Improving Sleep Quality Estimation in Children and Adolescents: A Comparative Study of Machine Learning and Deep Learning Techniques Utilizing Free-Living Accelerometer Data from Thigh-Worn Devices and EEG-Based Sleep Tracking

ESBEN HØEGHOLM LYKKE

University of Southern Denmark
eskovgaard@health.sdu.dk

ANDERS GRØNTVED

University of Southern Denmark
agroentved@health.sdu.dk

PETER LUND KRISTENSEN

University of Southern Denmark
plkristensen@health.sdu.dk

JESPER SCHMIDT-PERSSON

University of Southern Denmark
jespedersen@health.sdu.dk

SARA OVERGAARD SØRENSEN

University of Southern Denmark
sosorensen@health.sdu.dk

SOFIE RATH MORTENSEN

University of Southern Denmark
srmortensen@health.sdu.dk

JAN CHRISTIAN BRØND

University of Southern Denmark
jbrond@health.sdu.dk

2023-08-30

Abstract

Accurate assessment of sleep is vital in sleep research, but the gold standard, polysomnography, is costly and impractical for large-scale studiesand multiple consecutive days of assessment. An affordable alternative is using wearable accelerometers. While wrist and hip-worn devices are commonly used in sleep research, thigh-worn accelerometers have been relatively unexplored. Our study evaluated machine learning and deep learning models utilizing data from thigh-worn accelerometers to estimate sleep and sleep quality metrics, comparing them with an EEG-based sleep monitor. The dataset consisted of data from 585 days and nights, comprising accelerometry and EEG-based sleep estimates from children aged 4-17 years. We employed both sequential and multiclass model strategies on both raw and filtered data. The most effective model was XGBoost, which performed well when applied to 5-minute median filtered data, exhibiting small mean differences (bias) in sleep period time (0.2 minutes), total sleep time (-7.0 minutes), sleep efficiency (-1.1%), and wake after sleep onset (-0.9 minutes). Furthermore, the XGBoost model showed a robust correlation (0.66, 95% CI: 0.61 - 0.7) with total sleep time, indicating its potential. However, despite these promising results in bias, our study revealed limits of agreements (e.g., total sleep time LoA (95%CI): -95.5 (-105.2;-88) minutes to 81.4 (72.4;92.5) minutes) in accordance with previous research on hip- and wrist-worn devices. In conclusion, we present promising results in using machine learning technieques to estimate sleep quality metrics on a group level, however, accurately classifying awake periods during in-bed time remained challenging. Moreover, additional improvements are necessary to accurately assess individual sleep quality metrics based on thigh-worn accelerometry data due to the notable limits of agreement.

I. Introduction

A vast body of research highlights the critical role of sleep in maintaining both mental and physical health[1–4]. Consequently, accurate sleep assessment methods are crucial for tracking sleep patterns and improving our understanding of the sleep-health relationship. Furthermore, the ease of use and high acceptability of methods to assess sleep are essential to facilitate large-scale, longitudinal studies.

The traditional gold standard for objective sleep measurement, laboratory-based polysomnography (PSG), has been found to be impractical in large-scale observational- and experimental studies due to its high cost, need for professional administration, and susceptibility to rater bias[5,6], although recent advances has been made to automate the scoring of PSG data[7]. As an alternative, diaries have been used due to their cost-effectiveness and simplicity, although they are subject to recall bias and other limitations[8]. An innovative approach involves device-based measurement methods. These tools, which estimates a number of sleep metrics including sleep duration, are advantageous due to their reduced participant burden and elimination of potential recall biases. A prominent example of such tools is body-worn accelerometers, which offer a practical and affordable means of objectively assessing sleep patterns at home for extended periods. Accelerometers collect continuous, high-resolution data for several weeks without requiring recharging, further minimizing participant burden. Their use in sleep and wake classification began with a wrist movement-based algorithm developed in 1982, and validated using PSG[9]. This algorithm was refined in 1992[10], leading to the widely adopted Cole-Kripke model. With advancements in the field, a variety of techniques, including heuristic algorithms, machine learning models, regression, and deep learning, are now used to analyze data from hip and wrist-worn accelerometers[10–15].

While wrist and hip-worn devices have benefited from extensive methodological development, thigh-worn accelerometers have not seen the same level of advancement. Existing studies mainly focus on distinguishing sleep from wakefulness, with emphasis on defining 'waking time' and 'bedtime'[16–19]. Recent strides in estimating sleep duration using thigh-worn devices have been made, including the introduction of a promising algorithm and its comparison against PSG[20]. Despite these advancements, the application of machine learning techniques in this area is still unexplored. Considering the potential of thigh-worn accelerometers for accurate physical behavior assessment[21–23], there is a significant research gap. Therefore, future studies need 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.

The Zmachine® Insight+ (ZM) emerges as a valuable tool within this landscape. Favorably validated against PSG[24,25], the ZM provides comparable data without the high costs or the need for professional monitoring typically associated with PSG. Crucially, the ZM facilitates multi-night analysis in free-living conditions due to its ease of use[26], capturing the natural variations in sleep patterns. This makes it advantageous over single-night PSG, particularly as a gold standard data source in machine learning tasks, as it provides multiple nights of measurements without inter-rater bias. Despite these benefits, the ZM, like PSG, still poses a significant participant burden and cost, reinforcing the need for more accessible alternatives like accelerometers.

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 electroencephalography-based (EEG) sleep tracking device, which we, in this current study, considered as the criterion measure for assessing sleep. Furthermore, our secondary goal was to assess the developed models' performance in evaluating important sleep quality metrics, including sleep period time (SPT), total sleep time (TST), sleep efficiency (SE), latency until persistent sleep (LPS), and wake after sleep onset (WASO).

II. Methods

i. Dataset and Participants

The current study leverages data from the SCREENS trial[27,28], a study conducted from June 2019 to March 2021 in the Region of Southern Denmark, Southern Denmark, that evaluated the effect of limiting screen media usage within Danish families. For our analysis, we focused on data from child participants aged between 4 and 17 years within the SCREENS cohort (mean age 9.1 years). Our primary sources of

data were accelerometer readings from Axivity AX3 devices attached to the children's thighs, and EEG-derived sleep states and sleep quality metrics 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 participants of the SCREENS study were instructed to attach the device when they went to bed and remove the device upon leaving the bed. 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[24], while the Z-PLUS effectively differentiates sleep stages, as evidenced by its alignment with expert evaluations using PSG data[25]. In the current study, we treated all sleep stages (light sleep (N1 & N2), deep sleep (N3), and REM sleep) 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 the sleep quality metrics of interest and to simplify the learning process of the machine learning algorithms.

