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 metrics

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 metrics, including sleep period time (SPT), total sleep time (TST), sleep efficiency (SE), latency until persistent sleep (LPS), and wake after sleep onset (WASO). ~~By analyzing these additional metrics, we aimed to provide a comprehensive evaluation of the model's capability in assessing various aspects of sleep quality, allowing us to determine the overall effectiveness of our models in accurately estimating different parameters related to sleep duration and quality.~~

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 isolated data from child participants between the ages of 6 to 10 years within the SCREENS cohort. Our main 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, 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 aptness 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].

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), guaranteeing the ethical and secure management of participant information.

2.2. Data Preprocessing and Feature Extraction

In this study, raw data processing 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 developed 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 1). 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. These features provide a comprehensive understanding of the signal's attributes, facilitating a more nuanced analysis of the underlying patterns. All features are summarized in table ??? (overvej tabel til supp. material)

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 commonly occurring at the end of each cycle[28]. 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. 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 resulted in an anticipated, more age-appropriate count of awakenings per night, providing a more accurate depiction of children's sleep patterns (see Figure 2).

2.3. Algorithms, Training and Validation

In our study, we utilize two distinct modeling strategies to predict sleep patterns derived from thigh-mounted accelerometer data: a sequential ensemble and a multiclass bidirectional Long Short-Term Memory (biLSTM)[29] neural network. The sequential ensemble involves using four pairs of models, each pair consisting of the same algorithm, that predict in-bed and sleep time successively. Conversely, the biLSTM is a comprehensive model that predicts multiple classes simultaneously, effectively handling complex temporal patterns. By applying and comparing both, we aim to evaluate their relative effectiveness.

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

The biLSTM is a multiclass classifier capable of predicting the classes of out-bed-awake, in-bed-awake, and in-bed-asleep. The architecture of the biLSTM consisted of four layers, each with 128 hidden units. The choice of four layers and 128 hidden units in our model balanced complexity and efficiency: it was deep enough to learn intricate patterns yet feasible to train timely. Additionally, the bidirectional nature of the LSTM, doubling the hidden units at each time step, improved data comprehension and avoided overfitting.

something on sequence length and step size

The biLSTM model was chosen for sleep classification using accelerometer data from a thigh-mounted device due to its advantages in modeling temporal context, handling a large temporal scope, and enabling temporal inference over any feature without the need for hand-designed temporal features.

Previous studies have explored the use of LSTM models for sleep detection, showing promising results in capturing complex temporal patterns. Notable works by Sano et al.[30] and Chen et al.[31] have demonstrated the potential of LSTM models in improving sleep detection based on accelerometer data. However, further research is necessary to fully comprehend the strengths and limitations of this approach.

2.3.3. Model Training

For the models in sequence, we trained four pairs of classification models, each pair distinguishing between in-bed/out-of-bed and asleep/awake states. The dataset was randomly split into a training set (approximately 50% of the subjects) and a testing set (also approximately 50%), ensuring that samples from the same subject were never simultaneously used in both sets. Hyper-parameter optimization was performed using 10-fold Monte Carlo cross-validation, with the F1 score as the performance metric. The best-performing set of hyperparameters was then used to fit the models to the full training dataset, maximizing accuracy by leveraging all available data. Furthermore, the dataset for predicting sleep was highly imbalanced which can pose challenges when training machine learning models, as models may favor predicting the majority class. To address this issue, we employed oversampling techniques to ensure that each class had a roughly equal number of training samples. Specifically, we utilized the Synthetic Minority Over-sampling Technique (SMOTE)[32], which generates new samples by interpolating random samples with their nearest neighbors. In our study, we implemented SMOTE using the themis R package[33], resampling all classes to achieve a balanced distribution of training samples.

The biLSTM model was trained using the Adam optimizer, which is computationally efficient and adapts the learning rate during training. Cross-entropy loss function was employed for its suitability in multiclass classification with mutually exclusive classes. The softmax activation function was selected for the output layer to obtain a probability distribution over the classes. Data for training the biLSTM were randomly divided into a training, validation, and test tensors based on a 50/25/25 split. The model was evaluated using the F1 score on both the training and validation sets. Early stopping was implemented with a patience of 3 epochs, meaning training would stop if there was no improvement in the validation loss for 3 consecutive epochs.

