Deep Transfer Learning for Automated Single-Lead EEG Sleep Staging with Channel and Population Mismatches

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 05 January 2024
DOI 10.3389/fphys.2023.1287342

OPEN ACCESS

EDITED BY
Felicia Jefferson,
University of Nevada, Reno, United States

REVIEWED BY
Wolfgang Ganglberger,
Massachusetts General Hospital and Harvard Medical School, United States
Yanru Li,
Capital Medical University, China

*CORRESPONDENCE
Jaap F. Van Der Aar,
j.f.v.d.aar@tue.nl

RECEIVED 01 September 2023
ACCEPTED 08 December 2023
PUBLISHED 05 January 2024

CITATION
Van Der Aar JF, Van Den Ende DA, Fonseca P, Van Meulen FB, Overeem S, Van Gilst MM and Peri E (2024), Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches.
Front. Physiol. 14:1287342.
doi: 10.3389/fphys.2023.1287342

Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches

Jaap F. Van Der Aar[1,2]*, Daan A. Van Den Ende[2], Pedro Fonseca[1,2], Fokke B. Van Meulen[1,3], Sebastiaan Overeem[1,3], Merel M. Van Gilst[1,3] and Elisabetta Peri[1]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands, [2]Philips Research, Eindhoven, Netherlands, [3]Kempenhaeghe Center for Sleep Medicine, Heeze, Netherlands

Introduction: Automated sleep staging using deep learning models typically requires training on hundreds of sleep recordings, and pre-training on public databases is therefore common practice. However, suboptimal sleep stage performance may occur from mismatches between source and target datasets, such as differences in population characteristics (e.g., an unrepresented sleep disorder) or sensors (e.g., alternative channel locations for wearable EEG).

Methods: We investigated three strategies for training an automated single-channel EEG sleep stager: pre-training (i.e., training on the original source dataset), training-from-scratch (i.e., training on the new target dataset), and fine-tuning (i.e., training on the original source dataset, fine-tuning on the new target dataset). As source dataset, we used the F3-M2 channel of healthy subjects (N = 94). Performance of the different training strategies was evaluated using Cohen's Kappa (κ) in eight smaller target datasets consisting of healthy subjects (N = 60), patients with obstructive sleep apnea (OSA, N = 60), insomnia (N = 60), and REM sleep behavioral disorder (RBD, N = 22), combined with two EEG channels, F3-M2 and F3-F4.

Results: No differences in performance between the training strategies was observed in the age-matched F3-M2 datasets, with an average performance across strategies of κ = .83 in healthy, κ = .77 in insomnia, and κ = .74 in OSA subjects. However, in the RBD set, where data availability was limited, fine-tuning was the preferred method (κ = .67), with an average increase in κ of .15 to pre-training and training-from-scratch. In the presence of channel mismatches, targeted training is required, either through training-from-scratch or fine-tuning, increasing performance with κ = .17 on average.

Discussion: We found that, when channel and/or population mismatches cause suboptimal sleep staging performance, a fine-tuning approach can yield similar to superior performance compared to building a model from scratch, while requiring a smaller sample size. In contrast to insomnia and OSA, RBD data contains characteristics, either inherent to the pathology or age-related, which apparently demand targeted training.

KEYWORDS

polysomnography, sleep staging, single-channel, wearable EEG, fine-tuning, deep learning

1 Introduction

Sleep stage scoring is an essential component of both sleep research and the clinical diagnosis of sleep disorders. Conventionally, polysomnographic (PSG) recordings are manually scored by trained clinicians according to the American Academy of Sleep Medicine (AASM) guidelines (Troester et al., 2023). 30-s epochs are classified as either Wake, N1, N2, N3, or rapid-eye-movement (REM) sleep through visual inspection of multiple leads, including electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG). This manual process is labor-intensive, time-consuming, and suffers from a relatively large degree of inter-rater disagreement (Danker-Hopfe et al., 2004; Danker-Hopfe et al., 2009; Rosenberg and van Hout, 2013).

Therefore, in recent years, there has been a push towards automated sleep stage classification, enabled by advances in machine learning (Tsinalis et al., 2016; Supratak et al., 2017; Korkalainen et al., 2019; Mousavi et al., 2019; Phan et al., 2019; Seo et al., 2020; Guillot and Thorey, 2021). This development is further driven by the rise of new sleep monitoring technologies including wearable EEG. Wearable EEG facilitates the collection of more, longitudinal data that can be scored in a cost-efficient manner using automated sleep staging (Singh et al., 2015). Since wearable EEG usually uses less channels and potentially other electrode locations than the AASM standard, manual scoring is not always possible, while automated sleep stagers can be specifically trained on these channels (Finan et al., 2016; Garcia-Molina et al., 2018; Arnal et al., 2020). In particular, the use of deep neural network learning methods for automated sleep staging has yielded promising results, showing high agreement with manual scoring, similar or even superior to the inter-rater agreement between manual scorers (see, e.g., Fiorillo et al., 2019; Phan and Mikkelsen, 2022 for extensive reviews).

To reach expert-level performance using deep learning, hundreds of sleep recordings are typically required for model training (Biswal et al., 2018; Guillot and Thorey, 2021; Perslev et al., 2021). Databases such as the Montreal Archive of Sleep Studies (MASS; O'Reilly et al., 2014), the National Sleep Research Resource (NSRR; Zhang et al., 2018), and PhysioNet (Goldberger et al., 2000), together comprise thousands of sleep recordings and enable large scale training. However, training on publicly available data and subsequently employing the model on novel data of interest can be problematic when mismatches exist between source and target datasets, for example, with respect to population characteristics, or sensors.

Population mismatches can be present because population cohorts of public datasets generally contain either examples of healthy subjects, or mainly patients with common sleep disorders as obstructive sleep apnea (OSA) and insomnia. Patients with these sleep disorders can differ from healthy sleepers in sleep characteristics such as increased sleep fragmentation, and longer wake before sleep and after sleep onset (Mannarino et al., 2012; Baglioni et al., 2014). Many sleep disorders are much less prevalent, making data availability for training less abundant. Importantly, these sleep disorders can exhibit specific pathophysiological characteristics which are not learned by the model. The effects of such a population mismatch may, for example, appear in patients with REM sleep behavioral disorder (RBD). RBD patients are generally older, and the disorder is characterized by the absence of muscle atonia during REM sleep, causing dream-enacting behavior (Boeve et al., 2007; Sateia, 2014). Potentially, these characteristics contribute to the underperformance of automated sleep staging in RBD to healthy subjects and other sleep disorders, especially in REM sleep classification (Andreotti et al., 2018; Cooray et al., 2019).

Channel mismatches between the publicly available dataset and the novel dataset are the result when the electrode location of interest is not included in the publicly available data. For example, increasing interest in (prolonged) in-home sleep monitoring has led to the development of less obtrusive sleep monitoring technologies, including wearable EEG using dry frontopolar electrodes (Finan et al., 2016; Garcia-Molina et al., 2018; Arnal et al., 2020). The frontopolar location is not included in the majority of public available datasets, since it is not covered by AASM standards (Troester et al., 2023). Furthermore, compared to the standard wet electrodes in a PSG montage, dry electrodes in wearable systems result in lower signal quality and slightly different EEG signal information (Lopez-Gordo et al., 2014).

There are three different training strategies to employ an automated sleep stager on novel data, the new target dataset. In "pre-training", the model is only trained on the original source dataset, often public data (Guillot and Thorey, 2021). Hence, sleep staging performance in the target dataset is susceptible to the presence of data mismatches (Phan et al., 2019; Phan et al., 2019). On the other hand, "training-from-scratch" can be used to train a new model only on the target dataset (Biswal et al., 2018; Perslev et al., 2021). However, data availability in the target dataset can be too limited for sufficient training of a deep learning model. For instance, the prevalence of specific sleep disorders such as RBD is low, and the validation of new EEG monitoring technologies is often limited to a minimal number of healthy subjects (Finan et al., 2016; Mikkelsen et al., 2017; Bresch et al., 2018; Sterr et al., 2018; Arnal et al., 2020). Recently, "fine-tuning" has been proposed as a solution to overcome data mismatches and limited data availability. In this form of transfer learning, the model is first pre-trained on the source dataset and further fine-tuned on the (smaller) target dataset to learn its specific characteristics (Pan and Yang, 2010). Although transfer learning for automated sleep staging shows promising results, performance improvement is limited (Andreotti et al., 2018; He et al., 2023). Also, it remains unknown how fine-tuning can best be implemented, since ideal settings seem specific to the deep learning architecture (Phan et al., 2019).

