Laying the Foundation: Modern Transformers for Gold-Standard Sleep Analysis

William G. Coon
Research & Exploratory Development Department
Johns Hopkins University Applied Physics Laboratory
Laurel, USA
will.coon@jhuapl.edu

Mattson Ogg
Research & Exploratory Development Department
Johns Hopkins University Applied Physics Laboratory
Laurel, USA
mattson.ogg@jhuapl.edu

***Abstract*—Accurate sleep assessment is critical to the practice of sleep medicine and sleep research. Excitingly, the recent availability of large quantities of publicly available sleep data present opportunities for data-driven discovery efforts. Transformers are flexible neural network architectures that not only excel at classification tasks, but also can enable data-driven discovery through un- or self-supervised learning, which requires no human annotations to the input data. The first step for data-driven discovery in sleep data with transformers is to train a model on a supervised learning task (here, scoring sleep) to form the foundation for unsupervised extensions; it is not yet clear which model optimizations are ideal for this. To address this gap and lay the groundwork for data-driven discovery, we explored optimizations of a transformer-based model that learned the canonical 5-class sleep stage classification task. We examined different configurations of model size, input data size (sleep sequence length), input channels, and the use of multiple or single nights of sleep during training. We found that most configurations met or exceeded clinical criteria (accuracy/agreement above approximately 80–85%). Curiously, they achieve this when operating over as little as 10 minutes of input data at a time, which would preclude the use of valuable sleep cycle history that a human expert typically uses. Compared to most recurrent neural network analogues, the attention-based transformer was better able to distinguish REM sleep from neighboring classes using EEG alone. Finally, the full-size transformer model did not significantly outperform smaller versions, indicating that lightweight versions are sufficient for this task. In summary, we found that small transformer models can match and even outclass other state-of-the-art automated sleep scoring techniques, while also providing the basis for future data-driven discovery efforts using large sleep data sets, with or without human annotations.**