Figure 1 illustrates the selection criteria applied to the children's recordings from the SCREENS study. Only ZM recordings accompanied by complete accelerometer data and lasting between 7 and 14 hours were considered. Nights during which the ZM reported sensor issues were excluded. Consequently, a total of 585 nights from 151 children were included in the study, with a mean of 3.87 nights per child (SD = 1.86). The children whose recordings were considered had an average age of 9.4 years, with a standard deviation of 2.1. 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 classes during the nightly recordings.
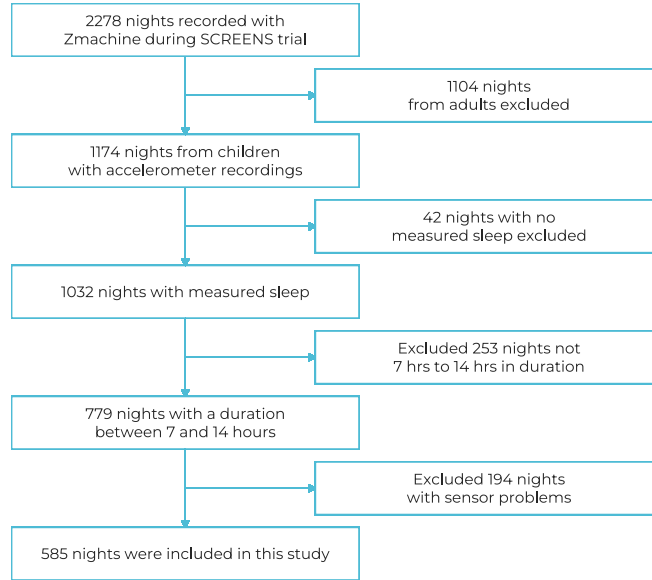

Figure 1: *Flowchart of eligible ZM recording nights included in the study.*

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.

ii.   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 similar to methods described by Skotte et al.[21]. Any non-wear data was removed using previously described methods[29] and data was resampled to 30-second epochs so every sample classified by the models corresponds to a 30-second epoch scored during the ZM recordings. Subsequently, we performed a feature

extraction process that yielded a set of 64 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 epochs. Furthermore, inspired by Walch et al.[30], 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. 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.
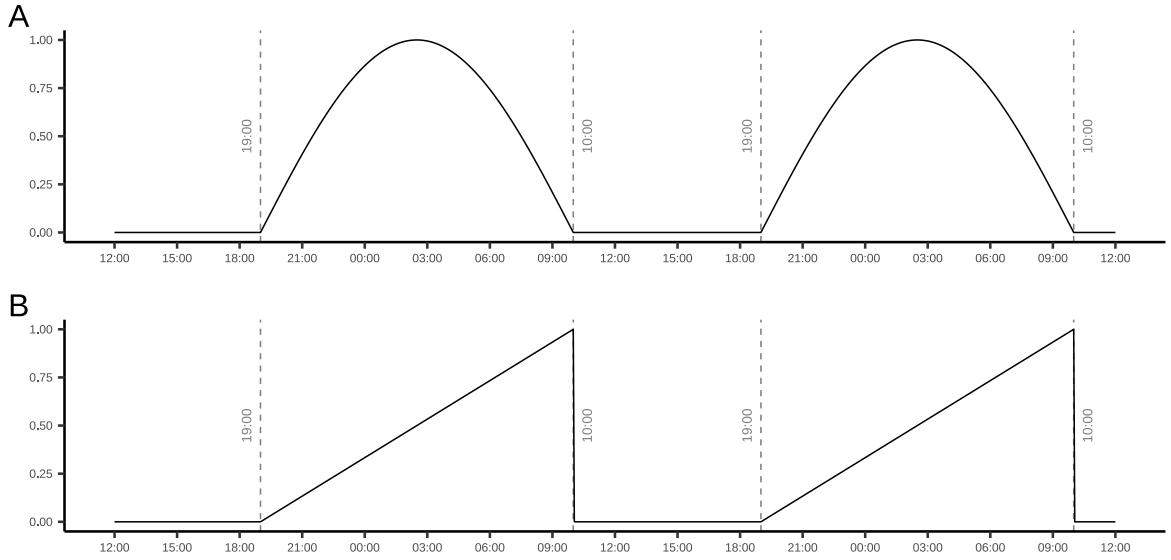

Figure 2: *Sensor-independent features of circadian rhythms across two consecutive nights. A) cosinus feature, B) linear feature.*

In addition to the engineered features, we chose to incorporate the median-filtered raw predictions from the ZM device into our modeling process. This choice was informed by the understanding that children typically experience around five to eight sleep cycles per night, with awakenings most likely at the end of each cycle[31]. In examining the raw ZM predictions, we observed a notable overestimation in the number of awakenings for the children in our study, surpassing expected counts based on typical sleep cycle patterns (refer to Figure 3). The average sleep efficiency determined by the ZM for our sample was 83%, which falls slightly below the recognized standards—85% is seen as good efficiency, and above 90% as ideal. In fact, prior research has indicated sleep efficiencies of over 90% in similar child cohorts[32,33]. This discrepancy suggests that the raw ZM predictions might be overestimating awake periods. 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 algorithms by obscuring the underlying patterns that the algorithms 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, which to the best of our knowledge provided a more accurate depiction of children's sleep patterns (see Table 1).

iii.   Algorithms

We employed two different model strategies to assess sleep patterns from thigh-mounted accelerometer data. The first model strategy was designed as a sequence of two models, each functioning as a binary
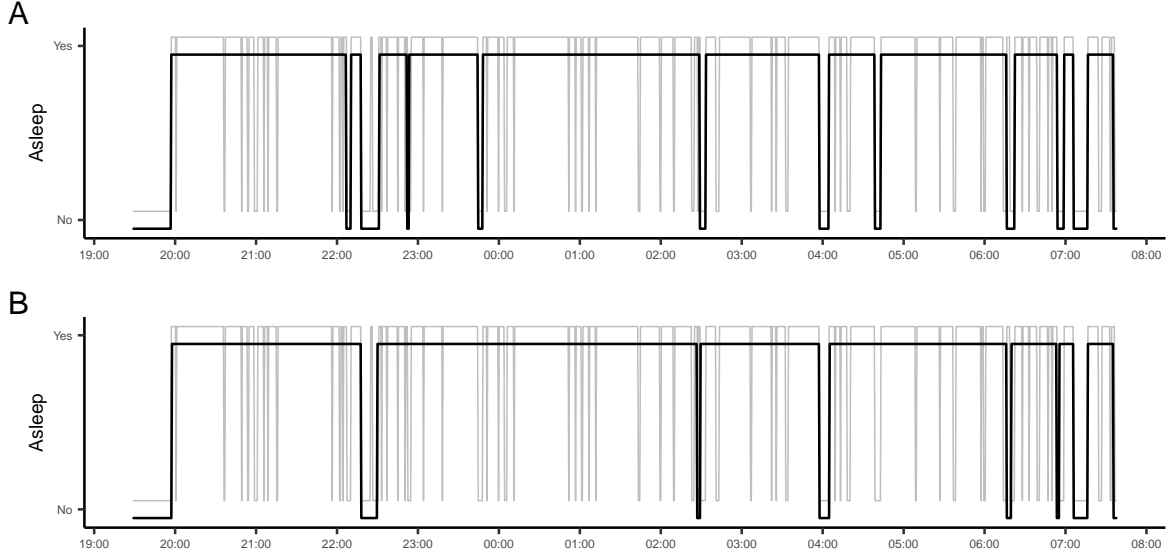
Figure 3: *The difference in number of awakenings between the raw ZM predictions vs. 5-minute, and 10-minute median filtered predictions for a random night (boy, 9 years). Grey line is the raw predictions, black line is the median filtered predictions. A: 5-minute median filter on raw ZM predictions, B: 10-minute median filter on raw ZM predictions.*