While each of the strategies (pre-training, training-from-scratch, and fine-tuning) has been used for training automated sleep stagers, no studies have been performed to systematically assess which method is favorable. The aim of this study was to evaluate which training strategy is preferred in the presence of data mismatches and limited data availability. For each strategy, we analyzed the performance on the combination of three age-matched populations (healthy subjects vs. OSA vs. insomnia) and one population with limited data availability (RBD), with two sets of EEG channels (F3-M2 vs. F3-F4). By comparing all strategies, populations, and channels in a systematic way, we could study the isolated effects of all these parameters. The TinySleepNet (Supratak and Guo, 2020), a previously published deep learning

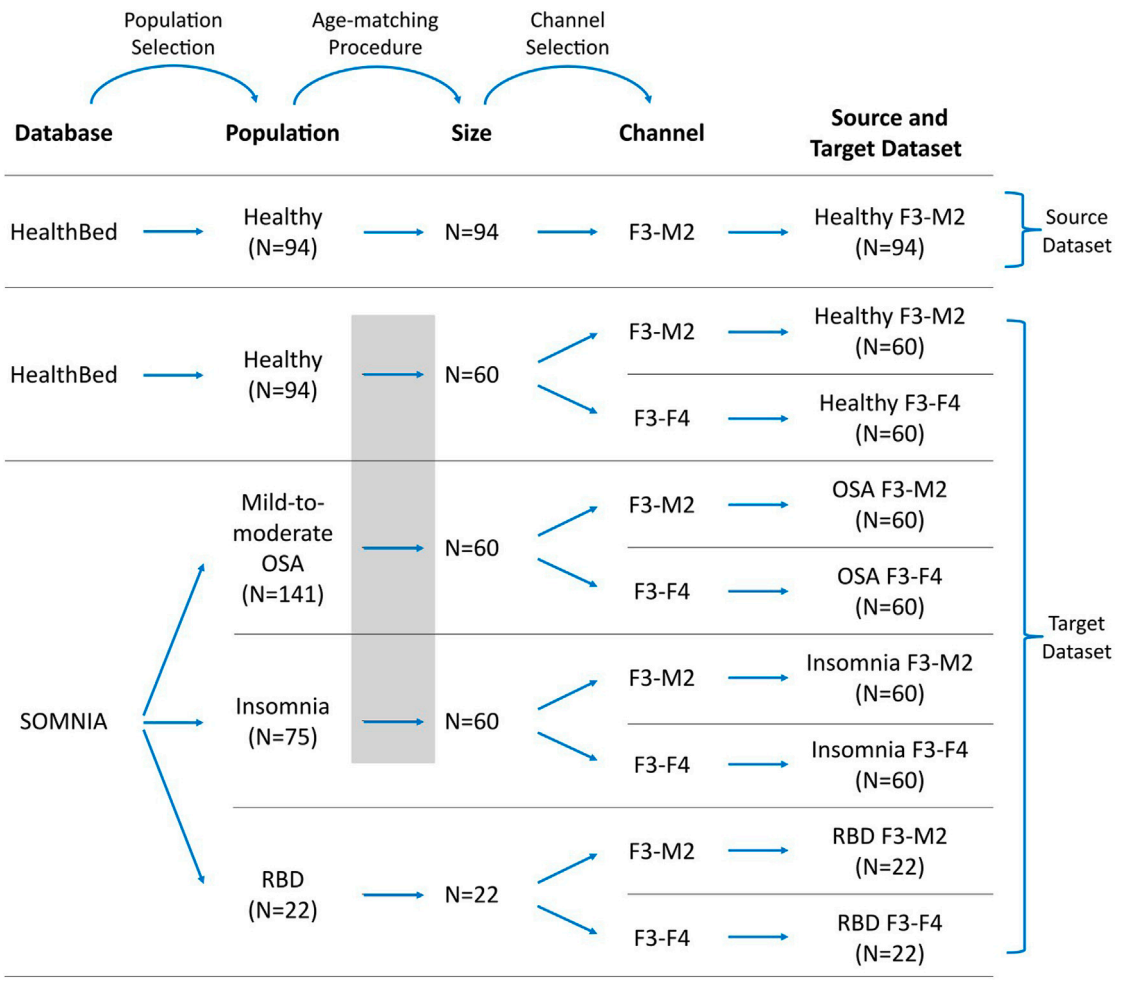

FIGURE 1
Workflow for defining the datasets. Datasets were selected from the Healthbed (van Meulen et al., 2023) and SOMNIA (van Gilst et al., 2019) databases. Population selection, an age-matching procedure, and channel selection resulted in one source dataset and eight target datasets. In target datasets, data availability can be limited, and/or can differ in characteristics to the source dataset due to population and channel mismatches. Grey box indicates the age-matching procedure for the healthy, OSA and insomnia target datasets.

model for single-channel EEG, was used as automated sleep stage model. Both EEG channels and the automated model were chosen for their potential implementation in wearable EEG systems.

2 Materials and methods

2.1 Data

2.1.1 Databases

The sleep-disordered participants were sampled from the PSG recordings of the Sleep and Obstructive Sleep Apnea Measuring with Non-Invasive Applications (SOMNIA) cohort recorded before January 2021 (van Gilst et al., 2019). All data was acquired at the Kempenhaeghe Center for Sleep Medicine (Heeze, the Netherlands) among individuals scheduled for an overnight PSG as part of the standard clinical routine. Trained clinicians manually scored the PSG recordings in accordance with AASM standards (Berry et al., 2017). The primary sleep diagnosis was coded according to the criteria specified in the International Classification of Sleep Disorders version 3 (ICSD-3).

The sleep recordings of healthy participants were obtained from the Healthbed database, which includes healthy adults without any known medical, psychiatric, or sleep disorders, recruited for an overnight PSG at Kempenhaeghe Center for Sleep Medicine (Heeze, the Netherlands) using the same setup as in the SOMNIA protocol (van Meulen et al., 2023).

The SOMNIA and Healthbed studies adhere to the guidelines of the Declaration of Helsinki, Good Clinical Practice, and current legal requirements. Both data collection studies were reviewed by the Maxima Medical Center medical ethical committee (Veldhoven, the Netherlands, reported under N16.074 and W17.128). The data analysis protocol was approved by the medical ethical committee of Kempenhaeghe Center for Sleep Medicine and the Philips Research Internal Committee for Biomedical Experiments.

2.1.2 Source and target datasets

A visual representation of how source and target datasets were defined, can be found in Figure 1.

We restricted our selection of source and target datasets to sleep recordings obtained from a single sleep center, thereby mitigating potential mismatches associated with inter-center variations, such as differences in sleep score training and in the PSG setup. For the source dataset, 94 sleep recordings of the F3-M2 EEG channel of the Healthbed database were used. These were all sleep recordings that were available from the Healthbed database.

For the target dataset the following selection procedure was used. We first included all 94 available Healthbed recordings. From the SOMNIA database we first included all patients with idiopathic, psychophysiological and/or chronic insomnia disorder, patients with mild-to-moderate OSA (apnea-hypopnea index (AHI) between 5 and 30, and patients with RBD. Patients with other diagnosed sleep disorder comorbidities were excluded. Since ageing is associated with an increasing variability, which can lower the generalization and thus can lower sleep staging performance (Guillot and Thorey, 2021), an age-matching algorithm was used to find the optimal grouping for the subjects of the healthy, OSA, and insomnia populations. Optimal grouping was defined as maximal subject inclusion and minimal age variance between the datasets, while keeping the age difference ≤5 years for each match. Age-matching resulted in 60 age-matched healthy, OSA, and insomnia subjects. In contrast, for the RBD target datasets no age-matching was performed. All 22 available subjects with RBD and without any other known sleep comorbidities, were included. The RBD datasets were defined as having limited data availability since, generally, a dataset size of 22 subjects is too small for sufficient training of a deep learning model including cross-validation.

Of each population, two target datasets were created comprising the F3-M2 and the F3-F4 EEG channel, resulting in a total of eight target datasets: healthy F3-M2, healthy F3-F4, OSA F3-M2, OSA F3-F4, insomnia F3-M2, insomnia F3-F4, RBD F3-M2, and RBD F3-F4.

2.1.3 Data preprocessing

A 5th order Butterworth bandpass filter between .2 and 49 Hz, and a 50 Hz notch filter were applied to the raw data to select the frequency range of interest and remove power line interference for all recordings. Afterwards, data was down sampled from 512 to 100 Hz.

2.2 Deep learning model

We used the TinySleepNet deep learning model (Supratak and Guo, 2020) for 5-stage (Wake/N1/N2/N3/REM) automated sleep staging. TinySleepNet is a computationally efficient version of the DeepSleepNet (Supratak et al., 2017). The model has shown similar or superior performance to inter-rater agreement in manual scoring and to similar automated models (Supratak et al., 2017; Korkalainen et al., 2019; Mousavi et al., 2019). Notably, using the training-from-scratch strategy, model performance has been evaluated on seven public datasets and two electrode derivations (F4-EOG/C4-EOG and Fpz-Cz). The model has shown good generalizability with performances ranging between $\kappa = .77$ and $\kappa = .82$ (Supratak and Guo, 2020). The model contains 1.3 M parameters and can process (raw) single-channel EEG data. The representational learning component has four consecutive convolutional neural network (CNN) layers interleaved with two max-pooling and drop-out layers. The sequential learning component consists of a single, unidirectional long short-term memory layer (LSTM) and a drop-out layer. For further details we refer to the original work (Supratak and Guo, 2020).

