From unsupervised to semi-supervised adversarial domain adaptation in EEG-based sleep staging

From unsupervised to semi-supervised adversarial domain adaptation in EEG-based sleep staging

Elisabeth R. M. Heremans[a], Huy Phan[b], Pascal Borzée[c], Bertien Buyse[c], Dries Testelmans[c], Maarten De Vos[a,d]

[a] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[b] Queen Mary University of London, London E1 4NS, U.K.

[c] UZ Leuven, Department of Pneumology, Herestraat 49, B-3000 Leuven, Belgium

[d] KU Leuven, Department of Development and Regeneration, Herestraat 49, B-3000 Leuven, Belgium

E-mail: elisabeth.heremans@kuleuven.be

Abstract. *Objective.* The recent breakthrough of wearable sleep monitoring devices results in large amounts of sleep data. However, as limited labels are available, interpreting these data requires automated sleep stage classification methods with a small need for labeled training data. Transfer learning and domain adaptation offer possible solutions by enabling models to learn on a source dataset and adapt to a target dataset. *Approach.* In this paper, we investigate adversarial domain adaptation applied to real use cases with wearable sleep datasets acquired from diseased patient populations. Different practical aspects of the adversarial domain adaptation framework are examined, including the added value of (pseudo-)labels from the target dataset and the influence of domain mismatch between the source and target data. The method is also implemented for personalization to specific patients. *Main results.* The results show that adversarial domain adaptation is effective in the application of sleep staging on wearable data. When compared to a model applied on a target dataset without any adaptation, the domain adaptation method in its simplest form achieves relative gains of 7%-27% in accuracy. The performance on the target domain is further boosted by adding pseudo-labels and real target domain labels when available, and by choosing an appropriate source dataset. Furthermore, unsupervised adversarial domain adaptation can also personalize a model, improving the performance by 1%-2% compared to a non-personal model. *Significance.* In conclusion, adversarial domain adaptation provides a flexible framework for semi-supervised and unsupervised transfer learning. This is particularly useful in sleep staging and other wearable EEG applications. (Clinical trial registration number: S64190.)

Keywords: domain adaptation, transfer learning, electroencephalography, deep learning, sleep stage classification

Submitted to: *J. Neural Eng.*

1. Introduction

Sleep disturbances can have a hugely negative impact on quality of life. They also play a key role in a variety of illnesses [1, 2]. Diagnosing these disturbances traditionally requires overnight monitoring of a patient and subsequent annotation of the recorded signals. The gold standard for such sleep assessments is based on in-hospital polysomnography (PSG) recordings. Every 30-second segment of such a PSG recording is manually classified as a particular sleep stage according to developed rules [3, 4]. This is referred to as sleep staging.

The recent breakthrough of low-cost wearable electroencephalography (EEG) recording devices gives rise to a new era for sleep research. These these devices enable remote long-term monitoring of patients and allow to conduct large-scale screenings across the population. The increasingly large volumes of EEG data collected through such wearables necessitate automated analysis of the recorded signals. This pressing need is aggravated further by the fact that trained clinicians have difficulties interpreting wearable EEG signals recorded from non-standard locations [5]. Deep neural networks have extensively been trained for automated sleep staging on large, manually labeled PSG datasets [2, 6, 7, 8, 9, 10, 11, 12]. However, in wearable EEG datasets, labels are typically scarce. This greatly limits the performance of these deep learning methods.

To compensate for the lack of labeled data, automated sleep staging methods for wearable EEG should benefit from exploiting large labeled PSG datasets with standard EEG modalities. Transfer learning allows transferring information learned from a large dataset (the source domain) to improve the sleep staging performance on a usually smaller dataset (the target domain). There is typically a mismatch between the two domains. This can be caused by differences in recording equipment or recording setups, i.e. the recorded EEG electrode positions. Moreover, the patient population can differ in both datasets, causing a change in the sleep architecture, and manual labels can differ between scorers. Transfer learning methods aim to overcome the mismatch caused by these differences. Previous studies have successfully applied these methods towards sleep staging [13, 14], and even towards personalized sleep staging, where the patient population of the target domain consists of just one person [15]. These studies adopted a fully supervised fine-tuning approach, in which models were pre-trained on the source domain and fine-tuned on the target domain with limited labeled data.

