Deep transfer learning for improving single-EEG arousal detection

Alexander Neergaard Olesen[*,1,2,3], *Member, IEEE*, Poul Jennum[3],
Emmanuel Mignot[2] and Helge B. D. Sorensen[1], *Senior Member, IEEE*

Abstract— **Datasets in sleep science present challenges for machine learning algorithms due to differences in recording setups across clinics. We investigate two deep transfer learning strategies for overcoming the channel mismatch problem for cases where two datasets do not contain exactly the same setup leading to degraded performance in single-EEG models. Specifically, we train a baseline model on multivariate polysomnography data and subsequently replace the first two layers to prepare the architecture for single-channel electroencephalography data. Using a fine-tuning strategy, our model yields similar performance to the baseline model (F1=0.682 and F1=0.694, respectively), and was significantly better than a comparable single-channel model. Our results are promising for researchers working with small databases who wish to use deep learning models pre-trained on larger databases.**

I. INTRODUCTION

A principal tool in the analysis of sleep is the polysomnography (PSG). Standard PSGs contain electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) from below the chin and lower limbs, electrocardiography, respiratory effort, and blood oxygenation, which is manually analysed by sleep experts according to guidelines published by the American Academy of Sleep Medicine [1].

Experts score sleep stages and annotate discrete events, such as arousals (short awakenings during sleep, $\leq 15\,\mathrm{s}$), limb movements, and decreased respiratory effort characterized by apneas (complete cessation of breathing), hypopneas (partial cessation of breathing), and desaturations (decreases in oxygen desaturation). Low inter-rater reliability has been reported for sleep stages scoring in multiple studies [2]–[4], arousals [5], and respiratory events [6], [7], prompting extensive research in automated methods for sleep analysis [8]–[14].

Designing reliable and robust systems for automated sleep analysis based on machine learning algorithms often require multiple heterogenous data sources of sufficient size. However, due to differences in clinical practice, very few datasets in sleep science are standardized with regards to recording setups despite guidelines from the AASM.

[†]Corresponding author: `aneol@dtu.dk`

[1]Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

[2]Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA 94304, USA.

[3]Danish Center for Sleep Medicine, University Hospital Copenhagen, 2600 Glostrup, Denmark

In these cases, we end up with a *channel mismatch problem*, in which the overlap between our source and target domains is small, and the domains are possibly disjointed. Recent studies have investigated the use of deep transfer learning to solve the channel mismatch problem when training and testing sleep stage classification models [15], [16]. The authors found that using a fine-tuning strategy significantly improved the performance of sleep stage scoring models when trained on various combinations of EEG and EOG channels.

We present results on using deep transfer learning to address the channel mismatch problem, when the source and target domains differ both in the number and type of channel modalities. Specifically, our source domain consists of multivariate PSG data comprising left and right central EEG, left and right EOG, and submental EMG recordings, while our target domain consists of only a single central EEG channel. We show that by employing a simple fine-tuning strategy on a pre-trained network stripped of the initial two layers, we can effectively reach the same level of F1 score as when using the full set of PSG data.

II. METHODS

Notation: We denote by $[\![a, b]\!]$ the set of integers $\{n \in \mathbb{N} \mid a \leq n \leq b\}$ with $[\![N]\!]$ being shorthand for $[\![1, N]\!]$, and by $n \in [\![N]\!]$ the nth sample in $[\![N]\!]$. A model for a given experiment is denoted by $\mathcal{M}_{(\cdot)}$, while an optimized model is superscripted with a star as $\mathcal{M}_{(\cdot)}^*$. A segment of PSG data is denoted by $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C, T is the number of channels and the duration of the segment in samples, respectively.