2.3.4. Model Validation [evt. bare "statistics"]

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. Furthermore, we present precision-recall curves and receiver operating characteristic curves, including the area under the curve, for each model and for each type of ZM prediction.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 evaluate our developed models' performance in assessing sleep quality metrics, we employed Bland-Altman plots and Pearson correlations. These methods were used to compare the sleep quality summaries calculated by our models with those obtained from EEG-based ZM sleep quality summaries. The sleep quality summaries included several metrics:

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

The analysis included children with an average age of 9.4 years (SD = 2.1). The raw ZM predictions covered 2,035,261 sleep epochs, or about 86% of the total recording duration.

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. The Sleep Period Time (SPT) were consistent across raw and filtered datasets (mean: 9.2 ± 2.1 hours) as this corresponds to the length of the ZM recording. Total Sleep Time (TST) and Sleep Efficiency (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. Latency to Persistent Sleep (LPS) also elevated, suggesting the filter smooths out brief awakenings at sleep onset, leading to a prolonged time to persistent sleep. The most significant change was seen in Wake After Sleep Onset (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. These results underscore the role of median filters on sleep metrics.

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	9.21 (2.15)	7.74 (1.93)	82.61 (11.97)	34.53 (27.87)	39 (33.56)	34.46 (11.33)
5-Min Median	9.21 (2.15)	8.1 (2.03)	86.35 (12.72)	36.26 (39.8)	30.63 (46.84)	4.43 (3.26)
10-Min Median	9.21 (2.15)	8.21 (2.06)	87.54 (12.89)	37.99 (48.7)	22.27 (55.38)	1.95 (2.01)

3.1. Performance on Epoch-To-Epoch Basis

The epoch-to-epoch evaluation of in-bed performance metrics, outlined in Table 2, demonstrates practically equivalent performance across all model types. While the Decision Tree model posted an F1 score of 94.4% and accuracy of 95.3%, the Logistic Regression and Feed-Forward Neural Network models each exhibited an F1 score of 95.0% and similar accuracies of 95.7% and 95.8%, respectively. 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. This underscores the consistency of performance among these models in classifying in-bed conditions.

Table 2: In-Bed Performance Metrics

	F1 Score (%)	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Decision Tree	94.36	95.27	93.12	95.64	96.86
Logistic Regression	94.99	95.74	95.03	94.94	96.26
Feed-Forward Neural Net	95.03	95.77	95.07	94.99	96.29
XGBoost	95.38	96.06	95.83	94.94	96.23

Table 3 details the performance of the sequential models on raw and median-filtered (5 and 10 minute) ZM predictions. 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 negative instances. 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 a concern, 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	72.94	93.24	48.36	86.34	67.15
Logistic Regression	71.05	93.72	43.88	82.72	70.93
Neural Network	71.76	93.77	45.13	83.59	70.83

XGBoost	76.18	92.80	57.98	91.32	62.84
5-Min Median					
Decision Tree	75.48	94.20	55.46	93.35	59.01
Logistic Regression	68.32	95.84	36.03	81.36	74.84
Neural Network	71.74	95.78	41.64	85.62	73.11
XGBoost	79.22	93.87	74.00	97.30	54.68
10-Min Median					
Decision Tree	76.28	94.74	58.06	94.86	57.47
Logistic Regression	67.96	96.54	34.29	81.87	76.35
Neural Network	70.95	96.06	39.54	86.48	71.37
XGBoost	80.87	94.90	75.76	97.72	57.60

The analysis of precision-recall and ROC curves across different models and ZM prediction types shows varying performance. In terms of precision-recall AUC, the Decision Tree model consistently outperforms others, indicating its superior predictive accuracy (see Figure 4). Conversely, the Neural Network model generally shows weaker performance. However, for ROC AUC, the XGBoost model consistently excels across all data types, indicating a strong ability to differentiate between classes, while the Neural Network model tends to underperform (see Figure 5). The F-measure (F1 score) shows variable performance across different configurations but generally, the Decision Tree model yields higher scores.

