# HARNESSING CARDIO-RESPIRATORY SLEEP STAGING UNDER UNCERTAINTY

**Jonathan F. Carter and Lionel Tarassenko**
Institute of Biomedical Engineering, University of Oxford
jcarter@robots.ox.ac.uk

## ABSTRACT

Automatic sleep stage classification from cardio-respiratory signals has emerged as a promising alternative to traditional polysomnography, which typically uses an extensive set of sensors including electrodes attached to the scalp. Despite impressive results to date, we argue that to harness the benefits of cardio-respiratory sleep staging, we require a greater focus on building models with calibrated uncertainty quantification. We describe how such models could enable important applications in sleep medicine, without necessarily requiring expert-level accuracy as measured by conventional metrics. Our work motivates further investigation into better-calibrated sleep staging models, to enable these applications.

## 1 INTRODUCTION

Given the importance of sleep monitoring and the ubiquity and ease of wearable devices such as smartwatches, prior work has investigated using cardio-respiratory signals to classify stages of sleep e.g. (Walch et al., 2019; Radha et al., 2021; Davidson et al., 2023). Using modalities such as the photoplethysmogram (PPG) or the electrocardiogram (ECG), these methods have achieved increasingly high levels of agreement with expert-annotated sleep stages using the American Academy of Sleep Medicine (AASM) scoring rules (Iber, 2007). However, their performance still lags behind that of conventional sleep monitoring i.e. polysomnography (PSG), which uses the electoencephalogram (EEG) as the main input source. At the time of writing, no device has been cleared by a regulatory body such as the US Food and Drug Administration (FDA) for automatic sleep stage classification (sleep staging) from non-EEG sensors.

In this paper, we argue that a greater focus on uncertainty quantification is required to leverage the benefits of cardio-respiratory sleep staging. In short, a model that is as accurate as a human expert 80% of the time could still have significant value, so long as it can indicate that it is uncertain the other 20% of the time. We describe how a well-calibrated[1] cardio-respiratory sleep staging model could be used as part of a screening workflow. We then perform simple experiments which illustrate how better uncertainty measures could be used to tailor the performance of sleep staging models for this purpose. Our work motivates further research designing sleep staging models that exhibit calibrated measures of uncertainty, which in turn, can have valuable applications in sleep medicine.

## 2 BACKGROUND AND RELATED WORK

### 2.1 AUTOMATED SLEEP STAGING

To reduce the manual effort of sleep assessment, many prior works have investigated automated sleep staging from both EEG (Phan & Mikkelsen, 2022) and non-EEG sensors (Imtiaz, 2021). Typically these methods jointly classify sleep stage sequences $y_{1:T}$ from a (possibly multivariate) input time series $\boldsymbol{x}_{1:kT}$ i.e. performing sequence–sequence classification. Here we use $k$ to account for the relative difference in sampling rate between output sleep stages, typically generated at 30-second intervals, and the input signals, which usually have much higher sampling rates.

[1] i.e. the output probabilities reflect the empirical probability of correctness.

These methods commonly produce an output distribution $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$ which assumes a factorized posterior over the sleep stages, i.e.:

$$p(y_{1:T}|\boldsymbol{x}_{1:kT}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{x}_{1:kT}) \tag{1}$$

The vast majority of sleep staging methods, including SleepTransformer (Phan et al., 2022) and SleepPPG-Net (Kotzen et al., 2023), state-of-the-art methods for EEG- and PPG-based sleep staging, assume this factorized output posterior. The arg max over the output probabilities from the model are typically then taken as the output sequence of sleep stage classifications:

$$\begin{aligned} \hat{y}_{1:T} &= \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\} \\ \hat{y}_t &= \arg\max_{y_t} p_\theta(y_t|\boldsymbol{x}_{1:kT}) \end{aligned} \tag{2}$$

These sequences are commonly displayed in hypnograms, such as in Figure 1. Visually, sleep hypnograms can convey important information to clinicians that can help to form a diagnosis.

