Estimating Sleep Quality Metrics Using Free-Living Accelerometer Data From Thigh-Worn Devices in Comparison to an EEG-Based Sleep Tracking Device

Esben Høegholm Lykke[a,*], Jan Christian Brønd[a]

[a] *University of Southern Denmark, Department of Sports Science and Clinical Biomechanics, Campusvej 55, Odense, 5230*

Abstract

LÆS IKKE! Det er volapyk på latin. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vitae dictum eros, ullamcorper elementum orci. Sed laoreet nulla neque, pulvinar fermentum mi iaculis at. Nulla ultricies nibh sit amet vestibulum rutrum. Nam pharetra nisl sed ipsum maximus suscipit. Duis metus nunc, ullamcorper eu mi rutrum, tempus ultricies ante. Nunc vitae lectus nisi. Aliquam efficitur ut eros ut pellentesque. Aenean blandit, nisl nec efficitur interdum, nisi ipsum fermentum dolor, at tempus sem turpis in lacus. Curabitur sollicitudin lectus sit amet velit pellentesque laoreet. Ut posuere diam lobortis nisi eleifend tincidunt. Ut at euismod sem, sed dignissim ligula. Aliquam lacinia massa libero, id eleifend velit pulvinar ac. Fusce volutpat elit eu nulla viverra, nec tempus orci pellentesque.

Keywords: Sleep, Accelerometry, EEG, Machine learning, Sleep quality

1. Introduction

An extensive array of research underlines the importance of sleep for both mental and physical health[1,2,3,4]. Consequently, accurate sleep assessment methods are essential for tracking sleep patterns, thereby enhancing our comprehension of the sleep-health relationship. Furthermore, ensuring high user-acceptability for these methods is essential in order to conduct large-scale studies over prolonged periods.

While laboratory-based polysomnography is considered the gold standard for objectively measuring sleep, its practicality in large-scale epidemiological studies is limited due to its high cost, the need for professional administration, and the substantial resources required for specialized equipment[5]. As an alternative, diaries is commonly used as low-cost and low-tech methods for sleep assessment in population research. However, relying solely on diary-based methods may introduce recall bias and other limitations[6]. A more feasible approach in large-scale epidemiological studies is the use of device-based measurement methods that can estimate sleep duration. This method offers the advantage of being less burdensome for participants and avoids potential biases associated with recall.

In this context, the Zmachine® Insight+ (ZM) emerges as a valuable tool. Validated against polysomnography with favorable results[7,8], the ZM provides comparable data without the high costs or need for professional monitoring associated with polysomnography. Furthermore, the ease of use of the ZM device makes it compliant for free-living use[9]. This allows for the analysis of multiple consecutive nights, as compared to single-night recordings from PSG, thereby capturing the important variations in sleep across multiple nights.

[*] Corresponding author
Email addresses: eskovgaard@health.sdu.dk (Esben Høegholm Lykke), jbrond@health.sdu.dk (Jan Christian Brønd)

The introduction of body-worn accelerometers has provided an effective and affordable alternative for objectively assessing sleep patterns in a home environment over extended periods. These accelerometers collect continuous, high-resolution data for several weeks without requiring recharging, thus minimizing participant burden. Initial applications of accelerometry for sleep and wake stage classification were based on wrist movements. The original algorithm, developed in 1982 using simple linear regression and validated with PSG[10], was later refined in 1992[11], leading to the widely used Cole-Kripke model. Subsequent research on wrist-worn accelerometer data has employed heuristic algorithms, advanced machine learning models, as well as regression and deep learning techniques[12,11,13,14,15,16].

Despite the well-developed field of accelerometer data analysis for sleep detection from wrist and hip-worn devices, the same level of advancement is not mirrored in studies utilizing thigh-worn accelerometers. Methods for assessing sleep using wrist and hip-worn accelerometers have greatly evolved over the years, employing an extensive range of techniques. These include heuristic algorithms, machine learning models, regression and deep learning techniques, all tailored to the specific signal characteristics of wrist and hip-worn devices[12,11,13,14,15,16,17]. However, for thigh-worn accelerometers, the landscape appears less mature, with only a handful of studies investigating sleep detection algorithms. The majority of the efforts are focused on delineating wakefulness from sleep, with particular emphasis on the definition of 'waking time' and 'bedtime'[18,19,20,21]. Furthermore, while strides have been made recently in estimating sleep duration with these devices, with the introduction of a promising algorithm and its comparison against PSG[22], the field is still in its infancy when it comes to employing machine learning techniques. Given the potential for accurate physical behavior assessment that thigh-worn accelerometers provide[23,24], a significant research gap exists. Therefore, there is a pressing need for future studies to develop techniques similar to those used for wrist and hip-worn accelerometers, with the ultimate goal of establishing a more holistic, accurate, and user-friendly method of sleep and physical activity tracking.

Our primary objective in this study was to evaluate a range of machine learning and deep learning models, utilizing the raw data collected from a tri-axial thigh-worn accelerometer to estimate in-bed and sleep time. To ensure the reliability and effectiveness of our models, we compared their outputs with an EEG-based sleep tracking device, which we considered the gold standard for measuring sleep. Furthermore, our secondary goal was to assess the developed models' performance in evaluating important sleep quality parameters, including sleep period time (SPT), total sleep time (TST), sleep efficiency (SE), latency until persistent sleep (LPS), and wake after sleep onset (WASO).

2. Methods

2.1. Dataset and participants

The current study leverages data from the SCREENS project[25], a study conducted from October 2018 to March 2019 in Middelfart, Southern Denmark, that evaluated the impact of screen media usage on Danish families. For our analysis, we focused on data from child participants aged between 6 and 10 years within the SCREENS cohort. Our primary sources of data were accelerometer readings from Axivity AX3 devices attached to the children's thighs, and electroencephalography (EEG) data derived from the ZM device. The Axivity AX3, an unobtrusive 3-axis accelerometer, was positioned midway between the hip and knee on the right anterior thigh, recording participant movement data.

Sleep state information was extracted using the ZM, a product of General Sleep Corporation. The ZM, which utilizes advanced EEG hardware and signal processing algorithms, employs three self-adhesive, disposable sensors placed outside the hairline for reliable EEG signal acquisition. The ZM uses two proprietary algorithms: Z-ALG and Z-PLUS. The Z-ALG is utilized for accurate sleep detection, showcasing its suitability for in-home monitoring[7], while the Z-PLUS effectively differentiates sleep stages, as evidenced by its alignment with expert evaluations using PSG data[8]. In the current study, we treated all sleep stages as a single category effectively deducing the output of the ZM to "awake" and "asleep" as the ability to distinguish sleep stages are not a necessity to derive sleep quality parameters of interest and to simplify the learning process of the models.

Figure 1 illustrates the selection criteria applied to the children's recordings from the SCREENS study. We included only ZM recordings that were accompanied by complete accelerometer data and lasted between 7 and 14 hours. Any night when the ZM reported sensor issues was excluded. The children whose recordings were considered had an average age of 9.4 years, with a standard deviation of 2.1. In their raw form, the ZM predictions encompassed 696,779 epochs, each 30 seconds long. Notably, approximately 84% of the total ZM recording duration was classified as sleep, resulting in an imbalance of the dataset.

Finally, we affirm that the SCREENS study received approval from the Regional Scientific Committee of Southern Denmark, and all data handling processes complied with the General Data Protection Regulation (GDPR), ensuring the ethical and secure management of participant information.

