NAP: Attention-Based Late Fusion for Automatic Sleep Staging

Alvise Dei Rossi
Faculty of Informatics
Università della Svizzera Italiana
Institute of Digital Technologies for Personalized Healthcare
University of Applied Sciences and Arts of Southern Switzerland
Lugano, Switzerland
alvise.dei.rossi@usi.ch

Julia van der Meer & Markus H. Schmidt & Claudio L.A. Bassetti
Sleep Wake Epilepsy Center, Department of Neurology, Inselspital, Bern University Hospital
University of Bern
Bern, Switzerland
{julia.vandermeer, markus.schmidt, claudio.bassetti}@insel.ch

Luigi Fiorillo & Francesca Faraci
Institute of Digital Technologies for Personalized Healthcare
University of Applied Sciences and Arts of Southern Switzerland
Lugano, Switzerland
{luigi.fiorillo, francesca.faraci}@supsi.ch

Abstract

Polysomnography signals are highly heterogeneous, varying in modality composition (e.g., EEG, EOG, ECG), channel availability (e.g., frontal, occipital EEG), and acquisition protocols across datasets and clinical sites. Most existing models that process polysomnography data rely on a fixed subset of modalities or channels and therefore neglect to fully exploit its inherently multimodal nature. We address this limitation by introducing **NAP** (**N**eural **A**ggregator of **P**redictions), an attention-based model which learns to combine multiple prediction streams using a tri-axial attention mechanism that captures temporal, spatial, and predictor-level dependencies. NAP is trained to adapt to different input dimensions. By aggregating outputs from frozen, pretrained single-channel models, NAP consistently outperforms individual predictors and simple ensembles, achieving state-of-the-art zero-shot generalization across multiple datasets. While demonstrated in the context of automated sleep staging from polysomnography, the proposed approach could be extended to other multimodal physiological applications.

Introduction

Polysomnography (PSG), the clinical gold standard for diagnosing sleep–wake disorders, records multiple physiological signals (e.g, EEG, EOG, EMG) using channel configurations that can vary considerably across clinical centers Phan & Mikkelsen (2022). These signals are typically segmented into 30-second windows, referred to as *sleep epochs*, and manually classified by trained clinicians into five sleep stages (Wake, N1, N2, N3, REM) Berry et al. (2017). The resulting *hypnogram* provides a comprehensive representation of sleep macrostructure, critical for identifying a wide range of sleep

EurIPS 2025 Workshop: Multimodal Representation Learning for Healthcare.

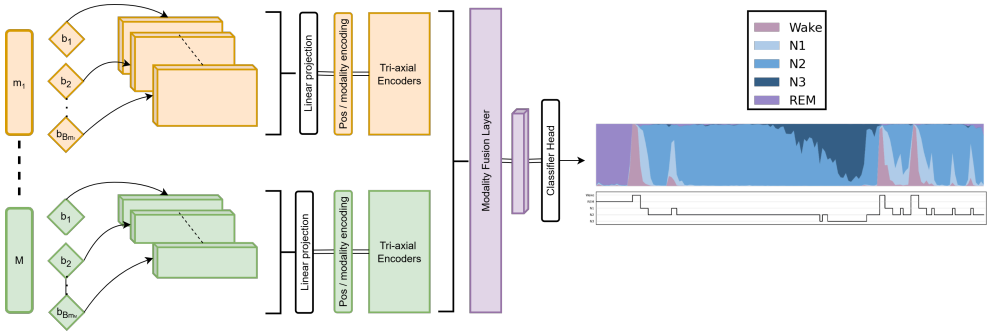

Figure 1: Overview of the Neural Aggregator of Predictions architecture. NAP flexibly integrates an arbitrary number of physiological modalities, different channel configurations, and base predictors.

disorders Ibáñez et al. (2018). Manual sleep staging, however, is time-consuming, and prone to high subjective bias. Leveraging large, annotated PSG datasets Zhang et al. (2018), researchers have increasingly pursued automated approaches to sleep staging Fiorillo et al. (2019), employing diverse modeling paradigms including convolutional Perslev et al. (2021), recurrent Phan et al. (2023), and attention-based Phan et al. (2022) networks. Models that operate on a subset of inputs, are typically suboptimal, since complementary modalities provide a more holistic view of sleep dynamics Phan & Mikkelsen (2022) and may offer better insights into a patient's health status Thapa et al. (2025). Combining information from multiple (sub-)networks that either fuse learned representations (*early fusion*) or aggregate model predictions (*late fusion*) Stahlschmidt et al. (2022) is therefore beneficial; however, effectively managing the heterogeneity of information remains challenging.

In this work, we focus on late fusion, which in automatic sleep staging is typically implemented via (soft-)voting across channels Perslev et al. (2021), modalities, or models Stephansen et al. (2018); Dei Rossi et al. (2025). Despite its simplicity and modularity, soft-voting assumes that averaging constitutes an adequate aggregation function and implicitly treats all contributors as equally reliable. Furthermore, it operates at epoch level, disregarding temporal dependencies that could otherwise be exploited to improve predictive accuracy. To overcome these limitations while preserving flexibility:

- We propose **NAP** (**N**eural **A**ggregator of **P**redictions), a lightweight attention-based meta-model that learns to aggregate predictions from pretrained single-channel models by explicitly capturing temporal, spatial/channel, model-level, and cross-modality dependencies.
- We generalize criss-cross attention Huang et al. (2019) beyond spatio-temporal dimensions Wang et al. (2024) to a tri-axial attention mechanism, serving as an effective fusion strategy.
- We extend dimension adaptive training Malekzadeh et al. (2021) to dynamically sample varying sequence lengths, numbers of channels, models, and modalities across batches.