classifier. This approach aimed to simplify the prediction task by decomposing the multiclass problem of classifying 'out-of-bed-awake', 'in-bed-awake', and 'in-bed-asleep' into two binary stages: first predicting 'in-bed' time, then 'sleep' time. The output from the first set of binary classifiers, which predicted in-bed time, was subjected to a 5-minute median filter to remove transient in-bed time blips. This process enabled us to establish a single continuous time interval that we identified as the SPT, the total time spent in bed attempting to sleep. The SPT then served as the input for the second stage of binary classifiers in the sequence, further enhancing their predictive accuracy for sleep time. We applied this sequential strategy using the following four machine learning algorithms:

1. Logistic Regression: 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: 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. To combat this, we used a maximum tree depth of 8.

3. Single-layer Feed-forward Neural Network: 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: 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.

In parallel, we also employed a multiclass algorithm as the second model strategy using a bidirectional Long Short-Term Memory (biLSTM)[34] neural network which also incorporates temporal aspects of the data. This network, which was designed to predict three distinct classes: 'out-of-bed-awake', 'in-bed-awake', and 'in-bed-asleep', was configured with four layers and 128 hidden units per layer. This balance between model complexity and training efficiency was intended to facilitate learning of intricate patterns

5

while ensuring feasible training times. The bidirectional nature of the LSTM enhanced data interpretation and reduced overfitting by doubling the hidden units at each time step. The LSTM model used sequences of tensors as input, with each sequence spanning 10 minutes and a step size of one epoch. As demonstrated by previous studies such as those by Sano et al.[35] and Chen et al.[36], LSTM models have shown great promise in sleep detection using accelerometer data, thanks to their ability to capture complex temporal patterns.

iv.   Model Training

We trained four pairs of models in sequence, with each pair distinguishing between in-bed/out-of-bed and asleep/awake states, respectively. We divided our dataset randomly into a training set and a testing set, with each containing roughly half of the subjects. The splitting of the data was ensured to not have samples from the same night simultaneously present in both sets. To optimize hyperparameters, we performed a 10-fold Monte Carlo cross-validation on a regular grid (i.e., for each hyperparameter, a range of values at evenly-spaced intervals was selected) 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 training data to estimate the model parameters. An imbalance was observed with the in-bed time determined in the initial step of the sequential model strategy which 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 favor predicting the majority class. To account for this imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE)[37]. SMOTE generates new samples by interpolating random samples with their nearest neighbors. We utilized the themis R package[38] to implement SMOTE, resulting in a balanced distribution of training samples across both classes.

The biLSTM model was trained to differentiate between three classes: 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. Again, 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.

v.   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

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$precision = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity}$$

where NPV is negative predictive value, F_1 is the F1 score, TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives.

In the context of our sequential model 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 model 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 the same metrics. To do this, we considered the multiclass output as to binary classifications, where the first was out-of-bed vs the rest and the second binary classification as in-bed-awake vs in-bed-asleep. 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[39,40].

To assess the performance of our models in deriving sleep quality metrics, 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. Given the nature of our dataset, which contains multiple observations per subject but not necessarily equal number of observations, we employed a bootstrap procedure to account for this added variability. We first calculated the mean difference (bias) and then defined the limits of agreement (LOA) as the mean difference plus or minus 1.96 times the standard deviation of these differences. Acknowledging the possibility of non-normality and potential skewness in our data, we chose to apply a bias-corrected and accelerated (BCa) bootstrap method[41]. This approach allowed us to better address potential bias in our estimates and the inherent intra-subject variability. Utilizing 10,000 bootstrap replicates, we estimated the 95% confidence intervals for both the bias and the LOA, thus ensuring robustness in our measurements. The sleep quality metrics included are defined as follows in accordance with the ZM definitions:

1. Sleep Period Time (SPT) - This refers to the total duration of time in bed with the intention to sleep, 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 the first 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)[42] and the Tidymodels[43] and Tidyverse[44] suite of packages were used as the core tools for model development and analyses. Python version 3.10.6[45] and PyTorch[46] were used to implement the biLSTM model.

III. Results

As reported in Table 1 the sleep quality metrics 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 $82.6 \pm 12.0\%$ to $86.4 \pm 12.7\%$ and $87.5 \pm 12.9\%$ respectively. LPS also increased, suggesting the filter removes brief awakenings at sleep onset, leading to a prolonged time to persistent sleep. A 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 (585 nights from 151 children). Values are represented as mean (SD). Hrs: hours, min: minutes.*

	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)

i. 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 different models. The XGBoost model provide the best performance with an F1 score of 95.4% and accuracy of 96.1%, although only outpacing the other models marginally.

Table 2: *Performance metrics of the classification of in-bed/out-of-bed time of the included models.*

	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
biLSTM	95.2	95.3	95.3	95.1	95.3

Table 3 details the performance of all sequential model types on raw and median-filtered (5 and 10 minute) ZM predictions for sleep/wake classification. For raw ZM predictions, the F1 scores, which are unweighted macro averages, range from 65.6% (biLSTM) to 76.2% (XGBoost). All models perform comparably, but the low specificity values (62.5% to 70.9%) suggest difficulty in correctly classifying awake epochs. Applying a 5-minute median filter improves the performance metrics. The XGBoost model tops the charts with an F1 score of 79.2% and NPV of 74.0%. However, specificity still remains low, with values between 54.7% (XGBoost) and 74.8% (Logistic Regression) across all models. With a 10-minute median filter, the metrics improve further. The XGBoost model still leads with an F1 score of 80.9% and an NPV of 75.9%. But, specificity remains low, ranging from 57.5% (Decision Tree) to 76.4% (Logistic Regression) across all models.

