.2014.7015904.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Syed Anas Imtiaz. A Systematic Review of Sensing Technologies for Wearable Sleep Staging. *Sensors*, 21(5):1562, January 2021. ISSN 1424-8220. doi: 10.3390/s21051562. URL https://www.mdpi.com/1424-8220/21/5/1562. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Leah A. Irish, Christopher E. Kline, Heather E. Gunn, Daniel J. Buysse, and Martica H. Hall. The Role of Sleep Hygiene in Promoting Public Health: A Review of Empirical Evidence. *Sleep medicine reviews*, 22:23–36, August 2015. ISSN 1087-0792. doi: 10.1016/j.smrv.2014.10.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4400203/.

Behrouz Jafari and Vahid Mohsenin. Polysomnography. *Clinics in Chest Medicine*, 31(2):287–297, June 2010. ISSN 1557-8216. doi: 10.1016/j.ccm.2010.02.005.

Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank Wilhelm, Tess Osborne, and Jessica Pykett. Detecting Moments of Stress from Measurements of Wearable Physiological Sensors. *Sensors (Basel, Switzerland)*, 19(17):3805, September 2019. ISSN 1424-8220. doi: 10.3390/

s19173805. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6749249/.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365.html.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, August 2021. ISSN 1554-3528. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758/s13428-020-01516-y.

Serena Moscato, Stella Lo Giudice, Giulia Massaro, and Lorenzo Chiari. Wrist Photoplethysmography Signal Quality Assessment for Reliable Heart Rate Estimate and Morphological Analysis. *Sensors (Basel, Switzerland)*, 22(15):5831, August 2022. ISSN 1424-8220. doi: 10.3390/s22155831. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9370973/.

Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M. Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital Medicine*, 3(1):1–15, March 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0244-4. URL https://www.nature.com/articles/s41746-020-0244-4. Number: 1 Publisher: Nature Publishing Group.

A Roebuck, V Monasterio, E Gederi, M Osipov, J Behar, A Malhotra, T Penzel, and GD Clifford. A review of signals used in sleep analysis. *Physiological measurement*, 35(1):R1–57, January 2014. ISSN 0967-3334. doi: 10.1088/0967-3334/35/1/R1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4024062/.

Mert Sevil, Mudassir Rashid, Iman Hajizadeh, Minsun Park, Laurie Quinn, and Ali Cinar. Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management. *IEEE Transactions on Biomedical Engineering*, 68 (7):2251–2260, July 2021. ISSN 1558-2531. doi: 10.1109/TBME.2020.3049109. Conference Name: IEEE Transactions on Biomedical Engineering.

Fred Shaffer and J. P. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, 5:258, September 2017. ISSN 2296-2565. doi: 10.3389/fpubh.2017.00258. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/.

Fabio Sigrist. Gaussian process boosting. *CoRR*, abs/2004.02653, 2020. URL https://arxiv.org/abs/2004.02653.

Sari Stenholm, Jenny Head, Mika Kivimäki, Linda L. Magnusson Hanson, Jaana Pentti, Naja H. Rod, Alice J. Clark, Tuula Oksanen, Hugo Westerlund, and Jussi Vahtera. Sleep Duration and Sleep Disturbances as Predictors of Healthy and Chronic Disease-Free Life Expectancy Between Ages 50 and 75: A Pooled Analysis of Three Cohorts. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 74(2):204–210, January 2019. ISSN 1758-535X. doi: 10.1093/gerona/gly016.

Hans Stuyck, Leonardo Dalla Costa, Axel Cleeremans, and Eva Van den Bussche. Validity of the Empatica E4 wristband to estimate resting-state heart rate variability in a lab-based context. *International Journal of Psychophysiology*, 182:105–118, December 2022. ISSN 0167-8760. doi: 10.1016/j.ijpsycho.2022.10.003. URL https://www.sciencedirect.com/science/article/pii/S0167876022002409.

Linh Trinh and Bach Ha. An incorporation of deep temporal convolutional networks with hidden markov models post-processing for sensor-based human activity recognition. In *The 11th International Symposium on Information and Communication Technology*, pages 96–102, Hanoi Vietnam, December 2022. ACM. ISBN 978-1-4503-9725-4. doi: 10.1145/3568562.3568610. URL https://dl.acm.org/doi/10.1145/3568562.3568610.

