

Article

Transfer Learning for Automatic Sleep Staging Using a Pre-Gelled Electrode Grid

Fabian A. Radke , Carlos F. da Silva Souto , Wiebke Pätzold  and Karen Insa Wolf 

Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, 26129 Oldenburg, Germany

* Correspondence: fabian.alexander.radke@idmt.fraunhofer.de

Abstract: Novel sensor solutions for sleep monitoring at home could alleviate bottlenecks in sleep medical care as well as enable selective or continuous observation over long periods of time and contribute to new insights in sleep medicine and beyond. Since especially in the latter case the sensor data differ strongly in signal, number and extent of sensors from the classical polysomnography (PSG) sensor technology, an automatic evaluation is essential for the application. However, the training of an automatic algorithm is complicated by the fact that the development phase of the new sensor technology, extensive comparative measurements with standardized reference systems, is often not possible and therefore only small datasets are available. In order to circumvent high system-specific training data requirements, we employ pre-training on large datasets with finetuning on small datasets of new sensor technology to enable automatic sleep phase detection for small test series. By pre-training on publicly available PSG datasets and finetuning on 12 nights recorded with new sensor technology based on a pre-gelled electrode grid to capture electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG), an F_1 score across all sleep phases of 0.81 is achieved (wake 0.84, N1 0.62, N2 0.81, N3 0.87, REM 0.88), using only EEG and EOG. The analysis additionally considers the spatial distribution of the channels and an approach to approximate classical electrode positions based on specific linear combinations of the new sensor grid channels.

Keywords: sleep staging; electrode grid; EEG; machine learning; transfer learning; home monitoring



Citation: Radke, F.A.; da Silva Souto, C.F.; Pätzold, W.; Wolf, K.I. Transfer Learning for Automatic Sleep Staging Using a Pre-Gelled Electrode Grid. *Diagnostics* **2024**, *14*, 909. <https://doi.org/10.3390/diagnostics14090909>

Academic Editor: Juan Rafael Orozco-Arroyave

Received: 29 February 2024

Revised: 18 April 2024

Accepted: 24 April 2024

Published: 26 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sleep stage classification is widely used for diagnostics for sleep-related disorders. Polysomnography (PSG) is used to diagnose such disorders. For PSG, patients spend one night in a sleep laboratory where electroencephalography (EEG) and other modalities are recorded. While disorders like sleep apnea can be diagnosed that way, there are multiple sleep disorders or sleep-related sleep disorders like insomnia that demand multiple recordings over a longer time span [1]. Therefore, it is not practical to make recordings for these patients, but it is more useful to monitor sleep at home, which requires self-administered EEG systems. Such systems also enable a long-term monitoring for detection of sleep-related disorders like Parkinson's and Alzheimer's disease [2]. Additionally, the PSG introduces inconvenience for the patients, as they need to come into a sleep laboratory and the measurement setup can have a bad influence on the patients' sleep. Certainly, the unfamiliar environment further worsens the sleep during the recording, known as the first night effect [3]. The gold standard for staging the recording is manual scoring by sleep clinicians according to AASM (American Academy of Sleep Medicine) or Rechtschaffen and Kales [4,5] guidelines, recorded at a sleep laboratory. However, this scoring process is time-consuming and prone to inter-rater variability [6]. To overcome these issues and improve the care of patient suffering sleep disorders, two solution approaches can be identified: addressing the environment issue, we propose the usage of an EEG setup for recording, that is self-applicable and allows EEG sleep monitoring at home at a high quality.

Different mobile solutions have been and are being developed to achieve an easy-to-use home monitoring system: the wet electrode grids cEEGrid [7,8] and trEEGrid [9], dry electrode solution like Dreem [10,11] or the pre-gelled HomeSleepTest [12]. The trEEGrid covers a wider range of electrode positions than the cEEGrid, see Figure 1, and it aims to be less intrusive and easier to use due to pre-gelled electrodes. The gel is used to obtain a better signal than with dry electrodes while at the same time being more comfortable to wear. It is shown based on a conceptual prototype of the trEEGrid that high quality hypnograms can be derived [9].

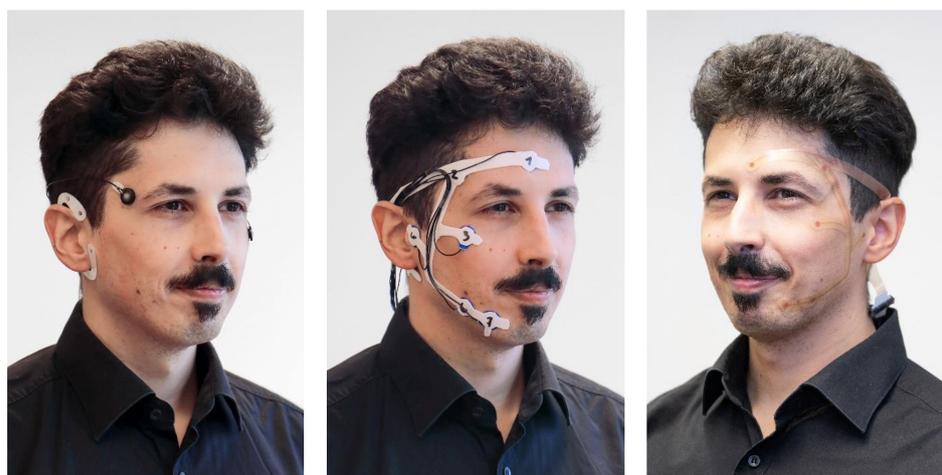


Figure 1. Development steps of the self-applicable, pre-gelled trEEGrid. **Left:** cEEGrid + EOG [13]. **Middle:** foam trEEGrid with labeled channels [9], used for recordings in this study. **Right:** trEEGrid prototype on a flexible circuit board. ©Fraunhofer IDMT/Anika Bödecker. Figure and caption are adapted from [9].