2.2. Data Preprocessing and Feature Extraction

In this study, data processing of the raw accelerometer data began with a low-pass filtration step using a 4th order Butterworth filter with a 5 Hz cut-off frequency to eliminate high-frequency noise. Following filtration, data were partitioned into overlapping 2-second intervals, each successive interval sharing a 50% overlap with the previous one similar to methods described by Skotte et al.[23]. Any non-wear data was remove using previously described methods[26] and data was resampled to 30-second epochs so every sample classified by the algorithms corresponds to a 30-second epoch scored during the ZM recordings. Subsequently, we performed a feature extraction process that yielded a set of 88 features, providing a robust characterization of the data. Extracted from accelerometer and temperature signals, these features include temporal elements that use both lag and lead values, capturing dynamic data trends by incorporating measurements from preceding and upcoming intervals. Furthermore, inspired by Walch et al.[27], we incorporated sensor-independent features to encapsulate circadian rhythms. These features offer unique insights not directly discernible from sensor outputs and are meant to approximate the changing drive of the circadian clock to sleep over the course of the night (see Figure 2). Furthermore, the feature set was enriched by including signal characteristics, which encompass vector magnitude, mean crossing rate, skewness, and kurtosis for each of the x, y, and z dimensions. All features are summarized in table ??? **in the supplementary matrials**. Subsequently, we merged the ZM and corresponding accelerometer recordings. Any overlapping time between the ZM and accelerometer data was treated as 'in-bed' time, with the remaining time considered 'out-of-bed'. This process yielded a dataset providing a around the clock temporal view of each participant's activity and sleep patterns.

In addition to the engineered features, we chose to incorporate the median-filtered raw predictions from the ZM device into our modeling process. This decision stemmed from the understanding that children typically undergo around five to eight sleep cycles per night, with awakenings most likely occurring at the end of each cycle[28]. Upon examining the raw ZM predictions, we noted a significant overestimation in the number of awakenings per night for the children in our study, exceeding what would be expected based on typical sleep cycle patterns. In particular, many of these brief awakenings could be considered as noise, which when present in the data, can potentially hinder the learning process of machine learning models by obscuring the underlying patterns that the models are trying to learn, leading to less accurate predictions. Consequently, we elected to train and evaluate our models using not only the raw ZM output, but also versions that were subjected to 5-minute and 10-minute median filters. This approach, by mitigating this noise, resulted in an anticipated, more age-appropriate count of awakenings per night, providing a more accurate depiction of children's sleep patterns (see Figure 3).

2.3. Algorithms, Training and Validation

In this study, we employed two distinct modeling strategies to analyze sleep patterns from thigh-mounted accelerometer data. We used a sequential strategy, comprising an ensemble of four pairs of models, each pair featuring the same algorithm. This strategy aimed to make the prediction task more manageable for the algorithms by breaking it down into a sequence of two binary classifications: first predicting 'in-bed' time, then 'sleep' time. Simultaneously, we also used a multiclass approach utilizing a bidirectional Long Short-Term Memory (biLSTM)[29] neural network.

2.3.1. Models in Sequence

To predict in-bed time and sleep time accurately, we employed an ensemble learning strategy based on sequential binary classification models. This approach involved constructing a sequence of models using multiple machine learning algorithms to improve predictive accuracy. The process began with an initial model predicting in-bed time, followed by a second model that utilized the output of the initial model to predict sleep time. This sequential approach was applied across all four algorithms detailed below, with each subsequent model leveraging the outputs of the previous models for improved predictions.

1. Logistic Regression (LREG): Logistic regression served as a simple and fast baseline model. However, due to its linear nature, it may struggle with capturing complex relationships and non-linear patterns present in the accelerometer data.
2. Decision Tree (TREE): Decision trees are capable of handling non-linear patterns and are easily interpretable. However, they are prone to overfitting, particularly when dealing with complex patterns that require simultaneous consideration of multiple features.
3. Single-layer Feed-forward Neural Network (SNN): Single-layer feed-forward neural networks can effectively capture non-linear relationships, even with their relatively simple structure. However, they tend to be more challenging to interpret compared to simpler models. Additionally, careful tuning of the network's architecture and training process is required to mitigate the risk of overfitting.
4. XGBoost (XGB): XGBoost is a powerful algorithm known for its ability to provide highly accurate predictions and handle complex, non-linear patterns in the data. It also incorporates built-in methods to prevent overfitting. However, training XGBoost models can be computationally intensive, and interpreting the predictions it generates can pose challenges.

2.3.2. Multiclass Model

In this study, we employed a biLSTM, a multiclass classifier, to predict three distinct states: out-of-bed-awake, in-bed-awake, and in-bed-asleep. The architecture of the biLSTM was set up with four layers, each equipped with 128 hidden units. This configuration was intentionally chosen to balance between model complexity and training efficiency: it provided the depth necessary for learning intricate patterns while remaining feasible for timely training. The bidirectional design of the LSTM served to enhance data interpretation and mitigate overfitting by doubling the hidden units at each time step. For input, we used tensors shaped as sequences, with each sequence spanning 10 minutes and a step size of one.This approach follows in the footsteps of previous studies that utilized LSTM models for sleep detection. These studies showcased the promising potential of LSTMs in capturing complex temporal patterns. Particularly, the works of Sano et al.[30] and Chen et al.[31] demonstrated the effectiveness of LSTM models in improving sleep detection using accelerometer data, underscoring the value of this modeling approach.

2.3.3. Model Training

For the models in sequence, we trained four pairs of classification models. Each pair was designed to distinguish between in-bed/out-of-bed and asleep/awake states, respectively. The dataset was randomly split into a training set and a testing set, each containing approximately 50% of the subjects. This division ensured that samples from the same night were never simultaneously present in both sets. To optimize hyperparameters, we performed a 10-fold Monte Carlo cross-validation on a regular grid, comprising 20 different combinations of hyperparameters. The F1 score served as the optimization metric. The best-performing set of hyperparameters was then used to fit the models to the full training dataset. This approach allowed us to maximize performance by leveraging all available data. Moreover, after extracting the in-bed time from the initial sequential models, the imbalance on the resulting dataset could cause biases during model training, as models may favor predicting the majority class. To rectify this, we employed the Synthetic Minority Over-sampling Technique (SMOTE)[32]. SMOTE generates new samples by interpolating random samples with their nearest neighbors. We utilized the themis R package[33] to implement SMOTE, resulting in a balanced distribution of training samples across both classes.

The biLSTM model was trained to differentiate between three states: out-of-bed-awake, in-bed-awake, and in-bed-asleep. The data used for training the biLSTM was randomly divided into training, validation, and test sets, based on a 50/25/25 split. We ensured that data from the same night was not present across different sets. The model was trained using the Adam optimizer, selected for its computational efficiency and adaptability of the learning rate during training. Given the multiclass classification task with mutually exclusive classes, we employed the cross-entropy loss function. To obtain a probability distribution over the classes, the softmax activation function was applied at the output layer. We evaluated the model's performance using the F1 score on both the training and validation sets. We implemented early stopping with a patience of 3 epochs, halting the training process if there was no improvement in the validation loss over three consecutive epochs.

2.3.4. Model Validation

In our study, we utilized standard evaluation metrics to assess the performance of each model on an epoch-to-epoch basis. These include accuracy ($accuracy = \frac{TP+TN}{TP+TN+FP+FN}$), sensitivity ($sensitivity = \frac{TP}{TP+FN}$), specificity ($specificity = \frac{TN}{TN+FP}$), precision ($precision = \frac{TP}{TP+FP}$), negative predictive value (NPV, $NPV = \frac{TN}{TN+FN}$), and F1 score ($F_1 = 2 * \frac{precision*sensitivity}{precision+sensitivity}$).