Metrics can also be derived from the sequence of sleep stages $y_{1:T}$, which give a more condensed view of a night's sleep. For example, low REM onset latency (ROL, (3)) can indicate the presence of narcolepsy (Mosko et al., 1984). Meanwhile, there are links between a decrease in total deep (N3) sleep time (4) and Alzheimer's disease (Lee et al., 2020).

$$\begin{aligned} m_{ROL} = f_{ROL}(y_{1:T}) &= t_R - t_S \\ \text{where} \quad t_R &= \min\{t \in \{1, \ldots, T\} : y_t = \text{REM}\} \\ \text{and} \quad t_S &= \min\{t \in \{1, \ldots, T\} : y_t \neq \text{Wake}\} \end{aligned} \tag{3}$$

$$m_{N3} = f_{N3}(y_{1:T}) = \sum_{t=1}^{T} \mathbf{1}(y_t = \text{N3}) \tag{4}$$

### 2.2 Uncertainty Quantification and Posterior Sampling

Phan et al. (2022) showed that the predictive entropy of an EEG-based model could be used to divide classifications into high and low-confidence sets, such that only low-confidence sleep labels are sent for human review, thereby reducing the manual annotation time. However, this 'human-in-the-loop' workflow, e.g. (Kang et al., 2021; Heremans et al., 2023), is only possible for an EEG-based approach, since there are no guidelines for scoring sleep stages from cardio-respiratory signals.

By sampling from the posterior distribution over sleep stages $Y_{1:T} \sim p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$, and then evaluating the resulting sleep metric $M_i = f_i(Y_{1:T})$, samples from the sleep metric distribution $p_\theta(m_i|\boldsymbol{x}_{1:kT})$ can also be drawn:

$$\mu_{\text{i}} = \mathbb{E}[M_{\text{i}}] = \mathbb{E}_{Y_{1:T} \sim p_\theta}[f_i(Y_{1:T})] \tag{5}$$

$$\text{Var}(M_{\text{i}}) = \mathbb{E}_{Y_{1:T} \sim p_\theta}[(f_i(Y_{1:T}) - \mu_{\text{i}})^2] \tag{6}$$

However, as noted by van Gorp et al. (2023), directly sampling output sequences from a factorized posterior (1) ignores the temporal structure of sleep stages.

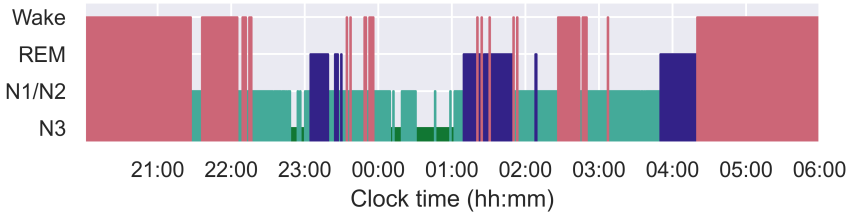



Figure 1: Example of an expert-labelled sleep hypnogram $y_{1:T}$. AASM scoring rules divide sleep into five stages: Wake, N1 (light), N2 (intermediate), N3 (deep) and rapid-eye-movement (REM) sleep. Due to low inter-scorer agreement (Danker-Hopfe et al., 2009), automatic sleep staging methods often merge N1 and N2 into a single class as shown.

## 3 WHY DO WE NEED CALIBRATED UNCERTAINTY?

As with any other contact sensor, data from cardio-respiratory signals such as the PPG suffer from noise (Charlton et al., 2023). However, given that these devices often only measure a single sensory modality, they have less redundancy. Polysomnography typically measures brain activity from multiple sites using the EEG, plus other complementary physiological measurements, such as muscle activity around the eyes and under the chin, which can all aid the classification of sleep stages. Therefore, when using cardio-respiratory signals, we should expect a greater proportion of the data to be unscorable due to signal noise and should design models that account for this.

Additionally, even though cardio-respiratory signals are known to encode for sleep stage information (Shinar et al., 2001; Hudgel et al., 1984), it may not always be possible to distinguish sleep stages solely from these signals, even in the absence of noise. AASM stages are a discrete model of sleep, whose states are ultimately defined by rules that are predominantly based on characteristic patterns of brain activity. For example, a 30-second epoch is labelled as N3 (deep) sleep where at least 20% of the epoch consists of slow-wave (0.5–2 Hz) EEG activity above 75 $\mu$V (Iber, 2007). Therefore, *exact* agreement between sleep stages from cardio-respiratory signals and expert-annotated stages from PSG relies on the assumption that these specific measurements of the central nervous system (CNS) e.g. EEG can be accurately predicted from measurements of the autonomic nervous system (ANS) e.g. PPG. This is before considering that expert-annotated sleep stages are **not** a ground truth, with high inter-expert disagreement (Danker-Hopfe et al., 2009; van Gorp et al., 2022).