Although supervised transfer learning techniques have proven useful, unsupervised techniques requiring only unlabeled wearable EEG data could be even more practical. Very recently, unsupervised domain adaptation techniques have found their way to EEG-based classification tasks [16, 17, 18, 19, 20]. Most of these techniques are based on domain-invariant feature learning, either by domain-adversarial training of neural networks [21], or using the maximum mean discrepancy (MMD) loss [22, 23]. A few studies focus on adversarial domain adaptation for sleep staging specifically [24, 25, 26]. These methods successfully cope with domain mismatch between different sleep databases without requiring any labels from the target domain. All these

studies exploit a common framework with some minor variations. The methods are demonstrated on traditional PSG sleep databases, with minimal differences between recorded channels of the source datasets and target datasets.

Building upon these previous studies, the present study makes the following contributions. First, instead of creating a novel domain adaptation method, this study assesses the performance of the common adversarial domain adaptation backbone in novel, real-world use cases. The target domains are three different databases with various non-traditional EEG derivations and real wearable EEG data. The data are acquired from different populations and include a realistic use case of elderly and diseased patients. Second, we investigate the added value of using pseudo-labels or real labels of the target domain, comparing a semi-supervised approach to fully unsupervised adversarial domain adaptation. To our best knowledge, the use of target labels was not yet investigated in domain adaptation studies for sleep staging, and the individual impact of pseudo-labels was not yet discussed. We hypothesize that a limited number of labels may help overcome the large mismatch between traditional PSG and wearable EEG datasets. Therefore, we augment the unsupervised domain adaptation framework used in previous studies to a semi-supervised framework allowing for labeled target data. We also add a separate target domain classifier to allow for more flexibility in this framework. Third, we evaluate the effect of the different causes for domain mismatch between a source domain and a target domain. To do so, we investigate alternative source domains to match with the target domains. Lastly, we demonstrate that the same framework can be used for personalization, applying the adversarial domain adaptation method to single subjects.

The remainder of this paper is structured as follows. Section 2 starts with an introduction to domain adaptation and transfer learning and explains the proposed adversarial domain adaptation framework with its different variations. Section 3 describes the datasets used to evaluate the methods. Section 4 discusses the experiments conducted to assess the performance of the transfer learning methods in different scenarios. Section 5 then shows the obtained results, which are discussed in section 6. Section 7 extracts some final conclusions and recommendations from this study.

2. Methods

2.1. Transfer learning and domain adaptation

Domain adaptation is generally referred to as a type of transfer learning. Transfer learning is defined as the transfer of information between two different but related machine learning problems [27]. It is aimed at addressing machine learning scenarios where a model is trained to perform a task on a source domain, but then applied to a potentially different task in a potentially different target domain. Formally, a domain consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$: $\mathcal{D} = \{\mathcal{X}, P(X)\}$. A task consists of a label space $\mathcal{Y}$ and predictive function $f(.)$

projecting the input space onto the label space: $\mathcal{T} = \{\mathcal{Y}, f(.)\}$. When a mismatch between two tasks or domains occurs ($\mathcal{T}_S \neq \mathcal{T}_T$ or $\mathcal{D}_S \neq \mathcal{D}_T$ with subscripts S and T denoting source and target), transfer learning is used to overcome this disparity, improving $f_T(.)$ using information from $\mathcal{T}_S$ and $\mathcal{D}_S$ [27, 28, 29].

Domain adaptation is a common type of transfer learning in which the task remains unchanged ($\mathcal{T} = \mathcal{T}_S = \mathcal{T}_T$), but the source domain differs from the target domain ($\mathcal{D}_S \neq \mathcal{D}_T$). In domain adaptation, labeled source data are usually assumed available. As such, we call a domain adaptation technique unsupervised, semi-supervised, or supervised depending on the use of labels from the target domain [28].

In the application to automated sleep staging on wearable EEG recordings, we select a standard EEG channel in a large public PSG database as the source domain, $\mathcal{D}_S$. Knowledge is transferred from this domain to a target domain, $\mathcal{D}_T$, consisting of a small dataset of non-standard EEG recordings. The main difference between the source and target dataset is in the domains, owing to the difference in electrode positions. The task is not profoundly changed between both datasets, as it always consists of classifying 30-second data segments into five sleep stages. As a result, this can be regarded predominantly as a domain adaptation problem. However, when the patient population or the annotator changes between the datasets, the predictive function $f(.)$ between the input data and output labels changes as well. Hence, the tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ are not actually identical, so we argue this should be regarded as a general transfer learning problem. When target and source tasks are different, this, in turn, implies that some labeled target data are required to infer the target predictive function $f_T(.)$, as argued by theoretical works [28, 29]. Therefore, while some of the discrepancies between the source and target can be overcome by using unlabeled target data, the performance on the target domain will be further improved with the addition of target domain labels.