Table 3: *Performance metrics of the sleep/wake classification of the included models.*

	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
biLSTM	65.6	86.1	75.0	98.1	26.9
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
biLSTM	70.3	87.3	81.8	98.4	33.9
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
biLSTM	70.9	89.2	61.0	94.6	42.4

A complete set of confusion matrices generated from data both containing the out-of-bed and in-bed time are presented in Figure 4. These matrices showcase the epoch-to-epoch performance of all sequential models in distinguishing between 'awake' and 'asleep' states, regardless of whether the subject is 'in-bed' or 'out-of-bed'. However, it's important to note that the binary nature of these sequential models means they cannot provide direct information about the classification of the 'in-bed-awake' state. In contrast, the biLSTM model, which also categorizes the 'in-bed-awake' state as a distinct class, appears to have less success in classifying this particular state.
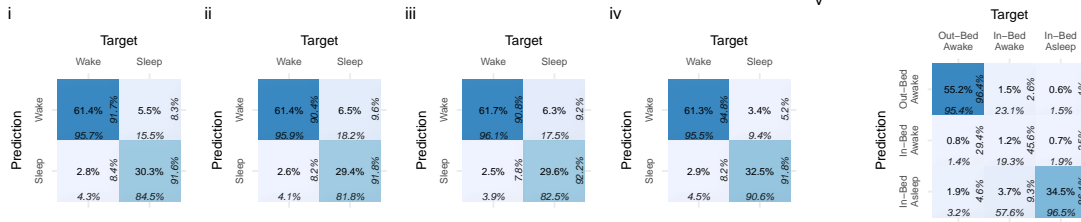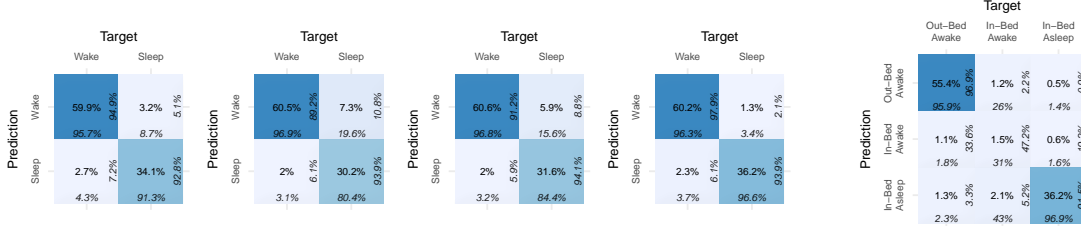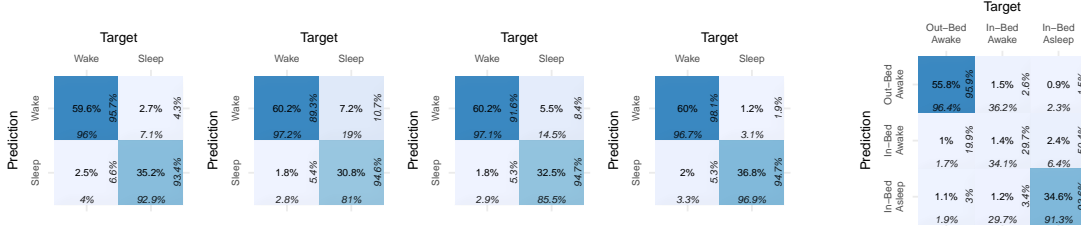

Figure 4: *Confusion matrices for binary prediction and multiclass prediction. The middle of each tile is the normalized count (overall percentage). The bottom number of each tile is the column percentage and the right side of each tile is the row percentage. i) decision tree, ii) logistic regression,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

ii. Evaluation of sleep quality metrics

Table 4 presents a comparative analysis of the included models used to predict various sleep quality metrics (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 table 1 in supplementary materials. 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 feed-forward neural network also exhibited similar bias as the decision tree and the logistic regression models,

9

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 in the differences across different sleep quality metrics 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 quality metrics 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 4: *Summary of bias, limits of agreement, and Pearson correlation for various sleep parameter predictions (SPT, TST,SE, LPS, WASO) using different machine learning and deep learning models (decision tree, logistic regression, feed-forward neural network, XGBoost) on raw ZM predictions, 5-minute and 10-minute median predictions. Each value is provided with its 95% confidence interval (CI).*