In the context of our sequential learning strategy, the initial models were tasked with the binary classification of in-bed vs. out-of-bed. For this task, we assessed performance using the F1-score, accuracy, sensitivity, specificity, and precision metrics. The second models in our sequential learning strategy focused on the binary classification of asleep vs. awake. For these models, we considered the same metrics, in addition to the negative predictive rate. The class imbalance in this case led us to compute the F1 score as an unweighted macro-average. Additionally, we evaluated the multiclass classifier, biLSTM, using macro-averaged F1-score, sensitivity, and precision. To further illustrate model performance, we provide confusion matrices for the full dataset, encompassing both in-bed and out-of-bed data. These matrices report relative counts, column percentages (the proportion of the true class accurately predicted), and row percentages (the proportion of predictions correctly classified). We considered both the in-bed/out-of-bed and awake/asleep scoring tasks as binary classification problems, designating in-bed and asleep as the positive labels and out-of-bed and awake as the negative labels in accordance with previous research[34,35].

To assess the performance of our models in deriving sleep quality parameters, we utilized Bland-Altman plots and Pearson correlations. The Bland-Altman method was employed specifically to determine the level of agreement between two measurement techniques. Considering our dataset contained multiple observations per subject, we integrated a bootstrap procedure to address this extra source of variability. We calculated the mean difference (bias) and defined the limits of agreement as the mean difference plus or minus 1.96 times the standard deviation of these differences. To ensure our measurements were robust and accounted for intra-subject variability, we estimated the 95% confidence intervals for both the bias and the limits of agreement using a bias-corrected and accelerated bootstrap method, utilizing 10,000 bootstrap replicates. The sleep quality parameteres included the following:

1. Sleep Period Time (SPT) - This refers to the total duration of the sleep period, which is defined as the time from the start to the end of the ZM recording.
2. Total Sleep Time (TST) - This is the time spent asleep within the SPT.
3. Sleep Efficiency (SE) - This is the ratio between TST and SPT, representing the proportion of the sleep period that was actually spent asleep.
4. Latency Until Persistent Sleep (LPS) - This metric represents the time it takes to transition from wakefulness to sustained sleep. It is calculated as the time from the beginning of the ZM recording until a period when 10 out of 12 minutes are scored as sleep.
5. Wake After Sleep Onset (WASO) - This refers to the time spent awake after initially falling asleep and before the final awakening. In our analysis, a period is counted as 'awake' only if it consists of 3 or more contiguous 30-second epochs which is also how the ZM summarizes WASO.

R version 4.3.0 (2023-04-21)[36] and the Tidymodels[37] and Tidyverse[38] suite of packages were used as the core tools for model development and analyses. Python version 3.10.6[39] and PyTorch[40] were used to implement the biLSTM model. All code used to perform the analysis and generate the figures in this paper are available in this repository.

3. Results

As reported in Table 1 the sleep quality parameters derived from ZM predictions were modified by the implementation of 5-minute and 10-minute median filters. SPT were consistent across raw and filtered datasets (mean: 9.2 ± 2.1 hours), corresponding to the length of the ZM recording. TST and SE increased in the filtered data, implying the filters categorize some wakefulness as sleep. Specifically, TST increased from a raw mean of 7.7 ± 1.9 hours to 8.1 ± 2.0 hours (5-minute filter) and 8.2 ± 2.1 hours (10-minute filter), while SE rose from 82.6 ± 12.0% to 86.4 ± 12.7% and 87.5 ± 12.9% respectively. LPS also elevated, suggesting the filter smooths out brief awakenings at sleep onset, leading to a prolonged time to persistent sleep. A significant change was seen in WASO, which dropped from 39.0 ± 33.6 minutes in raw data to 30.6 ± 46.8 minutes and 22.3 ± 55.4 minutes in the 5-minute and 10-minute filtered data, respectively. The number of awakenings was also considerably reduced with the application of filters. In the raw data, the average number of awakenings was 34.46 ± 11.33 per night, which reduced to 4.43 ± 3.26 and 1.95 ± 2.01 for the 5-minute and 10-minute filtered data sets respectively.

Table 1: Overview of characteristics of the ZM sleep quality summaries per night. Values are represented as mean (SD).

	SPT (hrs)	TST (hrs)	SE (%)	LPS (min)	WASO (min)	Awakenings (N)
Raw ZM Predictions	9.2 (2.1)	7.7 (1.9)	82.6 (12)	34.5 (27.9)	39 (33.6)	34.5 (11.3)
5-Min Median	9.2 (2.1)	8.1 (2)	86.4 (12.7)	36.3 (39.8)	30.6 (46.8)	4.4 (3.3)
10-Min Median	9.2 (2.1)	8.2 (2.1)	87.5 (12.9)	38 (48.7)	22.3 (55.4)	1.9 (2)

3.1. Performance on Epoch-to-Epoch Basis

The epoch-to-epoch evaluation of predicting in-bed time is outlined in Table 2, and demonstrates practically equivalent performance across all model types. The F1 score ranges from 94.4% (Decision Tree) to 95.4% (XGBoost), while accuracy ranges from 95.3% (Decision Tree) to 96.1% (XGBoost). Sensitivity, Precision, and Specificity also demonstrate consistent results across the models. The XGBoost model, despite recording the highest metrics with an F1 score of 95.4% and accuracy of 96.1%, outpaced the others only marginally.

Table 2: In-Bed Performance Metrics

	F1 Score (%)	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Decision Tree	94.4	95.3	93.1	95.6	96.9
Logistic Regression	95.0	95.7	95.0	94.9	96.3
Feed-Forward Neural Net	95.0	95.8	95.1	95.0	96.3
XGBoost	95.4	96.1	95.8	94.9	96.2

Table 3 details the performance of the second sequential models on raw and median-filtered (5 and 10 minute) ZM predictions. These are the predictions of the sequential models' ability to predict asleep/awake on only the extracted in-bed time. The F1 scores, which are unweighted macro averages, for raw ZM predictions range from 71.05% to 76.18%. The models perform comparably, but the low specificity values (62.84% to 70.93%) suggest difficulty in correctly classifying awake epochs. Applying 5-minute median filtering improves the performance metrics. The XGBoost model tops the charts with an F1 score of 79.22% and NPV of 74.00%. However, specificity still remains low, with values between 54.68% and 74.84% across all models. With 10-minute median filtering, the metrics improve further. The XGBoost model still leads with an F1 score of 80.87% and an NPV of 75.76%. But, specificity remains low, ranging from 57.47% to 76.35% across all models.

