A machine-learning sleep-wake classification model using a reduced number of features derived from photoplethysmography and activity signals

Douglas A. Almeida[1], Felipe M. Dias[1,2], Marcelo A. F. Toledo[1], Diego A. C. Cardenas[1], Filipe A. C. Oliveira [1,2], Estela Ribeiro[1], José E. Krieger[1] and Marco A. Gutierrez[1,2]

[1]Heart Institute, Clinics Hospital, University of Sao Paulo Medical School
Sao Paulo, SP – Brazil

[2]Polytechnique School, University of Sao Paulo – Sao Paulo, SP – Brazil

douglas.andrade@hc.fm.usp.br, f.dias@hc.fm.usp.br

marcelo.arruda@hc.fm.usp.br, diego.cardona@hc.fm.usp.br

filipe.acoliveira@hc.fm.usp.br, estela.ribeiro@hc.fm.usp.br

j.krieger@hc.fm.usp.br, marco.gutierrez@incor.usp.br

Abstract. *Sleep is a crucial aspect to overall health, impacting mental and physical well-being. The classification of sleep stages is an important step to assess sleep quality, and Photoplethysmography (PPG) has been demonstrated to be an effective signal for this task. Recent works in this area usually employ complex methods that may be unfeasible to be deployed in wearable devices. In this work, we present a XGBoost model for sleep-wake classification based on features extracted from PPG signal and activity counts. The performance of our method achieved a Sensitivity of 91.15 ± 1.16%, Specificity of 53.66 ± 1.12%, F1-score of 83.88 ± 0.56%, and Kappa of 48.0 ± 0.86%. Our method offers a significant improvement over other approaches as it uses a reduced number of features, making it suitable for implementation in wearable devices that have limited computational power.*

1. Introduction

Sleep plays a vital role in maintaining good health for individuals of all ages [Ramar et al. 2021]. Inadequate sleep of poor quality or duration can lead to a host of chronic health problems, including cardiovascular diseases, diabetes, and obesity [Knutson and Van Cauter 2008]. These highlight the importance of ensuring that individuals get adequate and high-quality sleep to maintain their health and well-being.

Polysomnography (PSG) is the gold standard exam for evaluating human sleep [Krystal and Edinger 2008]. The exam involves the simultaneous recording of multiple electrophysiological signals during sleep. After registration, a specialist analyzes the recorded signals and labels different sleep stages based on time windows of 30 s. Nevertheless, PSG is an expensive and time-consuming procedure and can also suffer from labeling errors by the specialist. Additionally, PSG exams are performed in an unfamiliar environment, with multiple electrodes attached, which can affect sleep quality.

With the rise of wearable devices, it is now possible to track several physiological signals, such as heart rate (HR), more cost-effectively and conveniently. Wearable devices are increasingly being used to track and evaluate sleep patterns by capturing Photoplethysmography (PPG) signals, raw triaxis accelerometer signal (ACC), and activity-based signals like Activity (ACT). PPG measures changes in the blood volume of vascular tissues, allowing measurements of vital parameters, such as heart rate, respiration rate, arterial oxygen saturation, and blood pressure [Mejía-Mejía et al. 2022]. On the other hand, the ACT is derived from the ACC signal processing and provides an estimation of rest and wakefulness periods to assess sleep patterns.

Sleep stages can be classified into five distinct categories: (i) wakefulness (W), (ii) non-REM stage 1 (N1), (iii) non-REM stage 2 (N2), (iv) non-REM stage 3 (N3), and (v) REM stage (R). While accurate classification of these five categories is important for monitoring sleep patterns, the identification of sleep-wake stages are sufficient to calculate important metrics such as: total sleep time, total wake time, sleep latency, sleep efficiency, and wakefulness after sleep onset [Shrivastava et al. 2014].