***Index Terms*—sleep, transformers, unsupervised learning, self-supervised learning**

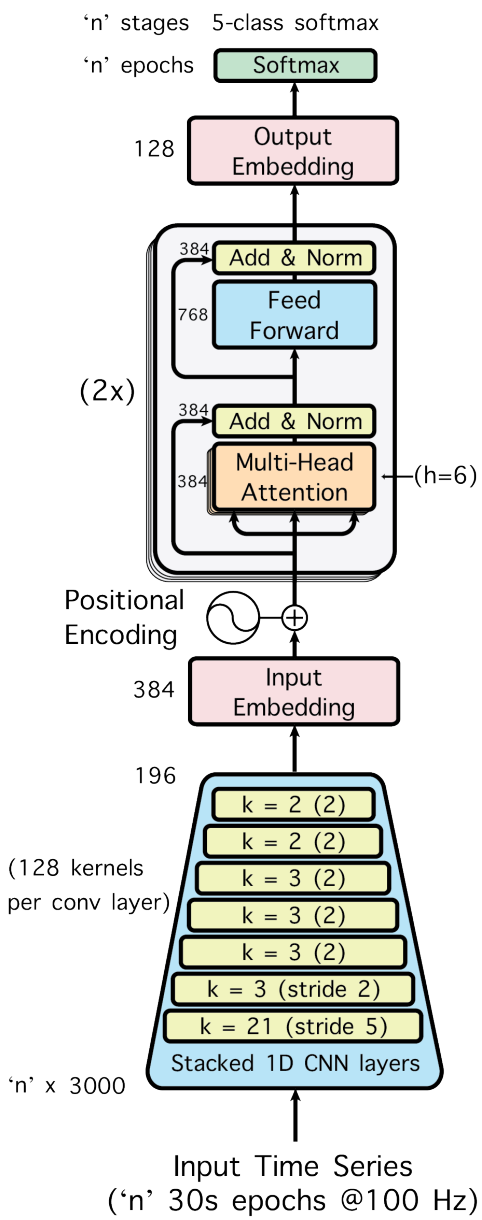
I. Introduction

Sleep is fundamental for good health and optimal performance, and its accurate assessment is critical to sleep medicine and research. Sleep assessment ("sleep scoring") has traditionally been done by human experts trained to classify sleep in segments of 30 seconds, into one of five stages according to patterns discernible principally in the electroencephalogram (EEG)[†] [2, 3]. Modern neural network architectures, supported by the recent availability of large quantities of publicly available sleep signals data, are beginning to supplant this tedious, manual approach, having met and exceeded human expert performance [1, 4, 5]. Today, the question is not whether automated approaches can serve instead of manual scoring, but rather how far they can be pushed to discover latent, heretofore unutilized health information from sleep data in a principled, data-driven way (two excellent recent examples are a study by Decat et al. [6], and a recent effort to update traditional sleep staging with a high-frequency classifier [7]).

Foundation models based on transformer architectures are particularly well-suited for data-driven discovery. They automatically learn compact representations of the data space that can be flexibly applied to many different ends (like classifying sleep stages or apnea state), or be interrogated to discover new information embedded within (for example, with naïve clustering approaches whose resultant features can then be correlated with health or examined for structure). With respect to sleep, these methods offer new ways to quantify and correlate sleep signals data with the many health states and conditions known to impact – or be impacted by – sleep.

The first step in generating a foundation model that can be applied to data-drive discovery in sleep signals data is to develop a transformer-based model trained on sleep stage classification, a supervised learning task in which the data labels (sleep stages) are provided by human annotators. Currently, it is not clear which model optimizations (ex., size, architecture, parameters, hyper-parameters) are ideal for this. To identify a candidate model, we explored optimizations of a transformer-based neural network model derived from acoustic speech

[†]traditionally sleep is scored using polysomnography (PSG), an aggregate of EEG, electro-oculography (EOG), and electromyography (EMG), but most information is in the EEG's brain signals and many automated techniques perform well using EEG alone [1]



Model		Large	Small
CNN Encoder	strides	(5,2,2,2,2,2,2)	
	kernel width	(21,3,3,3,3,2,2)	
	channels	512	128
Transformer	layers	4	2
	embedding dim	768	384
	inner FFN dim	1536	768
	layerdrop prob	0.05	
	attention heads	8	6
Average Pool 1D	n	(101 or 21)	
Embedding	dim	256	128
Output	dim	5	
Activation		gelu	

Figure 1. **Transformer-based sleep staging model.** The model is trained to map input time series data to a sequence of sleep stages. A stack of 1D CNNs first processes the time series, followed by linear projection (p_n), a series of transformer encoder layers, another projection (e_n) and an output layer (o_n) with one unit per sleep-stage label (Wake, N1, N2, N3, REM).

processing ("HuBERT") [8] and language modeling [9] – sequence learning tasks with similarities to interpreting sequences of sleep EEG – that learned the canonical 5-class sleep stage classification task. We examined different configurations of model size, input data size (sleep sequence length), input channels, and the use of multiple or single nights of sleep during training. We report our findings and discuss transitioning the trained foundation models to analyze sleep on wearable devices and to enable data-driven discovery in future work.

II. Methods

A. Data and Preprocessing

We compiled a large corpus of training data from the National Sleep Research Resource [10]. We obtained data from the Multi-Ethnic Study of Atherosclerosis (MESA [11]), the MrOS Sleep Study[12], the Sleep Heart Health Study [13], the Wisconsin Sleep Cohort [14], the Nationwide Children's Hospital Sleep DataBank (NCHSDB) [15], the Study of Osteopathic Fractures (SOF) [16], and the Cleveland Family Study (CFS) [17]. This comprised data from over 11,000 nights of sleep from over 9,500 individuals. We held out data from 915 participants (or approximately 10%) to monitor training. Two additional datasets were obtained for external validation and testing (i.e., after training using crossfold validation within a dataset, we evaluate final perfomance on a separate hold-out set whose data the model had not encountered during training). We used the Dreem Open Datasets (both the Healthy and OSA cohorts) [18] for external validation and the Sleep-EDF Database Expanded dataset [19] as a test set.

