PRETRAINING SLEEP STAGING MODELS WITHOUT PATIENT DATA

Niklas Grieger[1,2,3], **Siamak Mehrkanoon**[2] **& Stephan Bialonski**[1,3,*]
[1]Department of Medical Engineering and Technomathematics, FH Aachen University of Applied Sciences, 52428 Jülich, Germany
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
[3]Institute for Data-Driven Technologies, FH Aachen University of Applied Sciences, 52428 Jülich, Germany
[*]bialonski@fh-aachen.de

ABSTRACT

Analyzing electroencephalographic (EEG) time series can be challenging, especially with deep neural networks, due to the large variability among human subjects and often small datasets. To address these challenges, various strategies, such as self-supervised learning, have been suggested, but they typically rely on extensive empirical datasets. Inspired by recent advances in computer vision, we propose a pretraining task termed "frequency pretraining" to pretrain a neural network for sleep staging by predicting the frequency content of randomly generated synthetic time series. Our experiments demonstrate that our method surpasses fully supervised learning in scenarios with limited data and few subjects, and matches its performance in regimes with many subjects. We anticipate that our approach will be advantageous across a broad spectrum of applications where EEG data is limited or derived from a small number of subjects, including the domain of brain-computer interfaces.

1 INTRODUCTION

Deep neural networks have achieved significant advances in analyzing electroencephalographic (EEG) time series (Roy et al., 2019), ranging from brain-computer interfaces (Ko et al., 2021) to the intricacies of sleep stage scoring (Phan & Mikkelsen, 2022; Fiorillo et al., 2019). However, training neural networks requires large and diverse datasets that capture the considerable variety between individual subjects and their medical conditions (subject heterogeneity). Creating such datasets is challenging due to the typically limited amount of data per subject (data scarcity) and diverse measurement protocols used in different clinics, which can introduce additional variability in the data. Furthermore, acquiring large datasets is often expensive, complicated, or even intractable due to strict privacy policies and ethical guidelines. This hinders the advancement of deep neural networks for widespread application in real-world medical settings.

Efforts to mitigate the scarcity of large datasets have primarily followed two paths: (1) the development of network architectures that incorporate constraints mirroring the data's intrinsic characteristics, such as symmetries (Bronstein et al., 2021), and (2) enhancing model performance through the use of additional or cross-domain data to learn effective priors. Pertaining to the first path, a common feature in time series processing networks is the use of convolutional layers. These layers are designed to be translation-equivariant (Goodfellow et al., 2016), which ensures that a temporal shift in the input only affects the output by the same shift. For the second path, a variety of strategies have been proposed to learn useful priors from data, including data augmentation (Lashgari et al., 2020; He et al., 2021), transfer learning (Ebbehoj et al., 2022), self-supervised learning (Liu et al., 2023; Banville et al., 2021), and generative adversarial networks (GANs) (Habashi et al., 2023; Carrle et al., 2023). While all of these approaches have been demonstrated to be able to improve the performance of neural networks, they still rely on large empirical datasets for training.

Recent advances in computer vision have demonstrated that it is possible to learn effective priors exclusively from synthetic images, which has the potential to significantly reduce the need for large