As in many other fields of health care, the application of artificial intelligence has become feasible for automatic sleep staging. Various methods using deep neural networks have achieved annotation quality in the automatic sleep staging task that is comparable to the quality achieved by humans [14]. Most popular are sequence-to-sequence architectures made of epoch and sequence encoder structure [15–17], often consisting of recurrent neural networks (RNNs) such as long-short-term memory (LSTM) [18] and gated recurrent units (GRUs) [19]. Also fully convolutional networks [20], U-nets [21,22] and transformer architectures [23] have been proposed for automatic sleep staging. These methods use large databases, usually recorded with a classical PSG setup, of which most models utilize only a small number of channels. The use of these large databases is not suitable or possible in the case of specialized EEG setups, as the electrode setup differs both in the number of available channels and electrode positioning. This creates a domain mismatch. Application of transfer learning is a method for addressing the issue of domain mismatch in various fields, including health care, when artificial intelligence is employed [24]. This is also the case for automatic sleep staging, where different approaches to transfer learning were investigated [25]. To accommodate for channel mismatch between datasets for neural networks that predict sleep staging on a single channel, transfer learning was used successfully [26–28]. To deal with this and other domain mismatch issues, such as differences in the demographics between the training and inference data, transfer learning can be applied to automatic sleep staging [29,30]. In [30], a sequence-to-sequence architecture for automatic sleep staging is proposed and evaluated for transfer learning across different databases with varying electrode setups. In [17,31], the authors show the application of transfer learning for cEEGrid. Besides domain mismatch, the use of transfer learning was used to investigate the personalization of sleep staging algorithms [32,33].

This study evaluates the usage of the trEEGrid for automatic sleep staging and additionally investigates the influence of channel combinations to calculate virtual electrode

positions for this use case. Our main contributions are as follows: First, we apply sleep staging algorithms to recordings made with the trEEGrid. Second, we investigate the utilization of channel combinations for automatic sleep staging, which has been shown to be helpful for sleep expert scorers following a linear combination model to approximate classical electrode positions [9]. Thirdly, we use transfer learning, with recordings made with the trEEGrid as the target domain. Furthermore, we investigate the influence of channel combinations on the transfer learning for the trEEGrid. Experimental results demonstrate the applicability of the proposed method to automatic sleep staging aimed at at-home sleep monitoring applications in the future. We have limited ourselves here initially to data from healthy individuals in order to keep the analyses and results straightforward. For practical application, the transfer to sleep data with different pathologies is important in a further step.

2. Methods

Sleep staging is the task of assigning one of the five sleep stages “wake” (W), “rapid-eye-movement” (REM), “non-REM 1” (N1), “non-REM 2” (N2), “non-REM 3” (N3) defined by the AASM rules to segments of a PSG recording consisting of EEG and additional further sensors during sleep. The mentioned segments are usually 30 s long due to historical reasons. Throughout this work, we use only EEG and electrooculography (EOG) signals of a subject’s recorded night. The signals used in this study consist of multiple channels, containing voltages between a common reference electrode and the signal electrodes derived from the scalp.

To investigate the capabilities of the novel trEEGrid which provides EEG and EOG signals, we perform the task of automatic sleep staging on different sets of recordings. In the following section, we describe the source dataset used for transfer learning and the dataset recorded with the trEEGrid, as well the network architecture used to perform the automatic sleep staging and the metrics to quantify the resulting performance.

2.1. Data

We use multiple datasets in this study to investigate the usage of transfer learning for data recorded with the trEEGrid. As the source domain, we use the Montreal archive of sleep studies (MASS), cf. [34]. Recordings made with the trEEGrid represent the target domain in the experiment. Those datasets and the pre-processing are described in the following.

2.1.1. Montreal Archive of Sleep Studies

The MASS database of which the first cohort consists of five subsets was recorded in the center for Advanced Research in Sleep Medicine, Montreal, Canada. In this study, we utilize the third subset (SS3) to perform our experiments. SS3 consists of 30 s epochs, which were scored according to the AASM standards. The dataset consists of 62 recordings, each from 62 different healthy subjects, which accumulate to a total recording time of 21.8 days. The subject’s mean age is 42.5 years, with a standard deviation of 18.9 years. A total of 55% of the subjects are female and 45% are male. Recordings were made with an full PSG montage, of which we use several EEG and EOG channels. EEG channels available for all 62 subjects in SS3 are Fp1, Fp2, Fz, F3, F4, F7, F8, Cz, C3, C4, Pz, P3, P4, T3, T4, T5, T6, Oz, O1 and O2. Left and right horizontal EOG channels are also included. The recordings are sampled at a sampling rate of 100 Hz.

2.1.2. trEEGrid Database

This database was recorded to evaluate a prototype of a self-applicable grid that is able to record EEG, EOG and EMG data, cf. [9] and Figure 1. There are 20 datasets recorded with the trEEGrid prototype in combination with a wireless EEG amplifier (Smartering Sleep, mBrainTrain, Belgrade, Serbia) from young adults (mean age = 28.9 years) who reported to have no sleep disorders. In parallel to the recording with the trEEGrid prototype, a

classical PSG recording was made simultaneously with a commercial mobile PSG system (SOMNOscreen Plus, SomnoMedics, Randersacker, Germany). The commercial system included six EEG electrodes, two piezosensory belts for thorax and abdomen expansion and a finger clip sensor, which measured oxygen saturation and pulse. The two EMG electrodes of the PSG system were not used, since they competed with the space of the trEEGrid EMG channels R6 and R7. PSG electrodes were gold-plated cup electrodes mounted with adhesive gel.

Eight recordings had to be excluded due to technical problems either with the trEEGrid (six), with the Bluetooth connection to the amplifier (one) and with the PSG system caused by a power failure (one) so that 12 datasets remain for the sleep staging computations.

Both recordings were scored by an experienced sleep expert scorer according to the AASM rules. To perform the experiments, we use the labels made on the basis of the classical PSG recording. There is a total of five epochs in the 12 recordings, that could not have a label assigned, due to artifacts. The electrodes and the channels, respectively, are numbered from R1 to R7, as depicted in Figure 2. Channels R6 and R7 are aimed at EMG acquisition and will be ignored in the rest of this paper. Recordings are made with a sampling rate of 250 Hz. Recorded signals were bandstop filtered from 45 to 55 Hz and 60 to 65 Hz to reduce line noise and impedance current. Impedance current was used through the measurements to record impedance continuously for evaluation of the electrode grid prototype. Furthermore, a bandpass filter from 0.5 to 40 Hz was applied to the signals. All filters used are phase true Butterworth filters of fourth order.

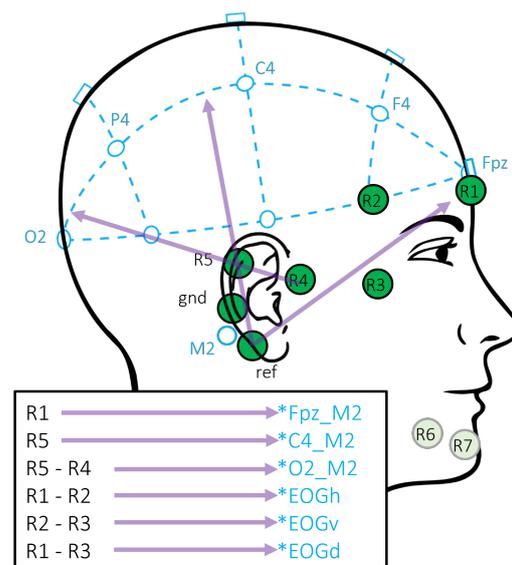


Figure 2. trEEGrid channel positions and their re-combinations. Re-referenced channels approximate classical PSG positions, as well as EOG and EMG, marked with an asterisk (*) to indicate the approximation.