Each sleep session (i.e., each EDF[‡] file) underwent the following pre-processing regime. We retained one central EEG recording (i.e., C3 or C4 from the International 10-20 System) and one (left or right) EOG recording (any sleep session without these channels explicitly marked was excluded). These channels were re-sampled to a 100 Hz sampling rate and then robustly re-scaled by subtracting the timeseries' medians and dividing by the difference between each's 25th and 75th percentiles (thus achieving an inter-quartile range (IQR) of 1 for each channel as in Perslev et al.[7]). Any values exceeding a value of ± 20 median absolute deviations were clipped to remain within that range. No further signal processing to the input data was applied. For the Sleep-EDF dataset, we used the Fpz-Cz and Horizontal EOG channels, and truncated the large awake periods to just the 30 minutes prior to when the participant was asleep.

Each 30-second epoch of EEG and EOG data, with their corresponding sleep labels, was then extracted based

[‡]European Data Format. EDF is an industry standard file format for polysomnographic records.

on the session's annotations. Any epochs that were not explicitly labeled with a sleep stage, or were labeled as artifacts, were removed. The remaining epochs were then grouped into sequences of 21 (10.5 minutes) or 101 (50.5 minutes) consecutive epochs[§]. These sequences of labels along with their corresponding time series data (at a 100 Hz sampling rate) served as the training labels and input data for the transformer model, respectively. We over-sampled the training and internal validation data by a factor of four when sequencing epochs (21-epoch examples were generated every 5 epochs and a 101-epoch examples were generated every 25 epochs). External validation and test data were not over-sampled (101 and 21-epoch examples were generated with no overlap).

B. Network Architecture and Training

Similar to work in the speech and audio domain[8], our model learns to map time series input to a sequence of labels, which in our case correspond to one of 5 sleep stages (W, N1, N2, N3 or REM). Transformers are an excellent choice for this application because of their ability to model long-distance temporal dependencies within the data.

The model architecture begins with a series of 1-dimensional convolution layers as a front end, to learn a set of filters directly from the input time series (as in previous work[4, 5]). A layer-norm was applied between the convolution and activation functions in each layer. This front end was followed by a linear projection layer, positional encoding, a stack of transformer encoder layers, temporal averaging (to align the higher temporal resolution of the convolution/transformer layers with epoch labels every 30-seconds), another linear projection layer, and, finally, an output layer. We use gelu activation functions after every convolution layer and within the transformer blocks. Convolution and linear layer weights were initialized with a Kaiming normal function. This common architecture is depicted in Figure 1, and more details on our variants of these models can be found in Figure 3.

We trained the larger variants of this model using input from both the EEG and EOG channels for a 101-epoch (or 50.5 minute) long input sequence (the 'large' model architecture described in Figure 1). The model was trained to minimize cross-entropy for 50 epochs (passes through the data) using an Adam optimizer. The learning rate ramped up linearly from 0.00001 to a peak of 0.0005 after the first ten passes, then ramped down by a factor of .75 over the proceeding epochs. We retained the weights for the model corresponding to the training pass that achieved the lowest validation loss during the training run. After training we evaluated our models on the external validation data and the Sleep-EDF test data.

[§]values were selected to vary the amount of sleep architecture history the model could assess at a time; the long sequence model captures most of a sleep cycle, while the short variant presumably cannot benefit from this informative background context.