A. Data

We collected PSGs from 1500 subjects in the MrOS Sleep Study [17]–[19] from the National Sleep Research Resource repository [20], [21]. From each PSG, we extracted left and right EEG, left and right EOG, and chin EMG. EEG and EOG channels were referenced to the contralateral mastoid process. For each PSG, we also extracted time-stamped arousal scorings containing starts and durations of scored arousal events. We did not exclude any PSGs from this study based on sleep duration, number of arousal events, or similar criteria.

B. Data partitioning

The 1500 PSGs were initially partitioned into three subsets TRAIN$_1$, EVAL$_1$, and TEST$_1$ containing 400, 100 and 1000 PSGs, respectively. Furthermore, we additionally partitioned TEST$_1$ into three smaller subsets TRAIN$_2$, EVAL$_2$, and TEST$_2$ containing 400, 100, and 500 PSGs, respectively.

TABLE I
NETWORK ARCHITECTURE OVERVIEW.

Module	Layer type	Kernel	Stride	Feature maps	Input size	Output size	Activation
$\mathbf{x}$	Input	—	—	—	$C \times T$	$1 \times C \times T$	—
ϕ_{mix}	2D convolution	$(C, 1)$	$(1, 1)$	C	$1 \times C \times T$	$C \times 1 \times T$	ReLU
$\varphi_{\mathrm{conv},1}$	2D convolution	$(1, c)$	$(1, s)$	$2f_0$	$C \times 1 \times T$	$2f_0 \times 1 \times T/s$	—
	Batch norm.	—	—	$2f_0$	$2f_0 \times 1 \times T/s$	$2f_0 \times 1 \times T/s$	ReLU
$\varphi_{\mathrm{conv},k}$	2D convolution	$(1, c)$	$(1, s)$	$f_0 2^k$	$f_0 2^{k-1} \times 1 \times T/s^{k-1}$	$f_0 2^k \times 1 \times T/s^k$	—
$k \in [\![2, k_{\max}]\!]$	Batch norm.	—	—	$f_0 2^k$	$f_0 2^k \times 1 \times T/s^k$	$f_0 2^k \times 1 \times T/s^k$	ReLU
φ_{rec}	bGRU	—	—	f'	$f' \times 1 \times T'$	$f' \times 2 \times T'$	—
ψ_{clf}	2D convolution	$(2, 1)$	$(1, 1)$	$(K+1)N_d$	$f' \times 2 \times T'$	$(K+1)N_d \times 1 \times T'$	Softmax over $K+1$
ψ_{loc}	2D convolution	$(2, 1)$	$(1, 1)$	$2N_d$	$f' \times 2 \times T'$	$2N_d \times 1 \times T'$	Linear
$\mathbf{z}$	Output, $\mathbf{p}$	—	—	—	$(K+1)N_d \times 1 \times T'$	$N_d \times T' \times (K+1)$	—
	Output, $\mathbf{y}$	—	—	—	$2N_d \times 1 \times T'$	$N_d \times T' \times 2$	—

$\mathbf{x}$, input containing PSG data; $\mathbf{z}$, output containing predicted arousal probabilities and associated start and duration predictions; ϕ_{mix}, non-linear mixing block; φ_{conv}, convolutional feature extraction block; φ_{rec} recurrent feature extraction block; ψ_{clf}, event classification block; ψ_{loc}, event localization block; C, number of input channels; T, number of samples in a segment of PSG data; c, temporal kernel size; s, temporal stride; f_0, base number of feature maps; $f' = f_0 2^{k_{\max}}$, maximum number of feature maps; $T' = T/s^{k_{\max}}$, reduced temporal dimension in samples; N_d, number of default event windows in segment; K, number of classes; ReLU, rectified linear unit; bGRU, bidirectional gated recurrent unit.