In [9], the benefit of linear combinations of different channels as approximation of classical electrode positions was investigated. The specific linear combinations are sketched in Figure 2. The same approach using re-referencing is used here to approximate the signal of the standard PSG setup based on the signals recorded with the trEEGrid.

2.2. Preprocessing

To align the recordings with each other, we performed resampling to a sampling rate of 60 Hz. Before the resampling, re-referencing may be performed. After that, we perform interquartile range normalization for each channel independently as follows

$$\mathbf{x}_k^{norm} = \frac{\mathbf{x}_k}{\text{IQR}(\mathbf{x}_k)} \quad (1)$$

where \mathbf{x} is a vector containing $t \in 1, \dots, T$ samples and $\text{IQR}(\cdot)$ is an operator that calculates the interquartile range of the signal \mathbf{x} for each channel k along the time t . Sections in the signal that exceed the interquartile range by twenty times were hard clipped within this range.

2.3. Model Architecture

In this work, we use the RobustSleepNet architecture proposed in [30], which is a variant of the sequence-to-sequence structure often used for automatic sleep staging. Transfer learning has been applied to this architecture, as it tackles the problem of electrode mismatch between datasets. The use of an attention mechanism to aggregate across the input channels enables the architecture to work efficiently in transfer learning with a different number of channels.

The network shown in Figure 3 is structured as follows: an epoch encoder, extracting a feature vector representing each epoch, a sequence encoder processing temporal dependencies among the sequence of epochs and finally a classification layer. The epoch encoder first normalizes the input features by mean and variance. Next, the RobustSleepNet employs a linear layer to reduce the frequency dimension. This linear layer shares the same weights across all channels. All input channels are then recombined using multi-head attention, so that the subsequent recurrent network receives a constant number of input channels. This two-layer bidirectional GRU network, operating along the time dimension, processes the information distributed within each epoch. Dropout is employed before and after the GRU network. To aggregate the GRU's output into a single vector representation for each epoch, an attention layer that accumulates information along the time dimension is employed.

The sequence encoder receives the feature vectors that were computed by the epoch encoder for each epoch and aims to model the inter-epoch dependencies. It consists of a bidirectional GRU network with two layers with skip connections. Skip connections are implemented by a linear layer, whose transformed representations are added to the output of the associated GRU layer. Also, after each GRU layer, dropout is employed. The sequence encoder consisting of GRU layers models the inter-epoch dependencies in the sleep explicitly.

The classification scores are calculated based on the output of the sequence encoder. There is a linear layer with an output dimension of five, followed by a softmax activation to achieve log likelihoods for each class. Each output dimension represents one out of the five sleep stages.

We use the network with the same parameters as described in [30]. The network consists of approximately 180 k trainable parameters.

2.4. Loss Function

We employ the cross entropy loss function defined by

$$\mathcal{L}(x, y) = \frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C x_{n,c} y_{n,c} \quad (2)$$

where $x_{n,c}$ denoted the log-likelihood and $y_{n,c}$ the label for class c , with $c \in [1, C]$ and epoch n . C represents the number of classes, which is five. Epochs that are not labeled as one of the sleep stages due to artifacts do not have a class label $c \in [1, C]$, so these epochs are excluded from loss computation. Each batch consists of N epochs, the product of batch size and context size.

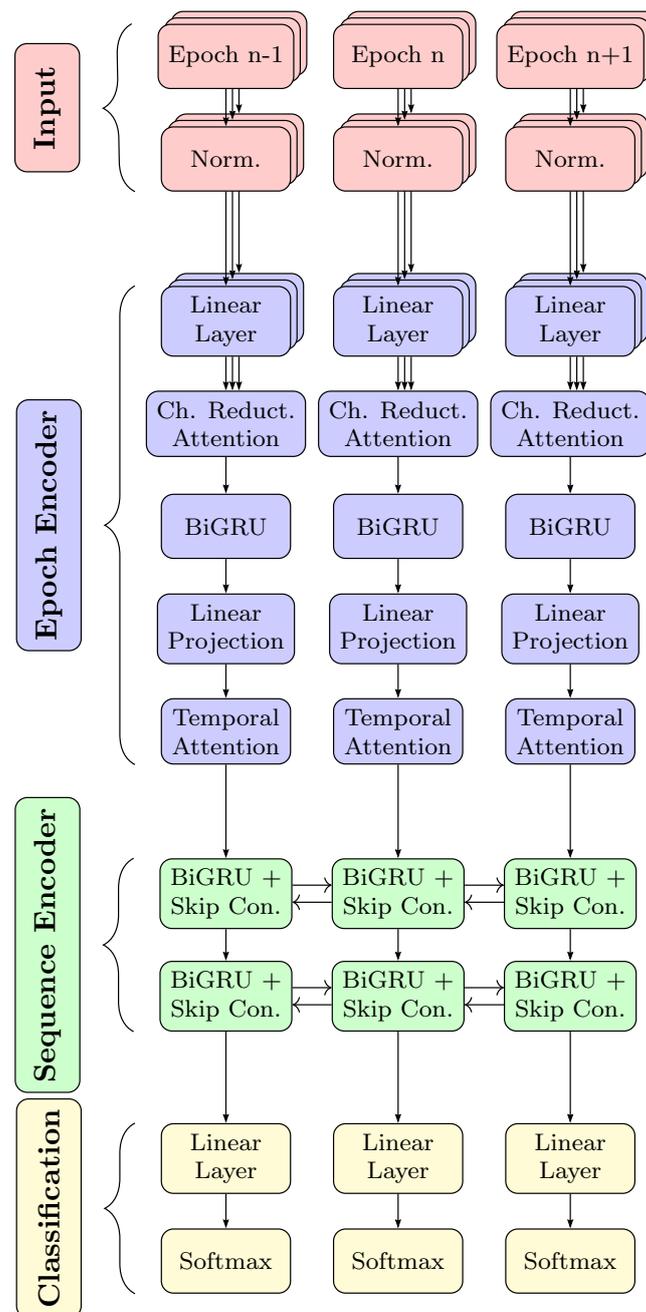


Figure 3. Overview of the employed neural network RobustSleepNet based on [30]. The center column shows the processing steps of epoch n . Building blocks left and right at epochs $n - 1$ and $n + 1$ share the same weights and are shown to show the inter-epoch dependency processing. As other sequence-to-sequence networks used for sleep staging, the network consists of an epoch encoder, sequence encoder and classifier. In the RobustSleepNet, the epoch encoder is capable of working with an arbitrary number of input channels. Depicted by the layered boxes, the input channels are processed independently first. An attention block merges the input channels for further processing. Finally, the softmax output gives the prediction of the sleep stage.