	Bias (95% CI)	Lower LOA (95% CI)	Upper 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 between the XGBoost model, trained on 5-minute median filtered ZM predictions, and the 5-minute median-smoothed ZM-derived sleep quality metrics. The Bland-Altman plot for the SPT and TST indicates a minimal average difference with the ZM, as evidenced by a bias close to zero. The scatterplot for SPT also demonstrates a positive trend, indicating a moderate linear correlation between the XGBoost model and the ZM-derived sleep quality metrics. The bias and LOA for TST are comparable to those observed for SPT, indicating a consistent level of agreement between the two methods. The scatterplot for TST also shows a slightly higher correlation, primarily driven by the absence of extreme outliers.Furthermore, the remaining three sleep quality metrics, SE, LPS, and WASO, exhibit heteroscedasticity in contrast to SPT and TST. A moderate positive linear correlation is observed between the XGBoost model and ZM-derived sleep quality metrics for SE, however, a poor correlation is observed for LPS and WASO.
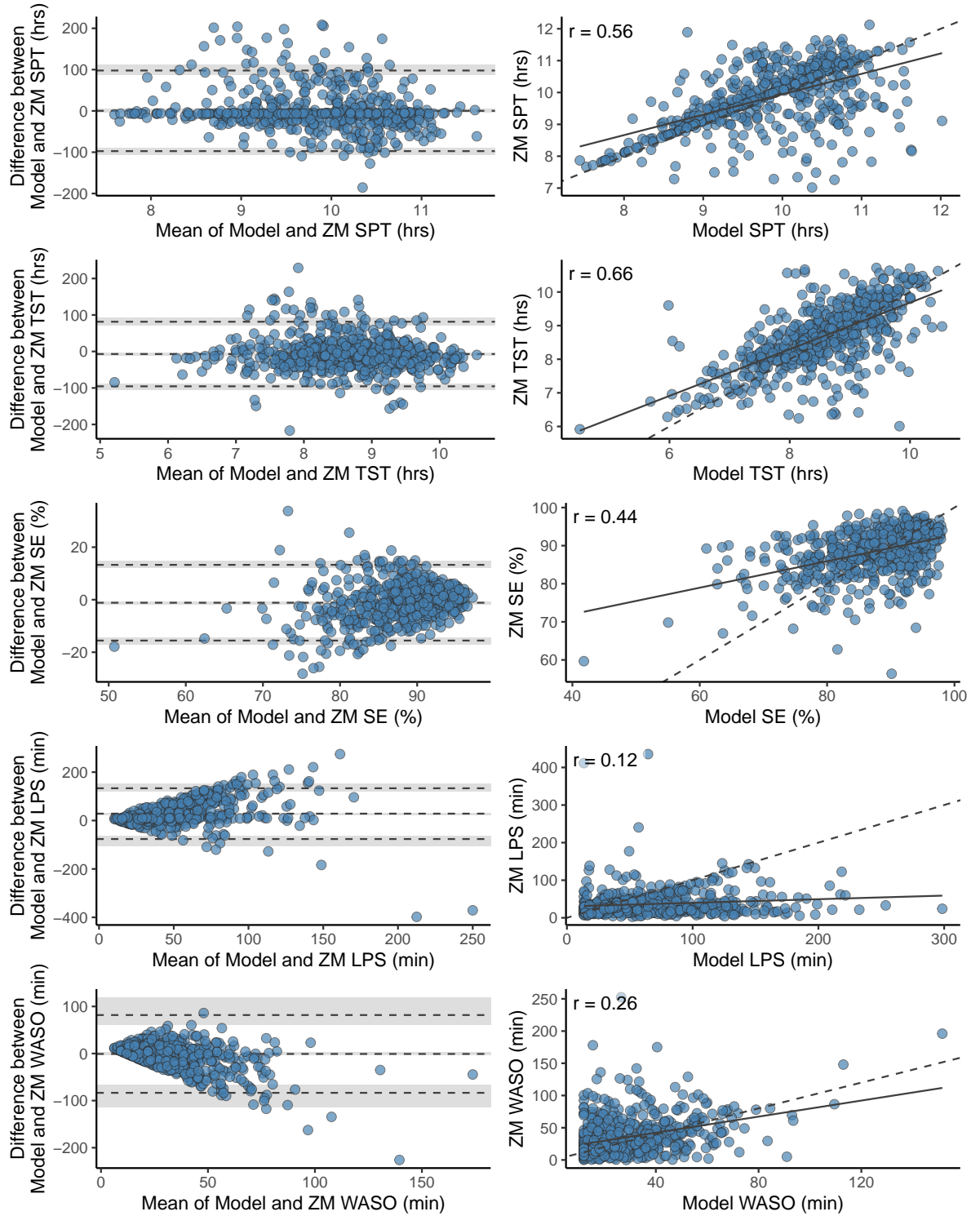
Figure 5: *Comparison of sleep quality metrics derived from the XGBoost model trained on the 5-minute smoothed ZM predictions. The left column displays Bland-Altman plots. Dashed lines represent the bias (the average difference between the two measurements) and LOA, with the 95% confidence intervals represented as the grayed areas. The right column displays scatter plots of XGBoost-derived vs ZM-derived sleep quality metrics. 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.*

IV. Discussion

To select the most optimal method for estimating sleep from thigh-worn accelerometers, we evaluated various models for predicting in-bed and sleep time and their derived sleep quality metrics. We trained and evaluated the models using raw and median-filtered sleep estimates from the ZM EEG-based sleep monitor. In general, all sequential models performed well at predicting in-bed time. More challenging was it to distinguish wake from sleep on the extracted in-bed time. Moreover, even though the multiclass biLSTM showed good performance across F1 score, precision and NPV, the derived sleep quality metrics were not on par with the XGBoost model which demonstrated the highest performance metrics across all evaluations, including epoch-to-epoch prediction and all sleep quality metrics. Despite this, all models showed low specificity values, indicating difficulty in correctly classifying awake epochs during time in bed. The application of 5-minute and 10-minute median filters improved the performance metrics of all models. Median filtering increase total sleep time and sleep efficiency, while reducing wake after sleep onset and the number of awakenings. The XGBoost model provided the smallest bias and highest correlation with all ZM sleep quality metrics.

Limited research exists regarding the epoch-to-epoch effectiveness of classifying in-bed time based on data from thigh-worn accelerometers. Nevertheless, Carlson and colleagues provided compelling insights. They demonstrated that a third-party algorithm, "ProcessingPal," and a proprietary one, "CREA," achieved accuracies of 91% and 86% respectively. These algorithms, evaluated against self-reported measures among adolescents and adults[16], produced F1 scores as high as 95% and 96%. These figures are consistent with the performance of our sequential models, which also achieved F1 scores and accuracy scores exceeding 95% in identifying in-bed time. In our study, in-bed time is equated with SPT. All models, with the exception of XGBoost, underestimated SPT. The biLSTM model showed the greatest underestimation, with a bias of -36 minutes, reflecting trends observed in previous research by Winkler et al. conducted in young- middle-aged and older adults[19]. They developed an algorithm that, despite a moderate correlation (Pearson correlation coefficient = .67) between their algorithmic results and diary-recorded waking times, overestimated waking wear time by more than 30 minutes, resulting in an underestimation of in-bed time[19]. This trend was further confirmed when Inan-Eroglu et al. examined Winkler et al.'s algorithm, revealing a underestimation of 9.8 minutes in bed time compared to self-reported measures in middle-aged adults[17]. In contrast, a study by Berg et al. reported a slight underestimation of in-bed time in a sample of middle-aged and older adults. They employed a unique approach with their algorithm, which relied on quantifying the number and duration of sedentary periods to determine time in bed, and active periods (standing or stepping) to identify wake times[18]. Finally. it is important to note that predictive performance in determining in-bed time does not necessarily translate to accurate predictions of broader sleep quality metrics. The crucial task of detecting awake periods during in-bed time, a key factor in assessing further derived sleep quality metrics, is not effectively captured by in-bed time predictions alone. Furthermore, the distinction between actual sleep and time spent in bed awake, often overlooked but vital in sleep research, is critical for a comprehensive understanding of sleep quality.