C. Preprocessing pipeline

All signals were resampled to $128\,\mathrm{Hz}$ using poly-phase filtering with a Kaiser window ($\beta = 5.0$) prior to subsequent processing. Extracted EEG and EOG signals were filtered with 2nd order Butterworth IIR bandpass filters with cutoff frequencies $0.3\,\mathrm{Hz}$ and $35\,\mathrm{Hz}$. Chin EMG was filtered with a 4th order Butterworth IIR highpass filter with a cutoff frequency of $10\,\mathrm{Hz}$. Filtered signals were subsequently standardized by

$$\mathbf{x}^{(i)} = \frac{\tilde{\mathbf{x}}^{(i)} - \boldsymbol{\mu}^{(i)}}{\boldsymbol{\sigma}^{(i)}}, \tag{1}$$

where $\tilde{\mathbf{x}}^{(i)} \in \mathbb{R}^{C \times T}$ is the raw matrix containing C input channels and T samples, and $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)} \in \mathbb{R}^C$ are the mean and standard deviation vectors for the i'th PSG, respectively.

D. Model setup

We expand upon previous work using similar models for sleep event detection [22]–[24]. Briefly, the model takes as input a tensor of PSG data $\mathbf{x} \in \mathbb{R}^{C \times T}$ and outputs

$$\mathbf{z} = (\mathbf{p}, \mathbf{y}) \in \mathbb{R}^{N_d \times T' \times (K+1)} \times \mathbb{R}^{N_d \times T' \times 2} \tag{2}$$

containing predicted arousal probabilities $\mathbf{p}$ and associated start and durations for predicted arousal events $\mathbf{y}$. The differentiable function underlying the model comprises a deep neural network architecture consisting of the following modules:

a) Input mixing module: Here, non-linear combinations of the input PSG data $\mathbf{x}$ are made using a non-linear mixing block $\phi_{\mathrm{mix}} : \mathbb{R}^{1 \times C \times T} \to \mathbb{R}^{C \times 1 \times T}$.

b) Feature extraction module: This module contains two components. The first is a convolutional feature extraction block $\varphi_{\mathrm{conv}} : \mathbb{R}^{C \times 1 \times T} \to \mathbb{R}^{f' \times 1 \times T'}$ consisting of $k_{\max}$ successions of convolutional, batch normalization, and rectified linear unit (ReLU) layers. Second is a recurrent feature extraction block $\varphi_{\mathrm{rec}} : \mathbb{R}^{f' \times 1 \times T'} \to \mathbb{R}^{f' \times 2 \times T'}$ with $f' = f_0 2^{k_{\max}}$ hidden units. The φ_{conv} block is responsible for bulk feature extraction and temporal decimation using strided convolutions, while φ_{rec} processes the raw features across the reduced temporal dimension using a bidirectional gated recurrent unit [25] with f' hidden units.

c) Event detection module: The output from φ_{rec} is processed by two separate blocks: $\psi_{\mathrm{clf}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{(K+1)N_d \times 1 \times T'}$ outputs the tensor $\mathbf{p}$ containing predicted arousal probabilities for each time point $t \in [\![T']\!]$ for each default event window. $\psi_{\mathrm{loc}} : \mathbb{R}^{f' \times 2 \times T'} \to \mathbb{R}^{2N_d \times T'}$ outputs the tensor $\mathbf{y}$ containing predicted start time and durations of arousal events. Both ψ_{clf} and ψ_{loc} are implemented using $(2, 1)$ convolutions rather than convolutions over the entire volume as in [22]–[24]. This serves a dual purpose: the first is to reduce the number of parameters to make the network more memory-efficient, while the second purpose is to allow the kernel and feature maps to be temporally invariant.

For a detailed description of the network architecture, see Table I.