2.5. Inference

During inference time, we sample a whole night, so that we concatenate all epochs into one sequence. This way, the RNN in the sequence encoder works along the whole sequence. This differs from the scoring method described in [30], in which a shifting context sequence

of length 21 was scored, with a stride of one and geometric averaging is employed on the scores across all classification of an epoch. We changed the inference method, as we found no improvement caused by this ensemble scoring method, despite the mismatch between training and inference setup.

2.6. Channel Setups

As described before, the dataset that we use to train and test the model were recorded on different EEG montages. To adapt the montages to each other, we choose to select the EEG channels in a way that the recorded signals are more similar. In particular, we take five different setups into consideration, three setups from the source domain (SS3, SS3^{grid}, SS3^{grid-PSG}) and two from the target domain (trEEGrid, trEEGrid^{PSG}). While the SS3 setup takes all available classical PSG EEG and EOG channels into account, the other setups aim to match the signals in the other domain, respectively. All setups are listed in Table 1. In the SS3^{grid-PSG} setup, classical electrode positions close to the electrode positions realized with the novel trEEGrid sensor were chosen. The trEEGrid setup uses all available channels of the trEEGrid sensor grid with the original reference close to the mastoid. For the trEEGrid^{PSG} setup, R1 and R5 are used, being close to the classical position Fpz and C4. Additionally, re-referencing was applied to approximate classical EEG electrode position O2, as sketched in Figure 2. In SS3^{grid}, we take a subset of MASS SS3 channels into consideration that corresponds to the channels approximated by the trEEGrid^{PSG} setup. As the MASS SS3 dataset does not contain an Fpz channel, it is approximated by the average of the channels Fp1 and Fp2. All setups based on the trEEGrid dataset also include a diagonal EOG channel. As this is not available in MASS SS3, we use the right horizontal EOG channel in there instead.

Table 1. Selection of channels used for each channel setup. In the case of the two trEEGrid setups, re-referenced channels are denoted as the difference of two channels.

Label	EEG Channels	EOG Channels
SS3	all 20	all 2
SS3 ^{grid}	Fz, F8, T4	Right Horiz
SS3 ^{grid-PSG}	(Fp1 + Fp2)/2, C4, O2	Right Horiz
trEEGrid	R1, R2, R4, R5	R1-R3
trEEGrid ^{PSG}	R1, R5, R5-R4	R1-R3

3. Experiments

3.1. Training Procedure

During training, the batch size was 12. Each training example of the batch consists of 21 consecutive epochs, resulting in 252 epochs in a training batch in total. To deal with the class imbalance problem, epochs were sampled in a weighted manner, similar to [21]. This means that, for each sequence in a batch, we first draw one epoch from the dataset. This epoch is drawn randomly, but weighted, to counteract the class imbalance. The drawn epoch is then given a random position in the sequence, so that the context is filled with the adjacent epochs. Some of the adjacent epochs will represent a different sleep stage than the balanced drawn epoch; nevertheless, this sampling method still mitigates the imbalance problem. One training iteration over the dataset is considered as finished, when there were as many training examples drawn as there are epochs in that dataset. Due to this method, iterating over the dataset does not necessarily contain all samples in one iteration. Nevertheless we did not ensure that an iteration contains every sample at least once and left the length of one iteration to the number of epochs in the dataset.

Cross validation was used to train on the target datasets, as well for training on the source dataset. The datasets were split so that each subject's epochs were present in only one of the training, validation or test set. For MASS SS3, 10-fold cross validation was used. As the trEEGrid dataset consists only of 12 recorded nights, we use a leave-one-night-out scheme, resulting in 12 cross validation folds. The sizes of training, validation and test sets

for the cross validation as well as the number of folds are depicted in Table 2. Percentages for the SS3 dataset are target proportions, which are varied across folds to accommodate that the total number of recordings is not divisible by the number of folds.

Table 2. Parameters used for cross validation setup for both datasets.

	MASS SS3	trEEGrid
number of recordings	62	12
number of folds	10	12
training	70%	9 recordings
validation	20%	2 recordings
test	10%	1 recording

We implemented the experiments using PyTorch in version 1.12.1 and pytorch-lightning in version 1.7.7. As an optimizer, Adam [35] with parameters (learning rate: 10^{-4} , β_1 : 0.9, β_2 : 0.999) and weight decay regularization (weight decay: 10^{-3}) was used. All weights and biases of the model layers were initialized with values drawn from a random uniform distribution with upper and lower bounds $(-\sqrt{l}, \sqrt{l})$, where l for each layer is defined as $l = \frac{1}{\text{number of input features}}$. The models were trained on an NVIDIA Quadro RTX 8000 GPU, on which a 10-fold cross validation run on the MASS SS3 subset takes approximately 3 h. We use early stopping as we end training after the validation loss did not decrease for 25 training iterations. Training was stopped after a maximum of 100 training iterations. The same seed to initialize all random number generators was used for all runs.

3.2. Transfer between Domains

For investigation, trainings on the source domain (MASS dataset) and target domain (trEEGrid) were performed. We show the learning-from-scratch (LFS) performance for both, the source and target domain, where the cross validation and the datasets was performed. The direct transferability between domains is quantified by applying a model trained on the source domain directly on the target domain, without any adaption to the new domain. We denote this as direct transfer (DT). Finally, we report the transfer learning performance, for which a mode, trained on the source domain data is finetuned (FT) on the target domain data. In this case, the weights for the sequence encoder and the classifier were fixed, so that only the epoch encoder becomes finetuned for the target domain. For DT and FT, the model that was trained on the source dataset and showed the lowest validation loss during cross validation was used.

4. Performance Metrics

To determine the algorithm's performance, we use the F_1 scores as well as Cohen's Kappa. The scores are always calculated across all folds of the cross validation. This is carried out by first calculating the F_1 scores and Cohen's Kappa on the test data for every fold. Then the mean and the standard deviation is determined across all folds for each metric.