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. However, Specificity values are not available for this configuration. 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	Sensitivity	Precision
Raw ZM Predictions	71.36	70.42	76.04
5-Min Median	75.99	74.62	78.53
10-Min Median	73.45	73.93	73.07

3.2. Evaluation of Sleep Quality Summaries

BA plots

pearson regression

scatterplots with identity line and best linear fit

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

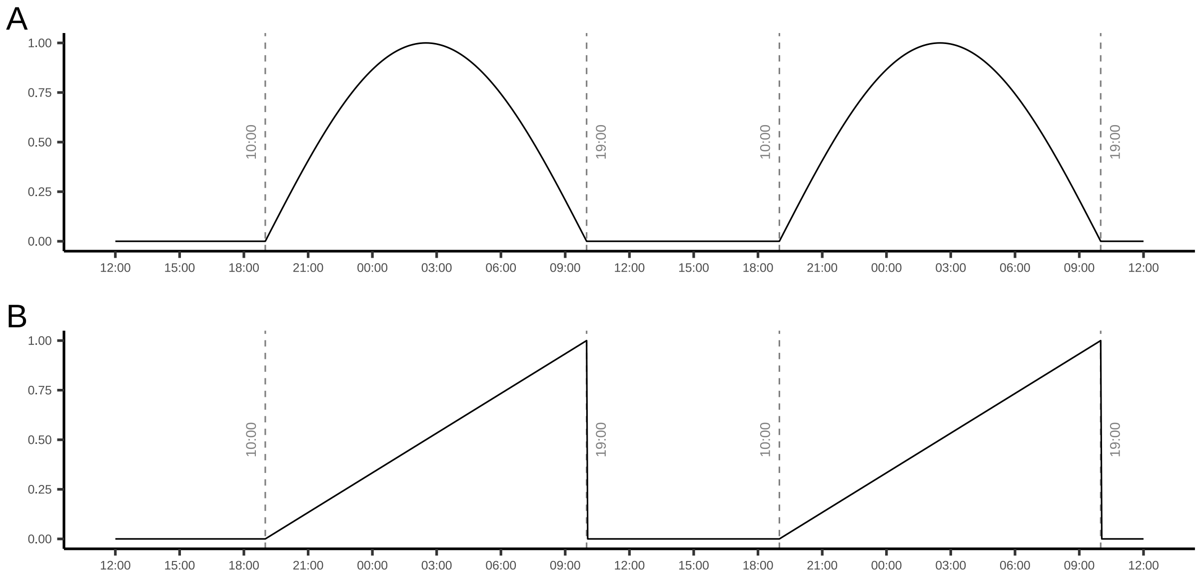

Figure 1: Sensor-independent features of circadian rhythms across two consecutive nights. A) cosinus feature, B) linear feature.