Elizabeth E. Van Voorhees, Paul A. Dennis, Lana L. Watkins, Tapan A. Patel, Patrick S. Calhoun, Michelle F. Dennis, and Jean C. Beckham. Ambulatory Heart Rate Variability Monitoring: Comparisons Between the Empatica E4 Wristband and Holter Electrocardiogram. *Psychosomatic*

Medicine, 84(2):210–214, March 2022. ISSN 1534-7796. doi: 10.1097/PSY.0000000000001010.

Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12):zsz180, December 2019. ISSN 0161-8105. doi: 10.1093/sleep/zsz180. URL https://doi.org/10.1093/sleep/zsz180.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10): 1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

Appendix A. Additional Study Details

A.1. Details of Literature Review

Supplementary Table 1: Public Datasets that include PSG and sleep stage labels for sleep tracking algorithm development.

Dataset	Year	No. participants	Length of study
MIT-BIH Polysomnographic Database	1999	16	One night
2018 PhysioNet/CinC Challenge	2018	1,985	One night
Sleep-EDF Database Expanded	2013	100	Two nights
Sleep Heart Health Study (SHHS)	2003	6,441	One (n=3,146) or two nights (n= 3,295)
MASS	2014	200	One night

Supplementary Table 2: Open-sourced datasets that include wearable sensors and PSG for sleep-staging algorithm development. All datasets were collected over a single night per participant.

Dataset	**Device**	**Sleep stage label**	**No. participants (% Male)**	**Mean Age, Years (STD)**	**Health Conditions**
Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography (2019)	Apple Watch	Yes	31 (32%)	29.42 (8.52)	Healthy
ECSMP (2021)	Empatica E4	No	67 (36%)	23.82 (1.93)	Healthy
MESA (2012)	Actigraphy	Yes	2,040 (46%)	68.0 (13.0)	Overweight (mean BMI: 27.9±7.3 kg/m2), cardiovascular diseases
Towards a Benchmark for Wearable Sleep Analysis with Inertial Wrist-worn Sensing Units (2014)	Custom-built 3D accelerometer	Yes	42 (52%)	Not reported; Range 24 - 86	Sleep disorders (Insomnia, narcolepsy, sleep apnea syndrome, restless leg syndrome)

The datasets listed in Supplementary Table 1 are the benchmark datasets used for sleep tracking. These datasets only provide PSG recordings and sleep stage labels.

Supplementary Table 2 summarizes the existing datasets containing raw or processed signals from wearable sensors or devices. To our knowledge, no E4 datasets exist in the public domain that match our total participant number, total hours of data recorded, or gold-standard sleep stage labels at high resolution (every 30 seconds).