While wearable devices cannot directly measure brain activity like EEG devices used in sleep stage analysis, they can capture PPG signals, which can be utilized to estimate HR [Mejía-Mejía et al. 2022]. HR is known to decrease during the transition from wakefulness to non-REM stages [Silvani 2008]. There are also indications that HR exhibits a slight decrease during the transition from non-REM to REM sleep stages [Silvani 2008, Habib et al. 2023]. Therefore, we hypothesize that only using the information of the HR and the Heart Rate Variability (HRV), extracted from the PPG signals, along with the ACT, derived from the ACC signals, is possible to accurately classify sleep and wake stages.

Recent works in the field of sleep-wake classification proposed to use both PPG and/or ACC signals. Most of them are based on features extracted of these signals. Table 1 displays a summary of the state-of-the-art on this topic. [Fonseca et al. 2017] used a Bayesian Classifier to predict sleep-wake stages based on a set of HRV features computed from interbeat intervals obtained from PPG signals along with measures from ACC. Likewise, [Eyal and Baharav 2017] used a similar approach as [Fonseca et al. 2017], without the measures from ACC. [Uçar et al. 2018] proposed to use k-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms based on PPG and HRV features. [Palotti et al. 2019] compared a cohort of classification algorithms to perform the sleep-wake classification based on classic machine learning or deep learning techniques using only ACC signals. Likewise, [Banfi et al. 2021] proposed to use only raw ACC signals using Convolutional Neural Networks (CNN) for this binary classification. [Habib et al. 2023] proposed a CNN derived from PPG raw signals of 10 subjects with sleep-disordered breathing, using a leave-one-out strategy on the sleep-wake classification with data augmentation. [Motin et al. 2023] used the same dataset as [Habib et al. 2023], extracting 72 features from the PPG signals, instead of using the raw PPG, using three different classifiers, KNN, SVM and a Random Forest (RF). Many of these methods have been evaluated using small datasets or rely on a large number of features, rendering the approach impractical for deployment on resource-constrained devices like wearables.

In this work we present a method to classify sleep and wake stages, comparing the results of three commonly used machine learning techniques: Logistic Regression (LR),

Table 1. Summary of state-of-the-art obtained results for the sleep-wake classification.

	Accuracy	Sensitivity	Specificity	F1-score	Kappa	Dataset	Method
[Fonseca et al. 2017]	91.5 ±5.1	58.2 ±17.3	92.9 ±2.0	-	0.55 ±0.14	Private (s = 101)	Bayesian classifier - PPG features and ACC
[Eyal and Baharav 2017]	84.3	38.1	91.7	-	0.31	Private (s = 88)	Bayesian classifier - PPG features
[Uçar et al. 2018]	79.23	78.0	80.0	79.0	0.58	Private (s = 10)	SVM - PPG features
[Uçar et al. 2018]	79.36	77.0	81.0	79.0	0.59	Private (s = 10)	KNN - PPG features
[Palotti et al. 2019]	81.8 ± 1.0	90.40 ±1.20	68.10 ±1.90	84.30 ±1.10	-	MESA (s = 1,817)	Extra Trees - 370 ACC features
[Palotti et al. 2019]	83.1 ± 1.0	91.40 ±1.10	69.90 ±2.00	85.50 ± 1.00	-	MESA (s = 1,817)	LSTM 100 - ACC raw signal
[Motin et al. 2019]	72.36	70.64	74.22	-	-	Private (s = 5)	Medium Gaussian SVM - PPG 17 features
[Motin et al. 2020]	81.10	81.06	82.50	81.74	-	Private (s = 10)	Cubic SVM - PPG 22 features
[Banfi et al. 2021]	-	89.20	92.0	90.9	0.782	Private (s = 81)	LightCNNA - ACC raw features
[Habib et al. 2023]	94.18 ±11.95	94.4	-	93.05 ±13.77	0.864 ±0.265	Private (s = 10)	CNN - PPG raw signal
[Motin et al. 2023]	83.75 ±0.85	87.79 ±1.10	73.63 ±2.45	80.01 ±1.88	-	Private (s = 10)	KNN - PPG 72 features
[Motin et al. 2023]	84.66 ±0.99	87.41 ±1.24	77.79 ±0.93	82.32 ±0.90	-	Private (s = 10)	SVM - PPG 72 features
[Motin et al. 2023]	85.22 ±0.62	87.86 ±1.48	77.67 ±3.26	82.45 ±1.62	-	Private (s = 10)	RF - PPG 72 features