2.2. Adversarial domain adaptation

In adversarial domain adaptation, the domain discrepancy between a source domain and a target domain is overcome through two competing objectives. A feature extractor F extracts features from the input data. These features are then processed by a classifier C predicting the class labels and by a domain discriminator D predicting whether samples are from the source domain or the target domain. The feature extractor tries to confuse the domain discriminator by producing domain-invariant features and works with the classifier to produce features that allow distinguishing different classes. Figure 1 shows this idea in a schematic representation. In the application of sleep staging, F is the sleep staging network, and C can simply be its classification layer. The features computed by the one-before-last layer of the network are thus fed to both the classification layer and domain discriminator.

Typically, in adversarial domain adaptation, the feature extractor and classifier are both shared between the source and the target domain ($F \equiv F_C \equiv F_T$, $C \equiv C_S \equiv C_T$). A shared F is not a necessity but rather a design choice [30] which is often made. This

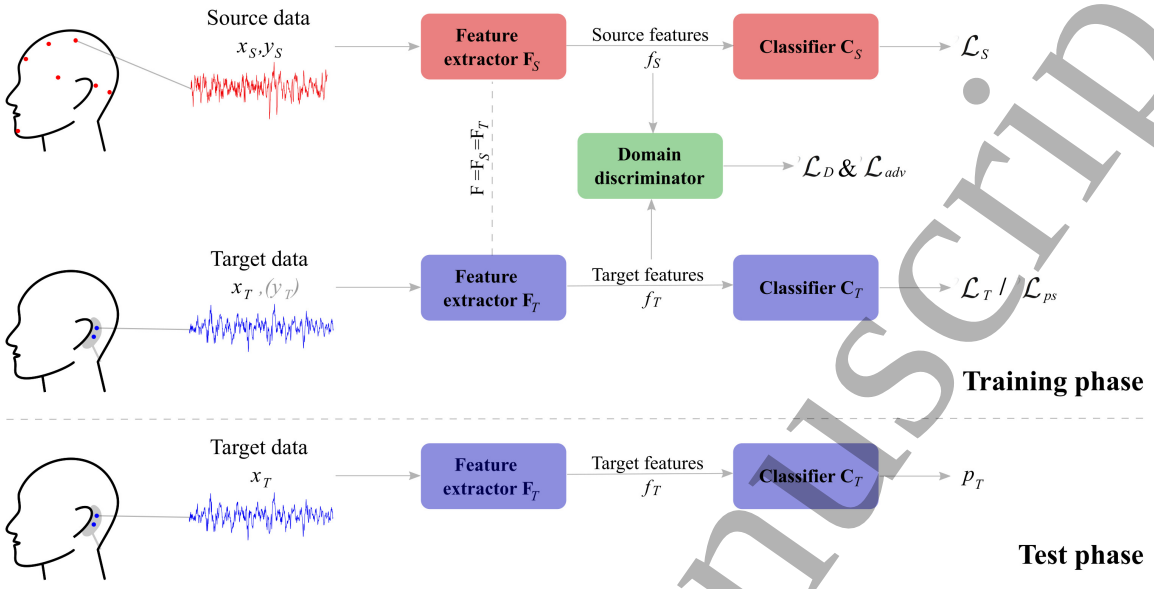

Figure 1. Schematic illustration of adversarial domain adaptation. Subscripts S and T stand for source and target domain, respectively. The feature extractors are always shared ($F_S \equiv F_T \equiv F$). The classifiers are shared in the basic method ($C_S \equiv C_T \equiv C$), but not in the pseudo-labels and target labels method. The loss function $\mathcal{L}_{ps}$ is used in the pseudo-labels method, $\mathcal{L}_T$ is used in the target labels method.

study also uses a shared feature extractor in line with previous studies [24, 26]. A shared C is necessary when there is no loss function to optimize a target classifier C_T independently. This is the case when no target domain (pseudo-)labels are used. When we do use (pseudo-)labels of the target domain, we can train an independent target classifier to learn the target task independently ($\mathcal{T}_S \neq \mathcal{T}_T$, see section 2.1).

Three versions of the adversarial domain adaptation technique are implemented: a basic method, a method with pseudo-labels, and a semi-supervised method with some labels from the target domain. In the pseudo-labels method and target labels method, a separate target classifier C_T is used. The following sections explain these three different adversarial domain adaptation methods in more detail, with a summary in table 1. Source code is available at https://github.com/elisabethRMH/adversarial_DA.