2.3 Training strategies

We tested three different training strategies; pre-training, training-from-scratch and fine-tuning, which are described in detail below. For each training strategy, a 10-fold cross-validation was employed, allowing to evaluate the performance of the model on all subjects in the different datasets. In each iteration, 80% was used as training, 10% as validation, and 10% as test set, except for fine-tuning, where the number of subjects used for training was lowered (further specified in the corresponding section). Hence, the same subject was never included in more than one set at the same time.

The best model of each cross-validation iteration was selected based on the highest accuracy and weighted average of the F1-score in the validation set. Each model was trained for a maximum of 200 iterations with early stopping if no performance improvement in the validation set was observed in the next 50 training iterations.

2.3.1 Pre-training

For pre-training, the TinySleepNet model was trained on all 94 F3-M2 EEG recordings of the healthy subjects (the source dataset). The model performance was then tested on each of the target datasets.

2.3.2 Training-from-scratch

For training-from-scratch, the TinySleepNet was trained on subjects from the target dataset, while performance was also tested on subjects from the target dataset. As for each strategy, the set split was performed within the target dataset at subject level, hence a given subject was part of either the training set, the validation set, or the hold-out test set. The above described 10-fold cross validation procedure was used to ensure each subject was represented once in the test set. This procedure was repeated for each of the eight target datasets.

2.3.3 Fine-tuning

For fine-tuning, the learning rate was lowered from $1e^{-4}$ to $1e^{-5}$ and only initial weights were loaded, without making any CNN layers non-trainable. These settings were derived from optimization experiments on separate Healthbed and SOMNIA data (i.e., not further used in the study).

Once fine-tuning parameters were optimized, the TinySleepNet model was first pre-trained on the source dataset. Afterwards a fine-tuning step was applied with a subset of the data from the target dataset. This procedure was repeated for each of the eight target datasets. The fine-tuning method was only tested in the sleep-disordered datasets (OSA, insomnia, RBD) and not on the healthy datasets, since fine-tuning a model on data on which the model was also pre-trained would be methodologically inappropriate.

TABLE 1 Distributions of the training, validation, and test sets in fine-tuning l_1, l_2, and l_3. Before fine-tuning, the model was first pre-trained on the source data. While train/validation/test set distributions were 80%/10%/10% in pre-training and training-from-scratch, the train data in fine-tuning systematically lowered. Set sizes of fine-tuning l_3 have been used in analyses where the training strategy was compared to pre-training and training-from-scratch. Distributions in absolute subject numbers with percentages in parentheses.

Name	Train/validation/test set distributions in OSA and insomnia datasets (N = 60)	Train/validation/test set distribution for RBD datasets (N = 22)
Fine-tuning l_1	6/6/6 (10%/10%/10%)	6/2/2 (30%/10%/10%)
Fine-tuning l_2	12/6/6 (20%/10%/10%)	12/2/2 (55%/10%/10%)
Fine-tuning l_3	24/6/6 (40%/10%/10%)	18/2/2 (80%/10%/10%)

For the OSA and insomnia datasets, additional training data of size n = 24 was used for fine-tuning, corresponding with 40% of the total target dataset, instead of the 80% using in pre-training and training-from-scratch. Since data availability was limited in RBD, additional training data of size n = 18 was used for fine-tuning, the maximum amount available.

To study how much additional training data from the target dataset is required for sufficient fine-tuning, we afterwards investigated three different conditions, referred to as fine-tuning l_3, fine-tuning l_2, and fine-tuning l_3. Fine-tuning l_3 corresponds with the dataset sizes described above, while additional training data was lowered to n = 12 in fine-tuning l_2, and to n = 6 in fine-tuning l_1. The absolute values and percentages used in fine-tuning are further specified in Table 1.

2.4 Statistics

All performances were analyzed using Cohen's Kappa coefficient of agreement (κ; Cohen, 1960) to compare the results of automatic scoring versus manual scoring on an epoch-per-epoch basis, for all stages (Wake, N1, N2, N3 and REM). In line with the general interpretation of the coefficient of agreement, $\kappa \geq .6$ was defined as threshold of substantial performance (Landis and Koch, 1977). For each target dataset in each training strategy, sleep stage specific accuracy was given via confusion matrices in the Supplementary Material.

Performance differences between the training strategies (pre-training, training-from-scratch, and fine-tuning) were evaluated using a within-subject design. Considering the large sample size for the age-matched datasets, and no violation of normality and equality of variance in the RBD datasets, parametric tests were performed. Specifically, we used repeated measures ANOVAs, followed by *post hoc* paired samples t-tests if significant, including a Bonferroni correction of $\alpha/3$ for multiple testing. Performance differences between the three different amounts of data in fine-tuning (l_1, l_2, and l_3) were analyzed with the same procedure.

To analyze population and channel differences, performances across applied methods were first aggregated by either population or channel. Performance was compared using a between-subject design and parametric tests, due to sufficient sample sizes. Specifically, for population, differences between the age-matched healthy, OSA and insomnia datasets were tested with a one-way ANOVAs, followed by *post hoc* independent t-tests if significant. A Bonferroni correction of $\alpha/3$ was applied to the significance threshold. For the channels, differences between the F3-M2 and F3-F4 databases were studied using an independent *t*-test.

3 Results

An overview of the demographic information and the PSG-derived sleep statistics of the datasets can be found in Table 2.

3.1 Training strategies

Figure 2 illustrates a boxplot of the performances on each target dataset using the three training strategies: pre-training, training-from-scratch, and fine-tuning. A detailed overview of the statistical differences between each training strategy can be found in Table 3.

In the healthy F3-M2 dataset, both training strategies achieved performance higher than $\kappa = .6$, the threshold of substantial agreement according to the general interpretation of the statistic (Landis and Koch, 1977). Pre-training performance ($\kappa = .84 \pm .06$) was slightly but significantly higher than training-from-scratch ($\kappa = .82 \pm .08$). In contrast, for the healthy F3-F4 dataset, pre-training ($\kappa = .53 \pm .16$) achieved an average performance lower than substantial agreement threshold, while training-from-scratch ($\kappa = .77 \pm .09$) obtained significantly higher performance.

In the OSA F3-M2 dataset, no difference between pre-training ($\kappa = .74 \pm .10$), training-from-scratch ($\kappa = .74 \pm .09$), and fine-tuning ($\kappa = .75 \pm .09$) was observed. In the OSA F3-F4 dataset, pre-training ($\kappa = .53 \pm .17$) on average performed under the threshold of substantial agreement, with significantly lower performance compared to training-from-scratch ($\kappa = .70 \pm .09$) and fine-tuning ($\kappa = .70 \pm .11$).

In the insomnia F3-M2 dataset, no difference between pre-training ($\kappa = .77 \pm .09$), training-from-scratch ($\kappa = .77 \pm .07$), and fine-tuning ($\kappa = .78 \pm .07$) was observed. In the insomnia F3-F4 dataset, pre-training ($\kappa = .56 \pm .13$) on average performed under the threshold of substantial agreement, with significantly lower performance compared to training-from-scratch ($\kappa = .73 \pm .10$) and fine-tuning ($\kappa = .73 \pm .11$).

In the RBD F3-M2 dataset, fine-tuning ($\kappa = .67 \pm .08$) significantly outperformed pre-training ($\kappa = .60 \pm .13$) and training-from-scratch ($\kappa = .57 \pm .12$), with training-from-scratch on average performing under the threshold of substantial agreement. Similarly, in the RBD F3-F4 dataset, performance was significantly higher in fine-tuning

10.3389/fphys.2023.1287342

TABLE 2 Demographic and sleep information of source and target datasets. Healthy datasets included healthy adults without any known medical, psychiatric, or sleep disorders. The sleep-disordered OSA, insomnia, and RBD datasets included adults without any other known sleep comorbidities.

	Healthy (n = 94)	Healthy (n = 60)	OSA (n = 60)	Insomnia (n = 60)	RBD (n = 22)
Dataset	Source	Target	Target	Target	Target
Age	35.9 ± 13.5	42.4 ± 11.3	44.1 ± 11.0	43.4 ± 12.9	65.3 ± 7.0
Age Range	[18–64]	[20–64]	[22–65]	[20–64]	[50–79]
Sex (m/f)	36/58	26/34	48/12	30/30	13/9
BMI (kg/m^2)	24.3 ± 3.2	25.2 ± 3.8	28.1 ± 5.4	25.6 ± 4.1	26.1 ± 4.7
TST (min)	431.7 ± 49.9	418.4 ± 52.6	414.1 ± 67.2	391.5 ± 68.6	393.8 ± 69.6
SOL (min)	11.3 ± 12.2	11.3 ± 13.7	18.5 ± 22.8	22.9 ± 17.1	13.5 ± 7.2
WASO (min)	33.7 ± 25.8	41.8 ± 28.1	42.7 ± 29.2	46.4 ± 36.6	59.6 ± 28.6
SE	87.9% ± 9.1%	85.9% ± 10.1%	82.1% ± 14.7%	80.1% ± 12.3%	79.2% ± 12.7%
Sleep Distribution	12% Wake	14% Wake	16% Wake	19% Wake	19% Wake
	8% N1	8% N1	11% N1	10% N1	13% N1
	43% N2	43% N2	43% N2	42% N2	41% N2
	19% N3	18% N3	16% N3	15% N3	13% N3
	18% REM	18% REM	14% REM	14% REM	13% REM

BMI, body mass index; AHI, apnea-hypopnea index; TST, total sleeping time; SE, sleep efficiency; WASO, wake after sleep onset; SOL, sleep onset latency.