Table 3: Sleep Performance Metrics

	F1 Score (%)	Precision (%)	NPV (%)	Sensitivity (%)	Specificity (%)
Raw ZM Predictions					
Decision Tree	72.9	93.2	48.4	86.3	67.1
Logistic Regression	71.0	93.7	43.9	82.7	70.9
Neural Network	71.8	93.8	45.1	83.6	70.8
XGBoost	76.2	92.8	58.0	91.3	62.8
5-Min Median					
Decision Tree	75.5	94.2	55.5	93.4	59.0
Logistic Regression	68.3	95.8	36.0	81.4	74.8
Neural Network	71.7	95.8	41.6	85.6	73.1
XGBoost	79.2	93.9	74.0	97.3	54.7
10-Min Median					
Decision Tree	76.3	94.7	58.1	94.9	57.5
Logistic Regression	68.0	96.5	34.3	81.9	76.4
Neural Network	71.0	96.1	39.5	86.5	71.4
XGBoost	80.9	94.9	75.8	97.7	57.6

The complete set of confusion matrices generated from data both containing the out-of-bed and in-bed time are presented in Figure 4. The figure shows favorable epoch-to-epoch performance across across all sequential models, however, it is evident that the biLSTM is less successful in classifying the in-bed-awake class which cannot be deduced from the confusion matrices from the sequential models. [**fejl i biLSTM confmat. Alle mats er ens**]

Table 4 presents the performance of the three-class biLSTM multiclassifier on raw and median-filtered (5-minute and 10-minute) ZM predictions. Raw ZM predictions achieve F1 Scores ranging from 71.36% to 76.04%, indicating overall good performance. Applying 5-minute median filtering improves the metrics further, resulting in F1 Scores ranging from 75.99% to 78.53%, demonstrating enhanced precision and sensitivity. [**Overvej flere metrics. Evt. per class... pt er det lidt svært at sammenligne på tværs af lstm og andre modeller ud over F1.**] With 10-minute median filtering, F1 Scores range from 73.45% to 73.93%, maintaining good performance but with limited information on Specificity. Further analysis is required to assess performance across all classes.

Table 4: Performance of the three-class biLSTM multicclassifier.

	F1 Score	Precision	NPV	Sensitivity	Specificity
Raw ZM Predictions	71.4	76.0	95.5	70.4	95.0
5-Min Median	76.0	78.5	96.4	74.6	96.2
10-Min Median	73.4	73.1	95.6	73.9	95.7

3.2. Evaluation of Sleep Quality Parameters

Table 5 presents a comparative analysis of the included models used to predict various sleep quality parameters (SPT, TST, SE, LPS, WASO) using the 5-minute median filtered ZM predictions. To see the full table including models developed from raw ZM predictions and 10-minute median filtered ZM predictions, see [**SUPP. MAT.**]. In terms of bias, the decision tree model consistently underestimated SPT, TST, and SE, and overestimated LPS and WASO in comparison to ZM. The logistic regression model had similar trends, with more pronounced underestimation in TST and overestimation in LPS. The eed-forward neural network also exhibited similar bias as the decision tree and the logistic regression models, but with a higher overestimation in WASO. On the other hand, the XGboost model showed least bias among all, especially in its 5-minute median predictions. Considering LOA, the decision tree had higher variability across different sleep quality parameters and filtering techniques, particularly for LPS and WASO, which indicates lower agreement with ZM. Other models had comparable LOA but with notable exceptions. For example, TST LOA for the logistic regression model was particularly wide in the 5-minute median predictions. Correlation-wise, the pearson coefficient, revealed that the XGboost model consistently had the highest correlation with ZM across all sleep qualityparameters and filtering methods Notably, the XGboost's 5-minute median predictions showed the strongest correlation (0.66) for TST among all models and filtering techniques.

Table 5: Summary of Bias, Limits of Agreement (LOA), and Pearson Correlation for various Sleep Parameter Predictions (SPT, TST, SE, LPS, WASO) using different Machine Learning Models (Decision Tree, Logistic Regression, Feed-Forward Neural Net, XGBoost) with Raw ZM Predictions, 5-Min and 10-Min Median as predictors. Each value is provided with its 95% Confidence Interval (CI).

	Bias (95% CI)	LOA (95% CI)	LOA (95% CI)	Pearson, *r* (95% CI)
5-Min Median - Decision Tree				
SPT (min)	-21.6 (-25.6;-17.6)	-117.5 (-125.6;-110.7)	74.2 (63.9;85.9)	0.54 (0.48;0.6)
TST (min)	-50.5 (-55.2;-46)	-161.4 (-175.8;-151.3)	60.4 (51.5;71.7)	0.48 (0.42;0.54)
SE (%)	-5.5 (-6.3;-4.7)	-23.9 (-26.4;-22.2)	12.9 (11.6;14.6)	0.22 (0.14;0.29)
LPS (min)	24.6 (19.7;29.1)	-88.8 (-115;-77.3)	138 (126.2;156.7)	0.06 (-0.02;0.14)
WASO (min)	9.9 (6.5;14)	-79.4 (-109;-63.1)	99.2 (80;136.1)	0.15 (0.07;0.22)
5-Min Median - Logistic Regression				
SPT (min)	-3.7 (-8;1)	-112.2 (-120.9;-105.2)	104.8 (94;117.4)	0.38 (0.3;0.44)
TST (min)	-139.7 (-146.9;-133)	-305.6 (-323.6;-291.8)	26.2 (16.1;38.6)	0.09 (0.01;0.17)
SE (%)	-23.2 (-24.3;-22.2)	-48.1 (-50.9;-46.1)	1.7 (0.1;3.8)	0.13 (0.05;0.21)
LPS (min)	58.1 (53.4;62.6)	-52.3 (-75;-40.1)	168.6 (155.9;187.7)	0.05 (-0.03;0.13)
WASO (min)	45.4 (41.7;49.7)	-50.7 (-74.4;-38.4)	141.5 (126.8;173)	0.19 (0.11;0.27)
5-Min Median - Feed-Forward Neural Net				
SPT (min)	-3.9 (-8.1;0.9)	-112.7 (-122;-105.2)	104.9 (94.1;118.4)	0.38 (0.3;0.44)
TST (min)	-126.5 (-132.8;-120.3)	-276.8 (-291.3;-264.7)	23.9 (14.8;33.9)	0.25 (0.17;0.32)
SE (%)	-20.9 (-21.9;-19.9)	-44.3 (-46.3;-42.5)	2.5 (1.1;4)	0.21 (0.13;0.29)
LPS (min)	35.3 (30.7;39.8)	-75.8 (-102.3;-63.4)	146.5 (134.4;166.9)	0.07 (-0.01;0.15)
WASO (min)	45 (41.2;49.2)	-51.8 (-76.4;-39.1)	141.7 (125.8;174.1)	0.21 (0.14;0.29)
5-Min Median - XGboost				
SPT (min)	0.2 (-3.7;4.5)	-97.4 (-106.2;-90.3)	97.8 (86.6;111)	0.56 (0.5;0.61)
TST (min)	-7 (-10.8;-3.3)	-95.5 (-105.2;-88)	81.4 (72.4;92.5)	0.66 (0.61;0.7)
SE (%)	-1.1 (-1.7;-0.5)	-15.6 (-17;-14.4)	13.3 (12.2;14.7)	0.44 (0.38;0.51)
LPS (min)	28.5 (23.9;32.6)	-76.4 (-104.2;-63.3)	133.4 (120.4;154.2)	0.12 (0.04;0.2)
WASO (min)	-0.9 (-3.9;3)	-83.4 (-113.1;-66)	81.7 (62;119.6)	0.26 (0.18;0.33)
5-Min Median - biLSTM				
SPT (min)	-36.1 (-41.7;-30)	-136.1 (-146.3;-126.9)	64 (51.1;78.6)	0.54 (0.45;0.62)
TST (min)	12.8 (7.4;18.3)	-80.1 (-89.8;-72.3)	105.8 (94.3;118.8)	0.63 (0.55;0.69)
SE (%)	8 (7.2;8.8)	-5.1 (-6.8;-3.8)	21.1 (19.5;23.1)	0.16 (0.04;0.27)
LPS (min)	-15.7 (-25.9;-7.5)	-169 (-230.7;-127.9)	137.6 (101.1;184.9)	0.09 (-0.02;0.2)
WASO (min)	-3 (-9.9;7.7)	-144.1 (-197.2;-107.2)	138.1 (90.8;211.4)	0.02 (-0.1;0.13)

Figure 5 shows the agreement of the XGBoost model trained on the 5-minute median filtered ZM predictions in estimating sleep quality parameters as evaluated 5-minute median-smoothed ZM derived sleep quality parameter. As can be seen on from the Bland-Altman plots, a large portion of the nights are in almost perfect agreement with the ZM and also a portion that can be considered outliers. This is true for all sleep quality parameters.