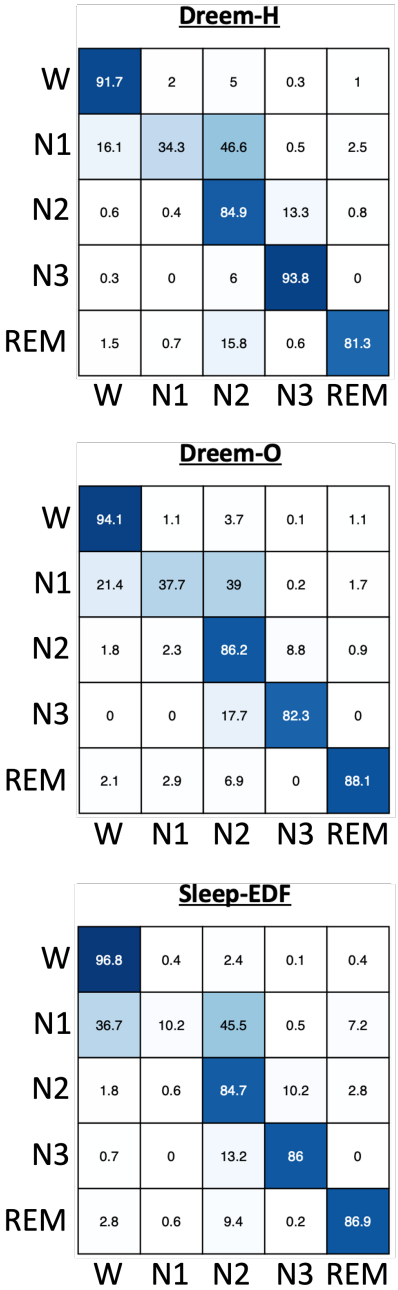


Figure 2. **Performance Summaries by Sleep Stage**. Confusion matrices show external validation performance from the top-performing 'large' model on three benchmark data sets.

To understand the influence of different parts of the model architecture and training regime, we conducted a series of ablation studies using the external validation data to compare different parameters. All models

used the same training hyperparameters described above except where noted. We tested a variety of configurations for the input data (EEG channel alone vs. both EEG+EOG together as input; 21-epoch vs. 101-epoch long input sequences; single nights of data per subject vs. multiple nights [where available]), and model architecture size (large vs. small models, see Figure 1 for details), for a total of 16 models varying in size from 3.9 million trainable parameters (23.4 MB on disk) to 28.4 million (129.2 MB on disk) trainable parameters. Models that took a 21-epoch input sequence were trained with a batch size of 64 and models that took the longer 101-epoch input sequence were trained with a batch size of 32.

III. Results

The performance results from all sleep staging transformer model variants are displayed in Figure 3; confusion matrices for the top-performing large model are shown in Figure 2. This variant achieved exceptionally high performance on the internal validation data (87.7%) and continued to perform well on the external validation data (83.3% and 85% on Dreem-H and Dreem-O, respectively). This variant also performed well on the Sleep-EDF corpus (80.5%), which uses a slightly different recording montage than the training data (Cz-Fpz instead of C3 or C4). These results compare favorably with the reliability of human annotators (82%), as well as with other state-of-the-art systems (83.2% to 88.8% on Dreem data reported by Hanna & Flöel, 2023[20]).

Most model and input configurations achieved high internal validation performance and were able to match human expert performance (>82% accuracy) on the Dreem validation data sets. Our best performing variant was the smaller architecture using EEG+EOG channels, multiple nights of sleep from participants where available, and longer 101-epoch input sequences. It is notable that this model slightly under-performed on our internal validation data relative to other models, but achieved better external validation performance. It is also notable that a minimal model (trained with the smallest input sequence length and model architecture) still performs well on the Dreem benchmark datasets (83.8%).