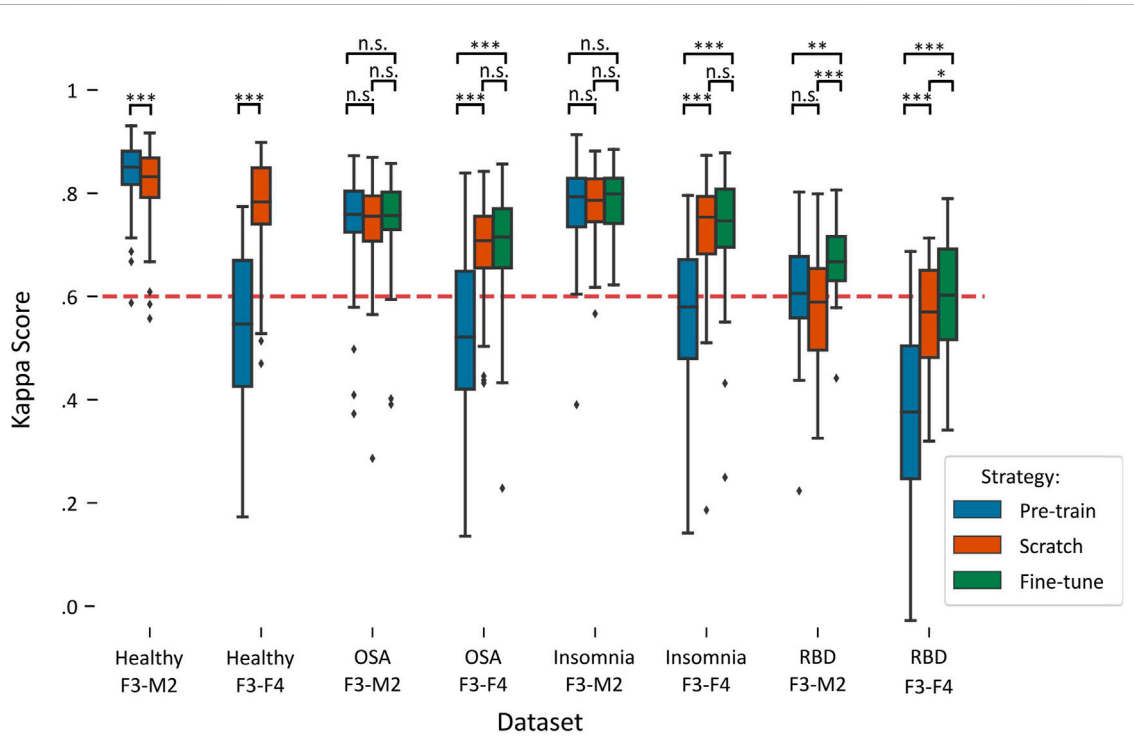

FIGURE 2
Boxplots of performances per training strategy in each target dataset. Red dashed line indicates the κ = .6 threshold of substantial agreement. Stars denote *p*-value of the test, where *, **, and *** denote $p < .017$, $p < .01$, and $p < .001$, respectively, while n.s. denotes "not significant" or $p > .017$. Training-from-scratch is abbreviated to "scratch".

($\kappa = .60 \pm .11$) than in pre-training ($\kappa = .37 \pm .18$) and training-from-scratch ($\kappa = .54 \pm .13$). Here, training-from-scratch significantly outperformed pre-training, although on average both performed under the threshold of substantial agreement.

Sleep stage specific performance results for each target dataset are detailed in Supplementary Figures S1–S4. In general, a decrease in performance across all sleep stages was observed in target datasets with lower Kappa scores. In addition, two notable observations can be made.

TABLE 3 Performances of each training strategy and statistical differences. For each dataset, average Kappa agreement and the percentage of subjects above the threshold of substantial agreement ($\kappa > .6$) is given, followed by the group statistic and *post hoc* analysis between training strategies if significant. For the healthy datasets, no group statistic is shown since comparison is between two training strategies. Training-from-scratch is abbreviated to "scratch".

Dataset	N	Method	Kappa Mean ± SD	Above substantial agreement (%)	Repeated measures ANOVA: *p*-value, effect size	Post-hoc comparison between	Paired samples t-test: *p*-value, effect size
Healthy F3-M2	60	Pre-train	$\kappa = .84 \pm .06$	98%		Pre-train & Scratch	$p < .001$, $d = .27$
		Scratch	$\kappa = .82 \pm .08$	97%			
Healthy F3-F4	60	Pre-train	$\kappa = .53 \pm .16$	35%		Pre-train & Scratch	$p < .001$, $d = 1.87$
		Scratch	$\kappa = .77 \pm .09$	93%			
OSA F3-M2	60	Pre-train	$\kappa = .74 \pm .10$	93%	$p = .26$, $\eta_p^2 = .002$		
		Scratch	$\kappa = .74 \pm .09$	95%			
		Fine-tune	$\kappa = .75 \pm .09$	95%			
OSA F3-F4	60	Pre-train	$\kappa = .53 \pm .17$	40%	$p < .001$, $\eta_p^2 = .29$	Pre-train & Scratch	$p < .001$, $d = 1.26$
		Scratch	$\kappa = .70 \pm .09$	87%		Pre-train & Fine-tune	$p < .001$, $d = 1.21$
		Fine-tune	$\kappa = .70 \pm .11$	83%		Scratch & Fine-tune	$p = .84$, $d = .02$
Insomnia F3-M2	60	Pre-train	$\kappa = .77 \pm .09$	98%	$p = .27$, $\eta_p^2 = .003$		
		Scratch	$\kappa = .77 \pm .07$	98%			
		Fine-tune	$\kappa = .78 \pm .07$	100%			
Insomnia F3-F4	60	Pre-train	$\kappa = .56 \pm .13$	43%	$p < .001$, $\eta_p^2 = .23$	Pre-train & Scratch	$p < .001$, $d = 1.39$
		Scratch	$\kappa = .73 \pm .10$	93%		Pre-train & Fine-tune	$p < .001$, $d = 1.42$
		Fine-tune	$\kappa = .73 \pm .11$	92%		Scratch & Fine-tune	$p = .53$, $d = .04$
RBD F3-M2	22	Pre-train	$\kappa = .60 \pm .13$	50%	$p = .002$, $\eta_p^2 = .11$	Pre-train & Scratch	$p = .43$, $d = .18$
		Scratch	$\kappa = .58 \pm .12$	45%		Pre-train & Fine-tune	$p = .001$, $d = .65$
		Fine-tune	$\kappa = .67 \pm .08$	91%		Scratch & Fine-tune	$p < .001$, $d = .89$
RBD F3-F4	22	Pre-train	$\kappa = .37 \pm .18$	14%	$p < .001$, $\eta_p^2 = .31$	Pre-train & Scratch	$p < .001$, $d = 1.06$
		Scratch	$\kappa = .54 \pm .13$	36%		Pre-train & Fine-tune	$p < .001$, $d = 1.51$
		Fine-tune	$\kappa = .60 \pm .11$	55%		Scratch & Fine-tune	$p = .01$, $d = .51$

First, in all RBD datasets, a relatively high degree of missed REM sleep is demonstrated, with REM recall ranging between 26.1% and 61.6% accuracy. Second, in the pre-training strategy on all F3-F4 dataset, 46.0%–61.4% of the N3 epochs were incorrectly classified as N2 sleep.

3.2 Population differences

In pre-training, no significant differences between the three populations were observed using a one-way ANOVA, $p = .09$, $\eta_p^2 = .01$.

In training-from-scratch, significant differences between the three populations were found, $p < .001$, $\eta_p^2 = .11$. Post-hoc independent t-tests showed higher performance in the aggregated healthy datasets ($\kappa = .80 \pm .09$) than in OSA ($\kappa = .72 \pm .09$), $p < .001$, $d = .86$, and insomnia ($\kappa = .75 \pm .09$), $p < .001$, $d = .50$. Also, outperformance of insomnia compared to OSA was shown, $p = .005$, $d = .37$.

In fine-tuning, higher performance in insomnia ($\kappa = .76 \pm .09$) than in OSA ($\kappa = .72 \pm .10$) was observed, $p = .007$, $d = .35$.