Random Forest (RF), and the eXtreme Gradient Boosting (XGBoost). These algorithm are based on features extracted from PPG and ACT signals with a reduced number of features for deployment feasibility on wearable devices. Additionally, a stratified analysis was conducted considering age and gender factors. Our method demonstrates advancement over existing approaches by reducing the feature set, thereby enabling implementation on computational-constrained wearable devices.

2. Materials and Methods

The proposed approach comprises five sequential steps, being them: Data Selection, Preprocessing, Windowing, Feature Extraction, and Classification. To begin, we describe the data utilized, specifically selecting subjects with both ACT and PPG signals. We merge the sleep stages to generate a Wake and Sleep dataset, forming the foundation for subsequent analysis. After that, we present the preprocessing and windowing procedures, detailing how the data is prepared for further processing. Additionally, we outline the feature extraction process. Furthermore, we describe our employed classifiers, providing information on our experimental setup. A flow diagram summarizing these steps is shown in Fig. 1.

2.1. Data Selection

The experiments were carried out on the MESA Sleep dataset, a subset of the Multi-Ethnic Study of Atherosclerosis (MESA) dataset [Chen et al. 2015]. The MESA Sleep dataset contains data collected from 2,237 subjects, including overnight Polysomnography exams along with their corresponding PPG signals, 7-day wrist-worn Actigraphy signals, sleep stage labels for every 30 s windows, and sleep questionnaires. For this study, only patients

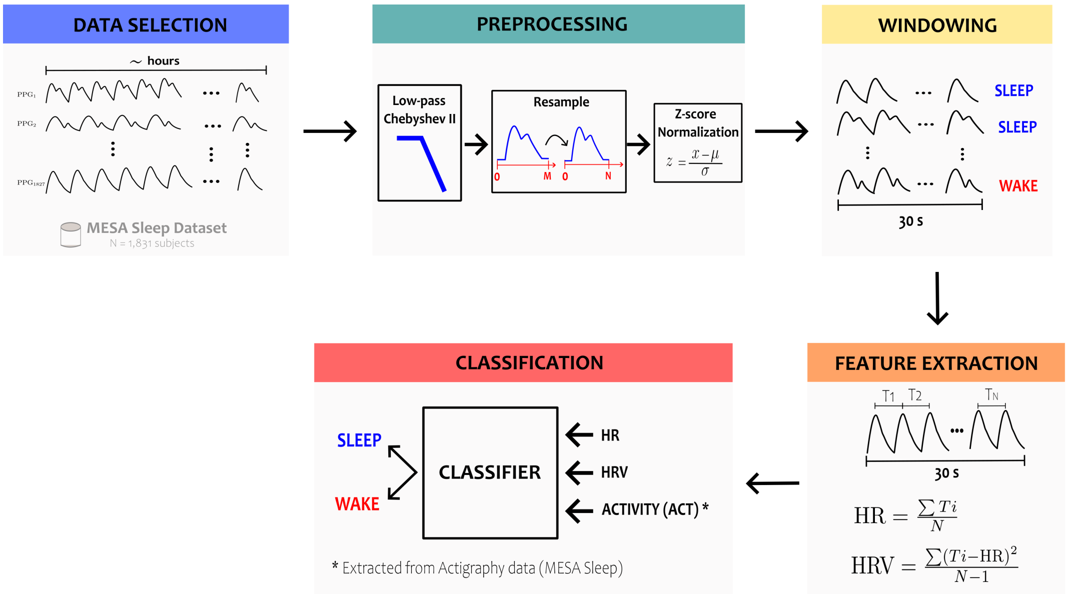

Figure 1. Overview of the proposed method for sleep-wake classification.

with both ACT and PPG signals were used, resulting in 1,831 patients. The metadata information about the employed dataset is shown in Table 2.