IV. Discussion

Overall, we found that these transformer model variants all proved adept at sleep-stage classification. They performed well testing on unseen data with a known character (i.e., internal validation, using data drawn from the same study as the training data[¶]) and also on unseen data with an unknown character (i.e., external validation, using data from a different study). Our results are nearly state-of-the-art using the Dreem and Sleep-EDF data sets for assessment (see comparisons in Hanna & Flöel, 2023 [20] and Yao & Liu, 2023 [21] for the Dreem and Sleep-EDF datasets, respectively). However, performance did drop on external validation data, and dropped even more when the recording montage differed (as for the Sleep-EDF data, which used a Cz-Fpz EEG derivation instead of the preferred C3-A2 or C4-A1). This is an important problem to solve because, this being the case, models will always need to be fine-tuned on the particular system or device on which they are deployed, which in turn implies a nontrivial amount of data collection beforehand. Ideally, one pretrained model would be able to perform equally well on traditional 10-20 locations and on nonstandard locations such as those used by commercial or at-home sleep monitoring wearables, without any fine-tuning to go from one to the other.

[¶]both the training data used for training the model, and the held out test data used to evaluate performance during model training, were drawn from the same source data corpus (for example, the MrOS study), but did NOT overlap.

Input Features	Input Length	Nights per Participant	Model Size	Internal Validation	Dreem-H	Dreem-O	Dreem-Avg
EEG+EOG	101	Multiple	Large	87.7	83.3	85.0	84.1
EEG+EOG	101	Multiple	Small	87.5	85.4	85.1	85.2
EEG+EOG	101	Single	Large	87.7	85.0	85.3	85.1
EEG+EOG	101	Single	Small	87.4	82.4	83.1	82.7
EEG+EOG	21	Multiple	Large	88	81.0	83.4	82.2
EEG+EOG	21	Multiple	Small	87.9	80.3	83.9	82.1
EEG+EOG	21	Single	Large	88.1	80.0	82.4	81.2
EEG+EOG	21	Single	Small	88	81.5	82.9	82.2
EEG	101	Multiple	Large	86.6	85.5	84.4	84.9
EEG	101	Multiple	Small	86.5	84.8	84.4	84.6
EEG	101	Single	Large	86.5	83.6	83.2	83.4
EEG	101	Single	Small	86.5	82.4	82.7	82.5
EEG	21	Multiple	Large	86.8	85.7	80.4	83.1
EEG	21	Multiple	Small	86.7	85.6	81.9	83.8
EEG	21	Single	Large	86.9	84.2	83.6	83.9
EEG	21	Single	Small	86.8	82.0	82.3	82.1

Figure 3. **Overall performance of all model variants**. Accuracies are shown as percentages.

It is interesting that our large model performed better on the Dreem-O data set than the Dreem-H data set. The former is composed of slightly older individuals diagnosed with sleep apnea, while the latter is made up of healthy younger sleepers. We speculate that this difference in age and apnea state may contribute to the observed performance difference.

We also note that smaller models generally performed as well as (or in some cases better than) comparable large models. This has important implications for deploying transformer-based sleep scoring models on mobile devices, for which storage space and processing power come at a premium (see Yao & Liu, 2023 [21] for some elegant work in this area). Encouragingly, the same high performance also held true for shorter input sequences,

which could reduce the amount of processing power required. Discouragingly, the memory storage burden was high even for the small models (containing millions of high-precision weights), making them challenging to adapt for mobile use. Future work will explore shrinking and compressing these models (e.g., by using quantization and reduced precision to reduce their size).

We were motivated to study transformer model optimizations for sleep stage classification because of their suitability as foundation models for data-driven discovery down the line. Future work will examine self-supervised and masked training regimes for these tasks, which eschew the bottleneck of needing human annotations and may be able to assimilate latent structure not captured by other sleep analysis techniques (including 5-stage scoring). Adopting an objective, data-driven mentality, new classification schemas could be revealed that more completely capture clinically informative facets of sleep. These could help address the recognized limitations of AASM sleep scoring guidelines [22], such as a "one-size-fits-all" lack of personalization (i.e., adjusting to the characteristics of individuals' unique sleep electrophysiology), age-wise scoring biases, disease state interactions, and other recognized phenomena such as the cyclic alternating pattern (CAP) [23]. In sum, this work lays a foundation that may unlock myriad other health applications in the future.