3.3 Channel differences

To study the effect of channel on performance, results of the three training strategies (pre-training, training-from-scratch and fine-tuning) in each age-matched population (healthy, OSA, and insomnia) were grouped by either the F3-M2 or the F3-F4 electrode channel. An independent *t*-test showed significantly higher performance using the F3-M2 channel ($\kappa = .78 \pm .09$) than using the F3-F4 channel ($\kappa = .66 \pm .16$), $p < .001$, $d = .96$.

3.4 Effect of set size in fine-tuning

Figure 3 illustrates the impact on performance when including more subjects from the target dataset for fine-tuning, referred to as fine-tuning l_1 (lowest amount), fine-tuning l_2 (medium amount), and fine-tuning l_3 (highest amount). A detailed overview of the statistical differences can be found in Table 4.

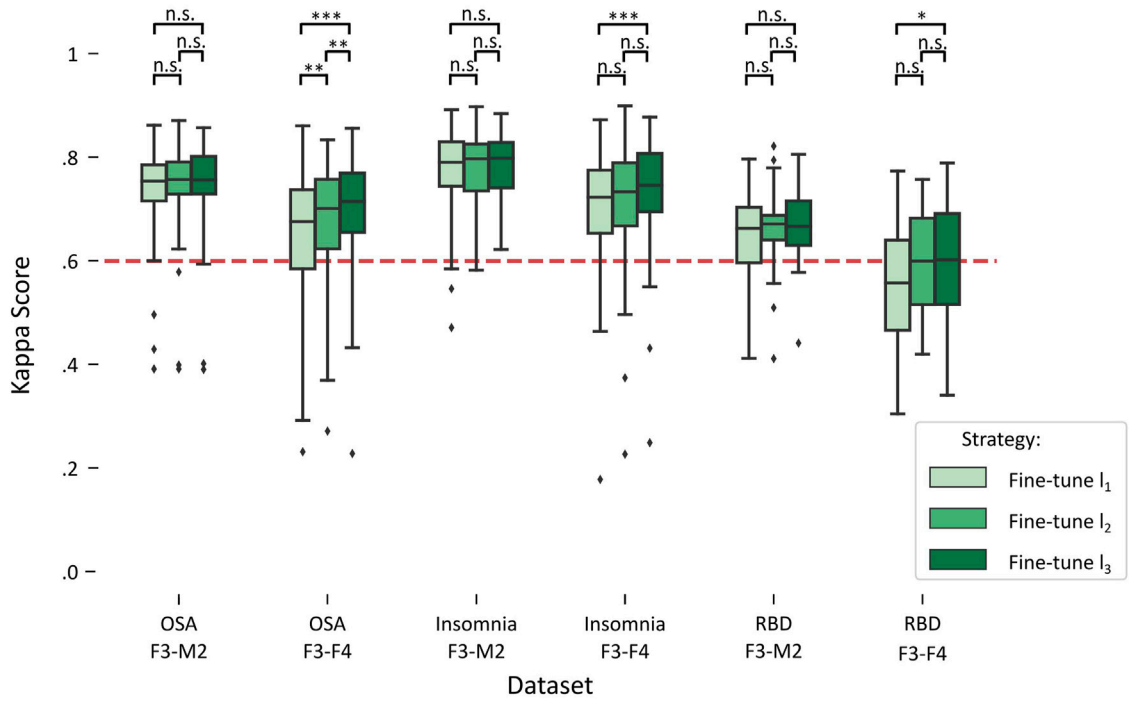

FIGURE 3
Boxplots of performances for fine-tune l_1, l_2, and l_3 in each target dataset. Red dashed line indicates the κ = .6 threshold of substantial agreement. Stars denote p-value of the test, where *, **, and *** denote $p < .017$, $p < .01$, and $p < .001$, respectively, while n.s. denotes "not significant" or $p > .017$.

For each of the datasets with the F3-M2 channel datasets (OSA F3-M2, insomnia F3-M2, and RBD F3-M2), a repeated measures ANOVA indicated no significant differences between l_1, l_2, and l_3. In contrast, significant differences were observed for each dataset with the F3-F4 channel (OSA F3-F4, insomnia F3-F4, and RBD F3-F4). Notably, in OSA and insomnia F3-F4 datasets, average performance above substantial agreement is reached in l_1. In RBD F3-F4, substantial agreement is realized in l_2.

3.5 Fine-tuning dynamics

The loss function for fine-tuning showed that optimal loss was achieved after 5 to 35 training iterations, depending on the target dataset and the number of subjects used for fine-tuning, after which overfitting occurred (see Figure 4A). A small number of subjects used for fine-tuning led to faster loss increase. In contrast, a logarithmic decrease in loss function was observed in training-from-scratch.

The model's accuracy exhibited similar differences between the training-from-scratch and fine-tuning methods. While in training-from-scratch, accuracy followed a logarithmic increase, optimal accuracy in fine-tuning was reached early, stabilized, and potentially decreased slightly as training continued (see Figure 4B).

For fine-tuning, learning rate was set to $1e^{-5}$ and only initial weights were loaded, without making any CNN layers non-trainable. Higher learning rates and an increasing number of non-trainable CNN layers exhibited similar loss functions, albeit faster overfitting occurred. Also, with higher learning rates, accuracy was lower and plateaued earlier.

4 Discussion

In this study we evaluated three strategies (pre-training, training-from-scratch, fine-tuning) for training an automated single-channel EEG sleep staging model when mismatches between the source and target dataset are present. Each strategy was tested on a total of eight target datasets, comprising of healthy subjects and patients with OSA, insomnia, or RBD; combined with the F3-M2 and F3-F4 EEG channels.

4.1 Model performance

First of all, our results illustrate the strong performance of the TinySleepNet automated sleep staging model (Supratak and Guo, 2020). Our Healthbed dataset yielded slightly higher agreement (κ = .84 and κ = .82 in pre-trained and trained-from-scratch healthy F3-M2 datasets, respectively) compared to previously reported performances on the sleep-EDF (κ = .77–.80) and MASS (κ = .77–.82) datasets (Supratak and Guo, 2020). Other automated single-channel EEG models have shown similar performance for sleep-EDF dataset (κ = .81; Phan et al., 2019) and MASS dataset (κ = .78–82; Phan et al., 2019; Mousavi et al., 2019; Seo et al., 2020).

TABLE 4 Fine-tune l_1, l_2, and l_3 performances and statistical differences. For each dataset, average Kappa agreement and the percentage of subjects above the threshold of substantial agreement ($\kappa > .6$) is given, followed by the group statistic and *post hoc* analysis between fine-tune l_1, l_2, and l_3 if significant.

Dataset	N	Fine-tune size	Kappa Mean ± SD	Substantial agreement (%)	Repeated measures ANOVA: *p*-value, effect size	Post-hoc comparison between	Paired samples t-test: *p*-value, effect size
OSA F3-M2	60	l_1	$\kappa = .74 \pm .09$	95%	$p = .17$, $\eta_p^2 = .003$		
		l_2	$\kappa = .74 \pm .09$	95%			
		l_3	$\kappa = .75 \pm .09$	95%			
OSA F3-F4	60	l_1	$\kappa = .65 \pm .13$	73%	$p < .001$, $\eta_p^2 = .03$	l_1 & l_2	$p = .003$, $d = .20$
		l_2	$\kappa = .67 \pm .12$	82%		l_1 & l_3	$p < .001$, $d = .39$
		l_3	$\kappa = .70 \pm .11$	83%		l_2 & l_3	$p = .007$, $d = .20$
Insomnia F3-M2	60	l_1	$\kappa = .77 \pm .09$	95%	$p = .05$, $\eta_p^2 = .005$		
		l_2	$\kappa = .77 \pm .08$	97%			
		l_3	$\kappa = .78 \pm .07$	100%			
Insomnia F3-F4	60	l_1	$\kappa = .70 \pm .11$	85%	$p < .001$, $\eta_p^2 = .01$	l_1 & l_2	$p = .20$, $d = .06$
		l_2	$\kappa = .71 \pm .12$	85%		l_1 & l_3	$p < .001$, $d = .28$
		l_3	$\kappa = .73 \pm .11$	92%		l_2 & l_3	$p < .001$, $d = .21$
RBD F3-M2	22	l_1	$\kappa = .65 \pm .09$	73%	$p = .28$, $\eta_p^2 = .009$		
		l_2	$\kappa = .66 \pm .09$	77%			
		l_3	$\kappa = .67 \pm .08$	91%			
RBD F3-F4	22	l_1	$\kappa = .56 \pm .12$	41%	$p = .02$, $\eta_p^2 = .03$	l_1 & l_2	$p = .04$, $d = .39$
		l_2	$\kappa = .60 \pm .10$	50%		l_1 & l_3	$p = .01$, $d = .38$
		l_3	$\kappa = .60 \pm .11$	55%		l_2 & l_3	$p = .99$, $d = .00$

The average performance in the OSA ($\kappa = .74$) and insomnia ($\kappa = .77$) datasets on the F3-M2 channel were lower compared to healthy individuals, with medium to large effect sizes ranging between $d = .35$ and $d = .86$ (Cohen, 1960). Since datasets were age-matched, sampled from the same database, and other known sleep comorbidities were excluded, underperformance with respect to healthy subjects can potentially be attributed to the sleep characteristics of the disorders. An earlier study with frontal (wearable) single-channel EEG automated sleep staging reported underperformance specifically in the differentiation of N1 sleep due to a lack of measured occipital activity (Lucey et al., 2016). This can particularly affect performance in OSA and insomnia due to the increased sleep fragmentation. Another possible explanation is that the dataset sizes used for training in this study have been too small to sufficiently capture the heterogeneity of sleep within in these sleep-disordered populations.