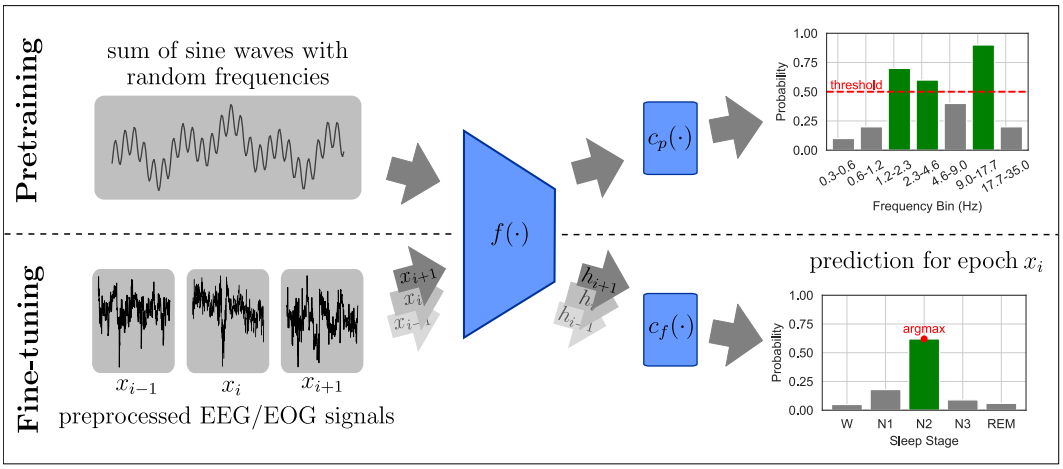

Figure 1: The training process consisted of a *pretraining* and a *fine-tuning* phase. In the pretraining phase, the feature extractor f was trained together with a classifier c_p to detect the frequency content of randomly generated synthetic time series signals (multi-label classification problem). In the fine-tuning phase, the pretrained feature extractor f extracted features h_i from individual epochs x_i of EEG and EOG signals. The features of a sequence of epochs (training sample) were then aggregated by a classifier c_f to predict the sleep stage of the middle epoch in the sequence (multi-class classification problem).

empirical datasets (Baradad et al., 2021; Kataoka et al., 2022). Synthetic images for image classification tasks were generated by simple random processes, such as iterated function systems to produce fractals (Kataoka et al., 2022) or random placement of geometric objects to cover an image canvas (Baradad et al., 2021). Deep neural networks pretrained on such data were demonstrated to learn useful priors for image classification tasks, yielding competitive performance comparable to pretraining on natural images on various benchmarks (Kataoka et al., 2022). This remarkable finding highlights the potential of synthetic datasets that can be generated without much computational resources and, theoretically, in unlimited amounts.

Inspired by these advances, we hypothesize that pretraining exclusively on synthetic time series data generated from simple random processes can also yield effective priors for sleep staging. Given the importance of frequencies for sleep stage scoring and other EEG-based applications (Berry et al., 2020; Motamedi-Fakhr et al., 2014), we introduce a pretraining method that centers on generating synthetic time series data with specific frequency content (see Fig. 1). During pretraining, deep neural networks learn to accurately predict the frequencies present in these synthetic time series. We observe that this conceptually simple pretraining task, which we call "frequency pretraining" (FPT), allows a deep neural network to detect sleep stages with better accuracy compared to fully supervised training when data from few subjects (few-subject regime) are available for fine-tuning (see Fig. 2). We consider pretraining techniques leveraging synthetic data, like the one we propose, as a promising area of research, offering the potential to develop models in sleep medicine and neuroscience that are particularly suited for scenarios involving small datasets. To facilitate testing and further advancements, we make the source code of our method publicly available (Grieger, 2024).

2 Methods

Fig. 1 presents an overview of our training scheme, which comprised two phases. In the *pretraining phase*, we generated synthetic time series signals and trained a convolutional feature extractor with a multi-layer perceptron classifier to predict the frequency content of these signals. In the *fine-tuning phase*, we utilized the pretrained feature extractor together with another classifier to perform sleep staging on EEG and EOG (electrooculography) signals. For more details on the data used for fine-tuning, the model architecture, and training procedure, please refer to appendix A.