E. Loss objective

The network parameters were optimized according to a three-component loss objective comprising a localization loss ℓ_{loc} and a positive and negative classification loss ℓ_+ and ℓ_-, respectively, such that

$$\ell = \ell_{\mathrm{loc}} + \ell_+ + \ell_-. \tag{3}$$

The localization loss was calculated using a Huber function

$$\ell_{\mathrm{loc}} = \frac{1}{N_{\pi \setminus \emptyset}} \sum_{i \in \pi \setminus \emptyset} h^{(i)} \tag{4}$$

$$\mathbf{h} = \begin{cases} 0.5(\mathbf{y} - \mathbf{t})^2, & \text{if } |\mathbf{y} - \mathbf{t}| < 1, \\ |\mathbf{y} - \mathbf{t}| - 0.5, & \text{otherwise,} \end{cases} \quad (5)$$

where $i \in \pi \backslash \emptyset$ indicates event windows with a non-empty arousal target. Contributions from the positive/negative classification losses were calculated using a focal loss function [26]:

$$\ell_+ = \frac{1}{N_{\pi \backslash \emptyset}} \sum_{i \in \pi \backslash \emptyset} -\alpha (1 - \mathbf{p})^\gamma \log(\mathbf{p}), \text{ and} \quad (6)$$

$$\ell_- = \frac{1}{N_{\pi = \emptyset}} \sum_{i \in \pi = \emptyset} -\alpha (1 - \mathbf{p})^\gamma \log(\mathbf{p}), \quad (7)$$

where $\alpha = 0.25$ and $\gamma = 2$. This serves to counter the class imbalance in a single data segment, which typically consists of many event windows with few positive examples.

F. Experimental setups

We investigated the channel mismatch problem with the following four experimental setups:

a) Full montage baseline (FM): In this experiment, we trained the event detection algorithm on TRAIN$_1$ using $C = 5$ channels: left/right central EEG, left/right EOG, and chin EMG. Convergence and the optimal detection threshold were assessed on EVAL$_1$ and performance was evaluated on TEST$_2$. The optimal baseline model was used as an initialization for the two transfer learning experiments described below.

b) Pretraining (PT): The optimal model $\mathcal{M}_{\mathrm{FM}}^*$ was used in this experiment as an initialization for $\mathcal{M}_{\mathrm{PT}}$. We adjusted the mixing module and first convolutional layer in the feature extraction module to account for the channel mismatch by replacing the convolutional and batch normalization layers, and subsequently trained these from scratch. The rest of the weights and bias terms were frozen to the optimized values from $\mathcal{M}_{\mathrm{FM}}^*$. The network was trained on TRAIN$_2$ with only $C = 1$ channels (left central EEG, C3). Convergence and optimal detection threshold were assesed on EVAL$_2$, while final performance was evaluated on TEST$_2$.

c) Fine-tuning (FT): Similar to PT, the optimal model $\mathcal{M}_{\mathrm{FM}}^*$ was used in this experiment as an initialization for $\mathcal{M}_{\mathrm{FT}}$. Also, the mixing module and first convolutional layer in the feature extraction module were likewise adjusted. However, all other layers in $\mathcal{M}_{\mathrm{FT}}$ were permitted to be further optmized by fine-tuning weights and bias terms during training. The model was trained using the same 400 PSGs from TRAIN$_2$ with the same $C = 1$ channel configuration as in PT.

d) Single EEG benchmark (SE): We benchmarked our two transfer learning experiments to a comparable situation in which an event detection model was trained on the same PSGs in TRAIN$_2$ using only the left central EEG (C3).

In all experimental runs, we optimized the loss objective in eq. (3) using the Adam optimization algorithm with a learning rate of $\alpha = 10^{-3}$ and the default parameter values $(\beta_1, \beta_2) = (0.9, 0.999)$ as suggested in [27]. We applied the same data sampling strategy as proposed in [24], in which a segment of data is sampled such that it contains at least 50% of a randomly sampled event across all PSGs. We used
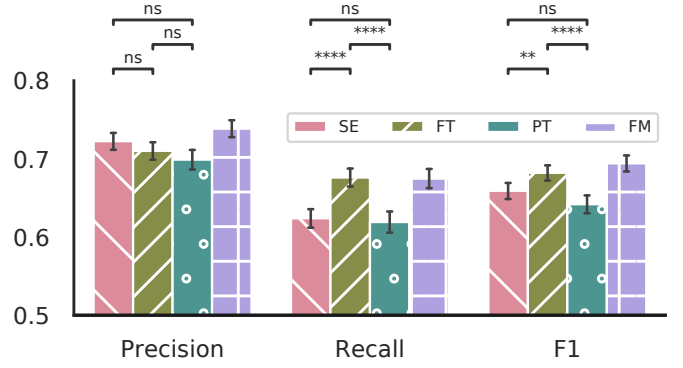