Methods

Model architecture

The NAP architecture, shown in Figure 1, is organized into four conceptually distinct modules: (i) a *base predictions generator*; (ii) a *tri-axial attention encoder*; (iii) a *modality fusion layer*; and (iv) a *classifier head*. We detail in the next paragraphs each module, focusing on a single instance within a batch, under the scenario where all available information within a PSG is utilized.

Base predictions generator. Let a PSG recording X be composed of a sequence of T contiguous sleep epochs, $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$, each associated with a ground-truth label $y_t \in \mathcal{S} = \{\text{Wake}, \text{N1}, \text{N2}, \text{N3}, \text{REM}\}$. We denote by M the number of physiological modalities present in X, where each modality m_k has C_{m_k} available channels, and for which we have access to B_{m_k} pretrained base predictors. For a modality m_k ($k = 1, \ldots, M$), channel c_j ($j = 1, \ldots, C_{m_k}$), and base predictor b_ℓ ($\ell = 1, \ldots, B_{m_k}$), we get the corresponding prediction $\left\{\hat{\mathbf{h}}_{(m_k, c_j, b_\ell), t} \;\middle|\; t = 1, \ldots, T\right\}$, where $\hat{\mathbf{h}}_{(m_k, c_j, b_\ell), t} \in \mathbb{R}^5$ is an epoch-wise probability distribution over $\mathcal{S}$, subsequently linearly

projected into a higher-dimensional feature space $\mathbb{R}^{d_{\text{model}}}$. The predictions represent the *hypnodensities* Stephansen et al. (2018), probabilistic representations of sleep stages over time.

Tri-axial attention encoder For a single modality m_k, the projected hypnodensity features are collected in a tensor $\hat{\mathbf{H}}_{(m_k)} \in \mathbb{R}^{T \times C_{m_k} \times B_{m_k} \times d_{model}}$. To encode both temporal order and modality identity, we add the standard relative positional encoding from Vaswani (2017), and a learnable embedding vector uniquely assigned to m_k. The resulting tensor is then processed by L stacked transformer encoder layers, each employing a tri-axial self-attention mechanism that extends the criss-cross attention paradigm Huang et al. (2019). Instead of computing a single joint attention map over all dimensions, the mechanism factorizes the standard multi-head attention into three pathways:

- **Spatial attention:** Attends over the channel axis C_{m_k} while holding the time and predictor dimensions fixed, capturing cross-channel dependencies.
- **Temporal attention:** Attends along the sequence length axis T while keeping channel and predictor dimensions fixed, enabling the model to learn temporal dependencies.
- **Blending attention:** Attends along the base predictor axis B_{m_k} while keeping channel and time dimensions fixed, facilitating the fusion of predictions from different base models.

To achieve this, the h attention heads are divided evenly across the three pathways, allowing each group of $h/3$ heads to specialize in modeling dependencies along a single axis. We apply pre-normalization of queries and keys Ba et al. (2016), and omit the bias term in attention computations Jiang et al. (2024). For the spatial pathway, the attention output is computed as:

$$Z_s^{(i)} = \text{Softmax}\left(\frac{\text{LN}(Q_s^{(i)})\,\text{LN}(K_s^{(i)})^\top}{\sqrt{d_\text{k}}}\right) V_s^{(i)}; \qquad Z_s = \text{Concat}\left(Z_s^{(1)}, \ldots, Z_s^{(h/3)}\right). \tag{1}$$

Analogous computations produce the temporal and blending pathway outputs Z_T and Z_B. The pathways outputs are then concatenated along the feature dimension. Finally, a feedforward network with residual connections and dropout is applied Vaswani (2017).

Modality fusion layer After the tri-axial encoders have processed each modality m_k independently, the outputs are concatenated along the feature dimension, yielding $\tilde{\mathbf{Z}} \in \mathbb{R}^{T \times N \times d_{\text{model}}}$, where $N = \sum_{k=1}^{M}(C_{m_k} \cdot B_{m_k})$ is the total number of prediction streams. To reduce $\tilde{\mathbf{Z}}$ to a compact, per-epoch feature vector, we employ an attention-based fusion mechanism Phan et al. (2022) that learns to weight the contributions of different channels, predictors, and modalities. For each time step t, the fusion layer computes a convex combination $\hat{\mathbf{z}}_t = \sum_{n=1}^{N} \alpha_{t,n}\, \tilde{\mathbf{z}}_{t,n}$, where $\alpha_{t,n} \in [0, 1]$ are normalized attention weights, obtained by projecting each $\tilde{\mathbf{z}}_{t,n}$ into an attention space of dimension d_A using a learned transformation, then scored by an epoch-level context vector:

$$\alpha_{t,n} = \frac{\exp(\tanh(W_A \mathbf{x}_{t,n} + b_A)^\top u_A)}{\sum_{j=1}^{N} \exp(\tanh(W_A \mathbf{x}_{t,j} + b_A)^\top u_A)}. \tag{2}$$

where $W_A \in \mathbb{R}^{d_\text{model} \times d_\text{A}}$, $b_A \in \mathbb{R}^{d_\text{A}}$, and $u_A \in \mathbb{R}^{d_\text{A}}$ are learnable parameters.

Classifier head The epoch representations are finally fed into a compact classifier head, comprised of a single hidden layer feedforward network with dropout, to produce sleep stage predictions. The NAP model is trained end-to-end using a cross-entropy loss against the ground-truth annotations.