4.1. F_1 Score

Reported F_1 scores for each class were calculated from the number of true positive (TP), false positive (FP) and false negative (FN) for each class, as

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

with precision = $\frac{TP}{TP+FP}$ and recall = $\frac{TP}{TP+FN}$. $F_{1,c}$ denotes the F_1 score for class c . The macro- F_1 score which gives a measure across all C sleep stages was obtained by averaging over the F_1 score of each class:

$$\text{Macro-}F_1 = MF_1 = \sum_{c=1}^C F_{1,c} \quad (4)$$

4.2. Cohen's Kappa

Cohen's Kappa is often used to measure the inter-rater agreement in the context of medical annotations. As we compare the performance of the automatic sleep staging to the human achievable performance, we also calculated Cohen's Kappa κ as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (5)$$

p_o stands for the observed probability that two related labels correspond to the same class, while p_e is the expected probability when randomly assigning labels. Cohen's Kappa was calculated across all sleep stages, denoted as κ .

5. Results

The automatic sleep staging was applied to the five different channel setups, of which the setups using the MASS SS3 dataset act as the source domain in the case of transfer learning. In Table 1, scores for LFS, DT and FT as well as the inter-scorer variability of our scorer [9] on the trEEGGrid dataset and according to [6] are listed.

Because the number of datasets has a large impact on classification performance in addition to differences in electrode montage, we conducted a preliminary examination of the effect of the number of nights used in the modeling. For that, we randomly sampled subjects of the MASS SS3 database to create artificially smaller datasets, starting with 10 subjects, going up to the whole MASS SS3 dataset. For each number of subjects, 20 trainings were performed on newly drawn subsets to assess the variation between runs. Each training consists of one split with again 70% train data, 20% validation data and 10% test data. We report the median macro- F_1 over the 20 repeated runs and its 25 and 75 percentiles. As shown in Figure 4, having a smaller dataset decreases the performance achieved by the trained network, as well as the variety in performance across multiple iterations. It can be seen that a dataset consisting of subjects in the low tens leads to a median macro- F_1 score of around 0.75. From a number of around 40 datasets, a macro- F_1 well above 0.80 is achieved.

To assess in a second step the differences between the five setups, the LFS performance was calculated. Next, the performance of transfer learning from the SS3 setups to the matching trEEGGrid setups was calculated. Performance metrics are listed in Table 3.

In the LFS scores, it can be seen that the algorithm benefits from multiple channels present in the SS3 dataset. Reducing the channels for the SS3^{grid} and SS3^{grid-PSG} setups deteriorates the performance slightly, compared to the SS3 setup. Automatic sleep staging performs worse on both trEEGGrid setups than on the SS3 setups, which can be expected due to the smaller dataset. Here, the trEEGGrid setup that uses the default trEEGGrid channels shows worse performance in both Cohen's Kappa and macro- F_1 scores than trEEGGrid^{PSG} setup, which has less channels in total.

The DT condition reveals the domain mismatch between the source and respective target datasets. So in direct transfer the average macro- F_1 score is in all setups in a range between 0.63 and 0.65 and worse than the LFS situation. This indicates that none of the models based on SS3 datasets succeed in the direct transfer to the new domain compared to the results achieved with the LFS approach even on small datasets. However, when considering Cohen's Kappa, models trained on a subset of available channels of MASS SS3 (SS3^{grid}, SS3^{grid-PSG}) tend to perform better than their counterparts trained on SS3.

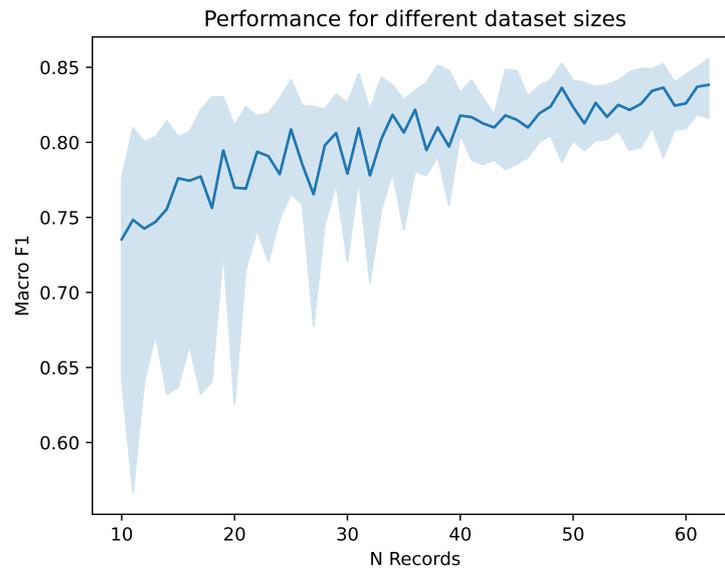


Figure 4. Macro- F_1 score achieved on different subsets of MASS SS3. The number of records drawn from MASS SS3 is listed on the x axis. Shown are the median and the quarter percentiles of the macro- F_1 score, over 20 repeated runs on randomly drawn subsets. Calculations were carried out with the SS3^{grid}-PSG setup.

Table 3. Performance measurements for Learning From Scratch (LFS), Direct Transfer (DT), Finetuning (FT), for the considered dataset and channels setups. Shown are the κ and macro- F_1 scores over all classes, as well as F_1 scores broken down for each class and their standard deviations. The best scores for each training type (LFS, DT, FT) are marked in bold.