Fig. 1. Performance metrics as evaluated on TEST$_2$ for each experimental setup. Metrics are shown as means with 95% confidence interval as error bars. Note the y-axis scaling. SE: single-EEG. FT: fine-tuning. PT: pre-training. FM: full montage. ns: not significant, **: $p_{\mathrm{adj}} \leq 10^{-2}$; ****: $p_{\mathrm{adj}} \leq 10^{-4}$.

TABLE II
PERFORMANCE METRICS ACROSS EXPERIMENTS.

Experiment	Precision	Recall	F1
FM	0.739 ± 0.122	0.675 ± 0.139	0.694 ± 0.115
SE	0.723 ± 0.124	0.624 ± 0.137	0.659 ± 0.117
FT	$\mathbf{0.710 \pm 0.128}$	$\mathbf{0.676 \pm 0.130}$	$\mathbf{0.682 \pm 0.110}$
PT	0.699 ± 0.141	0.619 ± 0.153	0.642 ± 0.129

Metrics are shown evaluated on TEST$_2$ as means $\pm$ standard deviation. Best performing transfer learning experiment is shown in bold. SE: single-EEG. FT: fine-tuning. PT: pre-training. FM: full montage.

a default event window size of $15\,\mathrm{s}$ with $50\,\%$ overlap as this was found previously to work well for arousal detection [24].

All experiments were implemented in PyTorch 1.2 [28].

G. Performance evaluation

Bipartite matching were used to match detected and true events during training and testing. At test time, detected events were subjected to non-maximum suppression based on an intersection-over-union (IOU) of at least 0.5 between detected and true events. We evaluated the performance of our experimental setups using precision, recall and F1 scores.

H. Statistical analysis

We used Kruskal–Wallis one-way analysis of variance tests for differences in performance metrics between groups (SE, FT and PT) with a significance level of $\alpha = 0.05$. Post-hoc testing was performed with Mann-Whitney U-tests for each pair-combination (SE/FT, SE/PT, and FT/PT) likewise with $\alpha = 0.05$. We accounted for multiple comparisons by adjusting p-values with Bonferroni corrections.

III. RESULTS AND DISCUSSION

We show the results of the transfer learning experiments (FT, PT) as well as the baseline and benchmark experiments (FM, SE) in Fig. 1 and Table II. Performance metrics were not calculated for 10 subjects in TEST$_2$, as these did not

have any scored arousals and are thus not reflected in Fig. 1 and Table II.

The baseline F1 performance is shown to be slightly lower than previously reported (0.694 ± 0.115 vs. 0.749 ± 0.105 [24]). However, our baseline model was trained on 400 subjects compared to 1485 in [24], which would account for the lower F1 score. By reducing the available input channels from $C = 5$ different modalities to $C = 1$ EEG channels as in the SE benchmark experiment, the F1 score drops to 0.659 ± 0.117, while the precision and recall scores likewise drop from 0.739 ± 0.122 to 0.723 ± 0.124, and 0.675 ± 0.139 to 0.624 ± 0.137, respectively.