Training protocol

We train NAP on inputs with varying dimensionality, pushing it to operate across different modality subsets, channel counts, and sequence lengths. We generate batches by randomly selecting a subset of dimensions along four axes: the number of time steps, the set of modalities, the number of channels, and the set of base predictors. Along the temporal axis, we uniformly sample K contiguous segments of the same random length from each of the B recordings within the batch. A subset of available

modalities is then randomly selected and, within each chosen modality, a random subset of channels and base predictors are sampled. This procedure yields batches where all samples share the same input shape although the shape may vary between batches. As a result, padding and masking are not required, improving computational efficiency. Furthermore, instead of performing a single gradient update per batch, we accumulate gradients over G distinct batches Malekzadeh et al. (2021). Further methodological and implementation details are reported in A.1 and A.2, respectively.

Experiments

We leverage the pre-trained single-channel models of SLEEPYLAND Dei Rossi et al. (2025) as base predictors. Currently, these models cover the EEG and EOG modalities, and include U-Sleep Perslev et al. (2021), DeepResNet Olesen et al. (2021), and SleepTransformer Phan et al. (2022) models.

Datasets The base models were pre-trained on several PSG datasets available from the National Sleep Research Resource Zhang et al. (2024). To prevent data leakage, NAP is trained on their hold out sets and on the BSWR dataset Aellen et al. (2024), unseen by the base models. Additional out-of-domain (OOD) datasets, never seen during the training of either the frozen base predictors or NAP itself, are used for evaluation: DOD-H & DOD-O Guillot et al. (2020), DCSM Perslev et al. (2021), SEDF-SC & SEDF-ST Kemp et al. (2000), and PHYS Ghassemi et al. (2018). Further details for all datasets are provided in Appendix A.3 and A.4. All recordings are resampled to 128 Hz and standardized using channel-wise robust scaling. Data splits are performed at the subject level.

Results

Table 1: Per-recording mean (SD) Macro F1 and per-stage F1 scores for the best individual ensemble model (soft-voting across all channels); the SOMNUS ensemble Dei Rossi et al. (2025) (soft-voting across all channels, modalities and models); and NAP. Metrics computed on the BSWR test set and all OOD datasets. ‡ indicates statistically significant ($\alpha < 0.05$) MF1 improvement over other methods.

Dataset	Model	MF1	$\mathbf{F1}_W$	$\mathbf{F1}_{N1}$	$\mathbf{F1}_{N2}$	$\mathbf{F1}_{N3}$	$\mathbf{F1}_{REM}$
BSWR	DeepResNet$_{EEG}$	.695(.120)	.828(.143)	.397(.172)	.793(.148)	.629(.270)	.848(.180)
	SOMNUS	.708(.120)	.836(.141)	.404(.178)	.804(.146)	.696(.280)	.864(.173)
	NAP	**.749(.117)**‡	**.856(.132)**	**.533(.164)**	**.809(.146)**	**.705(.260)**	**.864(.172)**
DCSM	DeepResNet$_{EEG}$	.797(.086)	.981(.027)	.507(.147)	.849(.096)	.779(.207)	.874(.149)
	SOMNUS	.803(.084)	.983(.023)	.505(.153)	**.858(.097)**	.783(.202)	.891(.146)
	NAP	**.815(.081)**‡	**.986(.020)**	**.550(.143)**	.848(.103)	**.802(.190)**	**.893(.145)**
DOD-H	U-Sleep$_{EEG}$	.816(.072)	.878(.085)	.526(.166)	.907(.051)	.851(.171)	.916(.073)
	SOMNUS	.828(.064)	**.886(.089)**	.564(.162)	**.912(.043)**	**.866(.161)**	**.930(.056)**
	NAP	**.834(.071)**	.876(.096)	**.616(.159)**	.904(.048)	.849(.165)	.927(.068)
DOD-O	U-Sleep$_{EEG}$	.777(.081)	.911(.066)	.496(.144)	.882(.070)	.693(.265)	.912(.084)
	SOMNUS	**.790(.084)**	**.914(.068)**	.516(.153)	**.882(.075)**	**.733(.266)**	.913(.071)
	NAP	.776(.093)	.874(.114)	**.523(.141)**	.864(.083)	.715(.265)	**.915(.068)**
PHYS	DeepResNet$_{EEG}$	.687(.097)	.744(.159)	.358(.153)	.832(.106)	.682(.247)	.837(.173)
	SOMNUS	.693(.099)	.743(.161)	.349(.157)	**.837(.107)**	.704(.248)	.847(.170)
	NAP	**.732(.095)**‡	**.780(.150)**	**.494(.141)**	.829(.109)	**.722(.237)**	**.849(.168)**
SEDF-SC	U-Sleep$_{EEG}$	.720(.090)	.981(.014)	.342(.130)	.814(.097)	.602(.287)	.845(.114)
	SOMNUS	.734(.083)	.982(.018)	.358(.138)	**.832(.083)**	**.611(.279)**	**.870(.094)**
	NAP	**.752(.084)**‡	**.983(.023)**	**.469(.127)**	.819(.090)	.605(.284)	.864(.102)
SEDF-ST	DeepResNet$_{EEG}$	.764(.074)	.814(.105)	.508(.158)	.863(.062)	.746(.232)	.891(.085)
	SOMNUS	.761(.074)	.803(.105)	.497(.148)	**.873(.057)**	.731(.237)	.902(.080)
	NAP	**.796(.079)**‡	**.847(.096)**	**.606(.162)**	.872(.059)	**.748(.234)**	**.905(.078)**

Table 1 presents the performance of the best single-modality model, SOMNUS, and NAP. Consistent with prior work Dei Rossi et al. (2025), we find that naïve averaging-base late fusion (SOMNUS) outperforms the strongest individual base model. However, attention-based aggregation proves more effective. Across OOD datasets, NAP delivers zero-shot MF1 gains in most cases (DCSM: $0.803 \rightarrow 0.815$, DOD-H: $0.828 \rightarrow 0.834$, PHYS: $0.693 \rightarrow 0.732$, SEDF-SC: $0.734 \rightarrow 0.752$, SEDF-ST: $0.761 \rightarrow 0.796$), indicating that the learned fusion strategy effectively generalize to unseen cohorts. MF1 improvements stem mostly from better recognition of the problematic N1 stage and, in some cases, the Wake stage, most often at the cost of a small decrease in N2 performance.