This dataset provides sleep stage labels for every 30 s window. The labeling process followed the AASM guidelines, which suggest the five sleep stage classes aforementioned. Since our goal is to detect only sleep or wake stages, non-REM stages 1-3 and REM stage were grouped into class "sleep" (S). For every 30 s window, the PPG signal was sampled at 256 Hz, and the corresponding ACT registered.

2.2. Preprocessing

We used the same preprocessing steps proposed by [Kotzen et al. 2022] in this study. The PPG data was filtered using a low-pass 8th-order Chebyshev Type II filter at 8 Hz, followed by downsampling from 256 Hz to 34 Hz using linear interpolation. Outlier values greater than or less than three standard deviations from the mean were clipped, and the data was normalized using z-score normalization.

2.3. Windowing

We partitioned the PPG signals into non-overlapping 30-second windows, as the PSG exam provides a sleep stage label for every 30 seconds of recording time. Due to the in-

Table 2. Metadata of the MESA dataset.

Parameter	Value
Subjects	$2,237$
with PSG	$2,056$
with actigraphy	$2,158$
with PSG and actigraphy	$1,831$
Age	69.6 ± 9.1
Male subjects	$1,039$

herent variable recording lengths for each subject in the MESA Sleep dataset, the number of windows differed among subjects. Figure 2 illustrates the windowing extraction step applied to the PPG signal of a single subject.

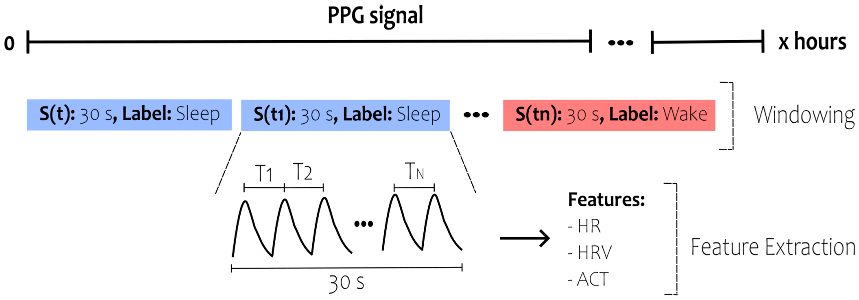

Figure 2. Windowing and Feature Extraction process example.

2.4. Feature Extraction

HR and HRV are widely recognized as important features associated with sleep stages and disorders, being both regulated by the sympathetic and parasympathetic nervous systems [Stein and Pu 2012]. HR was calculated for every window by taking the mean time difference between peaks, which were detected using the method proposed in [Bishop and Ercole 2018]. HRV was calculated as the standard deviation of the time difference between the peaks. Windows with HR greater than 180 beats per minute or HRV outside of two standard deviations from the mean HRV of the entire dataset were discharged. As consequence, all samples from four participants were excluded in this phase, resulting in $s = 1,827$ subjects. Our model also used as input the activity value provided by the MESA-sleep acquired with the Actiware-Sleep version 5.59 analysis software (Mini-Mitter Co, Inc, Bend, OR).

Therefore, for each i subject ($i = 1, 2, \ldots, s$), the features collected in each window were arranged as

$$\mathbf{X}_i = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}, \tag{1}$$

where s is the number of subjects, m is the number of features ($m = 3$), and n is the number of windows for each subject. Overall, we have a total of 2,050,280 windows: 63.93% (1,310,690) are labeled as sleep and the remaining 36.07% (739,590) as wake.

2.5. Classification

Our sleep-wake classification was performed using three commonly used machine learning techniques: LR, RF, and XGBoost. The models were implemented in Python using the default hyperparameters. A 10-fold cross-validation approach was employed to prevent bias in the training and testing split, with samples from the same patient grouped in the same fold to avoid intra-patient bias.