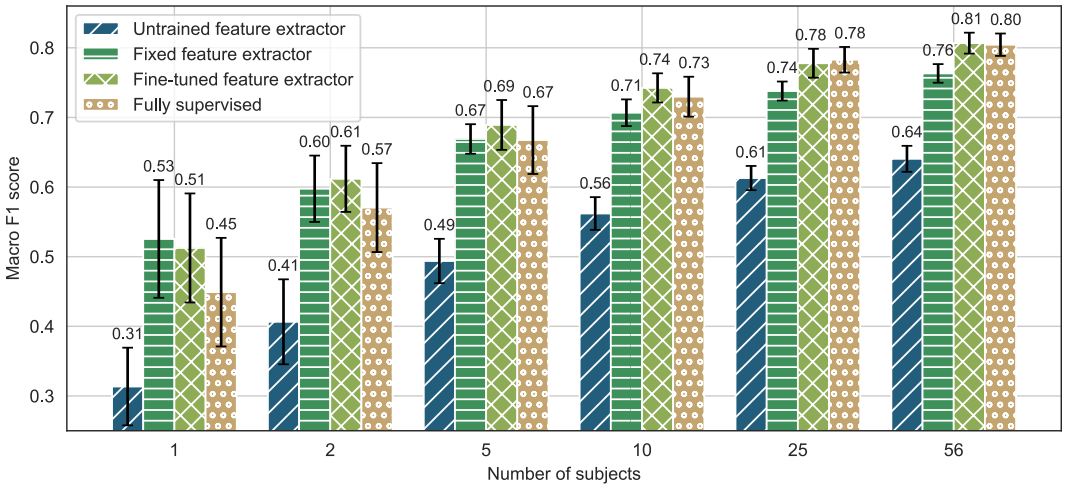

Figure 2: Average macro F1 scores of the different training configurations trained with data from a varying number of subjects. The bars indicate the mean of the macro F1 scores averaged over 15 trainings (3 repetitions of a 5-fold cross-validation) for each training configuration and number of subjects. Error bars show the standard deviation of the macro F1 scores.

Synthetic Data. For the pretraining phase (see Fig. 1), we defined a simple random process to generate synthetic time series signals. Each synthetic signal was a normalized time series composed of the sum of 30-second sine waves sampled at 100 Hz with random frequencies and phases. To sample the frequencies, we first divided the frequency range of 0.3–35 Hz recommended by the American Academy of Sleep Medicine (AASM) for filtering EEG and EOG signals into 20 bins with a base 2 logarithmic scale. We then randomly decided for each frequency bin (with a probability of 50% for each bin) whether it would be used to create the synthetic signal or not. Within each selected frequency bin, we randomly sampled the final frequencies of the sine waves.

When pretraining our neural networks, each training sample consisted of three synthetic signals, corresponding to three "channels" of sleep staging data, and an associated label vector. The label vector encoded the frequency bins from which the frequencies of the sine waves were drawn in a one-hot encoded format. Pretraining involved predicting all frequency bins encoded in this label vector, which made it a multi-label classification problem with 20 classes. We trained models on 100,000 of these synthetic samples.

Training Configurations. We created four training configurations to investigate the effectiveness of our pretraining method: (i) Fully Supervised, (ii) Fixed Feature Extractor, (iii) Fine-Tuned Feature Extractor, and (iv) Untrained Feature Extractor. In the Fully Supervised configuration, we skipped the pretraining step and trained (fine-tuned) the model from scratch using sleep staging data. In the Fixed Feature Extractor configuration, we pretrained the feature extractor using synthetic data and then fine-tuned only the classifier using sleep staging data (i.e., the feature extractor remained fixed). The Fine-Tuned Feature Extractor configuration was similar to the Fixed Feature Extractor configuration, except that we fine-tuned the full model (feature extractor and classifier) after pretraining. Finally, in the Untrained Feature Extractor configuration we randomly initialized the feature extractor using He initialization (He et al., 2015) and then fine-tuned only the classifier using sleep staging data.

3 Results

To assess the data efficiency of our pretraining method, we trained models from each of the training configurations with the data of a varying number of subjects (see x-axis in Fig. 2). The number of subjects in the training data only affected the fine-tuning step, as the pretraining step utilized the same amount of synthetic data regardless of the training data size.

Fig. 2 shows the sleep staging performance of the different training configurations (quantified by macro F1 scores) and how this performance depended on the number of subjects included in the training data. The pretrained feature extractors learned features that are more informative for sleep staging than those generated by a random feature extractor. This is especially evident in the low-data regime, where the performance gap between the untrained feature extractor and fixed feature extractor configurations was largest. Both configurations benefited from an increasing number of subjects in the training data, but the performance gap between the two configurations remained substantial even when trained with the data of all 56 subjects.