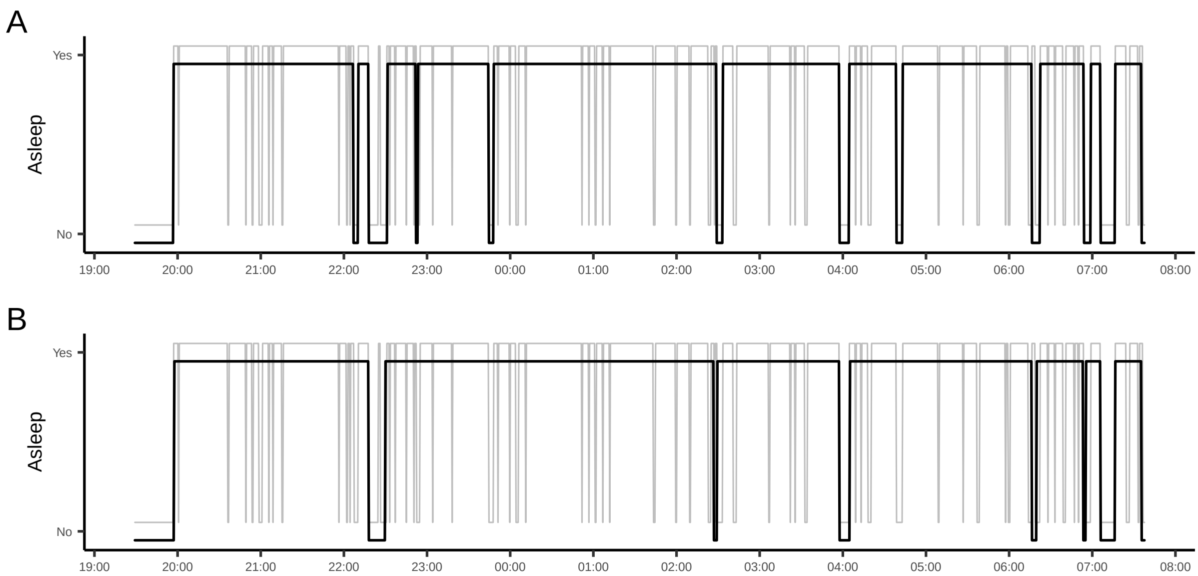

Figure 2: The difference in number of awakenings between the raw ZM predictions vs. 5-minute, and 10-minute median filtered predictions for a random night. Grey line is the raw predictions, black line is the median filtered predictions. A: 5-minute median filter on raw ZM predictions, B: 10-minute median filter on raw ZM predictions.

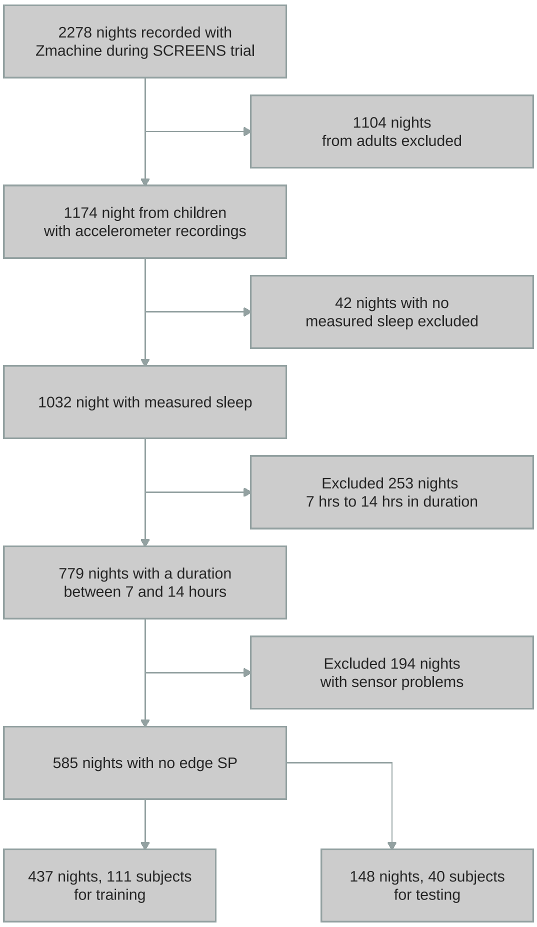

Figure 3: Flowchart of eligible nights included in the study - revise text boxes!