Numerous factors can affect CNS-ANS coupling (de Zambotti et al., 2018), which underpins cardio-respiratory sleep staging. Physical and mental health e.g. (Vanoli et al., 1995; Krystal, 2012), medications (Pagel & Parnes, 2001), and factors such as caffeine (Barry et al., 2008) and alcohol intake (Thakkar et al., 2015) may all affect the observed relationship between cardio-respiratory signals and sleep stages. We cannot expect to observe *all* combinations of factors that may affect this relationship during model training. Therefore, a model must be able to identify ambiguous or out-of-distribution inputs, to avoid producing incorrect outputs.

If the challenges outlined in this section can be overcome, the ubiquity, ease of use, and low cost of wearables means they could have a number of novel, valuable clinical applications, such as:

- Accelerating the discovery of treatments for sleep disorders such as insomnia, by enabling faster and cheaper clinical trials on larger study populations.
- Longitudinal monitoring of *known* sleep biomarkers to detect the presence of underlying conditions. For example, the well-known decrease in N3 sleep with Alzheimer's disease.
- The discovery of *novel* sleep biomarkers, using much larger study populations than can feasibly be collected using conventional PSG.

As we illustrate in the next section, a greater focus on uncertainty quantification, rather than simple summary statistics such as accuracy or Cohen's $\kappa$ (Cohen, 1960), can help to enable these.

## 4 HOW CAN WE USE CALIBRATED UNCERTAINTY?

Sleep staging with polysomnography can often be used to indicate or rule out the presence of specific sleep disorders. In Figure 2, we illustrate how calibrated uncertainty could enable an imperfect (measured by accuracy) but well-calibrated wearable-based sleep staging model $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$ to be used as an effective screening tool. If the results are conclusive i.e. the uncertainty in the outputs of interest are low, then these could be used to inform clinical decision making. Otherwise, the patient could be referred for alternative tests, such as a PSG exam. Even if the outputs are only conclusive 80% of the time, this could still significantly reduce (5x) the need for more time-consuming and expensive tests. An important consideration in the design of such a workflow is balancing the interpretability and granularity of information presented to the clinician. Alternative approaches have been proposed for presenting sleep staging uncertainty, such as hypnodensity charts (Bakker et al., 2023). However, if interpreting and processing the outputs results in significant additional work for the clinician, this could limit adoption (Greenhalgh et al., 2004).

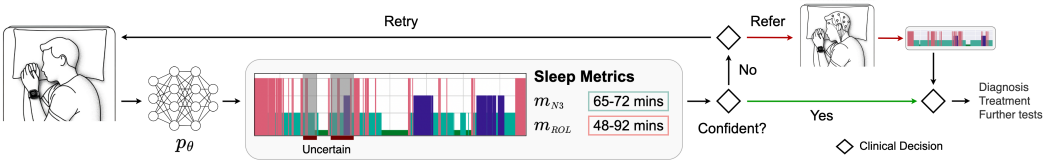
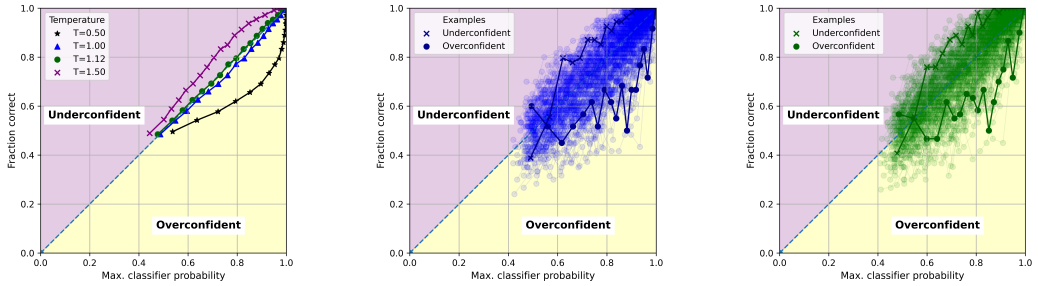



Figure 2: **Sleep staging under uncertainty using cardio-respiratory signals from wearables.** With a *calibrated* model $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$, simple and interpretable measures of uncertainty could be used to inform a clinician about the reliability of the outputs, such as confidence intervals on sleep metrics. If the uncertainty in the relevant outputs is low (e.g. $m_{N3}$ here), then these could be used to inform clinical decision-making. If not, the ubiquity and ease of wearables means that multiple nights of data could be used to reduce certain forms of uncertainty e.g. sensor noise from loose attachment. Otherwise, the patient could be referred for alternative tests, such as a full PSG exam.