We found statistically significant differences in F1 scores between SE, FT, and PT ($p = 3.189 \times 10^{-7}$). Post-hoc testing further revealed statistically significant differences between SE and FT ($p_{\mathrm{adj}} = 2.224 \times 10^{-3}$), and FT and PT ($p_{\mathrm{adj}} = 2.685 \times 10^{-7}$), but not between SE and PT ($p_{\mathrm{adj}} = 0.080$). We also found that recall scores differed between experimental setups ($p = 7.085 \times 10^{-13}$). Post-hoc testing showed statistically significant differences between SE, FT ($p_{\mathrm{adj}} = 5.180 \times 10^{-11}$), and FT and PT ($p_{\mathrm{adj}} = 1.440 \times 10^{-9}$), but not between SE and PT ($p_{\mathrm{adj}} = 1.000$). Lastly, we saw statistically significant differences in precision scores between experimental setups ($p = 0.033$), subsequent post-hoc testing did not reveal any statistical significant differences, when adjusting for multiple comparisons using the Bonferroni procedure (SE/FT, $p_{\mathrm{adj}} = 0.214$; FT/PT, $p_{\mathrm{adj}} = 1.000$; SE/PT, $p_{\mathrm{adj}} = 0.037$).

Our results show, that for some scenarios, we can learn information present in multi-variate PSG data and efficiently transfer that information to a target domain containing only a single EEG channel. Specifically, the performance of our fine-tuning strategy is high enough that the mean F1 scores across subjects are statistically insifignant, when comparing FT and FM setups (not shown).

Previous related work focused on the channel mismatch problem, when comparing different, but the same number of, channel modalities such as transferring EEG-based models to EOG-based target domains, and thus did not investigate how changing the model architecture might impact performance [15], [16]. In this work, we investigated transfer learning when the source and target domains only overlap by one input channel. This necessitates changing some parts of the underlying model architecture to accommodate the different number of input channels, and these changes might impact downstream feature extraction. We did not explore simply zeroing out a large number of input channels in this work, as this requires exhaustive search of which channel indices to zero out in the model based on the number of target input channels. Our strategy does not require this exhaustive search.

Our study applied a simple optimization strategy for the transfer learning experiments, which might limit the potential performance gain. This is especially relevant for the FT experiment. For example, one could experiment with with different learning rates and scheduling schemes for the initial layers and pre-trained layers, such that the initial layers were

trained with a higher relative learning rate to compensate for their lack of initial training.

Furthermore, we explored transfer learning for the channel mismatch problem in a single cohort of patient recordings. Future directions of this research will investigate scenarios, where both the source and target domains, and the datasets are different.

IV. CONCLUSIONS

We show in our experiments that a simple fine-tuning strategy can be employed to transfer learning from a model based on multi-variate PSG data to a configuration where only a single EEG lead is available for detecting arousals, and that the difference between single-EEG and multivariate PSG performance is negligible. Future work will explore the effects of various combinations of datasets on the impact of generalized event detection, when the source and target domains do not overlap completely.

ACKNOWLEDGMENT

REFERENCES

[1] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events*, 2.6. Darien, IL: American Academy of Sleep Medicine, 2020.

[2] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset.," *Sleep*, vol. 23, no. 7, pp. 901–8, 2000.

[3] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *J. Clin. Sleep Med.*, vol. 09, no. 01, pp. 81–87, 2013. DOI: 10.5664/jcsm.2350.

[4] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *J. Clin. Sleep Med.*, vol. 12, no. 6, pp. 885–894, 2016. DOI: 10.5664/jcsm.5894.

[5] M. H. Bonnet, K. Doghramji, T. Roehrs, *et al.*, "The Scoring of Arousal in Sleep: Reliability, Validity, and Alternatives," *J. Clin. Sleep Med.*, vol. 03, no. 02, pp. 133–145, 2007. DOI: 10.5664/jcsm.26815.

[6] U. J. Magalang, N.-H. Chen, P. A. Cistulli, *et al.*, "Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013. DOI: 10.5665/sleep.2552.

[7] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Respiratory Events," *J. Clin. Sleep Med.*, vol. 10, no. 04, pp. 447–454, 2014. DOI: 10.5664/jcsm.3630.