2.3. Basic method

The basic adversarial domain adaptation method is implemented as follows. The classifier and feature extractor are trained for the classification in the source domain using the standard categorical cross-entropy loss:

$$\mathcal{L}_S = -\mathbb{E}_{x_S \sim P_S} \sum_{c=1}^{N_c} y_S^c \log\, p_S^c, \tag{1}$$

where $p_S = C_S(F(x_S)) = C(F(x_S))$ is the class probability output of the network for a source sample, and y_S is the one-hot sleep stage label. c is the class number and N_c

the number of classes. The domain discriminator is trained with the standard binary cross-entropy loss to classify the domain of each sample:

$$\mathcal{L}_D = -\mathbb{E}_{x_S \sim P_S} \log\,(1 - d'_S) - \mathbb{E}_{x_T \sim P_T} \log\, d'_T, \tag{2}$$

where $d'_T = D(F(x_T))$ and $d'_S = D(F(x_S))$. The feature extractor F is trained with the adversarial loss to confuse the domain discriminator. The adversarial loss function corresponds to $\mathcal{L}_D$ with inverted labels:

$$\mathcal{L}_{adv} = -\mathbb{E}_{x_S \sim P_S} \log\, d'_S - \mathbb{E}_{x_T \sim P_T} \log\,(1 - d'_T). \tag{3}$$

The basic adversarial domain adaptation framework is governed by the following end-to-end optimization:

$$\min_{D} \mathcal{L}_D, \tag{4}$$

$$\min_{F,C} \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv}. \tag{5}$$

2.4. Pseudo-labels

Prior work on adversarial domain adaptation for EEG applications has already used multiple additions and tricks to improve upon the basic framework [24, 25, 26]. Pseudo-labels [25] and entropy minimization [24, 26] are techniques aimed at mitigating the lack of labels of the target domain. Both approaches include the addition of a categorical cross-entropy loss based on the target samples:

$$\mathcal{L}_{ps} = -\mathbb{E}_{x_T \sim P_T} \sum_{c=1}^{N_c} \tilde{y}_T^c \log\, p_T^c, \tag{6}$$

in which p_T is the class probability output of the network for a target sample, and $\tilde{y}_T^c$ is a substitute for the true label of this target sample. In entropy minimization, the true label is simulated by the classifier's output itself ($\tilde{y}_T^c = p_T^c$). Pseudo-labels are used as a more general term for an approximation of the true labels.

In this study, pseudo-labels are implemented as follows. Different from previous works on adversarial domain adaptation for sleep staging, a separate classifier is defined for the source domain (C_S) and the target domain (C_T). The source domain classifier is trained with $\mathcal{L}_S$ as defined in the loss function (1). The target domain classifier is trained with the loss function (6), where $\tilde{y}_T = C_S(F(x_T))$ is the probability output of the source classifier for the target samples, and $p_T = C_T(F(x_T))$ is the probability output of the target classifier for the target samples. This way, $\mathcal{L}_{ps}$ prevents the target classifier from straying too far from the source classifier while also favoring outputs with minimal entropy, i.e. with a higher probability for one class.

In conclusion, the pseudo-labels version of adversarial domain adaptation is governed by the following optimization procedure:

$$\min_{D} \quad \mathcal{L}_D, \tag{7}$$

$$\min_{F,C_S,C_T} \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{ps}\mathcal{L}_{ps}. \tag{8}$$

2.5. Target labels

As explained in section 2.1, the sleep staging task itself can differ between a source and target dataset. Learning the target task then requires some target labels. We thus investigate how much the target domain performance can be boosted by adding a limited amount of target labels. This semi-supervised approach is implemented by using labels of only two recordings of the target domain. C_T is trained with real target domain labels of a limited number of samples with the categorical cross-entropy loss function:

$$\mathcal{L}_T = -\mathbb{E}_{x_T \sim P_T} \sum_{c=1}^{N_c} y_T^c \log\ p_T^c. \tag{9}$$

This results in the following optimization framework for the adversarial domain adaptation with target labels:

$$\min_{D} \quad \mathcal{L}_D, \tag{10}$$

$$\min_{F, C_T, C_S} \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv} + \mathcal{L}_T. \tag{11}$$

Table 1. The three different adversarial domain adaptation methods.