2.5.1. Experimental Setup

Experiments were performed using a Foxconn HPC M100-NHI with an 8-GPU cluster of NVIDIA Tesla V100 16GB cards. The model was implemented in Python (3.8.10) with the support of the libraries scikit-learn (1.1.3), XGBoost (1.6.1), and scipy (1.8.1).

3. Results

Table 3 summarizes the overall performance of the proposed methods for the sleep-wake classification task. The results were compared to the literature by using five different evaluation metrics commonly used in sleep-wake classification studies: Accuracy (Ac), Sensitivity (Se), Specificity (Sp), F1-score, and Cohen's Kappa Coefficient (Kappa). Additionally, we used the same metrics to evaluate the classification provided by the actigraphy used in the MESA Sleep dataset as baseline.

Table 3. Obtained results for the sleep-wake classification.

	Method	Accuracy	Sensitivity	Specificity	F1-score	Kappa
Our model	*XGBoost*	77.62 ± 0.56	91.14 ± 1.15	53.66 ± 1.11	83.88 ± 0.56	0.480 ± 0.008
Our model	*LR*	74.62 ± 0.66	96.46 ± 0.33	35.91 ± 0.88	82.92 ± 0.55	0.370 ± 0.009
Our model	*RF*	73.80 ± 0.48	83.68 ± 0.93	56.30 ± 0.76	80.32 ± 0.54	0.413 ± 0.007
MESA Actigraphy		-	50.49	94.75	64.08	0.478
[Palotti et al. 2019]	LSTM 100	83.1 ± 1.0	91.40 ± 1.10	69.90 ± 2.00	85.50 ± 1.10	-
[Motin et al. 2023]	RF	85.22 ± 0.62	87.86 ± 1.48	77.67 ± 3.26	82.45 ± 1.62	-
[Habib et al. 2023]	CNN	94.18 ± 11.95	94.4	-	93.05 ± 13.77	0.86 ± 0.26

To assess possible biases in our method, we also performed a stratified analysis of our obtained results by age and gender. In Fig. 3, we show the obtained F1-score of our approach in four age groups (54 – 65, 66 – 75, 76 – 85, and 86+) and two gender groups (male and female).

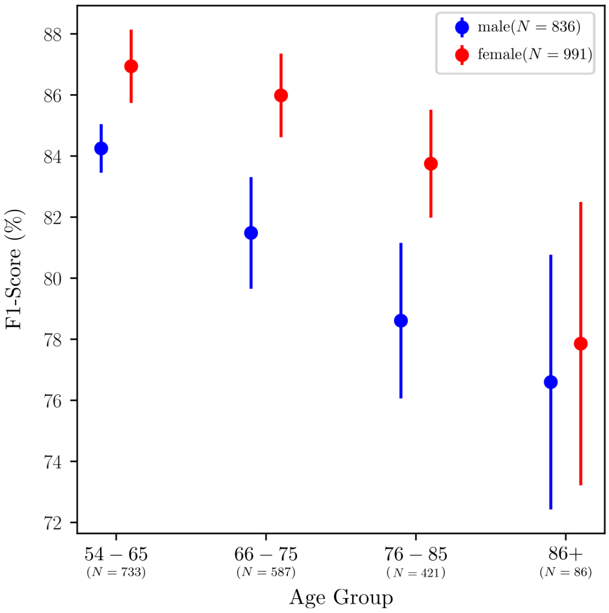

Figure 3. Stratified F1-score results for sleep-wake classification task for age and gender.

4. Discussion

Our proposed methodology used only three features: two obtained from the peak-to-peak interval from the PPG signal and one from the ACT value. The peak-to-peak derived features were extracted even in adverse acquisition scenarios and the actigraph value was easily extracted from the accelerometer. This is a major advantage compared to other reports that extract a complex set of features that can be compromised in noisy acquisitions. As shown in Table 3, the algorithm that best performed was the XGBoost. Apart from the Spe, our proposed method achieved higher metric values than the MESA Sleep actigraphy. This shows the gain of using PPG-derived physiological measures (HR and HRV).