[**Overvej at smide samtlige 75 plots i supp.mat.**]

4. Discussion

The present study embarked on an in-depth exploration of the utility and comparative performance of different machine learning models in the analysis of pediatric sleep data. The results offer promising insights into the role of machine learning in enhancing the precision of sleep quality metrics.

A significant finding was the impact of median filters on the sleep metrics. Implementing 5-minute and 10-minute median filters led to notable changes in key sleep parameters. In particular, the Total Sleep Time (TST), Sleep Efficiency (SE), and Latency to Persistent Sleep (LPS) increased, suggesting that the filters may categorize some instances of wakefulness as sleep and smooth out brief awakenings. This underscores the potential of median filters in refining sleep data and mitigating the potential misclassification of sleep states.

Interestingly, the most marked change was observed in Wake After Sleep Onset (WASO), with a substantial reduction in the filtered data. This reduction, coupled with the decrease in the number of awakenings, implies a significant role of median filters in sleep-wake cycle characterization, potentially enhancing the reliability of sleep assessment.

The machine learning models demonstrated high overall performance on an epoch-to-epoch basis, with the XGBoost model marginally outperforming the others. However, the models had more varied performance in classifying sleep conditions, as evidenced by the sleep performance metrics. The relatively lower Specificity values across all models suggest a common challenge in correctly classifying negative instances.

In the precision-recall and ROC curve analyses, the Decision Tree model showed superior precision-recall AUC, indicating strong predictive accuracy. In contrast, the XGBoost model consistently demonstrated a high ROC AUC, indicating a robust ability to differentiate between classes. However, it's worth noting that the Neural Network model generally showed weaker performance in these areas.

The three-class biLSTM multiclassifier showed overall good performance, further supporting the potential of machine learning in sleep study. However, the absence of Specificity values limits a complete assessment of its performance, calling for further exploration in future studies.

It is unclear whether Johansson et al.[22]] provide a method for detecting the in-bed time...

In conclusion, this study provides compelling evidence of the utility of machine learning models and median filters in enhancing the precision and reliability of pediatric sleep assessment. However, more research is needed to overcome the challenge of correctly classifying negative instances and to fully understand the performance of these models across all classes and conditions. The findings underscore the potential of machine learning in advancing sleep medicine, paving the way for more accurate, reliable, and personalized sleep assessment in pediatric populations.

References

[1] G. Ma, Sleep, health, and society, Sleep medicine clinics 12 (1), publisher: Sleep Med Clin PMID: 28159089 (03 2017). doi:10.1016/j.jsmc.2016.10.012.
URL https://pubmed.ncbi.nlm.nih.gov/28159089/

[2] N. Meyer, A. G. Harvey, S. W. Lockley, D.-J. Dijk, Circadian rhythms and disorders of the timing of sleep, The Lancet 400 (10357) (2022) 1061–1078, publisher: Elsevier PMID: 36115370. doi:10.1016/S0140-6736(22)00877-7.
URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)00877-7/fulltext

[3] M. K Pavlova, V. Latreille, Sleep disorders, The American Journal of Medicine 132 (3) (2019) 292–299, pMID: 30292731. doi:10.1016/j.amjmed.2018.09.021.

[4] S. Difrancesco, F. Lamers, H. Riese, K. R. Merikangas, A. T. F. Beekman, A. M. van Hemert, R. A. Schoevers, B. W. J. H. Penninx, Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study, Depression and Anxiety 36 (10) (2019) 975–986, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22949. doi:10.1002/da.22949.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/da.22949

[5] A. T. M. Van De Water, A. Holmes, D. A. Hurley, Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review, Journal of Sleep Research 20 (1pt2) (2011) 183–200, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2869.2009.00814.x. doi:10.1111/j.1365-2869.2009.00814.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2869.2009.00814.x

[6] C. M. Moore, S. J. Schmiege, E. E. Matthews, Actigraphy and sleep diary measurements in breast cancer survivors: Discrepancy in selected sleep parameters, Behavioral Sleep Medicine 13 (6) (2015) 472–490, pMID: 25117292 PMCID: PMC4326642. doi:10.1080/15402002.2014.940108.

[7] R. F. Kaplan, Y. Wang, K. A. Loparo, M. R. Kelly, R. R. Bootzin, Performance evaluation of an automated single-channel sleep–wake detection algorithm, Nature and Science of Sleep 6 (2014) 113–122, pMID: 25342922 PMCID: PMC4206400. doi:10.2147/NSS.S71159.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4206400/

[8] Y. Wang, K. A. Loparo, M. R. Kelly, R. F. Kaplan, Evaluation of an automated single-channel sleep staging algorithm, Nature and Science of Sleep 7 (2015) 101–111, pMID: 26425109 PMCID: PMC4583116. doi:10.2147/NSS.S77888.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4583116/

[9] J. Pedersen, M. G. B. Rasmussen, L. Olesen, P. L. Kristensen, A. Grøntved, Self-administered electroencephalography-based sleep assessment: compliance and perceived feasibility in children and adults, Sleep Science and Practice 5 (1) (2021) 8. doi:10.1186/s41606-021-00059-1.
URL https://doi.org/10.1186/s41606-021-00059-1

[10] J. B. Webster, D. F. Kripke, S. Messin, D. J. Mullaney, G. Wyborney, An activity-based sleep monitor system for ambulatory use, Sleep 5 (4) (1982) 389–399, pMID: 7163726. doi:10.1093/sleep/5.4.389.

[11] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, J. C. Gillin, Automatic sleep/wake identification from wrist activity, Sleep 15 (5) (1992) 461–469, pMID: 1455130. doi:10.1093/sleep/15.5.461.

[12] J. Palotti, R. Mall, M. Aupetit, M. Rueschman, M. Singh, A. Sathyanarayana, S. Taheri, L. Fernandez-Luque, Benchmark on a large cohort for sleep-wake classification with machine learning techniques, npj Digital Medicine 2 (1) (2019) 1–9, number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41746-019-0126-9.
URL https://www.nature.com/articles/s41746-019-0126-9

[13] E. Sazonov, N. Sazonova, S. Schuckers, M. Neuman, C. S. Group, Activity-based sleep-wake identification in infants, Physiological Measurement 25 (5) (2004) 1291–1304, pMID: 15535193. doi:10.1088/0967-3334/25/5/018.

[14] A. Sadeh, K. M. Sharkey, M. A. Carskadon, Activity-based sleep-wake identification: an empirical test of methodological issues, Sleep 17 (3) (1994) 201–207, pMID: 7939118. doi:10.1093/sleep/17.3.201.