Conclusion

NAP consistently outperforms both the averaging scheme of SOMNUS and all ensemble constituents, establishing a new state of the art in zero-shot automatic sleep staging across multiple datasets. While previous research has shown that robust zero-shot performance could be achieved by training on larger, diverse datasets, even matching or surpassing earlier in-domain approaches, a few exceptions were noted, namely for PHYS and SEDF Dei Rossi et al. (2025). NAP yields its largest improvements in such cases, suggesting that principled late fusion can close this gap. In contrast, for datasets where SOMNUS had already exceeded prior state of the art, NAP achieves comparable performance.

Although NAP is compatible with any number of modalities, here we considered only EEG and EOG due to the limited availability of pre-trained models Dei Rossi et al. (2025). As more unimodal sleep staging models are released, NAP can be extended to integrate them. Furthermore, while this work presented a framework for late fusion, it could be easily adapted for early or intermediate fusion of representations. Such an approach, a *Neural Aggregator of Representations*, would allow integration of features from modality or channel-specific encoders. Finally, while we demonstrated NAP in the context of automatic sleep staging, the underlying methodology can be applied to any domain that requires principled aggregation of predictive streams across diverse modalities and channels.

References

Florence M Aellen, Julia Van der Meer, Anelia Dietmann, Markus Schmidt, Claudio LA Bassetti, and Athina Tzovara. Disentangling the complex landscape of sleep–wake disorders with data-driven phenotyping: A study of the bernese center. *European journal of neurology*, 31(1):e16026, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jessie P Bakker, Ali Tavakkoli, Michael Rueschman, Wei Wang, Robert Andrews, Atul Malhotra, Robert L Owens, Amit Anand, Katherine A Dudley, and Sanjay R Patel. Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: a randomized controlled trial. *American journal of respiratory and critical care medicine*, 197(8):1080–1083, 2018.

Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Robin M Lloyd, Stuart F Quan, Matthew T Troester, and Bradley V Vaughn. Aasm scoring manual updates for 2017 (version 2.4), 2017.

Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011.

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, 2015.

Alvise Dei Rossi, Matteo Metaldi, Michal Bechny, Irina Filchenko, Julia van der Meer, Markus H Schmidt, Claudio LA Bassetti, Athina Tzovara, Francesca D Faraci, and Luigi Fiorillo. Sleepyland: trust begins with fair evaluation of automatic sleep staging models. *arXiv preprint arXiv:2506.08574*, 2025.

Janet A DiPietro, Radhika S Raghunathan, Hau-Tieng Wu, Jiawei Bai, Heather Watson, Francis P Sgambati, Janice L Henderson, and Grace W Pien. Fetal heart rate during maternal sleep. *Developmental Psychobiology*, 63(5):945–959, 2021.

Luigi Fiorillo, Alessandro Puiatti, Michela Papandrea, Pietro-Luca Ratti, Paolo Favaro, Corinne Roth, Panagiotis Bargiotas, Claudio L Bassetti, and Francesca D Faraci. Automated sleep scoring: A review of the latest approaches. *Sleep medicine reviews*, 48:101204, 2019.

Mohammad M Ghassemi, Benjamin E Moody, Li-Wei H Lehman, Christopher Song, Qiao Li, Haoqi Sun, Roger G Mark, M Brandon Westover, and Gari D Clifford. You snooze, you win: the

physionet/computing in cardiology challenge 2018. In *2018 Computing in Cardiology Conference (CinC)*, volume 45, pp. 1–4. IEEE, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Antoine Guillot, Fabien Sauvet, Emmanuel H During, and Valentin Thorey. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE transactions on neural systems and rehabilitation engineering*, 28(9):1955–1965, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 603–612, 2019.

Vanessa Ibáñez, Josep Silva, and Omar Cauli. A survey on sleep assessment methods. *PeerJ*, 6: e4849, 2018.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Harlin Lee, Boyue Li, Shelly DeForte, Mark L Splaingard, Yungui Huang, Yuejie Chi, and Simon L Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):421, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mohammad Malekzadeh, Richard Clegg, Andrea Cavallaro, and Hamed Haddadi. Dana: Dimension-adaptive neural architecture for multivariate sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–27, 2021.

Carole L Marcus, Reneé H Moore, Carol L Rosen, Bruno Giordani, Susan L Garetz, H Gerry Taylor, Ron B Mitchell, Raouf Amin, Eliot S Katz, Raanan Arens, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368(25): 2366–2376, 2013.

Hyatt Moore IV, Eileen Leary, Seo-Young Lee, Oscar Carrillo, Robin Stubbs, Paul Peppard, Terry Young, Bernard Widrow, and Emmanuel Mignot. Design and validation of a periodic leg movement detector. *PloS one*, 9(12):e114565, 2014.

Doris Moser, Peter Anderer, Georg Gruber, Silvia Parapatics, Erna Loretz, Marion Boeck, Gerhard Kloesch, Esther Heller, Andrea Schmidt, Heidi Danker-Hopfe, et al. Sleep classification according to aasm and rechtschaffen & kales: effects on sleep scoring parameters. *Sleep*, 32(2):139–149, 2009.