Compared to the best method from [Palotti et al. 2019], which also used the MESA Sleep dataset, we obtained comparable Se ($91.14 \pm 1.15\%$ vs $91.40 \pm 1.10\%$) and F1-score ($83.88 \pm 0.56\%$ vs $85.50 \pm 1.00\%$); however, the Spe value is lower ($53.66 \pm 1.11\%$ vs $69.90 \pm 2.00\%$). The lower Spe might be due to the imbalance between the classes. Adopting a proper strategy for dealing with unbalanced data might improve the observed results and need further investigation.

It is not possible to make a direct comparison with other works, since all of them are based on private datasets. [Motin et al. 2023] achieved comparable results as ours in terms of F1-score, however, they didn't used ACC signals and used a much larger set of PPG features than us. Moreover, their work are based only on 10 distinct subjects. Even though their results are promising, this reduced number of subjects may lead to limited generalizability, along with biased learning since there is a high interdependence among intra-subject heartbeats [Costa et al. 2023]. Likewise, [Habib et al. 2023] used a limited number of subjects, and employed a CNN on PPG raw signal, achieving the best F1-score to date. However, these deep learning approaches usually requires lots of data for training, computational resources and often have many hyperparameters.

In the stratified analysis shown in Fig. 3, the F1-score progressively decreases with age. Also, there was consistently better mean F1-score for females than for males, despite an increase in the standard deviation for both genders. This indicates that the reduction in the number of older patients affects both genders similarly. However, the cause of the higher mean F1-score for females cannot be determined based on the available data, as it could be due to a larger sample size or physiological differences between genders. Further research may be necessary to determine the underlying reasons for these results.

Furthermore, it was aforementioned that deep learning models could be cumbersome for deployment in wearable devices. However, the computational resource needed depends on a set of parameters that should be investigated. Thus, future works should address this by comparing the performance and computational resources of different models. Likewise, future works should also be done on external datasets to validate the generalizability of our proposed method. Additionally, the validation of wearable sleep monitoring devices would reduce the cost of sleep disorder compared to traditional polysomnography exams. This requires reliable and practical sleep stages prediction models easy to implement and capable of being integrated into commercial devices in line with the reported findings.

Moreover, this study has some limitations that need to be acknowledged. Firstly, the generalizability of our approach is limited as it relies solely on the MESA Sleep dataset, which comprises data from older individuals with an average age of 69 years, which may limit its applicability to other age groups.

5. Conclusion

We provided evidence for a simple method to classify sleep-wake states using only three features. Our method showed comparable results with more complex methods and superior results than using only the ACT measured by an actigraph, which is currently the wearable device of choice for sleep monitoring. The validation and integration of sleep-wake models into commercial devices hold promise for reducing the cost of diagnosing sleep disorders compared to traditional polysomnography exams, thereby making sleep health more accessible and convenient for a broader population.

Acknowledgments

This work was supported by Foxconn Brazil and Zerbini Foundation, as part of the research project “Remote Patient Monitoring System”.

References

Banfi, T., Valigi, N., di Galante, M., d’Ascanio, P., Ciuti, G., and Faraguna, U. (2021). Efficient embedded sleep wake classification for open-source actigraphy. *Scientific reports*, 11(1):1–12.

Bishop, S. M. and Ercole, A. (2018). Multi-scale peak and trough detection optimised for periodic and quasi-periodic neuroscience data. In *Intracranial Pressure & Neuromonitoring XVI*, pages 189–195. Springer.

Chen, X., Wang, R., Zee, P., Lutsey, P. L., Javaheri, S., Alcántara, C., Jackson, C. L., Williams, M. A., and Redline, S. (2015). Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888.