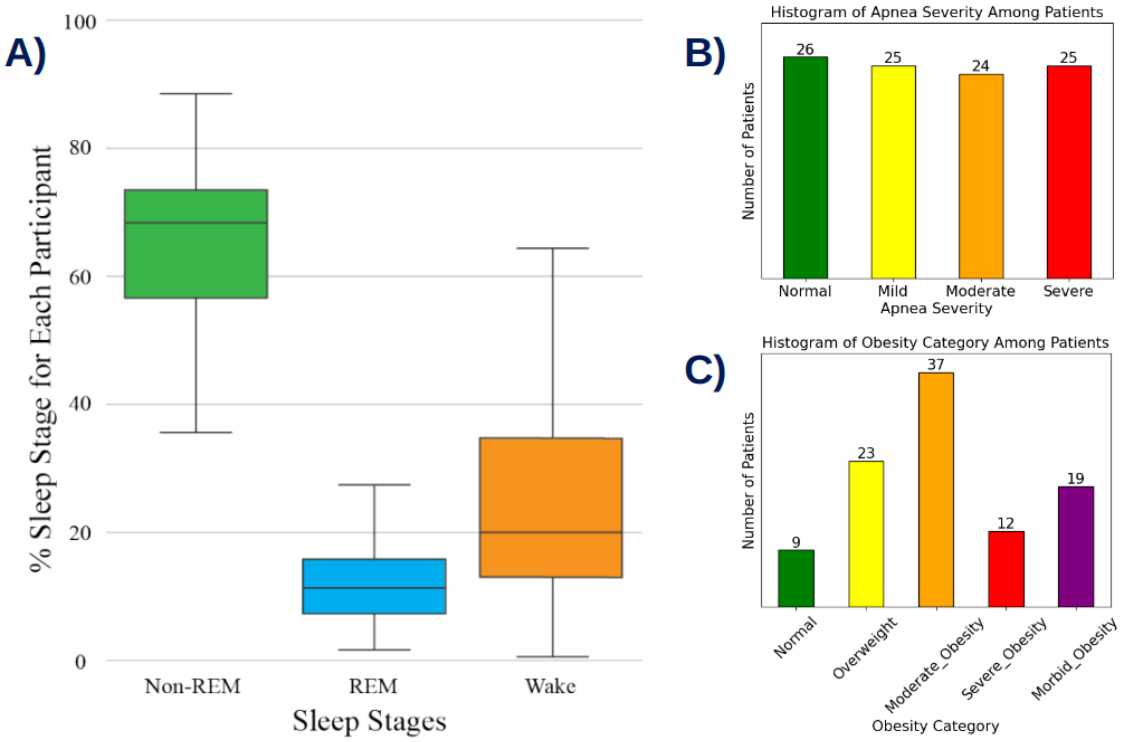
A.2. Study protocol details

The sleep study protocol at Duke Sleep Disorders Center is designed to detect and monitor patients' apnea events during sleep. The protocol requires 6 hours of sleep data to identify any abnormal sleep behavior, and all patients follow a 10PM - 6AM sleep schedule. Patients are asked to abstain from caffeine after 12 PM on the day of PSG recording to ensure standard protocol. There is no explicit monitoring of caffeine intake on the arrival of the participant in the sleep lab.

Each participant stayed in a hotel-like room for a night at the sleep lab, in different rooms on the same floor. After they checked in for their study at the sleep lab, the clinician would ask for their willingness to join our data collection. Due to the limited number of devices we possessed, we were only able to recruit at

most 4 participants each night. There were no selection criteria for the potential participants since we were aiming for people with sleep apnea. The collected wearable data and clinician-labeled sleep stage were joined together by the datetime of the data point and actual time information was replaced by the duration of the data point to the time 0 (first data point) of the data. The purpose of the time shifting was to preserve privacy. The released data used resampled data. Each participant's wearable data and wearable data were joined together in a dataframe and data was resampled to the highest sampling frequency (64Hz) in the wearable data. All the data were resampled to match the highest sampling frequency and this resampled, time-shifted dataframe was released.

During each study, overnight polysomnography (PSG) is performed using the Nihon Khoden Polysmith (version 11) Data Management System (DMS). A total of sixteen channels were recorded. Six EEG channels are recorded using Grass disc electrodes placed according to the standard 10/20 electrode placement system to assess sleep stages (C4-M1, F4-M1, O2-M1, Fp1-O2, T3-Cz, Cz-T4). Airflow is measured using an oronasal thermal sensor and a nasal air pressure transducer microphone for snoring. Thoracic and abdominal respiratory effort is measured using inductance plethysmography. Axial EMG activity is recorded from the mentalis muscle. Leg movements were recorded using 3M adhesive red dot electrodes placed over the tibialis anterior muscles of both legs. Electrooculogram and electrocardiogram tracings were recorded. The oxygen saturation is recorded using a finger probe connected to the Nihon Khoden Polysmith DMS. The amplifiers and other hardware vary in age from 2006-2022.



Supplementary Figure 1: A) Boxplot of percentage epochs for sleep stages NREM, REM and Wake for each participant. B) Histogram of number of participants with different apnea severity levels. C) Histogram of number of participants with different obesity categories

A.3. Study Demographics

In addition to the demographic characteristics we described in the main text, participants' key conditions and relevant medical histories can be found in our published dataset. Supplementary Figures Figure 1 show histograms of the percentages of NREM sleep, REM sleep and wake epochs for each participant, the numbers of participants in different apnea severity levels, and the number of participants in different obesity or weight categories.

A.4. Additional Experimental details

Details of Computing Resources The GPUs used to accelerate deep learning model trainings are RTX 4090 24G. The CPU used was AMD Ryzen 9 7900X 12-Core Processor.

Results of Experiments on participants with no sleep apnea We also conducted the same experiments aligned in the methods section using 5-fold cross-validation, on the 22 subjects with AHI < 5 and less than 20% epochs with artifacts. We see that while LSTM post-processing is still helpful in improving performance, as evident by the consistent better scores from the models with LSTM post-processing. However, adding Apnea Severity or Obesity as random effects to the LightGBM models does not improve algorithm performance significantly any more. This is reasonable because we are training and testing on a subpopulation with no sleep apnea.