[15] V. T. v. Hees, S. Sabia, K. N. Anderson, S. J. Denton, J. Oliver, M. Catt, J. G. Abell, M. Kivimäki, M. I. Trenell, A. Singh-Manoux, A novel, open access method to assess sleep duration using a wrist-worn accelerometer, PLOS ONE 10 (11) (2015) e0142533, publisher: Public Library of Science. doi:10.1371/journal.pone.0142533.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0142533

[16] K. Sundararajan, S. Georgievska, B. H. W. te Lindert, P. R. Gehrman, J. Ramautar, D. R. Mazzotti, S. Sabia, M. N. Weedon, E. J. W. van Someren, L. Ridder, J. Wang, V. T. van Hees, Sleep classification from wrist-worn accelerometer data using random forests, Scientific Reports 11 (1) (2021) 24, number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41598-020-79217-x.
URL https://www.nature.com/articles/s41598-020-79217-x

[17] M. R. Patterson, A. A. S. Nunes, D. Gerstel, R. Pilkar, T. Guthrie, A. Neishabouri, C. C. Guo, 40 years of actigraphy in sleep medicine and current state of the art algorithms, npj Digital Medicine 6 (1) (2023) 1–7, number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41746-023-00802-1.
URL https://www.nature.com/articles/s41746-023-00802-1

[18] J. A. Carlson, F. Tuz-Zahra, J. Bellettiere, N. D. Ridgers, C. Steel, C. Bejarano, A. Z. LaCroix, D. E. Rosenberg, M. A. Greenwood-Hickman, M. M. Jankowska, L. Natarajan, Validity of two awake wear-time classification algorithms for activpal in youth, adults, and older adults, Journal for the Measurement of Physical Behaviour 4 (2) (2021) 151–162, publisher: Human Kinetics Section: Journal for the Measurement of Physical Behaviour. doi:10.1123/jmpb.2020-0045.
URL https://journals.humankinetics.com/view/journals/jmpb/4/2/article-p151.xml

[19] E. Inan-Eroglu, B.-H. Huang, L. Shepherd, N. Pearson, A. Koster, P. Palm, P. A. Cistulli, M. Hamer, E. Stamatakis, Comparison of a thigh-worn accelerometer algorithm with diary estimates of time in bed and time asleep: The 1970

british cohort study, Journal for the Measurement of Physical Behaviour 4 (1) (2021) 60–67, publisher: Human Kinetics Section: Journal for the Measurement of Physical Behaviour. doi:10.1123/jmpb.2020-0033.
URL https://journals.humankinetics.com/view/journals/jmpb/4/1/article-p60.xml

[20] J. D. van der Berg, P. J. B. Willems, J. H. P. M. van der Velde, H. H. C. M. Savelberg, N. C. Schaper, M. T. Schram, S. J. S. Sep, P. C. Dagnelie, H. Bosma, C. D. A. Stehouwer, A. Koster, Identifying waking time in 24-h accelerometry data in adults using an automated algorithm, Journal of Sports Sciences 34 (19) (2016) 1867–1873, publisher: Routledge _eprint: https://doi.org/10.1080/02640414.2016.1140908 PMID: 26837855. doi:10.1080/02640414.2016.1140908.
URL https://doi.org/10.1080/02640414.2016.1140908

[21] E. A. H. Winkler, D. H. Bodicoat, G. N. Healy, K. Bakrania, T. Yates, N. Owen, D. W. Dunstan, C. L. Edwardson, Identifying adults' valid waking wear time by automated estimation in activpal data collected with a 24 h wear protocol, Physiological Measurement 37 (10) (2016) 1653, publisher: IOP Publishing. doi:10.1088/0967-3334/37/10/1653.
URL https://dx.doi.org/10.1088/0967-3334/37/10/1653

[22] P. J. Johansson, P. Crowley, J. Axelsson, K. Franklin, A. H. Garde, P. Hettiarachchi, A. Holtermann, G. Kecklund, E. Lindberg, M. Ljunggren, E. Stamatakis, J. Theorell Haglöw, M. Svartengren, Development and performance of a sleep estimation algorithm using a single accelerometer placed on the thigh: an evaluation against polysomnography, Journal of Sleep Research 32 (2) (2023) e13725, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.13725. doi:10.1111/jsr.13725.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.13725

[23] J. Skotte, M. Korshøj, J. Kristiansen, C. Hanisch, A. Holtermann, Detection of Physical Activity Types Using Triaxial Accelerometers, Journal of Physical Activity and Health 11 (1) (2014) 76–84, publisher: Human Kinetics, Inc. Section: Journal of Physical Activity and Health. doi:10.1123/jpah.2011-0347.
URL https://journals.humankinetics.com/view/journals/jpah/11/1/article-p76.xml

[24] D. Arvidsson, J. Fridolfsson, M. Börjesson, L. B. Andersen, . Ekblom, M. Dencker, J. C. Brønd, Re-examination of accelerometer data processing and calibration for the assessment of physical activity intensity, Scandinavian Journal of Medicine & Science in Sports 29 (10) (2019) 1442–1452, pMID: 31102474. doi:10.1111/sms.13470.

[25] M. G. B. Rasmussen, J. Pedersen, L. Olesen, S. Brage, H. Klakk, P. L. Kristensen, J. C. Brønd, A. Grøntved, Short-term efficacy of reducing screen media use on physical activity, sleep, and physiological stress in families with children aged 4–14: study protocol for the screens randomized controlled trial, BMC Public Health 20 (1) (2020) 380. doi:10.1186/s12889-020-8458-6.
URL https://doi.org/10.1186/s12889-020-8458-6

[26] E. L. Skovgaard, M. A. Roswall, N. Pedersen, K. T. Larsen, A. Grøntved, J. C. Brønd, Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings, Scientific Reports 13 (1) (2023) 2496, number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41598-023-29666-x.
URL https://www.nature.com/articles/s41598-023-29666-x

[27] O. Walch, Y. Huang, D. Forger, C. Goldstein, Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device, Sleep 42 (12) (2019) zsz180. doi:10.1093/sleep/zsz180.
URL https://doi.org/10.1093/sleep/zsz180

[28] B. C. Galland, B. J. Taylor, D. E. Elder, P. Herbison, Normal sleep patterns in infants and children: a systematic review of observational studies, Sleep Medicine Reviews 16 (3) (2012) 213–222. doi:10.1016/j.smrv.2011.06.001.

[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.
URL https://doi.org/10.1162/neco.1997.9.8.1735

[30] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, R. W. Picard, Multimodal ambulatory sleep detection using lstm recurrent neural networks, IEEE journal of biomedical and health informatics 23 (4) (2019) 1607–1617, pMID: 30176613 PMCID: PMC6837840. doi:10.1109/JBHI.2018.2867619.

[31] Z. Chen, M. Wu, W. Cui, C. Liu, X. Li, An attention based cnn-lstm approach for sleep-wake detection with heterogeneous sensors, IEEE journal of biomedical and health informatics 25 (9) (2021) 3270–3277, pMID: 32749983. doi:10.1109/JBHI.2020.3006145.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.
URL https://www.jair.org/index.php/jair/article/view/10302

[33] E. Hvitfeldt, themis: Extra Recipes Steps for Dealing with Unbalanced Data, r package version 1.0.1 (2023).
URL https://CRAN.R-project.org/package=themis

[34] M. F. Hjorth, J.-P. Chaput, C. T. Damsgaard, S.-M. Dalskov, K. F. Michaelsen, I. Tetens, A. Sjödin, Measure of sleep and physical activity by a single accelerometer: Can a waist-worn actigraph adequately measure sleep in children?, Sleep and Biological Rhythms 10 (4) (2012) 328–335, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1479-8425.2012.00578.x. doi:10.1111/j.1479-8425.2012.00578.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1479-8425.2012.00578.x

[35] C. A. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, W. C. Dement, Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients, Sleep Medicine 2 (5) (2001) 389–396, pMID: 14592388. doi:10.1016/s1389-9457(00)00098-8.

[36] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2023).
URL https://www.R-project.org/

[37] M. Kuhn, H. Wickham, Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.