To the best of our knowledge, Johansson and colleagues[20] are the only researchers who have reported epoch-to-epoch performance metrics for sleep scoring using thigh-worn accelerometers, beyond just "waking time" and "in-bed time." They achieved a mean sensitivity of 0.84, specificity of 0.55, and accuracy of 0.80, using a single-night evaluation dataset of 71 adult subjects. Despite our models achieving a sensitivity above 97%, they, like Johansson et al.'s algorithm, struggled with detecting in-bed awake epochs. This is reflected in the low specificity scores, ranging from 54.7% to 76.4%, reported in our study. The challenge of low specificity is not unique to methods using data collected from thigh-worn devices. Conley et al.'s meta-analysis[47] reported similar findings when estimating sleep using wrist-worn accelerometers among healthy adults, with a mean sensitivity, accuracy, and specificity of 0.89, 0.88, and 0.53, respectively. Furthermore, Patterson and colleagues[48] recently summarized the performance of various heuristic algorithms, machine learning, and deep learning models used to predict sleep. They found the mean sensitivity and specificity to be 93% (SD = 2.8) and 60% (SD = 11.1) respectively. These findings underscore the challenge of automating the detection of in-bed awake periods. Interestingly, despite low specificity values for most of our models and configurations, we observed an overestimation from several of our models of LPS and WASO, contrasting with most previous research[11,47]. This overestimation of wake epochs is evident from the low NPV scores, indicating that only a small proportion of the wake predictions are actually correct. This discrepancy may be driven by the SMOTE process

used to balance the dataset. If the synthetic "wake" samples created by SMOTE are not representative of the true "wake" data, the models might learn to incorrectly classify certain "sleep" epochs as "wake". This could lead to an overestimation of LPS and WASO, as the models are incorrectly identifying more periods of wakefulness during the sleep period.

The use of the SMOTE technique likely improved the performance of our models by addressing the class imbalance in our data. However, this technique also introduced synthetic"wake" samples that may not be fully representative of true wake data. This could potentially lead some models to overestimate the wake class. Interestingly, the biLSTM model, which was not trained on SMOTE-processed data, was the only one to overestimate TST and SE. On the other hand, the XGBoost model, which was trained on data subjected to the SMOTE process, was able to handle the synthetic "wake" samples better than the other models, and it did not overestimate TST to the same degree. The Bland-Altman statistics for the XGBoost model trained on the 5-minute median filtered ZM predictions showed a mean difference of -7 minutes for TST and -1.1% for SE, with limits of agreement ranging from -95.5 to 81.4 minutes and from -15.6% to 13.3% respectively. This suggests that the XGBoost model was able to maintain a balance between sensitivity and specificity, and it was not overly influenced by the synthetic "wake" samples. The XGBoost model's success with the SMOTE dataset may be due to its ability to handle non-representative synthetic samples. XGBoost's gradient boosting mechanism allows it to iteratively learn from the errors of previous models, which can help it to better distinguish between true wake data and non-representative synthetic wake samples created by SMOTE. This iterative learning process could make XGBoost more robust to the inaccuracies introduced by the synthetic samples, leading to better overall performance.

Typically, sleep detection methods are applied in two contexts: either to night recordings or to 24-hour recordings. In night recordings, it is possible to derive sleep quality metrics like SE and LPS because the SPT is already known because it is inferred from the length of the recording[47,48]. On the other hand, when sleep detection methods are applied to 24-hour recordings, most methods do not have the ability to infer the SPT without sleep diaries[49]. Consequently, these methods are unable to generate sleep quality metrics that rely on the SPT[50,51]. To overcome this limitation, we have incorporated models that can differentiate between in-bed awake time and in-bed asleep from out-bed awake time over a 24-hour recording. This approach allows our models to estimate all commonly used sleep quality metrics. Van Hees et al.[52] have proposed an algorithm to determine SPT from data collected by wrist-worn devices. This algorithm was recently validated by Plekhanova and her team[53]. By combining this algorithm with other methods, further sleep quality metrics can be inferred based on the identified SPT. Van Hees et al.[52] reported good agreements and low mean differences compared to self-report and PSG on SPT, findings later confirmed by Plekhanova and colleagues. However, they also observed poor agreement with LPS and Wake After Sleep Onset (WASO). They found low reliability with PSG, indicating difficulties in detecting wakefulness during in-bed time. These challenges parallel those we experienced in our study.

In our evaluation of sleep quality metrics, we found that LPS had the largest mean error relative to absolute time allocated to LPS. This suggests that the initial epochs of Sleep Period Time (SPT) are particularly challenging to classify correctly. This is also supported by the poor Pearson correlations between LPS derived from model predictions and the ZM. The XGBoost model, which was the best performer among all models, overestimated LPS by an average of 26.4 minutes for models trained on raw ZM predictions, 28.5 minutes for models trained on 5-minute filtered ZM predictions, and 34.5 minutes for models trained on 10-minute filtered ZM predictions. This level of discrepancy is comparable to the mean error of sleep latency of 23 minutes reported by Johansson et al.[20]. Johansson et al. suggest that the discrepancy with the gold standard is likely due to the multifaceted nature of the sleep state, which is a complex physiological process. Short awakenings or sleep episodes may not necessarily correspond to noticeable changes in thigh movement, making them difficult to detect and accurately classify. These results align with several methods for wrist-worn devices reviewed by Conley and colleagues[47]. They reported correlations between accelerometer and PSG sleep onset latency (equivalent to LPS) from 10 studies with a mean correlation of 0.2 (ranging from -0.69 to 0.69), indicating the inherent difficulty in estimating LPS using accelerometry alone.

Our study's XGBoost model demonstrated relatively narrower LOAs for TST, SE, and WASO, with ranges of -95.5 to 81.4 min, -15.6 to 13.3%, and -83.4 to 81.7 min, respectively when compared with other models such as the Van Hees algorithm[14], Oakley rsc (rescored)[11], and LSTM-50[11] evaluated in the Patterson et al. study[48]. Furthermore, comparing the LOAs between our XGBoost model and the

algorithm developed for thigh-worn devices by Johansson et al. study[20], our XGBoost model showed narrower LOAs for TST , SE, LPS , and WASO, but not SPT. Generally, all methods, both from this study and from the reviewed literature, exhibit wide LOAs suggesting that there is high variability in the derived sleep quality metrics, and accelerometry cannot be used interchangeably as an alternative to the EEG-based ZM or PSG to measure sleep on an individual level. In the current study, the presence of extreme outliers seem to drive the widening of the LOAs. These findings imply that the current methods, are only reasonably accurate for assessing sleep quality metrics at a group level and caution should be exercised when applying the models and methods to individual-level sleep assessments. Therefore, further improvements and refinements are needed to enhance the validity of these models for individual sleep assessments.