Alexander Neergaard Olesen, Poul Jørgen Jennum, Emmanuel Mignot, and Helge Bjarup Dissing Sorensen. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep*, 44(1):zsaa161, 2021.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):1–12, 2021.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, 2022.

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, 2022.

Huy Phan, Kristian P Lorenzen, Elisabeth Heremans, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, Mathias Baumert, Kaare B Mikkelsen, and Maarten De Vos. L-seqsleepnet: Whole-cycle long sequence modeling for automatic sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4748–4757, 2023.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Stuart F Quan, Cynthia S Chan, William C Dement, Alan Gevins, James L Goodwin, Daniel J Gottlieb, Sylvan Green, Christian Guilleminault, Max Hirshkowitz, Pamela R Hyde, et al. The association between obstructive sleep apnea and neurocognitive performance—the apnea positive pressure long-term efficacy study (apples). *Sleep*, 34(3):303–314, 2011.

Susan Redline, Peter V Tishler, Tor D Tosteson, John Williamson, Kenneth Kump, Ilene Browner, Veronica Ferrette, and Patrick Krejci. The familial aggregation of obstructive sleep apnea. *American journal of respiratory and critical care medicine*, 151(3):682–687, 1995.

Carol L Rosen, Emma K Larkin, H Lester Kirchner, Judith L Emancipator, Sarah F Bivins, Susan A Surovec, Richard J Martin, and Susan Redline. Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity. *The Journal of pediatrics*, 142(4):383–389, 2003.

Carol L Rosen, Dennis Auckley, Ruth Benca, Nancy Foldvary-Schaefer, Conrad Iber, Vishesh Kapur, Michael Rueschman, Phyllis Zee, and Susan Redline. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the homepap study. *Sleep*, 35(6): 757–767, 2012.

Adam P Spira, Terri Blackwell, Katie L Stone, Susan Redline, Jane A Cauley, Sonia Ancoli-Israel, and Kristine Yaffe. Sleep-disordered breathing and cognition in older women. *Journal of the American Geriatrics Society*, 56(1):45–50, 2008.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

Jens B Stephansen, Alexander N Olesen, Mads Olsen, Aditya Ambati, Eileen B Leary, Hyatt E Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications*, 9(1):5229, 2018.

Rahul Thapa, Magnus Ruud Kjær, Bryan He, Ian Covert, Hyatt Moore, Umaer Hanif, Gauri Ganjoo, M Brandon Westover, Poul Jennum, Andreas Brink-Kjær, et al. A multimodal sleep foundation model developed with 500k hours of sleep recordings for disease predictions. *medRxiv*, 2025.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.

Terry Young, Mari Palta, Jerome Dempsey, Paul E Peppard, F Javier Nieto, and K Mae Hla. Burden of sleep apnea: rationale, design, and major findings of the wisconsin sleep cohort study. *WMJ: official publication of the State Medical Society of Wisconsin*, 108(5):246, 2009.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Ying Zhang, Matthew Kim, Michael Prerau, Daniel Mobley, Michael Rueschman, Kathryn Sparks, Meg Tully, Shaun Purcell, and Susan Redline. The national sleep research resource: making data findable, accessible, interoperable, reusable and promoting sleep science. *Sleep*, 47(7):zsae088, 2024.

A Supplementary material

A.1 Dynamic batch sampling

The following algorithm determines the dimensions of a single batch.

Input: $M^{\max}$, $\{C_{m_k}^{\max}\}_{k=1}^{M_{\max}}$, $\{B_{m_k}^{\max}\}_{k=1}^{M_{\max}}$
Output: Batch dimensions ;
$\{T, M, \{C_{m_k}\}_{k=1}^{M}, \{B_{m_k}\}_{k=1}^{M}\}$

$T \sim \mathcal{U}\{20, 80\}$; // sequence length
$M \sim \mathcal{U}\{1, M_{\max}\}$; // modalities

for $k \leftarrow 1$ **to** M **do**
  $C_{m_k} \sim \mathcal{U}\{1, C_{m_k}^{\max}\}$; // channels
  $B_{m_k} \sim \mathcal{U}\{1, B_{m_k}^{\max}\}$; // base predictors
end

Based on the returned dimensions, the specific modalities, channels, and base learners are then uniformly sampled from the available options within the observations belonging to the batch.

A.2 Implementation details

This section reports model architecture specifications and training protocol.

In the tri-axial attention encoder module, we use an embedding dimension of $d_{\text{model}} = 24$, with $h = 6$ attention heads, 2 per attention pathway, and a feed-forward dimension of $d_{\text{ff}} = 4 \cdot d_{\text{model}}$, following the original Transformer formulation Vaswani (2017). The encoder stack comprises 4 such layers, followed by the modality fusion layer with an attention size of $d_A = 2 \cdot d_{\text{model}}$, and finally a classifier head consisting of a single hidden layer with 16 neurons. Across all components, we use the GeLU activation function Hendrycks & Gimpel (2016) and apply dropout with a rate of $p = 0.1$.

During training, each batch includes $B = 8$ recordings. For every recording, we randomly sample $K = 4$ segments of the same random length. Gradients are accumulated over $G = 4$ forward–backward passes, resulting in an effective batch of $G \cdot B \cdot K = 128$ segments per optimization step, with tensor dimensions varying across the accumulation steps. Optimization is performed using AdamW Loshchilov & Hutter (2017) with a learning rate of $\eta = 10^{-3}$, and early stopping is applied based on validation macro-F1 score over BSWR validation set, with a patience of 15 epochs.