(2020).
URL https://www.tidymodels.org

[38] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, Journal of Open Source Software 4 (43) (2019) 1686. doi:10.21105/joss.01686.

[39] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.
URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

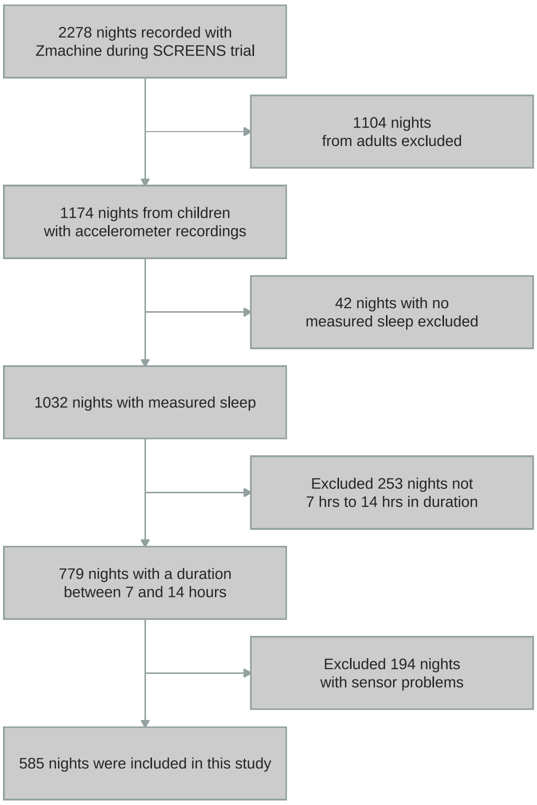
2278 nights recorded with
Zmachine during SCREENS trial
1104 nights
from adults excluded
1174 nights from children
with accelerometer recordings
42 nights with no
measured sleep excluded
1032 nights with measured sleep
Excluded 253 nights not
7 hrs to 14 hrs in duration
779 nights with a duration
between 7 and 14 hours
Excluded 194 nights
with sensor problems
585 nights were included in this study

Figure 1: Flowchart of eligible ZM recording nights included in the study

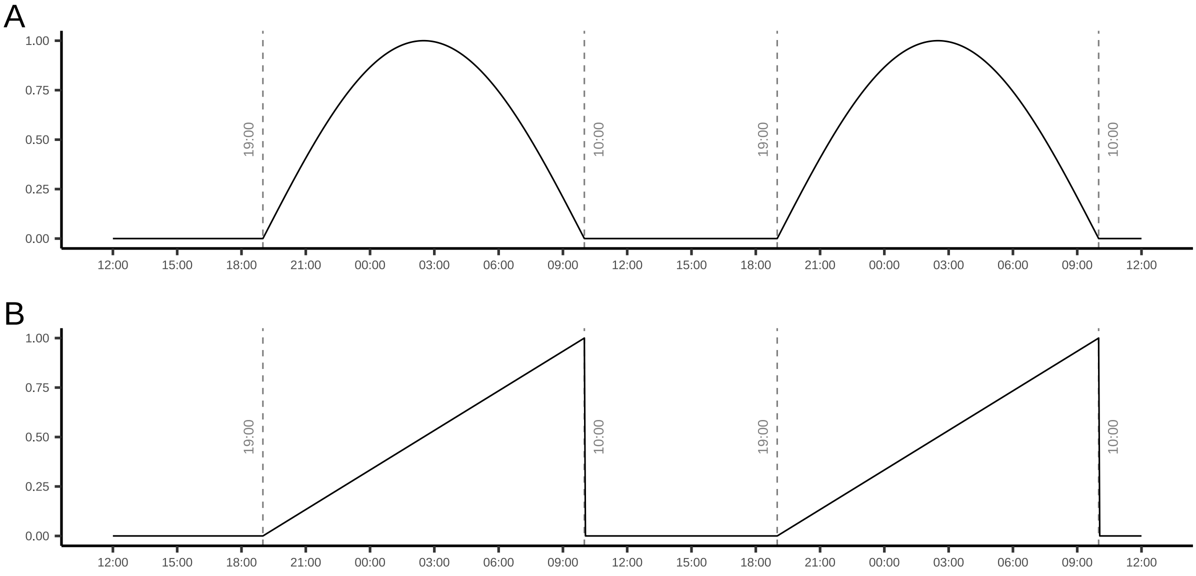

Figure 2: Sensor-independent features of circadian rhythms across two consecutive nights. A) cosinus feature, B) linear feature.

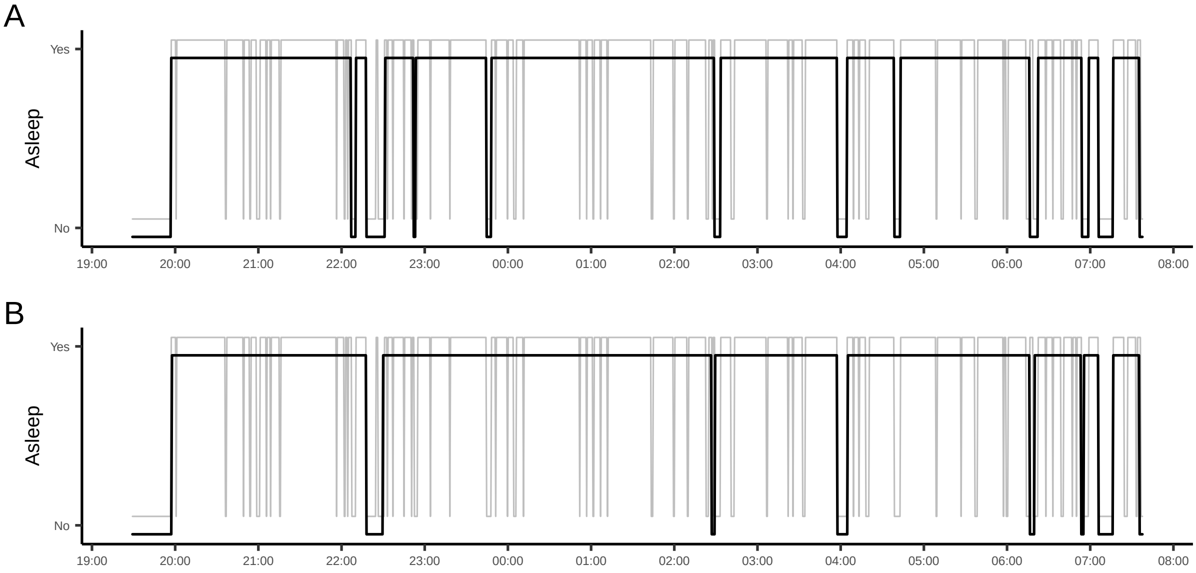

Figure 3: The difference in number of awakenings between the raw ZM predictions vs. 5-minute, and 10-minute median filtered predictions for a random night. Grey line is the raw predictions, black line is the median filtered predictions. A: 5-minute median filter on raw ZM predictions, B: 10-minute median filter on raw ZM predictions.