Supplementary Table 3: 5-fold cross-validation results for 22 subject who had no apnea and had less than 20% epochs with artifacts

Random Effects	LSTM post-processing	F1 Score	AUROC	AUPRC	Accuracy	Cohen's Kappa
None	No	0.755 ± 0.060	0.853 ± 0.049	0.848 ± 0.056	0.757 ± 0.059	0.492 ± 0.084
Obesity	No	0.760 ± 0.045	0.864 ± 0.048	0.857 ± 0.063	0.764 ± 0.040	0.506 ± 0.059
Apnea Severity	No	0.758 ± 0.039	0.859 ± 0.050	0.853 ± 0.062	0.760 ± 0.033	0.499 ± 0.048
None	Yes	0.763 ± 0.057	0.874 ± 0.044	0.876 ± 0.047	0.763 ± 0.064	0.502 ± 0.099
Obesity	Yes	0.775 ± 0.035	0.886 ± 0.038	0.887 ± 0.040	0.784 ± 0.026	0.538 ± 0.043
Apnea Severity	Yes	**0.776 ± 0.033**	**0.883 ± 0.041**	**0.886 ± 0.039**	**0.784 ± 0.026**	**0.543 ± 0.048**

Appendix B. Dataset Statements & Documents

B.1. Data Hosting, Licensing, and Maintenance Plan

Due to the sensitive nature of the dataset, we release our highest-resolution data with restricted credentialed access. Therefore, we leverage the PhysioNet platform for data hosting and licensing, and maintenance. We have submitted our dataset and it is currently under review by the PhysioNet Team. A link to view the current dataset is available at: PhysioNet/DREAMT. The dataset link may also be updated, and the most updated version will be available at DREAMT_FE
Host: The PhysioNet platform with Credentials Access.
License: PhysioNet Credentialed Health Data License 1.5.0

B.2. Dataset Description

For each participant, the dataset consists of six raw signals from the E4 wristband (BVP, ACC_X, ACC_Y, ACC_Z, EDA, TEMP), two derived signals (HR and IBI), and the ground-truth sleep stage labels determined by trained technicians (REM, Stages 1-3 NREM, Wake). Each participant's data is stored in a comma separated values (CSV) file, where each column represents one of the E4 wristband signals, or the ground truth sleep stage labels, and each row represents one time point. Additionally, a column with the timestamp for each row is provided. Each of the signals was upsampled (with repeated values, no imputation) to the maximum sampling frequency, 64Hz. Therefore, each row represents 1/64th of a second. The inherent sampling frequencies for each signal are as follows:

- **TIMESTAMP** (64 Hz): Timestamp shifted and started with 0, with frequency of 64 Hz.
- **BVP** (64 Hz): Blood volume pulse derived from the photoplethysmography (PPG) sensor.
- **IBI**: Inter-beat interval is the time interval between individual beats of the heart, derived from the photoplethysmography (PPG) sensor.
- **EDA** (4 Hz): Electrodermal activity from the galvanic skin response sensor.
- **TEMP** (4 Hz): Skin temperature from the infrared thermopile sensor.
- **ACC** (32 Hz): Triaxial accelerometry with each axis named **ACC_X**, **ACC_Y**, and **ACC_Z**.
- **HR** (1 Hz): Heart rate is estimated from the BVP signal.
- **Sleep_Stage**: The technician-annotated sleep labels derived from PSG are recorded every 30 seconds.

Here's the overview of the dataset:

- *E4_aggregate/*
- *features_df/*
- *participant_info.csv*

The folder titled '*data*' contains 100 csv files. Each file contains the recorded signals and corresponding sleep stages. All the signals and sleep stage labels were upsampled to 64 Hz in the file, with repeated values.

Each *.csv* file has the following columns: TIMESTAMP, BVP, IBI, EDA, TEMP, ACC_X, ACC_Y, ACC_Z, HR, Sleep_Stage

participant_info.csv: The file contains information on all participants, such as age and gender.

Ethics:

Our dataset contains raw wearable sensor data collected during sleep along with expert-annotated sleep stage labels originating from PSG. Our dataset can support the development of robust sleep tracking algorithms using wearable data.

We ensured well-informed consent regarding safety and privacy before data collection. The released data has been completely anonymized and all identifiable information has been removed, in compliance with the IRB guidelines. Any identifiable information cannot be viewed by anyone outside the study team.

Data collected from human participants:

The study protocol was approved by relevant Institutional Review Boards (IRBs). Human participants signed a consent form before participating in the study.

Clinical trial data: N/A

Data collected from animals: N/A

Acknowledgments:

We acknowledge the support of Shirah Pokusa with recruiting participants and Duke Sleep Disorders Center for providing the sleep stage annotations.

Conflicts of interest: The author(s) have no conflicts of interest to declare.