We run inference considering one recording at a time, employing all available modalities, channels, and base predictors. Inference is performed with a segment length of 35 sleep epochs, consistent with Perslev et al. (2021); Dei Rossi et al. (2025).

NAP and the overall pipeline are implemented in Pytorch 2.8.0. and trained with a single Nvidia T4 GPU and worker, with 64 GB of RAM.

A.3 Datasets

This section provides a comprehensive description of all datasets used in our experiments.

A.3.1 Training datasets

We consider during the training phase the BSWR dataset, described first, and NSRR datasets Zhang et al. (2018, 2024), more specifically their hold-out sets as defined in SLEEPYLAND Dei Rossi et al. (2025), to avoid any overlapping data in the training of base predictors and meta-models.

BSWR. The *Bern Sleep-Wake Registry* (BSWR) Aellen et al. (2024) is a private dataset which comprises a total of 8,410 PSG recordings ($\approx 67'000$ hours) from patients aged 0–91 years, collected during routine clinical practice. This dataset uniquely covers the full spectrum of sleep-wake disorders, including cases with multiple comorbidities and non-sleep-related conditions. Only a small fraction of participants ($< 1\%$) are healthy controls, while the majority are patients diagnosed with one or more sleep disorders or cases with uncertain diagnoses. Among the recorded disorders, sleep-related breathing disturbances are the most prevalent, followed by central hypersomnolence disorders, insomnia, parasomnias, and sleep-related movement disorders. A smaller subset of patients present circadian rhythm disorders or isolated symptoms without a definitive clinical classification. We consider EEG signals (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2) and EOG signals (E2-M1, E1-M2), sampled at 200 Hz. All recordings are manually annotated by certified sleep experts following the American Academy of Sleep Medicine (AASM) guidelines Berry et al. (2017). The dataset is partitioned into training, validation, and test splits, with splits performed by considering subject identifiers, using a 90/5/5 ratio.

Ethical approval : The secondary usage of the BSWR dataset was approved by the local ethics committee (Kantonale Ethikkommission Bern [KEK]-Nr. 2022-00415), ensuring compliance with the Human Research Act (HRA) and Ordinance on Human Research with the Exception of Clinical Trials (HRO). All methods were carried out in accordance with relevant guidelines and regulations. Written informed consent was obtained from participants as of the introduction of the general consent process at Inselspital in 2015. Data were maintained with confidentiality throughout the study.

NSRR Datasets.

The National Sleep Research Resource (NSRR) is an NHLBI-supported data repository designed to promote open sharing of large-scale sleep research data Zhang et al. (2018, 2024). Established in 2014, NSRR provides access to polysomnography, actigraphy, and questionnaire-based datasets collected from diverse cohorts and clinical studies. By enabling secondary analyses, algorithm development, and signal processing research, NSRR aims to advance sleep and circadian science. The repository currently hosts tens of thousands of high-quality sleep records. More info: `https://sleepdata.org/pages/about`.

ABC. The Apnea, Bariatric surgery, and CPAP study includes 132 recordings from 49 patients with severe OSA and morbid obesity (BMI 35–45) Bakker et al. (2018). EEG (F3-M2, F4-M1, C3-M2, C4-M1, O1-M1, O2-M2) and EOG (E1-M2, E2-M1) were acquired at 256 Hz, band-pass filtered, and scored according to AASM criteria. More info: `https://clinicaltrials.gov/ct2/show/NCT01187771`. We consider 35 recordings from the hold out set of SLEEPYLAND.

APOE. The Sleep Disordered Breathing, apolipoprotein E, and Lipid Metabolism dataset is a study investigating genetic associations with sleep-disordered breathing, comprising 712 PSGs from untreated participants stratified by ApoE genotype Moore IV et al. (2014). EEG (C3-M2, C4-M1, O2-M1, O1-M2, C3-M1, C4-M2, O2-M2, O1-M1, F1-M2, F2-C4, F2-T4, FP1-C3, FP1-C3, FP2-C4, Fz-M1, Fz-M2, T3-O1 T4-O2) and EOG (ROC-M1, LOC-M2) were recorded at 256 Hz, and scored according to AASM criteria. More info: `https://doi.org/10.25822/6ssj-2157`. We consider 150 recordings from the hold out set of SLEEPYLAND.

APPLES. The Apnea Positive Pressure Long-term Efficacy Study is a multi-center randomized clinical trial on positive airway pressure for OSA, with 1094 PSGs Quan et al. (2011). EEG signals (C3-M2, C4-M1, O2-M1, O1-M2) and EOG signals (ROC-M1, LOC-M2) are recorded at 128 Hz, initially scored according to Rechtschaffen and Kales scoring rules (R&K) and then re-aligned to AASM Moser et al. (2009). More info: `https://clinicaltrials.gov/study/NCT00051363?tab=results`. We consider 150 recordings from the hold out set of SLEEPYLAND.

CCSHS. The Cleveland Children's Sleep and Health Study includes 515 PSGs Rosen et al. (2003) from three different cohorts in Cleveland, Ohio, USA. EEG (C3-A2, C4-A1) and EOG (ROC-A1, LOC-A2) were recorded at 128 Hz, and manually scored according to AASM rules. More info:

https://doi.org/10.25822/cg2n-4y91. We consider 128 recordings from the hold out set of SLEEPYLAND.

CFS. The Cleveland Family Study is a family-based study on OSA Redline et al. (1995). SLEEPYLAND used 730 PSGs from 144 families, with splits respecting family membership. EEG (C3-A2, C4-A1) and EOG (ROC-A1, LOC-A2) signals were recorded at 128 Hz and scored according to AASM rules. More info: https://doi.org/10.25822/jmyx-mz90. We consider 185 recordings from the hold out set of SLEEPYLAND.