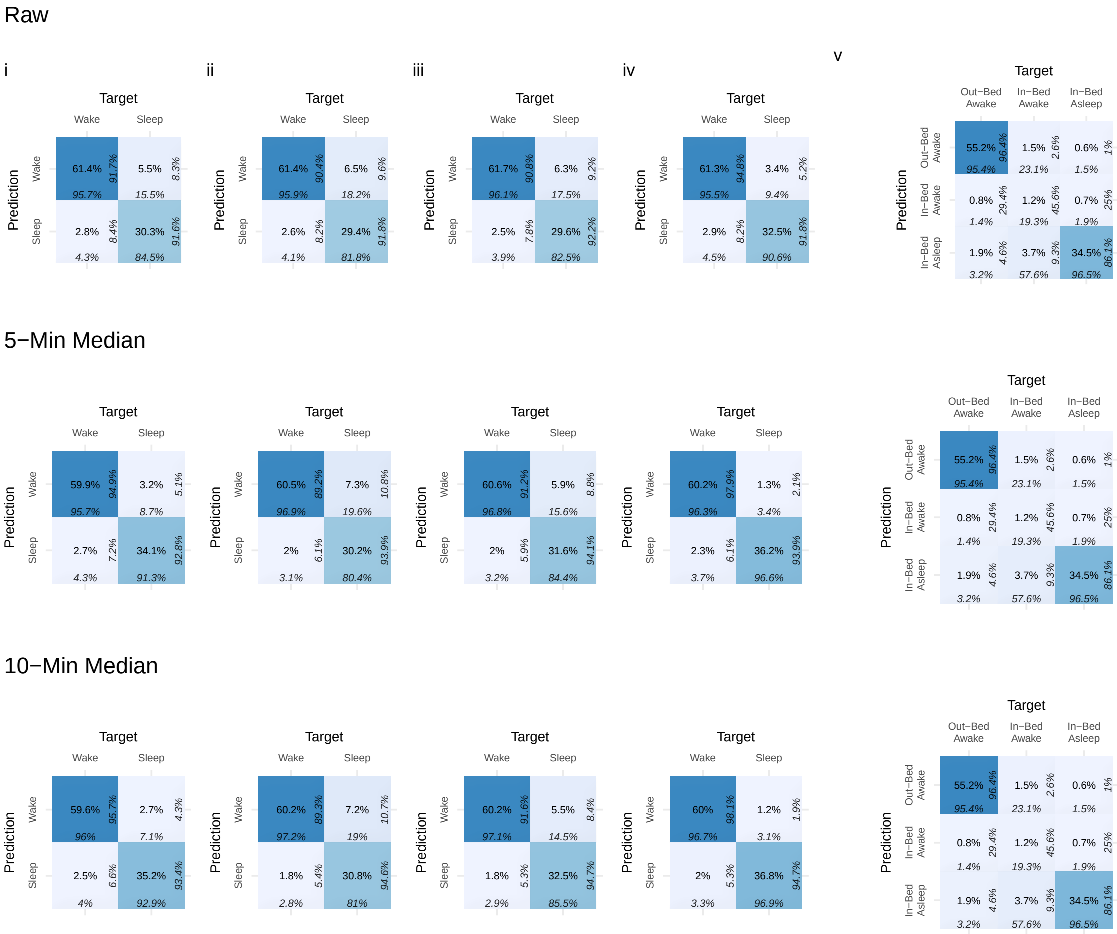

Figure 4: Confusion matrices for binary sleep prediction. The middle of each tile is the normalized count (overall percentage) and, beneath it, the count. The bottom number is the column percentage (target). At the right side of each tile is the row percentage (prediction). i) decision tree, ii) logistic regression, iii) feed-forward neural net, iv) xgboost, and v) biLSTM.

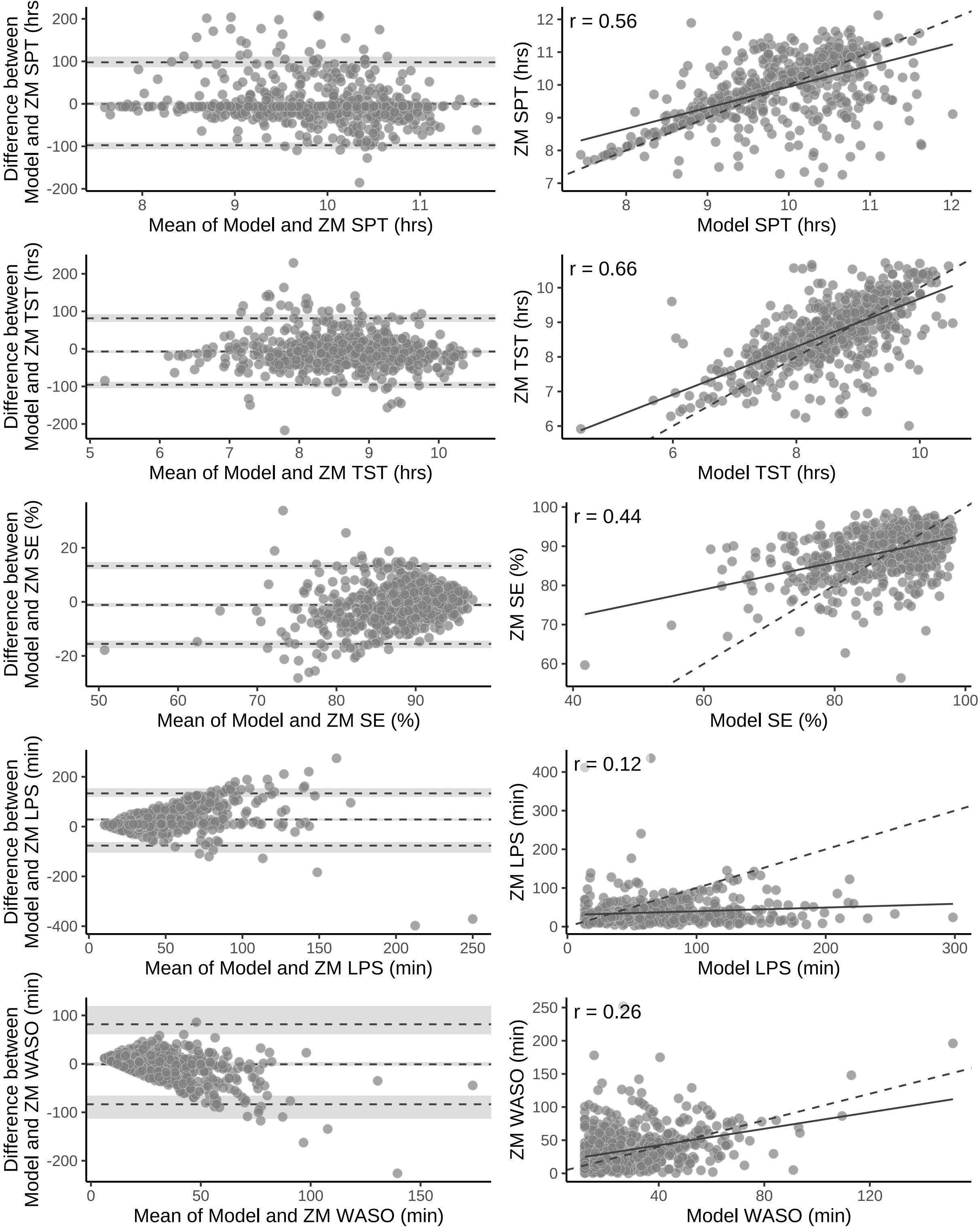

Figure 5: Comparison of sleep quality parameters derived from the XGboost model trained on the 5-minute smoothed ZM predictions. The left column displays Bland-Altman plots. Dashed lines represent the bias (the average difference between the two measurements) and limits of agreement, with the 95% confidence intervals represented as the grayed areas. The right column displays scatter plots of XGboost-derived vs ZM-derived sleep quality parameters. The dashed line represents the identity line, while the full-drawn line represents the best linear fit. Pearson's correlations are annotated in the upper left corner.