Acknowledgment

This study was supported by the Johns Hopkins University Applied Physics Laboratory under an internal research & development grant. We are also grateful for the herculean efforts behind the National Sleep Research Resource, which provided sleep data for our analyses and was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

References

[1] Huy Phan and Kaare Mikkelsen. Automatic sleep staging of eeg signals: recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, 2022. doi:10.1088/1361-6579/ac6049.

[2] Kales A Rechtschaffen A (eds). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. BIS/BRI University of California, Los Angeles, 1968.

[3] Michael H Silber, Sonia Ancoli-Israel, Michael H Bonnet, Sudhansu Chokroverty, Madeleine M Grigg-Damberger, Max Hirshkowitz, Sheldon Kapen, Sharon A Keenan, Meir H Kryger, Thomas Penzel, et al. The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, 3(02):121–131, 2007. doi:https://doi.org/10.5664/jcsm.26814.

[4] Henri Korkalainen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac O Afara, Sami Myllymaa, Juha Töyräs, and Timo Leppänen. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE journal of biomedical and health informatics*, 24(7):2073–2081, 2019.

[5] William G Coon and Naresh M Punjabi. Automatic sleep staging using a small-footprint sensor array and recurrent-convolutional neural networks. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1144–1147. IEEE, 2021. doi:10.1109/NER49283.2021.9441432.

[6] Nicolas Decat, Jasmine Walter, Zhao H Koh, Piengkwan Sribanditmongkol, Ben D Fulcher, Jennifer M Windt, Thomas Andrillon, and Naotsugu Tsuchiya. Beyond traditional sleep scoring: Massive feature extraction and data-driven clustering of sleep time series. *Sleep Medicine*, 98:39–52, 2022. doi:doi.org/10.1016/j.sleep.2022.06.013.

[7] 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):72, 2021. doi:https://doi.org/10.1038/s41746-021-00440-5.

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi:10.1109/TASLP.2021.3122291.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. doi:doi.org/10.48550/arXiv.1810.04805.

[10] 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. doi:doi.org/10.1093/jamia/ocy064.

[11] 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. doi:doi.org/10.5665/sleep.4732.

[12] 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 (MrOS) Study Group. Association of sleep characteristics and cognition in older community-dwelling men: the mros sleep study. *Sleep*, 34(10):1347–1356, 2011. doi:doi.org/10.5665/SLEEP.1276.

[13] 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. doi:doi.org/10.1093/sleep/20.12.1077.

[14] 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. doi:PMID: 19743755 PMCID: PMC2858234.

[15] 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. doi:doi.org/10.1038/s41597-022-01545-6.

[16] 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.

doi:10.1111/j.1532-5415.2007.01506.x.

[17] 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. doi:doi.org/10.1164/ajrccm.151.3.7881656.

[18] 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. doi:10.1109/TNSRE.2020.3011181.

[19] Bob Kemp, Aeilko Zwinderman, B Tuk, H Kamphuisen, and J Oberyé. Sleep-edf database expanded. *Physionet org*, 2018. doi:doi.org/10.13026/C2X676.

[20] Jevri Hanna and Agnes Flöel. An accessible and versatile deep learning-based sleep stage classifier. *Frontiers in Neuroinformatics*, 17:1086634, 2023. doi:doi.org/10.3389/fninf.2023.1086634.

[21] Zongyan Yao and Xilin Liu. A CNN-transformer deep learning model for real-time sleep stage classification in an energy-constrained wireless device. In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1–4. IEEE, 2023. doi:ieeexplore.ieee.org/abstract/document/10123825/.

[22] Sari-Leena Himanen and Joel Hasan. Limitations of rechtschaffen and kales. *Sleep medicine reviews*, 4(2):149–167, 2000. doi:https://doi.org/10.1053/smrv.1999.0086.

[23] Mario Giovanni Terzano and Liborio Parrino. Origin and significance of the cyclic alternating pattern (cap). *Sleep medicine reviews*, 4(1):101–123, 2000. doi:https://doi.org/10.1053/smrv.1999.0083.