In RBD, agreement levels of $\kappa = .67$ (F3-M2 channel) and $\kappa = .60$ (F3-F4 channel) were obtained using fine-tuning, an average increase of $\kappa = .15$ compared with pre-training and an average increase of $\kappa = .08$ compared with training-from-scratch. Despite reaching or exceeding substantial agreement according to the general interpretation of the Kappa statistic (Landis and Koch, 1977), performance still falls below that obtained for the other tested sleep disorders. Although, to the best of our knowledge, studies evaluating human inter-rater agreement for sleep scoring in RBD are lacking, relatively low agreement has been reported in presence of Parkinson's disease ($\kappa = .61$; Danker-Hopfe et al., 2004), a neurodegenerative disease strongly associated with RBD (Schenck, Boeve and Mahowald, 2013). Automated sleep scoring performance in RBD was higher than previously described in a cohort of 22 RBD subjects, with performances of $\kappa = .45$ before and $\kappa = .56$ after subject-specific fine-tuning (i.e., fine-tuning on each patient's first night PSG, tested on the second night; Andreotti et al., 2018).

4.2 Training strategies

Several conclusions can be drawn from the study regarding the preferred training strategy in presence of population mismatches. In the age-matched OSA F3-M2 and insomnia F3-M2 datasets, no difference in performance between the three training strategies was observed, implying it is inconsequential whether the model was trained on the healthy F3-M2 source dataset (pre-training), target dataset (training-from-scratch) or both (fine-tuning). These results suggest a similarity in data characteristics between the healthy source and the OSA and insomnia target when datasets are solely mismatched for these sleep disorders, without the presence of age-related population mismatches. Hence, sleep of OSA and insomnia patients is considered abnormal, and abnormal

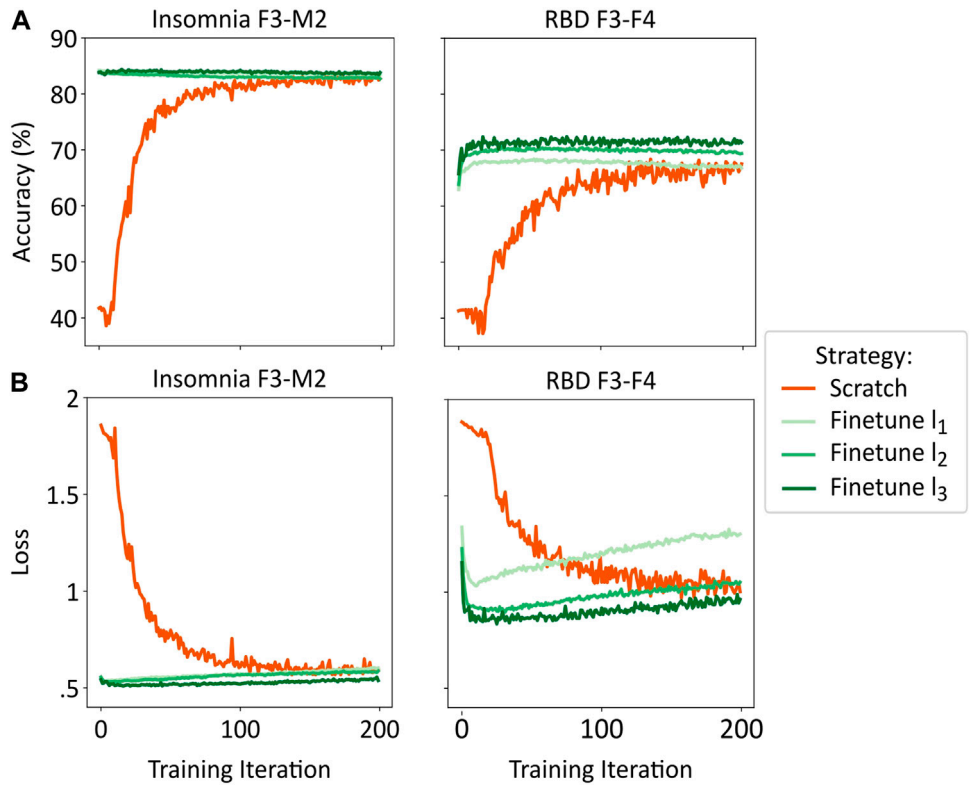

FIGURE 4
Examples of accuracy **(A)** and loss function **(B)** for training-from-scratch (abbreviated to "scratch"), fine-tuning l_1, l_2, and l_3 for the best (insomnia F3-M2, left) and worst (RBD F3-F4, right) performing validation sets. Averaged across the ten models of the cross-validation.

characteristics including increased sleep fragmentation are likely decreasing sleep staging performance. However, these are not characteristics that cause population mismatches and thus require targeted training.

In contrast, for the RBD F3-M2 dataset, differences between the training strategies were found. Underperformance in pre-training suggests specific characteristics in RBD that the model needs to train on, either caused by the older age of RBD patients and/or by sleep characteristics inherent to RBD. Previously reported lower agreement in RBD (κ = .54) in comparison to age-matched healthy subjects (κ = .73) indicate likely not all underperformance in RBD is age-related (Cooray et al., 2019). Notably, for each of the training strategies, the missed classification of REM sleep is especially problematic, but lower agreement in RBD is observed across all sleep stages (see Supplementary Figure S4). These findings are in line with earlier studies (Andreotti et al., 2018; Cooray et al., 2019). We hypothesize that the frontopolar EEG channel can capture the characteristic elevated muscle activity during REM sleep in RBD, complicating the correct classification of REM sleep when muscle tone is present. Additionally, the generally lower performance could be attributed to microstructural changes and decreased sleep stability (Christensen et al., 2016; Cesari et al., 2021) that are associated with RBD. However, future research on RBD sleep staging is needed for the further characterization of the lower performance. Underperformance in training-from-scratch can likely be explained because insufficient RBD data is available to obtain robust performance. This practical challenge can emerge given that RBD is a less prevalent sleep disorder. Hence, when differences in data characteristics are present and target data availability is limited, fine-tuning is the preferred training strategy. Medium to large effect sizes (d = .65–.87; Cohen, 1960) emphasize the strong advantages of fine-tuning for RBD data.

Results in the F3-F4 target datasets suggest that, when channel mismatches with the source dataset are present, pre-training (on the F3-M2 channel) is not sufficient. Training on the target is necessary through either training-from-scratch or fine-tuning. Notably, fine-tuning delivers similar classification performance to training-from-scratch, while requiring only half the amount of target data. The threshold of substantial agreement is already exceeded with fine-tuning on 12 (6 train +6 validation) target subjects in OSA and insomnia. Again, for the RBD F3-F4 dataset specifically, fine-tuning is preferred since the dataset is too small for training-from-scratch. Here, substantial agreement is reached with 14 (12 train +2 validation) target subjects.

4.3 Transfer learning

Making several layers of a model non-trainable is perceived as a common method for transfer learning (e.g., Shin et al., 2016;

Tajbakhsh et al., 2016). However, for TinySleepNet, we achieved best performance by fine-tuning the full model, hence making no layers non-trainable and only loading in the weights of the pre-trained model. This irregularity is possibly explained by the low computational costs and straight-forward design of the model, making it more flexible to tune to new data characteristics. Furthermore, early stopping and a lowered learning rate (from $1e^{-4}$ to $1e^{-5}$) were necessary to prevent the model from overfitting on the target data, especially when larger mismatches with the source dataset were present and/or when target data availability was limited. Optimal model performance in fine-tuning was reached after 5 to 35 training iterations, while requiring at least 80 training iterations when using pre-training or training-from-scratch as training strategy, emphasizing the lower computational costs for transfer learning. Our findings are similar to observations for fine-tuning in channel mismatches using the SeqSleepNet model, where full fine-tuning of the model showed superior sleep staging performance for transferring from the C4-A1 channel in the MASS dataset to the Fpz-Cz channel in the sleep-EDF dataset (Phan et al., 2019). However, in the DeepSleepNet, an architecture more similar to the TinySleepNet, fine-tuning only the softmax layer for channel changes seems to yield highest performance (Phan et al., 2019). These results suggest fine-tuning is a model and mismatch specific process, and no one-size-fits-all strategy is (yet) known.