[8] A. N. Olesen, P. Jennum, P. Peppard, E. Mignot, and H. B. D. Sorensen, "Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA: IEEE, 2018, pp. 1–4. DOI: 10.1109/EMBC.2018.8513080.

[9] J. B. Stephansen, A. N. Olesen, M. Olsen, *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, p. 5229, 2018. DOI: 10.1038/s41467-018-07229-3.

[10] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018. DOI: `10.1109/TNSRE.2018.2813138`.

[11] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018. DOI: `10.1093/jamia/ocy131`.

[12] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2019. DOI: `10.1109/TNSRE.2019.2896659`.

[13] L. Carvelli, A. N. Olesen, A. Brink-Kjær, *et al.*, "Design of a deep learning model for automatic scoring of periodic and non-periodicleg movements during sleep validated against multiple human experts," *Sleep Medicine*, DOI: `10.1016/j.sleep.2019.12.032`.

[14] A. Brink-Kjaer, A. N. Olesen, P. E. Peppard, *et al.*, "Automatic Detection of Cortical Arousals in Sleep and their Contribution to Daytime Sleepiness," 2019. arXiv: `1906.01700 [q-bio.NC]`.

[15] H. Phan, O. Y. Chén, P. Koch, *et al.*, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," pp. 1–11, 2019. arXiv: `1907.13177 [cs.LG]`.

[16] H. Phan, O. Y. Chen, P. Koch, A. Mertins, and M. D. Vos, "Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain: IEEE, 2019, pp. 1–5. DOI: `10.23919/EUSIPCO.2019.8902977`.

[17] J. B. Blank, P. M. Cawthon, M. L. Carrion-Petersen, *et al.*, "Overview of recruitment for the osteoporotic fractures in men study (MrOS)," *Contemporary Clinical Trials*, vol. 26, no. 5, pp. 557–568, 2005. DOI: `10.1016/j.cct.2005.05.005`.

[18] E. Orwoll, J. B. Blank, E. Barrett-Connor, *et al.*, "Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men," *Contemporary Clinical Trials*, vol. 26, no. 5, pp. 569–585, 2005. DOI: `10.1016/j.cct.2005.05.006`.

[19] T. Blackwell, K. Yaffe, S. Ancoli-Israel, *et al.*, "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*, vol. 59, no. 12, pp. 2217–2225, 2011. DOI: `10.1111/j.1532-5415.2011.03731.x`.

[20] D. A. Dean, A. L. Goldberger, R. Mueller, *et al.*, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016. DOI: `10.5665/sleep.5774`.

[21] G.-Q. Zhang, L. Cui, R. Mueller, *et al.*, "The National Sleep Research Resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018. DOI: `10.1093/jamia/ocy064`.

[22] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "A Deep Learning Architecture to Detect Events in EEG Signals During Sleep," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark: IEEE, 2018, pp. 1–6. DOI: `10.1109/MLSP.2018.8517067`.

[23] S. Chambon, V. Thorey, P. Arnal, E. Mignot, and A. Gramfort, "DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal," *Journal of Neuroscience Methods*, vol. 321, pp. 64–78, 2019. DOI: `10.1016/j.jneumeth.2019.03.017`.

[24] A. N. Olesen, S. Chambon, V. Thorey, P. Jennum, E. Mignot, and H. B. D. Sorensen, "Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany: IEEE, 2019, pp. 556–561. DOI: `10.1109/EMBC.2019.8856570`.

[25] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," arXiv: `1409.1259 [cs.CL]`.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. DOI: `10.1109/TPAMI.2018.2858826`.

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pp. 1–15, 2015. arXiv: `1412.6980 [cs.LG]`.

[28] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advanced in Neural Information Processing Systems 32*, 2019, pp. 8024–8035. arXiv: `1912.01703 [cs.LG]`.