Type	Source Domain	Target Domain	κ	MF_1	$F_{1,W}$	$F_{1,N1}$	$F_{1,N2}$	$F_{1,N3}$	$F_{1,REM}$
LFS	SS3	SS3	0.82 ± 0.02	0.84 ± 0.02	0.91 ± 0.05	0.65 ± 0.05	0.91 ± 0.01	0.82 ± 0.04	0.90 ± 0.02
LFS	SS3 ^{grid}	SS3 ^{grid}	0.80 ± 0.02	0.82 ± 0.02	0.89 ± 0.04	0.61 ± 0.02	0.90 ± 0.02	0.80 ± 0.06	0.88 ± 0.03
LFS	SS3 ^{grid} -PSG	SS3 ^{grid} -PSG	0.78 ± 0.02	0.81 ± 0.02	0.87 ± 0.03	0.61 ± 0.06	0.89 ± 0.01	0.79 ± 0.04	0.89 ± 0.02
LFS	trEEGrid	trEEGrid	0.63 ± 0.14	0.66 ± 0.10	0.70 ± 0.12	0.46 ± 0.15	0.73 ± 0.09	0.77 ± 0.20	0.66 ± 0.25
LFS	trEEGrid ^{PSG}	trEEGrid ^{PSG}	0.74 ± 0.05	0.77 ± 0.04	0.77 ± 0.11	0.60 ± 0.10	0.76 ± 0.07	0.87 ± 0.04	0.85 ± 0.08
DT	SS3	trEEGrid	0.56 ± 0.14	0.63 ± 0.10	0.74 ± 0.14	0.48 ± 0.11	0.71 ± 0.11	0.66 ± 0.21	0.56 ± 0.24
DT	SS3	trEEGrid ^{PSG}	0.57 ± 0.11	0.63 ± 0.08	0.72 ± 0.13	0.47 ± 0.11	0.71 ± 0.11	0.68 ± 0.18	0.60 ± 0.20
DT	SS3 ^{grid}	trEEGrid	0.59 ± 0.14	0.63 ± 0.11	0.73 ± 0.15	0.41 ± 0.14	0.72 ± 0.10	0.68 ± 0.21	0.62 ± 0.24
DT	SS3 ^{grid}	trEEGrid ^{PSG}	0.60 ± 0.12	0.64 ± 0.11	0.71 ± 0.13	0.40 ± 0.14	0.72 ± 0.11	0.71 ± 0.19	0.65 ± 0.27
DT	SS3 ^{grid} -PSG	trEEGrid	0.62 ± 0.14	0.64 ± 0.10	0.67 ± 0.17	0.34 ± 0.12	0.73 ± 0.11	0.71 ± 0.21	0.76 ± 0.11
DT	SS3 ^{grid} -PSG	trEEGrid ^{PSG}	0.62 ± 0.16	0.65 ± 0.13	0.71 ± 0.14	0.35 ± 0.16	0.71 ± 0.13	0.74 ± 0.18	0.73 ± 0.20
FT	SS3	trEEGrid	0.71 ± 0.15	0.74 ± 0.12	0.77 ± 0.08	0.59 ± 0.15	0.80 ± 0.07	0.75 ± 0.19	0.80 ± 0.19
FT	SS3	trEEGrid ^{PSG}	0.70 ± 0.12	0.75 ± 0.08	0.85 ± 0.06	0.58 ± 0.08	0.77 ± 0.10	0.80 ± 0.11	0.76 ± 0.19
FT	SS3 ^{grid}	trEEGrid	0.69 ± 0.13	0.75 ± 0.10	0.80 ± 0.09	0.61 ± 0.15	0.77 ± 0.10	0.77 ± 0.17	0.79 ± 0.17
FT	SS3 ^{grid}	trEEGrid ^{PSG}	0.78 ± 0.09	0.81 ± 0.05	0.84 ± 0.11	0.62 ± 0.09	0.81 ± 0.10	0.87 ± 0.06	0.88 ± 0.07
FT	SS3 ^{grid} -PSG	trEEGrid	0.70 ± 0.17	0.73 ± 0.12	0.79 ± 0.14	0.51 ± 0.14	0.80 ± 0.10	0.77 ± 0.21	0.79 ± 0.18
FT	SS3 ^{grid} -PSG	trEEGrid ^{PSG}	0.76 ± 0.09	0.79 ± 0.04	0.86 ± 0.06	0.61 ± 0.13	0.79 ± 0.10	0.83 ± 0.07	0.88 ± 0.08
Scorer [6]	-	SIESTA [36]	0.76	-	-	-	-	-	-
Our Scorer [9]	-	trEEGrid ^{PSG}	0.70	0.74	0.78	0.51	0.78	0.85	0.80

In the case of FT condition, in every case the performance improves compared to DT, in both Cohen's Kappa and macro- F_1 score. More improvement is observed when SS3^{grid} and SS3^{grid}-PSG were used for pre-training. As in LFS, the trEEGrid^{PSG} setup as target domain tends to outperform the trEEGrid setup. This is even the case if pre-training happened on SS3^{grid}, where larger channel mismatch is expected. Compared to the results of DT, it can be observed for FT also in macro- F_1 score that the pre-training on SS3 performs worse on both

target domains, trEEGrid and trEEGrid^{PSG}. Thus, a more generalized model in terms of input channels does not give an advantage here. Finetuning to the trEEGrid^{PSG} from both SS3^{grid} and SS3^{grid-PSG} performs best, with almost equal performance in terms of Cohen's Kappa and macro- F_1 score. In these cases, the automatic sleep staging performs better on the data recorded with the trEEGrid than from a human scorer on the same data [9] and in terms of Cohen's Kappa it achieves an accuracy as scorer on classical recordings following the AASM rules.

6. Discussion

This study investigates the use of transfer learning for sleep staging based on a small dataset of only 12 recordings obtained with a prototype of a self-applied electrode grid, so-called trEEGrid. We trained RobustSleepNet [30], an established neural network for this task, on both the MASS SS3 database and the dataset recorded with the trEEGrid prototype. The influence of the channel setup in both domains on the performance during the transfer was investigated. Also, the influence of the size of the dataset on the expected performance and the uncertainty was considered. This demonstrated a deterioration in the system performance with a reduction in the number of records in the dataset, accompanied by a higher uncertainty between multiple runs. Sources for this uncertainty are the random components in the training procedure, such as the random initialization of the trainable network parameters, the random selection of training records from the MASS SS3 dataset and the random order in which training samples are compiled in batches during the training process.

For smaller datasets, a decreasing performance and a higher uncertainty are generally to be expected. However, the results show that it also depends on the selected channels and their combination. In the results, it can be seen that LFS scores are highest for the SS3 setup, the largest dataset. SS3^{grid} and SS3^{grid-PSG} show worse performance, which can be explained by the fact that these setups have fewer channels. However, if the variance is taken into account, the three setups based on MASS S3 are very similar to each other. Reported performance measures are always influenced by the random nature of the training procedures of neural networks, so that differences must not be over-interpreted and the variance must be taken into account.

Both trEEGrid datasets have also a limited number of channels but additionally a much smaller number of records (12 nights). The dataset trEEGrid achieves the worst result of the LFS runs. However, by combining the channels with the aim of approximating classic electrode positions, better results are achieved with trEEGrid^{PSG}, even though one channel less is included in the model. By adding the linear combination R5-R4, which approximates the classical channel O2 in the trEEGrid^{PSG} setup, new independent information is available to the model, compared to the information of channel R1, R2, R4 and R5 where (R1 and R2) and (R4 and R5) have a similar orientation and thus similar information. A more specific investigation of this form of signal approximation has been carried out in relation to sleep spindles [9]. The larger variance values for the performance data must be taken into account here as well.