## 5 Experiments

We investigate the uncertainty quantification properties of the current state-of-the-art model for sleep staging from the photoplethysmogram, SleepPPG-Net (Kotzen et al., 2023) using the MESA dataset (Chen et al., 2015), following the setup described in the original paper.

### 5.1 Model Calibration




(a) Total calibration. (b) Per-night ($T = 1$). (c) Per-night ($T_{\text{Opt}} = 1.12$).

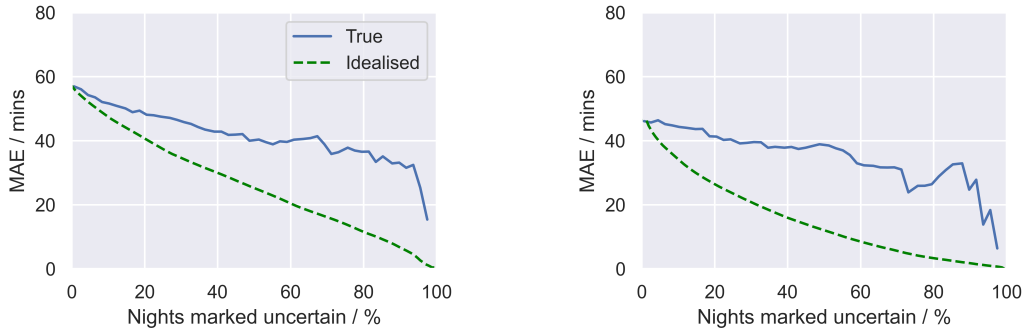
Figure 3: Reliability diagrams for MESA-Test before and after temperature scaling.

Figure 3 shows reliability diagrams for the SleepPPG-Net model before and after applying temperature scaling (Guo et al., 2017) to improve calibration. When aggregated over all sleep labels in the test set, Figure 3a gives a false impression that the model is well-calibrated. In practice, we find that this aggregation masks the true behaviour, that the sleep stage classifications for most nights are either highly under or overconfident, and which temperature scaling is unable to remedy, as shown in Figures 3b and 3c.

### 5.2 Performance–Yield Trade-off

Next, we investigate the ability to mark outputs as uncertain using simple heuristics derived from the posterior $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$, which we believe is a desirable property to enable the workflow described in the previous section. For a well-calibrated posterior, we expect that nights where the posterior variance of sleep metrics (6) is lower should, on average, give lower errors in the resulting estimate (5). So, by marking outputs as uncertain when this variance is above a threshold i.e. $\text{Var}(M_i) > \delta_i$, we should be able to reduce errors at the expense of yield i.e. the proportion of nights marked as certain.

Figure 4 shows error–yield curves for $m_{N3}$ and $m_{ROL}$ as we sweep over thresholds $\delta_i$ for each metric using SleepPPG-Net, which results in only modest error reductions in both sleep metrics.




(a) $\mathrm{Var}(M_{N3}) < \delta_{N3}$ (b) $\mathrm{Var}(M_{ROL}) < \delta_{ROL}$

Figure 4: Error–yield curves using SleepPPG-Net on MESA-Test for (a) $m_{N3}$ and (b) $m_{ROL}$, for varying thresholds $\delta_i$ on the max. posterior variance. Dashed lines indicate the best-case behaviour of the model for the given level of accuracy i.e. if the model knew exactly when it was least accurate, by sorting nights from highest to lowest error and marking as uncertain in perfect order.

Orthogonal to improving raw accuracy statistics i.e. the intercepts of Figure 4, better model calibration could greatly improve this two-dimensional view of performance. This is illustrated by the dashed lines, which show idealised, best-case behaviour for SleepPPG-Net. By improving the utility of sleep staging models on these error-yield curves, this can, in turn, improve their utility for applications such as the screening workflow described in Section 4.

## 6 Conclusions

In this paper, we have discussed how calibrated uncertainty quantification could enable important applications of cardio-respiratory sleep staging without requiring expert-level accuracy as measured by conventional metrics, and highlighted deficiencies in existing state-of-the-art models. We believe there are three complementary avenues that together can address these deficiencies:

1. Further research into architectures that do not employ a factorized posterior e.g. U-Flow (van Gorp et al., 2023), to enable proper sampling from $p_\theta(y_{1:T}|\boldsymbol{x}_{1:kT})$.
2. Improving model calibration, by building on recent advances in Bayesian deep learning e.g. (Lakshminarayanan et al., 2017; Liu et al., 2020).
3. Using labels from multiple experts during training **and** evaluation (Fiorillo et al., 2023), to account for inter-expert subjectivity (Danker-Hopfe et al., 2009).