In this study, we used the ZM as the reference method, rather than PSG, which is considered the gold standard for sleep measurement. This choice may contribute to discrepancies between our models and the ZM, as without a true gold standard, it's difficult to determine the source of disagreement. However, we believe that the use of ZM, which allows for multiple consecutive nights of recording in free-living[26], is valuable. This approach captures intra-individual variances in sleep, which is impractical with PSG. It also enabled us to include more nights in our study typically compared to those relying on PSG. For instance, the widely used Newcastle dataset[14] only contains data from 28 participants. However, upon examining the ZM outputs, we found that the raw predictions were not optimal for developing machine learning models due to a seemingly low signal-to-noise ratio (see Figure 3). The ZM itself mitigates this issue by applying certain filtering processes when generating sleep quality metrics. For example, epochs contributing to WASO must be in contiguous epochs of 3, and sleep only counts towards sleep quality metrics if 10 out of 12 minutes are scored as sleep. To improve the prospect of our machine learning algorithms, we applied median filters to the ZM raw predictions. This did in fact alter the derived sleep quality metrics. Notably, the mean WASO decreased from 39 minutes in the raw predictions to 30.6 minutes in the 5-minute median filtered predictions, and further decreased to 22.3 minutes in the 10-minute filtered predictions. The application of 5-minute and 10-minute median filters also led to increases in TST, SE, and LPS. This suggests that the filters may categorize some instances of wakefulness as sleep and remove brief awakenings. Despite these changes, the overall sleep quality profile derived from the median-filtered predictions is still comparable to that from the raw predictions, justifying our approach.

The current study boasts several strengths, including the capacity to distinguish in-bed awake and asleep time from out-of-bed time, thereby allowing for the extraction of vital sleep quality metrics. Furthermore, the research benefits from evaluating multiple nights per subject, providing valuable information into intra-subject sleep variability. However, certain limitations exist. The use of ZM, which isn't recognized as a gold standard, could potentially compromise our findings' validity. Future research should consider using PSG as a reference for methods similar to ours, despite its limitations, for a more accurate comparison. Moreover, our models weren't validated using an external dataset, a process that would have showcased their broader applicability. Hence, our conclusions remain confined primarily to children.

In conclusion, our study contributes to the ongoing efforts to improve sleep estimation methods using thigh-worn accelerometers. We evaluated different machine learning and deep learning models for predicting in-bed and sleep times and their corresponding sleep quality metrics. While the sequential models generally demonstrated excellent performance in predicting in-bed time, they faced challenges in accurately distinguishing between sleep and wake epochs during in-bed time. Among all models and configurations evaluated, the XGBoost model exhibited the best performance, including epoch-to-epoch predictions and sleep quality metrics. Our research also highlighted the current limitations of sleep detection methods, such as challenges in effectively detecting wake periods during in-bed time and the need for further improvements to increase the precision of individual sleep assessments. We believe our work lays the groundwork for future research to further refine and improve the performance of these models, contributing to a more precise and accurate evaluation of sleep patterns and quality using thigh-worn accelerometers.

V. DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

VI.   Code Availability

All code used to perform the analysis and generate the figures in this paper are available in this repository. Additionally, we provide a freely available tool which allows users to employ the developed XGBoost models for estimating sleep quality metrics based on accelerometer data in this repository.

VII.   Acknowledgements

VIII.   Author Contributions

E.H.L and J.C.B designed the study. P.L.K, JS.P., S.O.S, S.R.M, and A.G. conducted the data collection. E.H.L prepared the manuscript and performed all analyses and model developments. E.H.L. and J.C.B. interpreted the results. All authors validated the methodology and approved the final manuscript.

IX.   Disclosure Statement

None declared.

References

1. Ma G. Sleep, Health, and Society. *Sleep medicine clinics.* 2017;12(1). doi:10.1016/j.jsmc.2016.10.012

2. Meyer N, Harvey AG, Lockley SW, Dijk DJ. Circadian rhythms and disorders of the timing of sleep. *The Lancet.* 2022;400(10357):1061-1078. doi:10.1016/S0140-6736(22)00877-7

3. K Pavlova M, Latreille V. Sleep Disorders. *The American Journal of Medicine.* 2019;132(3):292-299. doi:10.1016/j.amjmed.2018.09.021

4. Difrancesco S, Lamers F, Riese H, et al. Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. *Depression and Anxiety.* 2019;36(10):975-986. doi:10.1002/da.22949

5. Van De Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review. *Journal of Sleep Research.* 2011;20(1pt2):183-200. doi:10.1111/j.1365-2869.2009.00814.x

6. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine.* 2022;18(1):193-202. doi:10.5664/jcsm.9538

7. Gaiduk M, Serrano Alarcón Á, Seepold R, Martínez Madrid N. Current status and prospects of automatic sleep stages scoring: Review. *Biomedical Engineering Letters.* 2023;13(3):247-272. doi:10.1007/s13534-023-00299-3

8. Moore CM, Schmiege SJ, Matthews EE. Actigraphy and Sleep Diary Measurements in Breast Cancer Survivors: Discrepancy in Selected Sleep Parameters. *Behavioral Sleep Medicine.* 2015;13(6):472-490. doi:10.1080/15402002.2014.940108

9. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. *Sleep.* 1982;5(4):389-399. doi:10.1093/sleep/5.4.389

10. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep.* 1992;15(5):461-469. doi:10.1093/sleep/15.5.461

11. Palotti J, Mall R, Aupetit M, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine.* 2019;2(1):1-9. doi:10.1038/s41746-019-0126-9

12. Sazonov E, Sazonova N, Schuckers S, Neuman M, CHIME Study Group. Activity-based sleep-wake identification in infants. *Physiological Measurement.* 2004;25(5):1291-1304. doi:10.1088/0967-3334/25/5/018

13. Sadeh A, Sharkey KM, Carskadon MA. Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep.* 1994;17(3):201-207. doi:10.1093/sleep/17.3.201

14. Hees VT van, Sabia S, Anderson KN, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. *PLOS ONE.* 2015;10(11):e0142533. doi:10.1371/journal.pone.0142533

15. Sundararajan K, Georgievska S, Lindert BHW te, et al. Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports.* 2021;11(1):24. doi:10.1038/s41598-020-79217-x

16. Carlson JA, Tuz-Zahra F, Bellettiere J, et al. Validity of Two Awake Wear-Time Classification Algorithms for activPAL in Youth, Adults, and Older Adults. *Journal for the Measurement of Physical Behaviour.* 2021;4(2):151-162. doi:10.1123/jmpb.2020-0045

17. Inan-Eroglu E, Huang BH, Shepherd L, et al. 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.* 2021;4(1):60-67. doi:10.1123/jmpb.2020-0033

18. Berg JD van der, Willems PJB, Velde JHPM van der, et al. Identifying waking time in 24-h accelerometry data in adults using an automated algorithm. *Journal of Sports Sciences.* 2016;34(19):1867-1873. doi:10.1080/02640414.2016.1140908