CHAT. The Childhood Adenotonsillectomy Trial includes 1638 PSGs from 1232 children (age range: 5–10) post-adenotonsillectomy-surgery with mild-to-moderate OSA across six U.S. centers Marcus et al. (2013). EEG (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2, T4-M1, T3-M2) and EOG signals (E2-M1, E1-M2) were recorded at ≥ 200 Hz, and scored according to AASM rules. More info: https://clinicaltrials.gov/study/NCT00560859. We consider 199 recordings from the hold out set of SLEEPYLAND.

HOMEPAP. The Home Positive Airway Pressure dataset is a multi-site U.S. study on home PAP therapy Rosen et al. (2012), with 246 PSGs considered in SLEEPYLAND. We consider the EEG signals (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2, T4-M1, T3-M2) and EOG signals (E2-M1, E1-M2), originally recorded at 200 Hz, and scored according to AASM scoring rules. More info: https://clinicaltrials.gov/ct2/show/NCT00642486. We consider 62 recordings from the hold out set of SLEEPYLAND.

MESA. The Multi-Ethnic Study of Atherosclerosis includes 2056 PSGs from adults aged 45–84 across four ethnic groups Chen et al. (2015). EEG signals (Fz-Cz, C4-M1, CzOz) and EOG signals (E2-Fpz, E1-Fpz) were recorded at 256Hz, low-pass filtered at 100 Hz, and scored by sleep experts according to the AASM rules. More info: https://doi.org/10.25822/n7hq-c406. We consider 150 recordings from the hold out set of SLEEPYLAND.

MNC. The Mignot Nature Communications dataset comprises ≈ 1000 PSGs used in Stephansen et al. (2018). Sub-cohorts include CNC (78 PSGs, of which we consider 20 for NAP training), DHC (83 PSGs, of which we consider 22 for NAP training), and SSC (767 PSGs, of which we consider 150 for NAP training). EEG (C3-M2, C3, C4-M1, C4, Cz, F3-M2, F3, F4-M1, F4, O1-M2, O1, O2-M1, O2) and EOG signals (E1-M2 E1 E2-M1 E2) were recorded at 128Hz, and manually scored by sleep experts according to the AASM rules. More info: https://doi.org/10.25822/00tc-zz78.

MROS. A subset of the Osteoporotic Fractures in Men study Blackwell et al. (2011), with 3930 PSGs from older men (> 65 years), most affected by sleep disorders. EEG (C4-A1, C3-A2) and EOG signals (ROC-A1, LOC-A2) were recorded at 256 Hz, and scored according to AASM rules. More info: https://doi.org/10.25822/kc27-0425. We consider 195 recordings from the hold out set of SLEEPYLAND.

MSP. The Maternal Sleep in Pregnancy dataset DiPietro et al. (2021) is comprised of 105 overnight PSGs from women at week 36 of pregnancy, without previously identified sleep disorders. EEG (C3-M2, C4-M1, F3-M2, F4-M1, O1-M2, O2-M1) and EOG signals (LOC, ROC) were recorded at 256Hz and scored according to the AASM manual. More info: https://sleepdata.org/datasets/msp. We consider 27 recordings from the hold out set of SLEEPYLAND.

NCHSDB. The Nationwide Children's Hospital Sleep DataBank consists of 3950 pediatric PSGs (age range: 0–18) Lee et al. (2022). EEG (FP1, FP2, FZ, CZ, PZ, OZ, FPZ, P3-M2, P4-M1, F3-M2, F4-M1, F4-M2, C3-M2, C4-M1, C4-M2, T3-M2, T4-M1, O1-M2, O2-M1, F4, O1, O2) and EOG (E1, E2) and EOG signals (E1-M2, E2-M1, E1, E2) were recorded at 256 Hz. Recordings were manually scored following AASM criteria. More info: https://sleepdata.org/datasets/nchsdb. We consider 161 recordings from the hold out set of SLEEPYLAND.

SHHS. The Sleep Heart Health Study is a large dataset that comprises 8444 PSGs from 5797 adults (≥ 40 years), most of which suffering from sleep disorders, across two visits Quan et al. (1997). EEG (C3-A2, C4-A1) and EOG signals (ROC-A1, LOC-A2) were recorded at sampling frequencies of 125 Hz and 50 Hz, respectively. Recordings were initially R&K scored and subsequently re-aligned to AASM scoring rules. More info: https://clinicaltrials.gov/ct2/show/NCT00005275. We consider 221 recordings from the hold out set of SLEEPYLAND.

SOF. We consider a subset of the Study of Osteoporotic Fractures Spira et al. (2008), with 453 PSGs from older women. We consider only EEG (C3-A2, C4-A1) and EOG signals (ROC-A1, LOC-A2),

which were recorded at 128 Hz, initially R&K scored, and re-aligned with AASM criteria. More info: `https://doi.org/10.25822/e1cf-rx65`. We consider 114 recordings from the hold out set of SLEEPYLAND.

WSC. Wisconsin Sleep Cohort is an ongoing longitudinal study investigating the causes, consequences, and natural history of sleep disorders; SLEEPYLAND considers 2569 in-laboratory PSGs across four visits Young et al. (2009). EEG signals (F3-M1, F3-M2, F3-AVG, F4-M1, F4-M2, F4-AVG, Fz-M1, Fz-M2, Fz-AVG, Cz-M1, Cz-M2, Cz-AVG, C3-M1, C3-M2, C3-AVG, C4-M1, C4-M2, C4-AVG, Pz-M1, Pz-M2, Pz-AVG, Pz-Cz, O1-M1, O1-M2, O1-AVG, O2-M1, O2-M2, O2-AVG) and EOG signals (E1, E2) are included, recorded at 100 Hz and 200 Hz, respectively. Recordings are scored by sleep experts according to AASM criteria. More info: `https://sleepdata.org/datasets/wsc`. We consider 347 recordings from the hold out set of SLEEPYLAND.