Costa, T. B. D. S., Dias, F. M., Cardenas, D. A. C., Toledo, M. A. F. D., Lima, D. M. D., Krieger, J. E., and Gutierrez, M. A. (2023). Blood pressure estimation from photoplethysmography by considering intra- and inter-subject variabilities: Guidelines for a fair assessment. *IEEE Access*, 11:57934–57950.

Eyal, S. and Baharav, A. (2017). Sleep insights from the finger tip: How photoplethysmography can help quantify sleep. In *2017 Computing in Cardiology (CinC)*, pages 1–4.

Fonseca, P., Weysen, T., Goelema, M. S., Møst, E. I., Radha, M., Lunsingh Scheurleer, C., van den Heuvel, L., and Aarts, R. M. (2017). Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults. *Sleep*, 40(7):zsx097.

Habib, A., Motin, M. A., Penzel, T., Palaniswami, M., Yearwood, J., and Karmakar, C. (2023). Performance of a convolutional neural network derived from ppg signal in classifying sleep stages. *IEEE Transactions on Biomedical Engineering*, 70(6):1717–1728.

Knutson, K. L. and Van Cauter, E. (2008). Associations between sleep loss and increased risk of obesity and diabetes. *Annals of the New York Academy of Sciences*, 1129(1):287–304.

Kotzen, K., Charlton, P. H., Salabi, S., Amar, L., Landesberg, A., and Behar, J. A. (2022). Sleepppg-net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*.

Krystal, A. D. and Edinger, J. D. (2008). Measuring sleep quality. *Sleep Medicine*, 9:S10–S17. The Art of Good Sleep Proceedings from the 5th International Sleep Disorders Forum: Novel Outcome Measures of Sleep, Sleep Loss and Insomnia.

Mejía-Mejía, E., Allen, J., Budidha, K., El-Hajj, C., Kyriacou, P. A., and Charlton, P. H. (2022). 4 - photoplethysmography signal processing and synthesis. In Allen, J. and Kyriacou, P., editors, *Photoplethysmography*, pages 69–146. Academic Press.

Motin, M. A., Karmakar, C., Palaniswami, M., and Penzel, T. (2020). Photoplethysmographic-based automated sleep–wake classification using a support vector machine. *Physiol. Meas.*, 41:075013.

Motin, M. A., Karmakar, C., Palaniswami, M., Penzel, T., and Kumar, D. (2023). Multi-stage sleep classification using photoplethysmographic sensor. *Royal Society Open Science*, 10(4):221517.

Motin, M. A., Karmakar, C. K., Penzel, T., and Palaniswami, M. (2019). Sleep-wake classification using statistical features extracted from photoplethysmographic signals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5564–5567. IEEE.

Palotti, J., Mall, R., Aupetit, M., Rueschman, M., Singh, M., Sathyanarayana, A., Taheri, S., and Fernandez-Luque, L. (2019). Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ digital medicine*, 2(1):1–9.

Ramar, K., Malhotra, R. K., Carden, K. A., Martin, J. L., Abbasi-Feinberg, F., Aurora, R. N., Kapur, V. K., Olson, E. J., Rosen, C. L., Rowley, J. A., et al. (2021). Sleep is essential to health: an american academy of sleep medicine position statement. *Journal of Clinical Sleep Medicine*, 17(10):2115–2119.

Shrivastava, D., Jung, S., Saadat, M., Sirohi, R., and Crewson, K. (2014). How to interpret the results of a sleep study. *Journal of community hospital internal medicine perspectives*, 4(5):24983.

Silvani, A. (2008). Physiological sleep-dependent changes in arterial blood pressure: Central autonomic commands and baroreflex control. *Clinical and Experimental Pharmacology and Physiology*, 35(9):987–994.

Stein, P. K. and Pu, Y. (2012). Heart rate variability, sleep and sleep disorders. *Sleep Medicine Reviews*, 16(1):47–66.

Uçar, M. K., Bozkurt, M. R., Bilgin, C., and Polat, K. (2018). Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques. *Neural Computing and Applications*, 29:1–16.