When comparing the fixed feature extractor to the fully supervised configuration, our pretraining scheme again appears to be most beneficial in the low-data regime. The fixed feature extractor configuration outperformed the fully supervised configuration by 0.08 in the average macro F1 score when trained with data from only one subject. This performance gap narrowed as more subjects were included in the training data, until both configurations achieved comparable macro F1 scores of 0.67 when trained with data from five subjects. Training with more than five subjects resulted in the fully supervised configuration outperforming the fixed feature extractor configuration.

Fine-tuning the feature extractor after pretraining appeared to combine the advantages of the fixed feature extractor configuration in the low-data regime and the fully supervised configuration in the high-data regime. When trained with the data of only one subject, the fine-tuned feature extractor configuration achieved similar performance to the fixed feature extractor configuration and outperformed the fully supervised configuration. When fine-tuned with the full training data, the fine-tuned feature extractor configuration was on par with the fully supervised configuration and outperformed the fixed feature extractor configuration. Overall, the fine-tuned feature extractor configuration achieved similar or better performance than the other training configurations across all numbers of subjects in the training data.

4 Discussion

Our results confirm the effectiveness of our pretraining scheme, particularly in *few-subject* regimes. Pretrained models outperformed fully supervised models when trained with a reduced number of subjects (see Fig. 2). This supports observations made in the field of Self-Supervised Learning (SSL) that pretrained models generally have better data efficiency than fully supervised ones (Banville et al., 2021; Eldele et al., 2023). In contrast to SSL methods, however, our pretraining scheme improves data efficiency without requiring empirical data. We hypothesize that this effect is caused by the priors that the model learns during pretraining. These priors could prevent overfitting to a small number of training samples, particularly those from minority classes (e.g., N1), or subject-specific features, which is especially problematic in situations with very little training data. As expected, we observed that all of our training configurations improved with a larger training dataset (see Fig. 2). This aligns with the prevalent view in the literature that deep learning models for sleep staging need substantial amounts of diverse data to perform well (Phan & Mikkelsen, 2022; Alvarez-Estevez, 2023; Fiorillo et al., 2019; 2023). When trained with the full training data, pretrained models performed comparably to fully supervised models, achieving macro F1 scores similar to those of other deep learning approaches for sleep staging (Phan & Mikkelsen, 2022; Gaiduk et al., 2023).

There are several opportunities for future work that could build upon our findings. One promising direction is to explore the pretraining task in more detail, for example, by investigating the impact of changing the frequency range that is used to generate the synthetic signals during pretraining. In addition, we suggest exploring models with greater capacity and less inductive bias, such as transformer models (Vaswani et al., 2017; Brandmayr et al., 2022), which we expect to benefit even more from our pretraining method. Pretraining such models with synthetic data may alleviate their need for large amounts of training data (Dosovitskiy et al., 2021). Another avenue for future research is to investigate the generalizability of our method to more diverse datasets. Initiatives such as the sleepdata.org platform (Zhang et al., 2018) and the Temple University data corpus (Obeid & Picone, 2016) provide a wide range of datasets that could be used for this purpose. Finally, it could be insightful to compare our approach with recent SSL methods (Liu et al., 2023) and data augmentation strategies that employ synthetic EEG generators (Lashgari et al., 2020; Habashi et al.,

2023). To enable such comparisons and to facilitate future research in this direction, we make our code available online (Grieger, 2024).

Our method presents a novel solution to address important issues that affect current deep learning models in the EEG time series domain, without requiring large amounts of patient data. We expect our approach to be advantageous in various applications where EEG data is scarce or derived from a limited number of subjects, such as brain-computer interfaces (Ko et al., 2021) or neurological disorder detection (van Dijk et al., 2022).

ACKNOWLEDGMENTS

We are grateful to M. Reißel and V. Sander for providing us with computing resources.