4.4 Single-channel (wearable) EEG

There is an ongoing debate on the preferred channel for single-channel EEG automated sleep staging, including the suggested use of Fpz-Cz, Pz-Oz (see, e.g., Supratak et al., 2017), and Fp1-Fp2 (Radha et al., 2014). In this study, we chose to include leads present in the PSG setup which are as close as possible to the frontopolar locations often used in wearable EEG. For the development of these technologies, it is valuable to understand how public PSG databases can be leveraged through pre-training and fine-tuning. Especially since validation of wearable (single-channel) EEG is often limited to small and homogeneous recordings of healthy subjects because of costs, time, and ethical regulations (Finan et al., 2016; Garcia-Molina et al., 2018; Arnal et al., 2020). Also, in contrast to similar models (Supratak et al., 2017; Korkalainen et al., 2019; Mousavi et al., 2019), TinySleepNet can potentially be implemented for real-time sleep stage classification in wearable EEG, due to the model's architecture. With the fine-tuning approach used in this study, we have shown that such a sleep staging model can also be trained for less prevalent disorders with specific characteristics, such as RBD. Furthermore, the use of wearable EEG for prolonged monitoring and the collection of longitudinal data can enable new possibilities to develop personalized sleep staging models, wherein (subject-specific) fine-tuning can play a critical role (Andreotti et al., 2018; Phan et al., 2020).

Our results consistently show underperformance of the F3-F4 channel (κ = .66) when compared with the F3-M2 channel (κ = .78) with a large effect (d = .96; Cohen, 2013), possibly because the F3 and F4 locations share more signal characteristics of interest, which are subtracted when re-referencing the channels. Two studies have been performed in the sleep-disordered population using frontal channels in wearable EEG, both showing κ = .67 agreement levels (Lucey et al., 2016; Levendowski et al., 2017). It should be noted that also lower inter-rater reliability for manual scoring of the recordings is reported (κ = .69 in Garcia-Molina et al., 2019; κ = .70 in Levendowski et al., 2017; κ = .78 in Popovic et al., 2014), suggesting an upper limit to the sleep staging information extracted from wearable frontal electrodes.

Furthermore, it is imperative to acknowledge that sleep characteristics may exhibit variation across channels, and thus can cause channel mismatches. A prime example is the reduced N3 classification performance observed in the F3-F4 datasets when the model is pre-trained on F3-M2, in contrast to the results obtained with training-from-scratch or fine-tuning on F3-F4 (see Supplementary Figures S1–S4). These findings suggest a distinct manifestation of the characteristic N3 slow wave sleep in the two channels.

4.5 Limitations

Some limitations of the current study should be considered. First, while typically hundreds of sleep recordings are required to reach expert-level performance using deep learning (Biswal et al., 2018; Guillot and Thorey, 2021; Perslev et al., 2021), in the current study, 94 recordings for pre-training, and 60 recordings for training-from-scratch, have been used. Potentially, larger sample sizes can achieve better model training and higher performance, especially in the sleep-disordered populations which can be characterized by increased variability. However, the goal of the current study was to investigate different training strategies in the presence of data mismatches. Hence, only data from one sleep center was used, allowing to isolate population and channel mismatches, and avoiding potential additional mismatches from PSG setup differences or differences in manual sleep score training.

Second, TinySleepNet discerns from other related automated models for the potential implementation into wearable EEG due to its low computation costs and unidirectional LSTM, which allows for real-time classification. However, implementation here is only theorized, and should be tested in future research. Since the current study is limited to gold-standard EEG recordings only, it remains unknown how fine-tuning performance is affected when recordings from dry electrodes with higher signal-to-noise ratios are used as target.

Last, one should consider the generalizability of this study carefully. Although other single-channel EEG deep learning classification models (Tsinalis et al., 2016; Supratak et al., 2017; Korkalainen et al., 2019; Mousavi et al., 2019; Phan et al., 2019; Seo et al., 2020; Guillot and Thorey, 2021) have related architectures, their differences potentially result in model-specific transfer learning dynamics (Phan et al., 2019). Similarly, the preferred applied methods that have been investigated in this work are possibly specific to the database, available amount of data, and mismatches between the source and target datasets.

5 Conclusion

In this work, we have shown that the preferred training strategy for automated single-channel automated sleep stager depends on the presence of data mismatches, type of mismatch, and the availability of data. OSA and insomnia target datasets show no population mismatches when the model is pre-trained on healthy individuals. In contrast, RBD sleep recordings contain characteristics, either inherent to the pathology or age-related, which demand targeted model training. Targeted training is also needed when source and target datasets differences cause channel mismatches. In the presence of these data mismatches, fine-tuning can yield similar to superior performance than training-from-scratch, with a significantly reduced dataset size.

Data availability statement

The Healthbed (van Meulen et al., 2023) and SOMNIA (van Gilst et al., 2019) datasets, collected by Kempenhaeghe Center for Sleep Medicine. Further information on data availability can be obtained in the original works. The TinySleepNet model (Supratak and Guo, 2020) is available on https://github.com/akaraspt/tinysleepnet.

Ethics statement

This studies involving humans used the retrospective data of the HealthBed and SOMNIA studies, acquired by Kempenhaeghe Center for Sleep Medicine. The data collection protocol of these studies was approved by the Maxima Medical Center medical ethical committee (Veldhoven, the Netherlands), reported under N16.074 (HealthBed) and W17.128 (SOMNIA). All participants provided informed consent for participation and reuse of their data for secondary data analysis. The data analysis protocol for the current study was approved by the institutional review board of Kempenhaeghe Center for Sleep Medicine as well as the Philips Research Internal Committee for Biomedical Experiments.

Author contributions

JA: Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft. DE: Conceptualization, Methodology, Supervision, Writing–review and editing. PF: Conceptualization, Methodology, Supervision, Writing–review and editing. FM: Data curation, Writing–review and editing. SO: Conceptualization, Methodology, Supervision, Writing–review and editing. MG: Conceptualization, Methodology, Supervision, Writing–review and editing. EP: Conceptualization, Methodology, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was performed within the IMPULS framework of the Eindhoven MedTech Innovation Center (e/MTIC, incorporating Eindhoven University of Technology, Philips Research, and Sleep Medicine Center Kempenhaeghe), including a PPS-supplement from the Dutch Ministry of Economic Affairs and Climate Policy.

Acknowledgments

We thank the Advanced Sleep Monitoring Lab at the Technical University of Eindhoven, and in particular Iris Huijben, for invaluable discussion and comments. We thank Kempenhaeghe Center for Sleep Medicine for data acquisition.

Conflict of interest

At the time of writing, JA, DE, and PF were employed and/or affiliated with Royal Philips, a commercial company and manufacturer of consumer and medical electronic devices, commercializing products in the area of sleep diagnostics and sleep therapy. Philips had no role in the study design, decision to publish or preparation of the manuscript. SO received an unrestricted research grant from UCB Pharma and participated in advisory boards for UCB Pharma, Takeda, Jazz Pharmaceuticals and Bioprojet, all paid to institution and all unrelated to the present work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2023.1287342/full#supplementary-material

References

Andreotti, F., Phan, H., Cooray, N., Lo, C., Hu, M. T., and De Vos, M. (2018). "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in 2018 40th annual international conference of the IEEE engineering in medicine and biology society July 17-21, 2018, China, (IEEE), 171–174. doi:10.1109/EMBC.2018.8512214

Arnal, P. J., Thorey, V., Debellemaniere, E., Ballard, M. E., Bou Hernandez, A., Guillot, A., et al. (2020). The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep* 43 (11), zsaa097. doi:10.1093/sleep/zsaa097

Baglioni, C., Regen, W., Teghen, A., Spiegelhalder, K., Feige, B., Nissen, C., et al. (2014). Sleep changes in the disorder of insomnia: a meta-analysis of polysomnographic studies. *Sleep. Med. Rev.* 18 (3), 195–213. doi:10.1016/j.smrv.2013.04.001

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., et al. (2017). AASM scoring manual updates for 2017 (version 2.4). *J. Clin. Sleep Med.* 13 (5), 665–666. doi:10.5664/jcsm.6576

Biswal, S., Sun, H., Goparaju, B., Westover, M. B., Sun, J., and Bianchi, M. T. (2018). Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inf. Assoc.* 25 (12), 1643–1650. doi:10.1093/jamia/ocy131

Boeve, B. F., Silber, M. H., Saper, C. B., Ferman, T. J., Dickson, D. W., Parisi, J. E., et al. (2007). Pathophysiology of REM sleep behaviour disorder and relevance to neurodegenerative disease. *Brain* 130 (11), 2770–2788. doi:10.1093/brain/awm056

Bresch, E., Großekathöfer, U., and Garcia-Molina, G. (2018). Recurrent deep neural networks for real-time sleep stage classification from single channel EEG. *Front. Comput. Neurosci.* 12, 85. doi:10.3389/fncom.2018.00085

Cesari, M., Christensen, J. A., Muntean, M. L., Mollenhauer, B., Sixel-Döring, F., Sorensen, H. B., et al. (2021). A data-driven system to identify REM sleep behavior disorder and to predict its progression from the prodromal stage in Parkinson's disease. *Sleep. Med.* 77, 238–248. doi:10.1016/j.sleep.2020.04.010

Christensen, J. A. E., Jennum, P., Koch, H., Frandsen, R., Zoetmulder, M., Arvastson, L., et al. (2016). Sleep stability and transitions in patients with idiopathic REM sleep behavior disorder and patients with Parkinson's disease. *Clin. Neurophysiol.* 127 (1), 537–543. doi:10.1016/j.clinph.2015.03.006

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20 (1), 37–46. doi:10.1177/001316446002000104

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. USA: Academic Press.