For DT, the results show worse performance, which is expected due to domain mismatch. However, the matching channel setups SS3^{grid} to trEEGrid and SS3^{grid-PSG} to trEEGrid^{PSG} do not help in direct transfer. This hints at a still remaining channel mismatch, which can arise due to differences in the exact positions between the trEEGrid electrodes and the EEG cap used in the recording of the MASS SS3 database. Although the channels for SS3^{grid} and SS3^{grid-PSG} were selected to correspond with the channels represented in trEEGrid and trEEGrid^{PSG} setups, respectively, these channels remain distinct in their placement and, consequently, in their recorded signal.

The FT for the trEEGrid datasets as the target domain shows no improvement over the LFS if the base dataset SS3 is used as source domain. The performance for the target domain trEEGrid also remains limited, even if the sub-datasets SS3^{grid} or SS3^{grid-PSG} are used as source domain. A performance gain is achieved if trEEGrid^{PSG} is selected as the target

domain. Here, for both source domains SS3^{grid} and SS3^{grid-PSG} finetuning to trEEGrid^{PSG} results in highest scores in terms of both Cohen's Kappa and F_1 metrics. Finetuning on the trEEGrid dataset improves the performance relative to LFS and DT, but cannot achieve the same performance as trEEGrid^{PSG}. Again, this is a hint concerning the benefit of using the linear combination R5-R4.

It is important to note that while the approach is promising, its current application in healthcare settings may be limited because the datasets used consisted exclusively of healthy participants. Sleep architecture can vary widely due to a number of factors, including age and various sleep disorders. Therefore, using only data from healthy individuals may not provide a sufficient basis for clinical applications. For example, it has been shown that a model trained on recordings from healthy participants suffers when applied to patients with obstructive sleep apnea because this disease pattern results in higher sleep fragmentation [11]. It would be beneficial to include assessments using sleep data from patients with different pathologies. Depending on the prevalence of the specific pathology, only a small number of recordings may be available; the transfer learning approach could also be advantageous here.

7. Conclusions

Using a sequence-to-sequence network, we showed the performance of a sleep staging algorithm finetuned on a small dataset of recordings made with the trEEGrid, a prototype of a self-applicable electrode grid capable of recording EEG, EOG and EMG at home. The used network architecture, previously used for transfer learning on multiple larger datasets is applied here on a smaller dataset. We can see that finetuning helps to improve the performance on this very small dataset, consisting of 12 records. Furthermore, it is demonstrated that the approximation of the source dataset towards the target dataset helps to improve the performance for finetuning. This is relevant not only in the use of automated sleep staging for new sensor technology for which there is a limited amount of observational data possible, but also for special disease conditions for which few datasets are available. The datasets used in this study are from healthy adults. Further analysis needs to be conducted on the application to data with different sleep pathologies. Transfer learning could also be a way of improving model performance here. Finally, transfer learning based on small datasets of the target domain is of interest for classification models that are trained individually for specific subjects to enable long-term observation. Our results also indicate that while a more generalized pre-training can aid in the direct transfer of a trained model to a new domain, it does not improve the preconditions when finetuning is necessary. We demonstrated that using re-referenced channels to estimate signals that are not directly measurable with the trEEGrid improves automatic sleep staging, particularly when trEEGrid recording is utilized to approximate classical PSG setups. As a result, automatic sleep staging outperforms human sleep scorers when scoring the same data and performs on par with human scorers on classical PSG recordings.

Author Contributions: Conceptualization and methodology, F.A.R., K.I.W. and C.F.d.S.S.; software, validation, formal analysis and investigation, F.A.R.; data curation, C.F.d.S.S. and W.P.; writing—original draft preparation, F.A.R.; writing—review and editing, K.I.W., W.P., C.F.d.S.S. and F.A.R.; visualization, W.P. and F.A.R.; project administration and funding acquisition, K.I.W. All authors have read and agreed to the published version of the manuscript.

Funding: The Oldenburg Branch for Hearing, Speech and Audio Technology HSA of the Fraunhofer IDMT is funded in the program (Vorab) by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics Committee, University of Oldenburg, Oldenburg, Germany (protocol code Drs.EK/2020/028 and date of approval 12.5.2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The availability of the data presented in this study is limited due to General Data Protection Regulations (GDPR).

Acknowledgments: We would like to express our gratitude to Nils Strodthoff for his valuable contributions, exchange and critical discussion throughout the process of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PSG	Polysomnography
EEG	Electroencephalography
EOG	Electrooculography
EMG	Electromyography
LFS	Learning from scratch
DT	Direct transfer
FT	Finetuning