19. Winkler EAH, Bodicoat DH, Healy GN, et al. Identifying adults' valid waking wear time by automated estimation in activPAL data collected with a 24 h wear protocol. *Physiological Measurement.* 2016;37(10):1653. doi:10.1088/0967-3334/37/10/1653

20. Johansson PJ, Crowley P, Axelsson J, et al. Development and performance of a sleep estimation algorithm using a single accelerometer placed on the thigh: An evaluation against polysomnography. *Journal of Sleep Research.* 2023;32(2):e13725. doi:10.1111/jsr.13725

21. Skotte J, Korshøj M, Kristiansen J, Hanisch C, Holtermann A. Detection of Physical Activity Types Using Triaxial Accelerometers. *Journal of Physical Activity and Health.* 2014;11(1):76-84. doi:10.1123/jpah.2011-0347

22. Arvidsson D, Fridolfsson J, Börjesson M, et al. Re-examination of accelerometer data processing and calibration for the assessment of physical activity intensity. *Scandinavian Journal of Medicine & Science in Sports.* 2019;29(10):1442-1452. doi:10.1111/sms.13470

23. Brønd JC, Grøntved A, Andersen LB, Arvidsson D, Olesen LG. Simple Method for the Objective Activity Type Assessment with Preschoolers, Children and Adolescents. *Children (Basel, Switzerland).* 2020;7(7):72. doi:10.3390/children7070072

24. Kaplan RF, Wang Y, Loparo KA, Kelly MR, Bootzin RR. Performance evaluation of an automated single-channel sleep–wake detection algorithm. *Nature and Science of Sleep.* 2014;6:113-122. doi:10.2147/NSS.S71159

25. Wang Y, Loparo KA, Kelly MR, Kaplan RF. Evaluation of an automated single-channel sleep staging algorithm. *Nature and Science of Sleep.* 2015;7:101-111. doi:10.2147/NSS.S77888

26. Pedersen J, Rasmussen MGB, Olesen LG, Kristensen PL, Grøntved A. Self-administered electroencephalography-based sleep assessment: Compliance and perceived feasibility in children and adults. *Sleep Science and Practice.* 2021;5(1):8. doi:10.1186/s41606-021-00059-1

27. Rasmussen MGB, Pedersen J, Olesen LG, et al. 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.* 2020;20(1):380. doi:10.1186/s12889-020-8458-6

28. Pedersen J, Rasmussen MGB, Sørensen SO, et al. Effects of Limiting Recreational Screen Media Use on Physical Activity and Sleep in Families With Children: A Cluster Randomized Clinical Trial. *JAMA pediatrics.* 2022;176(8):741-749. doi:10.1001/jamapediatrics.2022.1519

29. Skovgaard EL, Roswall MA, Pedersen NH, Larsen KT, Grøntved A, Brønd JC. Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings. *Scientific Reports.* 2023;13(1):2496. doi:10.1038/s41598-023-29666-x

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

31. Galland BC, Taylor BJ, Elder DE, Herbison P. Normal sleep patterns in infants and children: A systematic review of observational studies. *Sleep Medicine Reviews.* 2012;16(3):213-222. doi:10.1016/j.smrv.2011.06.001

32. Pedersen MJ, Leonthin H, Mahler B, Rittig S, Jennum PJ, Kamperis K. Two nights of home polysomnography in healthy 7-14-year-old children – Feasibility and intraindividual variability. *Sleep Medicine.* 2023;101:87-92. doi:10.1016/j.sleep.2022.10.027

33. Palm L, Persson E, Elmqvist D, Blennow G. Sleep and Wakefulness in Normal Preadolescent Children. *Sleep.* 1989;12(4):299-308. doi:10.1093/sleep/12.4.299

34. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Computation.* 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

35. Sano A, Chen W, Lopez-Martinez D, Taylor S, Picard RW. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE journal of biomedical and health informatics.* 2019;23(4):1607-1617. doi:10.1109/JBHI.2018.2867619

36. Chen Z, Wu M, Cui W, Liu C, Li X. An Attention Based CNN-LSTM Approach for Sleep-Wake Detection With Heterogeneous Sensors. *IEEE journal of biomedical and health informatics.* 2021;25(9):3270-3277. doi:10.1109/JBHI.2020.3006145

37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.* 2002;16:321-357. doi:10.1613/jair.953

38. Hvitfeldt E. *Themis: Extra Recipes Steps for Dealing with Unbalanced Data.*; 2023. https://CRAN.R-project.org/package=themis.

39. Hjorth MF, Chaput JP, Damsgaard CT, et al. Measure of sleep and physical activity by a single accelerometer: Can a waist-worn Actigraph adequately measure sleep in children? *Sleep and Biological Rhythms.* 2012;10(4):328-335. doi:10.1111/j.1479-8425.2012.00578.x

40. Kushida CA, Chang A, Gadkary C, Guilleminault C, Carrillo O, Dement WC. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Medicine.* 2001;2(5):389-396. doi:10.1016/s1389-9457(00)00098-8

41. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science.* 1996;11(3):189-228. doi:10.1214/ss/1032280214

42. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2023. https://www.R-project.org/.

43. Kuhn M, Wickham H. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.*; 2020. https://www.tidymodels.org.

44. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software.* 2019;4(43):1686. doi:10.21105/joss.01686

45. Van Rossum G, Drake FL. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace; 2009.

46. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc.; 2019:8024-8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

47. Conley S, Knies A, Batten J, et al. Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: A systematic review and meta-analysis. *Sleep Medicine Reviews.* 2019;46:151-160. doi:10.1016/j.smrv.2019.05.001

48. Patterson MR, Nunes AAS, Gerstel D, et al. 40 years of actigraphy in sleep medicine and current state of the art algorithms. *npj Digital Medicine.* 2023;6(1):1-7. doi:10.1038/s41746-023-00802-1

49. Girschik J, Fritschi L, Heyworth J, Waters F. Validation of self-reported sleep against actigraphy. *Journal of Epidemiology.* 2012;22(5):462-468. doi:10.2188/jea.je20120012

50. Doherty A, Jackson D, Hammerla N, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE.* 2017;12(2):e0169649. doi:10.1371/journal.pone.0169649

51. Anderson KN, Catt M, Collerton J, et al. Assessment of sleep and circadian rhythm disorders in the very old: The Newcastle 85+ Cohort Study. *Age and Ageing.* 2014;43(1):57-63. doi:10.1093/ageing/aft153

52. Hees VT van, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports.* 2018;8(1):12975. doi:10.1038/s41598-018-31266-z

53. Plekhanova T, Rowlands AV, Davies MJ, Hall AP, Yates T, Edwardson CL. Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. *Journal of Sleep Research.* 2023;32(3):e13760. doi:10.1111/jsr.13760