By designing models that combine the capacity and inductive biases necessary to achieve high accuracy, with calibrated measures of uncertainty that better account for the time-series nature of the problem, we believe this can lead to valuable, new applications in sleep medicine.

## Acknowledgements

This work was supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1] and funded by Oxehealth Ltd. We gratefully acknowledge the National Sleep Research Resource for granting access to the MESA dataset.

## References

Jessie P Bakker, Marco Ross, Andreas Cerny, Ray Vasko, Edmund Shaw, Samuel Kuna, Ulysses J Magalang, Naresh M Punjabi, and Peter Anderer. Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*, 46(2):zsac154, February 2023. ISSN 0161-8105, 1550-9109. doi: 10.1093/sleep/zsac154. URL https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsac154/6628222.

Robert J. Barry, Adam R. Clarke, Stuart J. Johnstone, and Jacqueline A. Rushby. Timing of caffeine's impact on autonomic and central nervous system measures: Clarification of arousal effects. *Biological Psychology*, 77(3):304–316, March 2008. ISSN 0301-0511. doi: 10.1016/j.biopsycho.2007.11.002. URL https://www.sciencedirect.com/science/article/pii/S0301051107001895.

Peter H. Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A. Behar, Fei Chen, Gari D. Clifford, David A. Clifton, Harry J. Davies, and Cheng Ding. The 2023 wearable photoplethysmography roadmap. *Physiological measurement*, 44(11):111001, 2023. URL https://iopscience.iop.org/article/10.1088/1361-6579/acead2/meta. Publisher: IOP publishing.

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

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644, 1552-3888. doi: 10.1177/001316446002000104. URL http://journals.sagepub.com/doi/10.1177/001316446002000104.

Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, Bernd Saletu, Andrea Schmidt, and Georg Dorffner. Inter-rater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research*, 18(1):74–84, 2009. ISSN 1365-2869. doi: 10.1111/j.1365-2869.2008.00700.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2869.2008.00700.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2869.2008.00700.x.

Shaun Davidson, Cristian Roman, Jonathan Carter, Mirae Harford, and Lionel Tarassenko. Sleep Staging Using Wearables and Deep Neural Networks. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, October 2023. doi: 10.1109/BHI58575.2023.10313497. URL https://ieeexplore.ieee.org/document/10313497. ISSN: 2641-3604.

Massimiliano de Zambotti, John Trinder, Alessandro Silvani, Ian M. Colrain, and Fiona C. Baker. Dynamic coupling between the central and autonomic nervous systems during sleep: A review. *Neuroscience & Biobehavioral Reviews*, 90:84–103, July 2018. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2018.03.027. URL https://www.sciencedirect.com/science/article/pii/S0149763417306577.

Luigi Fiorillo, Davide Pedroncelli, Valentina Agostini, Paolo Favaro, and Francesca Dalia Faraci. Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring. *Sleep*, 46(5):zsad028, May 2023. ISSN 0161-8105. doi: 10.1093/sleep/zsad028. URL https://doi.org/10.1093/sleep/zsad028.

Trisha Greenhalgh, Glenn Robert, Fraser Macfarlane, Paul Bate, and Olivia Kyriakidou. Diffusion of Innovations in Service Organizations: Systematic Review and Recommendations. *The Milbank Quarterly*, 82(4): 581–629, December 2004. ISSN 0887-378X, 1468-0009. doi: 10.1111/j.0887-378X.2004.00325.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.0887-378X.2004.00325.x.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Elisabeth R. M. Heremans, Nabeel Seedat, Bertien Buyse, Dries Testelmans, Mihaela van der Schaar, and Maarten De Vos. U-PASS: an Uncertainty-guided deep learning Pipeline for Automated Sleep Staging, June 2023. URL http://arxiv.org/abs/2306.04663. arXiv:2306.04663 [cs, eess].

D. W. Hudgel, R. J. Martin, B. Johnson, and P. Hill. Mechanics of the respiratory system and breathing pattern during sleep in normal humans. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, 56(1):133–137, January 1984. ISSN 0161-7567. doi: 10.1152/jappl.1984.56.1.133.