References

- Miettinen, T.; Myllymaa, K.; Westeren-Punnonen, S.; Ahlberg, J.; Hukkanen, T.; Toyras, J.; Lappalainen, R.; Mervaala, E.; Sipila, K.; Myllymaa, S. Success Rate and Technical Quality of Home Polysomnography With Self-Applicable Electrode Set in Subjects With Possible Sleep Bruxism. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1124–1132. [CrossRef] [PubMed]
- Figorilli, M.; Marques, A.R.; Vidal, T.; Delaby, L.; Meloni, M.; Pereira, B.; Lambert, C.; Puligheddu, M.; Durif, F.; Fantini, M.L. Does REM sleep behavior disorder change in the progression of Parkinson's disease? *Sleep Med.* **2020**, *68*, 190–198. [CrossRef] [PubMed]
- Tamaki, M.; Bang, J.W.; Watanabe, T.; Sasaki, Y. Night Watch in One Brain Hemisphere during Sleep Associated with the First-Night Effect in Humans. *Curr. Biol.* **2016**, *26*, 1190–1194. [CrossRef] [PubMed]
- Berry, R.B.; Quan, S.F.; Abreu, A.R.; Bibbs, M.L.; DelRosso, L.; Harding, S.M.; Mao, M.; Plante, D.T.; Pressman, M.R.; Troester, M.M.; et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*; Version 2.6; American Academy of Sleep Medicine: Darien, IL, USA, 2020.
- Rechtschaffen, A.; Kales, A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Electroencephalogr. Clin. Neurophysiol.* **1969**, *26*, 644. [CrossRef]
- Danker-Hopfe, H.; Anderer, P.; Zeitlhofer, J.; Boeck, M.; DORN, H.; Gruber, G.; Heller, E.; Loretz, E.; Moser, D.; Parapatics, S.; et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J. Sleep Res.* **2009**, *18*, 74–84. [CrossRef] [PubMed]
- Bleichner, M.G.; Debener, S. Concealed, Unobtrusive Ear-Centered EEG Acquisition: cEEGrids for Transparent EEG. *Front. Hum. Neurosci.* **2017**, *11*, 163. [CrossRef] [PubMed]
- Debener, S.; Emkes, R.; Vos, M.D.; Bleichner, M. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Sci. Rep.* **2015**, *5*, 16743. [CrossRef] [PubMed]
- Da Silva Souto, C.F.; Pätzold, W.; Paul, M.; Debener, S.; Wolf, K.I. Pre-gelled Electrode Grid for Self-Applied EEG Sleep Monitoring at Home. *Front. Neurosci.* **2022**, *16*, 883966. [CrossRef]
- Arnal, P.J.; Thorey, V.; Debellemaniere, E.; Ballard, M.E.; Bou Hernandez, A.; Guillot, A.; Jourde, H.; Harris, M.; Guillard, M.; van Beers, P.; et al. The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep* **2020**, *43*, zsa097. [CrossRef]
- Guillot, A.; Sauvet, F.; During, E.H.; Thorey, V. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 1955–1965. [CrossRef]
- Somnomedics. Home Sleep Test. Insomnie—Diagnostik einfach zu Hause. Available online: <https://somnomedics.de/de/home-sleep-test/> (accessed on 21 June 2023).
- Da Silva Souto, C.F.; Pätzold, W.; Wolf, K.I.; Paul, M.; Matthiesen, I.; Bleichner, M.G.; Debener, S. Flex-Printed Ear-EEG Sensors for Adequate Sleep Staging at Home. *Front. Digit. Health* **2021**, *3*, 688122. [CrossRef] [PubMed]
- Fiorillo, L.; Puiatti, A.; Papandrea, M.; Ratti, P.L.; Favaro, P.; Roth, C.; Bargiotas, P.; Bassetti, C.L.; Faraci, F.D. Automated sleep scoring: A review of the latest approaches. *Sleep Med. Rev.* **2019**, *48*, 101204. [CrossRef] [PubMed]
- Phan, H.; Andreotti, F.; Cooray, N.; Chen, O.Y.; Vos, M.D. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 400–410. [CrossRef] [PubMed]
- Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1998–2008. [CrossRef]

17. Phan, H.; Lorenzen, K.P.; Heremans, E.; Chén, O.Y.; Tran, M.C.; Koch, P.; Mertins, A.; Baumert, M.; Mikkelsen, K.; Vos, M.D. L-SeqSleepNet: Whole-cycle Long Sequence Modelling for Automatic Sleep Staging. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4748–4757. [[CrossRef](#)] [[PubMed](#)]
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
19. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
20. Sors, A.; Bonnet, S.; Mirek, S.; Vercueil, L.; Payen, J.F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed. Signal Process. Control* **2018**, *42*, 107–114. [[CrossRef](#)]
21. Perslev, M.; Jensen, M.H.; Darkner, S.; Jennum, P.J.; Igel, C. U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging. *arXiv* **2019**, arXiv:1910.11162.
22. Perslev, M.; Darkner, S.; Kempfner, L.; Nikolic, M.; Jennum, P.J.; Igel, C. U-Sleep: Resilient high-frequency sleep staging. *npj Digit. Med.* **2021**, *4*, 72. [[CrossRef](#)]
23. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; Vos, M.D. SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 2456–2467. [[CrossRef](#)] [[PubMed](#)]
24. Chato, L.; Regentova, E. Survey of Transfer Learning Approaches in the Machine Learning of Digital Health Sensing Data. *J. Pers. Med.* **2023**, *13*, 1703. [[CrossRef](#)]
25. Waters, S.H.; Clifford, G.D. Comparison of deep transfer learning algorithms and transferability measures for wearable sleep staging. *BioMed. Eng. OnLine* **2022**, *21*, 66. [[CrossRef](#)] [[PubMed](#)]
26. Phan, H.; Chén, O.Y.; Koch, P.; Mertins, A.; Vos, M.D. Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019.
27. Jadhav, P.; Rajguru, G.; Datta, D.; Mukhopadhyay, S. Automatic sleep stage classification using time–frequency images of CWT and transfer learning using convolution neural network. *Biocybern. Biomed. Eng.* **2020**, *40*, 494–504. [[CrossRef](#)]
28. ElMoaqet, H.; Eid, M.; Ryalat, M.; Penzel, T. A Deep Transfer Learning Framework for Sleep Stage Classification with Single-Channel EEG Signals. *Sensors* **2022**, *22*, 8826. [[CrossRef](#)]
29. Banluesombatkul, N.; Ouppaphan, P.; Leelaarporn, P.; Lakhan, P.; Chaitusaney, B.; Jaimcharyatam, N.; Chuangsuwanich, E.; Chen, W.; Phan, H.; Dilokthanakul, N.; et al. MetaSleepLearner: A Pilot Study on Fast Adaptation of Bio-signals-Based Sleep Stage Classifier to New Individual Subject Using Meta-Learning. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 1949–1963. [[CrossRef](#)]
30. Guillot, A.; Thorey, V. RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 1441–1451. [[CrossRef](#)]
31. Phan, H.; Chén, O.Y.; Koch, P.; Lu, Z.; McLoughlin, I.; Mertins, A.; Vos, M.D. Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 1787–1798. [[CrossRef](#)] [[PubMed](#)]
32. Mikkelsen, K.; Vos, M.D. Personalizing deep learning models for automatic sleep staging. *arXiv* **2018**, arXiv:1801.02645.
33. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; Kidmose, P.; Vos, M.D. Personalized automatic sleep staging with single-night data: A pilot study with Kullback-Leibler divergence regularization. *Physiol. Meas.* **2020**, *41*, 064004. [[CrossRef](#)]
34. O’Reilly, C.; Gosselin, N.; Carrier, J.; Nielsen, T. Montreal Archive of Sleep Studies: An open-access resource for instrument benchmarking and exploratory research. *J. Sleep Res.* **2014**, *23*, 628–635. [[CrossRef](#)] [[PubMed](#)]
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Klösch, G.; Kemp, B.; Penzel, T.; Schlögl, A.; Rappelsberger, P.; Trenker, E.; Gruber, G.; Zeitlhofer, J.; Saletu, B.; Herrmann, W.M.; et al. The SIESTA project polygraphic and clinical database. *IEEE Eng. Med. Biol. Mag.* **2001**, *20*, 51–57. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.