REFERENCES

Diego Alvarez-Estevez. Challenges of applying automated polysomnography scoring at scale. *Sleep Med. Clin.*, 18:277–292, 2023. ISSN 1556-407X. doi: 10.1016/j.jsmc.2023.05.002.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *J. Neural Eng.*, 18:046020, 2021. doi: 10.1088/1741-2552/abca18.

Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Annu. Conf. on Neural Information Processing Systems, NeurIPS*, pp. 2556–2569, virtual, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/14f2ebeab937ca128186e7ba876faef9-Abstract.html.

Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, Susan M. Harding, Robin M. Lloyd, Carole L. Marcus, and Bradley V. Vaughn. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6*. American Academy of Sleep Medicine, Darien, Illinois, 2020.

Georg Brandmayr, Manfred Martin Hartmann, Franz Fürbass, Gerald Matz, Matthias Samwald, Tilmann Kluge, and Georg Dorffner. Relational local electroencephalography representations for sleep scoring. *Neural Networks*, 154:310–322, 2022. doi: 10.1016/J.NEUNET.2022.07.020.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. doi: 10.48550/arxiv.2104.13478.

Friedrich Philipp Carrle, Yasmin Hollenbenders, and Alexandra Reichenbach. Generation of synthetic EEG data for training algorithms supporting the diagnosis of major depressive disorder. *Front. Neurosci.*, 17, 2023. ISSN 1662-453X. doi: 10.3389/fnins.2023.1219133.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. on Learning Representations, ICLR*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. Transfer learning for non-image data in clinical research: A scoping review. *PLOS Digital Health*, 1:e0000014, 2022. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000014.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Self-supervised learning for label-efficient sleep stage classification: A comprehensive evaluation. *IEEE T. Neur. Sys. Reh.*, 31:1333–1342, 2023. ISSN 1558-0210. doi: 10.1109/tnsre.2023.3245285.

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 Med. Rev.*, 48:101204, 2019. doi: 10.1016/j.smrv.2019.07.007.

Luigi Fiorillo, Giuliana Monachino, Julia van der Meer, Marco Pesce, Jan D. Warncke, Markus H. Schmidt, Claudio L. A. Bassetti, Athina Tzovara, Paolo Favaro, and Francesca D. Faraci. U-Sleep's resilience to AASM guidelines. *npj Digit. Medicine*, 6, 2023. doi: 10.1038/S41746-023-00784-0.

Maksym Gaiduk, Ángel Serrano Alarcón, Ralf Seepold, and Natividad Martínez Madrid. Current status and prospects of automatic sleep stages scoring: Review. *Biomed. Eng. Lett.*, 13:247–272, 2023. ISSN 2093-985X. doi: 10.1007/s13534-023-00299-3.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Massachusetts, 2016. ISBN 978-0-262-03561-3. URL `https://www.deeplearningbook.org/`.

Niklas Grieger. Source code of the model presented in Grieger et al., "Pretraining sleep staging models without patient data". `https://github.com/dslaborg/frequency-pretraining`, 2024.

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 T. Neur. Sys. Reh.*, 28:1955–1965, 2020. doi: 10.1109/tnsre.2020.3011181.

Ahmed G. Habashi, Ahmed M. Azab, Seif Eldawlatly, and Gamal M. Aly. Generative adversarial networks in EEG analysis: An overview. *J. NeuroEng. Rehabil.*, 20, 2023. ISSN 1743-0003. doi: 10.1186/s12984-023-01169-w.

Chao He, Jialu Liu, Yuesheng Zhu, and Wencai Du. Data augmentation for deep neural networks model in EEG classification task: A review. *Front. Hum. Neurosci.*, 15, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.765525.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE Int. Conf. on Computer Vision, ICCV 2015*, pp. 1026–1034, Santiago, Chile, 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.123.

Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. *Int. J. Comput. Vis.*, 130:990–1007, 2022. doi: 10.1007/S11263-021-01555-8.

Wonjun Ko, Eunjin Jeon, Seungwoo Jeong, Jaeun Phyo, and Heung-Il Suk. A survey on deep learning-based short/zero-calibration approaches for EEG-based brain–computer interfaces. *Front. Hum. Neurosci.*, 15, 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.643386.