A.3.2 Evaluation datasets

The following datasets are used exclusively in inference; neither SLEEPYLAND base predictors nor NAP were trained on recordings from these datasets, enabling evaluation of zero-shot performance.

Table 2: Summary statistics of evaluation datasets, reporting the number of PSG recordings, average participant age (mean $\pm$ standard deviation), and gender distribution, where available.

Dataset	**# PSGs**	**Age (years)**	**F/M (%)**
DCSM	255	–	–
DOD-H	25	35.3 ± 7.5	$24/76$
DOD-O	55	45.6 ± 16.5	$36/64$
PHYS	994	55.2 ± 14.3	$33/67$
SEDF-SC	153	58.8 ± 22.0	$53/47$
SEDF-ST	44	40.2 ± 17.7	$68/32$

DOD. The *Dreem Open Datasets* consist of two subsets, **DOD-H** and **DOD-O** Guillot et al. (2020). DOD-H includes 25 recordings from healthy adults (19 males, 6 females) aged 18–65 years, collected at the Fatigue and Vigilance Unit of the French Armed Forces Biomedical Research Institute (IRBA), Bretigny-Sur-Orge, France. We use EEG channels (C3-M2, C4-M1, F3-F4, F3-M2, F3-O1, F4-O2, O1-M2, O2-M1) along with left and right EOG signals, sampled at 512 Hz. DOD-O contains 55 PSG recordings from patients diagnosed with obstructive sleep apnea (35 males, 20 females) aged 39–62 years, collected at the Stanford Sleep Medicine Center. EEG signals include (C3-M2, C4-M1, F4-M1, F3-F4, F3-M2, F3-O1, F4-O2, FP1-F3, FP1-M2, FP1-O1, FP2-F4, FP2-M1, FP2-O2) and left/right EOG. Recordings are sampled at 250 Hz. Following Guillot et al. (2020), all signals undergo preprocessing: a Butterworth IIR band-pass filter $[0.4, 18]$ Hz is applied, recordings are resampled to 100 Hz, clipped, and scaled by dividing by 500 to mitigate extreme amplitude variations. Sleep stages are scored by five physicians across three independent centers using AASM guidelines.

DCSM. The *Danish Centre for Sleep Medicine* (DCSM) dataset Perslev et al. (2021) consists of 255 PSG recordings from patients referred for suspected or nonspecific sleep-related disorders. No demographic metadata is provided. We include EEG channels (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2, T4-M1, T3-M2) and EOG channels (E2-M1, E1-M2), sampled at 256 Hz. A band-pass filter between 0.3 Hz and 70 Hz is applied. All recordings are scored manually by certified clinicians according to AASM criteria. Additional dataset details are available at `https://sid.erda.dk/wsgi-bin/ls.py?share_id=fUH3xbOXv8`.

SEDF. The *Sleep-EDF Expanded* dataset Goldberger et al. (2000); Kemp et al. (2000) consists of two subsets, **SEDF-SC** and **SEDF-ST**. SEDF-SC (Sleep Cassette) is comprised of 153 PSG recordings from 78 healthy participants aged 25–101 years. SEDF-ST (Sleep Telemetry) includes 44 recordings from 22 subjects. For our experiments, we use EEG channels (Fpz-Cz, Pz-Oz) and an EOG channel (ROC-LOC), sampled at 100 Hz. Original annotations, scored according to Rechtschaffen and Kales

criteria, were re-aligned to match the AASM scoring standard. Additional details are available at `https://doi.org/10.13026/C2C30J`.

PHYS. The dataset from the *PhysioNet/Computing in Cardiology Challenge 2018* Goldberger et al. (2000); Ghassemi et al. (2018) includes 1,985 overnight PSG recordings, of which we use 994 labeled sessions in our experiments. EEG channels (F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2) and one EOG channel (E1-M2) are considered. Recordings are sampled at 200 Hz and manually annotated following AASM guidelines. Full documentation can be found at `https://physionet.org/content/challenge-2018/1.0.0/`.

A.4 Evaluation against medical consensus

We adopt the multi-annotator evaluation framework introduced in Guillot et al. (2020) for DOD datasets. Each recording in DOD is independently annotated by $S = 5$ experienced sleep technologists, allowing model performance to be evaluated relative to both individual scorers and collective consensus.

Given S scorers, let $y_s^t \in \{0, 1, 2, 3, 4\}$ denote the label assigned by scorer s to epoch t and $\hat{y}_s^t \in \{0, 1\}^5$ its one-hot encoding. For scorer s, we define the agreement of the remaining scorers at epoch t as:

$$\hat{z}_s^t = \frac{\sum_{i \neq s} \hat{y}_i^t}{\max\left(\sum_{i \neq s} \hat{y}_i^t\right)}. \tag{3}$$

The *soft-agreement* of scorer s over a recording is:

$$\text{Soft-Agreement}_s = \frac{1}{T} \sum_{t=1}^{T} \hat{z}_s^t[y_s^t], \tag{4}$$

which measures how often the scorer aligns with the collective judgment, weighted by inter-scorer agreement. Reliable scorers are defined as those with the highest soft-agreement scores for a given recording. The discrete consensus hypnogram is obtained by majority voting across scorers, with ties resolved using the most reliable scorer.