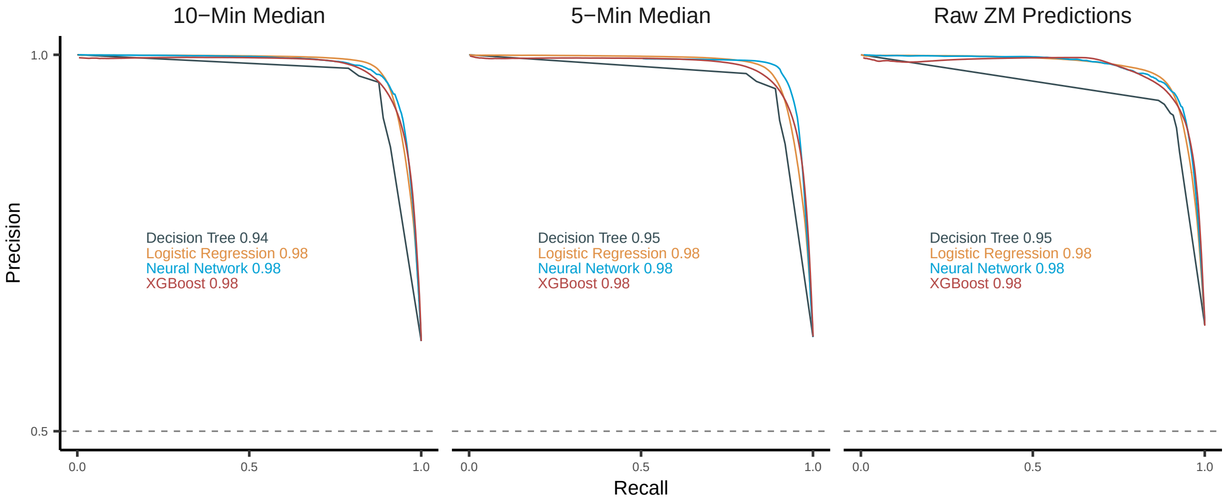

Figure 4: Precision-recall curves of the models evaluated across the different ZM predictions, including raw ZM predictions, as well as 5-minute and 10-minute median smoothing of the ZM raw predictions. The x-axis of the plot represents the proportion of true wake epochs that were correctly classified as wake, while the y-axis represents the proportion of all epochs labeled as wake by the classifier that were classified correctly. The area under the curve values are displayed as color-coded text in the plot indicate the area under the Precision-Recall curve for each model and condition.

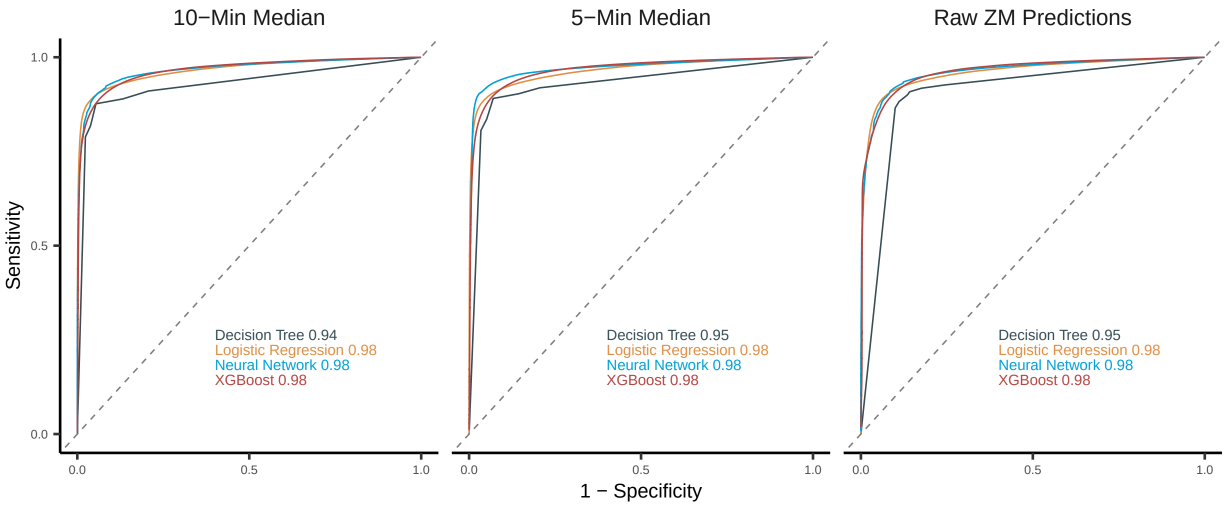

Figure 5: Receiver operating characteristic curves of the models evaluated across the different ZM predictions, including raw ZM predictions, as well as 5-minute and 10-minute median smoothing of the ZM raw predictions. The x-axis of the plot represents the proportion of true asleep epochs that were incorrectly classified as awake, while the y-axis represents the proportion of all epochs labeled as awake by the classifier that were correctly classified. The area under the curve values displayed are displayed as color-coded text in the plot to indicate the area under the receiver operating characteristic curve for each model and condition.