C. Iber. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification. 2007. URL https://cir.nii.ac.jp/crid/1370004237604151044.

Syed Anas Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5): 1562, 2021. Publisher: MDPI.

Dae Y. Kang, Pamela N. DeYoung, Justin Tantiongloc, Todd P. Coleman, and Robert L. Owens. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *npj Digital Medicine*, 4(1):1–9, September 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00515-3. URL https://www.nature.com/articles/s41746-021-00515-3. Number: 1 Publisher: Nature Publishing Group.

Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar. SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 27(2):924–932, February 2023. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2022.3225363. URL https://ieeexplore.ieee.org/document/9965588/.

Andrew D. Krystal. Psychiatric disorders and sleep. *Neurologic clinics*, 30(4):1389–1413, 2012. URL https://www.neurologic.theclinics.com/article/S0733-8619(12)00059-X/abstract. Publisher: Elsevier.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Yee Fun Lee, Dmitry Gerashchenko, Igor Timofeev, Brian J. Bacskai, and Ksenia V. Kastanenka. Slow Wave Sleep Is a Promising Intervention Target for Alzheimer's Disease. *Frontiers in Neuroscience*, 14, 2020. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2020.00705.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness, October 2020. URL http://arxiv.org/abs/2006.10108. Number: arXiv:2006.10108 arXiv:2006.10108 [cs, stat].

Sarah S. Mosko, David S. Shampain, and Jon F. Sassin. Nocturnal REM Latency and Sleep Disturbance in Narcolepsy. *Sleep*, 7(2):115–125, September 1984. ISSN 0161-8105, 1550-9109. doi: 10.1093/sleep/7.2.115. URL https://academic.oup.com/sleep/article-lookup/doi/10.1093/sleep/7.2.115.

J. F. Pagel and Bennett L. Parnes. Medications for the treatment of sleep disorders: an overview. *Primary care companion to the Journal of clinical psychiatry*, 3(3):118, 2001. URL https://www.psychiatrist.com/wp-content/uploads/2021/02/25094_medications-treatment-sleep-disorders-overview.pdf. Publisher: Physicians Postgraduate Press, Inc.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, April 2022. ISSN 0967-3334. doi: 10.1088/1361-6579/ac6049. URL https://dx.doi.org/10.1088/1361-6579/ac6049.

Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, August 2022. ISSN 1558-2531. doi: 10.1109/TBME.2022.3147187. URL https://ieeexplore.ieee.org/abstract/document/9697331.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M. Aarts. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *npj Digital Medicine*, 4(1):1–11, September 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00510-8. URL https://www.nature.com/articles/s41746-021-00510-8.

Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod. Automatic detection of slow-wave-sleep using heart rate variability. In *Computers in Cardiology 2001. Vol.28*, pp. 593–596, September 2001. doi: 10.1109/CIC.2001.977725.

Mahesh M. Thakkar, Rishi Sharma, and Pradeep Sahota. Alcohol disrupts sleep homeostasis. *Alcohol*, 49(4):299–310, June 2015. ISSN 0741-8329. doi: 10.1016/j.alcohol.2014.07.019. URL https://www.sciencedirect.com/science/article/pii/S0741832914201157.

Hans van Gorp, Iris A M Huijben, Pedro Fonseca, Ruud J G van Sloun, Sebastiaan Overeem, and Merel M van Gilst. Certainty about uncertainty in sleep staging: a theoretical framework. *Sleep*, 45(8), August 2022. ISSN 0161-8105, 1550-9109. URL https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsac134/6604464.

Hans van Gorp, Merel M. van Gilst, Pedro Fonseca, Sebastiaan Overeem, and Ruud J. G. van Sloun. Aleatoric Uncertainty Estimation of Overnight Sleep Statistics Through Posterior Sampling Using Conditional Normalizing Flows. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10096894. URL https://ieeexplore.ieee.org/abstract/document/10096894. ISSN: 2379-190X.

Emilio Vanoli, Philip B. Adamson, null Ba-Lin, Gian D. Pinna, Ralph Lazzara, and William C. Orr. Heart Rate Variability During Specific Sleep Stages. *Circulation*, 91(7):1918–1922, April 1995. doi: 10.1161/01.CIR.91.7.1918. URL https://www.ahajournals.org/doi/full/10.1161/01.cir.91.7.1918. Publisher: American Heart Association.

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