Elnaz Lashgari, Dehua Liang, and Uri Maoz. Data augmentation for deep-learning-based electroencephalography. *J. Neurosci. Meth.*, 346:108885, 2020. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2020.108885.

Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23, 2023. ISSN 1424-8220. doi: 10.3390/s23094221.

Shayan Motamedi-Fakhr, Mohamed Moshrefi-Torbati, Martyn Hill, Catherine M. Hill, and Paul R. White. Signal processing techniques applied to human sleep EEG signals - A review. *Biomed. Signal Process. Control.*, 10:21–33, 2014. doi: 10.1016/J.BSPC.2013.12.003.

Iyad Obeid and Joseph Picone. The Temple University hospital EEG data corpus. *Front. Neurosci.*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196.

Huy Phan and Kaare Mikkelsen. Automatic sleep staging of EEG signals: Recent development, challenges, and future directions. *Physiol. Meas.*, 43:04TR01, 2022. ISSN 1361-6579. doi: 10.1088/1361-6579/ac6049.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: A systematic review. *J. Neural Eng.*, 16:051001, 2019. doi: 10.1088/1741-2552/ab260c.

Akara Supratak and Yike Guo. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *42nd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society, EMBC 2020*, pp. 641–644, Montreal, QC, Canada, 2020. IEEE. doi: 10.1109/EMBC44109.2020.9176741.

Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Sci. Data*, 9, 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01409-z.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Annu. Conf. Neural Information Processing Systems, NeurIPS*, pp. 5998–6008, Long Beach, CA, USA, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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. *J. Am. Med. Inform. Assn.*, 25:1351–1358, 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

A Methods Details

Sleep Staging Data. During the fine-tuning phase (see Fig. 1), we used two publicly available datasets: the DODO (55 recordings) and DODH (25 recordings) datasets (Guillot et al., 2020). All recordings were annotated with sleep stages (Wake, N1, N2, N3, REM) by five sleep experts following the AASM guidelines (Berry et al., 2020) and we used the consensus annotations of these experts for training and testing. In our experiments, we focused on the following EEG and EOG derivations that were available in all recordings: C3_M2, F3_M2, EOG1. We filtered these signals between 0.3 and 35 Hz, downsampled them to 100 Hz, and normalized each individual epoch. For training, we combined the DODO and DODH datasets (80 recordings in total) and performed 5-fold cross-validation. Each fold contained 14 training recordings and 2 validation recordings, which were used for hyperparameter tuning and early stopping.

Model. We based our model on the TinySleepNet architecture, a conceptually simple deep neural network for sleep staging that has previously demonstrated competitive results (Supratak & Guo, 2020). This architecture consists of a convolutional feature extractor that extracts features from individual epochs and a classifier that aggregates these feature across multiple epochs for sleep staging. We modified this classifier slightly by replacing the unidirectional LSTM layer with a bidirectional LSTM layer and used it to predict sleep stages during fine-tuning. During pretraining, we switched the classifier to a multi-layer perceptron classifier with two layers (80 neurons with ReLU activation and 20 neurons with sigmoid activation, respectively) to predict frequency bins.

Training. In the pretraining phase, we generated synthetic time series signals by summing sine waves with random frequencies, and then trained models to identify the frequency bins from which these frequencies were drawn for a given signal (see Fig. 1 and section 2). Each frequency bin was represented by a single output neuron of the model, with output values greater than 0.5 indicating that the corresponding frequency bin was used to generate the input signal.

In the fine-tuning phase, we trained models to predict sleep stages based on sequences of sleep staging data (see Fig. 1). Each sequence (training sample) consisted of 11 epochs, and the feature

extractor of our model generated features for each individual epoch. The LSTM-based classifier then aggregated these features and predicted the sleep stage of the middle epoch in the input sequence. To track model performance during fine-tuning, we recorded macro F1 scores on the training and validation data after each training epoch.