Cooray, N., Andreotti, F., Lo, C., Symmonds, M., Hu, M. T., and De Vos, M. (2019). Detection of REM sleep behaviour disorder by automated polysomnography analysis. *Clin. Neurophysiol.* 130 (4), 505–514. doi:10.1016/j.clinph.2019.01.011

Danker-hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., et al. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen and Kales and the new AASM standard. *J. sleep Res.* 18 (1), 74–84. doi:10.1111/j.1365-2869.2008.00700.x

Danker-Hopfe, H., Kunz, D., Gruber, G., Klösch, G., Lorenzo, J. L., Himanen, S. L., et al. (2004). Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J. sleep Res.* 13 (1), 63–69. doi:10.1046/j.1365-2869.2003.00375.x

Finan, P. H., Richards, J. M., Gamaldo, C. E., Han, D., Leoutsakos, J. M., Salas, R., et al. (2016). Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers. *J. Clin. Sleep Med.* 12 (11), 1443–1451. doi:10.5664/jcsm.6262

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., et al. (2019). Automated sleep scoring: a review of the latest approaches. *Sleep. Med. Rev.* 48, 101204. doi:10.1016/j.smrv.2019.07.007

Garcia-Molina, G., Tsoneva, T., Jasko, J., Steele, B., Aquino, A., Baher, K., et al. (2018). Closed-loop system to enhance slow-wave activity. *J. neural Eng.* 15 (6), 066018. doi:10.1088/1741-2552/aae18f

Garcia-Molina, G., Tsoneva, T., Neff, A., Salazar, J., Bresch, E., Grossekathofer, U., and Aquino, A. (2019). "Hybrid in-phase and continuous auditory stimulation significantly enhances slow wave activity during sleep," in 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC) 23-27 July 2019, USA, (IEEE).

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101 (23), e215–e220. doi:10.1161/01.CIR.101.23.e215

Guillot, A., and Thorey, V. (2021). RobustSleepNet: transfer learning for automated sleep staging at scale. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 29, 1441–1451. doi:10.1109/TNSRE.2021.3098968

He, Z., Tang, M., Wang, P., Du, L., Chen, X., Cheng, G., et al. (2023). Cross-scenario automatic sleep stage classification using transfer learning and single-channel EEG. *Biomed. Signal Process. Control* 81, 104501. doi:10.1016/j.bspc.2022.104501

Korkalainen, H., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., Duce, B., et al. (2019). Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE J. Biomed. health Inf.* 24 (7), 2073–2081. doi:10.1109/JBHI.2019.2951346

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics* 33, 159–174. doi:10.2307/2529310

Levendowski, D. J., Ferini-Strambi, L., Gamaldo, C., Cetel, M., Rosenberg, R., and Westbrook, P. R. (2017). The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *J. Clin. Sleep Med.* 13 (6), 791–803. doi:10.5664/jcsm.6618

Lopez-Gordo, M. A., Sanchez-Morillo, D., and Valle, F. P. (2014). Dry EEG electrodes. *Sensors* 14 (7), 12847–12870. doi:10.3390/s140712847

Lucey, B. P., Mcleland, J. S., Toedebusch, C. D., Boyd, J., Morris, J. C., Landsness, E. C., et al. (2016). Comparison of a single-channel EEG sleep study to polysomnography. *J. sleep Res.* 25 (6), 625–635. doi:10.1111/jsr.12417

Mannarino, M. R., Di Filippo, F., and Pirro, M. (2012). Obstructive sleep apnea syndrome. *Eur. J. Intern. Med.* 23 (7), 586–593. doi:10.1016/j.ejim.2012.05.013

Mikkelsen, K. B., Villadsen, D. B., Otto, M., and Kidmose, P. (2017). Automatic sleep staging using ear-EEG. *Biomed. Eng. online* 16 (1), 111–115. doi:10.1186/s12938-017-0400-5

Mousavi, S., Afghah, F., and Acharya, U. R. (2019). SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one* 14 (5), e0216456. doi:10.1371/journal.pone.0216456

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. data Eng.* 22 (10), 1345–1359. doi:10.1109/TKDE.2009.191

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., and Igel, C. (2021). U-Sleep: resilient high-frequency sleep staging. *NPJ Digit. Med.* 4 (1), 72. doi:10.1038/s41746-021-00440-5

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. (2019). SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 27 (3), 400–410. doi:10.1109/TNSRE.2019.2896659

Phan, H., Chén, O. Y., Koch, P., Lu, Z., McLoughlin, I., Mertins, A., et al. (2020). Towards more accurate automatic sleep staging via deep transfer learning. *IEEE Trans. Biomed. Eng.* 68 (6), 1787–1798. doi:10.1109/TBME.2020.3020381

Phan, H., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. (2019). "Deep transfer learning for single-channel automatic sleep staging with channel mismatch," in 2019 27th European signal processing conference (EUSIPCO) 2-6 Sept. 2019, USA, (IEEE).

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., Kidmose, P., et al. (2022). Personalized automatic sleep staging with single-night data: a pilot study with Kullback-Leibler divergence regularization. *Physiol. Meas.* 41, 064004. doi:10.1088/1361-6579/ab921e

Popovic, D., Khoo, M., and Westbrook, P. (2014). Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. *J. sleep Res.* 23 (2), 211–221. doi:10.1111/jsr.12105

Radha, M., Garcia-Molina, G., Poel, M., and Tononi, G. (2014). "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal," in 2014 36th annual international conference of the IEEE engineering in medicine and biology society 26-30 Aug. 2014, USA, (IEEE).

Rosenberg, R. S., and Van Hout, S. (2013). The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J. Clin. sleep Med.* 9 (1), 81–87. doi:10.5664/jcsm.2350

Sateia, M. J. (2014). International classification of sleep disorders-third edition: highlights and modifications. *Chest* 146 (5), 1387–1394. doi:10.1378/chest.14-0970

Schenck, C. H., Boeve, B. F., and Mahowald, M. W. (2013). Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series. *Sleep. Med.* 14 (8), 744–748. doi:10.1016/j.sleep.2012.10.009

Seo, H., Back, S., Lee, S., Park, D., Kim, T., and Lee, K. (2020). Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomed. signal Process. control* 61, 102037. doi:10.1016/j.bspc.2020.102037

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. imaging* 35 (5), 1285–1298. doi:10.1109/TMI.2016.2528162

Singh, J., Badr, M. S., Diebert, W., Epstein, L., Hwang, D., Karres, V., et al. (2015). American Academy of Sleep Medicine (AASM) position paper for the use of telemedicine for the diagnosis and treatment of sleep disorders: an American Academy of Sleep Medicine Position Paper. *J. Clin. Sleep Med.* 11 (10), 1187–1198. doi:10.5664/jcsm.5098

Sterr, A., Ebajemito, J. K., Mikkelsen, K. B., Bonmati-Carrion, M. A., Santhi, N., Della Monica, C., et al. (2018). Sleep EEG derived from behind-the-ear electrodes (cEEGrid) compared to standard polysomnography: a proof of concept study. *Front. Hum. Neurosci.* 12, 452. doi:10.3389/fnhum.2018.00452

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 25 (11), 1998–2008. doi:10.1109/TNSRE.2017.2721116

Supratak, A., and Guo, Y. (2020). "TinySleepNet: an efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in 2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC) 20-24 July 2020, China, (IEEE).

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., et al. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. imaging* 35 (5), 1299–1312. doi:10.1109/TMI.2016.2535302

Troester, M., Quan, S., and Berry, R. (2023). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*.

Tsinalis, O., Matthews, P. M., Guo, Y., and Zafeiriou, S. (2016). *Automatic sleep stage scoring with single-channel EEG using convolutional neural networks*.

van Gilst, M. M., van Dijk, J. P., Krijn, R., Hoondert, B., Fonseca, P., van Sloun, R. J., et al. (2019). Protocol of the SOMNIA project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring. *BMJ open* 9 (11), e030996. doi:10.1136/bmjopen-2019-030996

van Meulen, F. B., Grassi, A., van den Heuvel, L., Overeem, S., van Gilst, M. M., van Dijk, J. P., et al. (2023). Contactless camera-based sleep staging: the HealthBed study. *Bioengineering* 10 (1), 109. doi:10.3390/bioengineering10010109

Zhang, G. Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., et al. (2018). The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inf. Assoc.* 25 (10), 1351–1358. doi:10.1093/jamia/ocy064