	Classifiers	Loss functions	Hyperparameters
Basic	Shared (C)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}$	$\lambda_{adv} = 0.01$
Pseudo-labels	Unshared ($C_S \neq C_T$)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}, \mathcal{L}_{ps}$	$\lambda_{adv} = 0.01, \lambda_{ps} = 0.01$
Target labels	Unshared ($C_S \neq C_T$)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}, \mathcal{L}_T$	$\lambda_{adv} = 0.01$

3. Data

3.1. Source domain: PSG datasets

3.1.1. MASS PSG An appropriate choice for the source domain in transfer learning experiments is a standard EEG channel of a large public PSG dataset. In all experiments but one, we use the Montreal Archive of Sleep Studies (MASS) [31] for this purpose. MASS is a large, public PSG database gathered from three hospital-based sleep laboratories. It consists of 200 recordings of 103 women and 97 men between 18 and 76 years of age. Sleep stages were manually labeled according to either the AASM standards [32] or the R&K guidelines [4]. To make the dataset homogeneous, we combine the six sleep stages of the R&K rules into the five sleep stages of the AASM standard (W, N1, N2, N3, and REM) and expand all segments to 30 seconds, as in [6]. The C4-A1 channel of this database is selected as the source domain for all our experiments unless otherwise mentioned. This EEG derivation is a good choice for the source domain, as it is commonly used in automated sleep staging methods [6, 7, 9, 24, 25, 26] and recommended in sleep scoring guidelines [3, 32].

3.1.2. Leuven PSG In one experiment, a different source database is used to show the influence of changing the source domain on the transfer learning process. This source database is extracted from a large PSG dataset, consisting of 218 recordings of patients with suspected sleep apnea [33]. The dataset was recorded at the sleep laboratory of the University Hospitals Leuven (UZ Leuven) and annotated according to AASM standards [32]. From this PSG dataset, we select the C4-A1 channel of 38 recordings corresponding to patients over 60 years of age. This dataset is further referred to as the Leuven PSG database.

3.2. Target domain: wearable EEG datasets

Three different target domains are used to investigate the transfer learning scenarios in various setups reflecting possible use cases and wearable EEG configurations.

3.2.1. Surrey – cEEGrid The first target domain is the Surrey - cEEGrid dataset [34, 5]. It was recorded using the cEEGrid array, a wearable EEG device that records from multiple EEG channels around the ear with a flexible electrode strip [35, 36]. Simultaneous measurements of cEEGrid-EEG and PSG were collected from 12 healthy adults. A sony Z1 Android smartphone and wireless SMARTING amplifier (mBrainTrain, Belgrade, Serbia) were used to record the signals at a 250 Hz sampling rate. Manual sleep staging was performed on the PSG. In this study, the right-ear front-versus-back cEEGrid derivation is used.

3.2.2. Dreem - Headband The Dreem - Headband dataset [37] is a dataset recorded from 25 adult volunteers with self-reported quality of sleep varying between no complaints and sub-threshold insomnia symptoms [38]. EEG signals were recorded with the Dreem Headband, a reduced-montage wearable dry-EEG device recording from five frontal and occipital EEG electrodes. This dataset also includes simultaneous PSG measurements, which are scored with a consensus based on the manual labels of five sleep experts [37]. In the present study, the F7-F8 derivation of the Dreem Headband is used as a target domain.

3.2.3. Leuven – crosshead behind-the-ear The extended Leuven - crosshead behind-the-ear (Leuven-CBTE-46) sleep dataset consists of 46 recordings from the sleep laboratory of the University Hospitals Leuven (UZ Leuven). It is an extended version of the dataset described in [39]. The study was conducted in accordance with the Declaration of Helsinki, and the protocol with registration number S64190 / B3222020000148 was approved on 08.11.2018 by the Ethics Committee Ethische Commissie Onderzoek UZ/KU Leuven. The population consists of elderly patients with suspicion of sleep apnea. Data from these patients were recorded simultaneously with the full PSG and a crosshead behind-the-ear EEG. This additional EEG channel was recorded using an extra EEG electrode behind the right ear, referenced to A1. Manual labeling of the sleep

stages was performed on the PSG. In this study, the crosshead behind-the-ear channel is used as a simulation of a wearable behind-the-ear EEG.

4. Experiments

A number of experiments were performed to investigate multiple factors influencing the performance of adversarial domain adaptation. In each of these experiments, the feature extractor and classifier were first initialized with weights pre-trained on the source domain. Then, one of the adversarial domain adaptation methods from section 2 was applied. Source and target domain samples were fed to the feature extractor, and the feature extractor was trained to match the two domains. The classification performance on the target domain was investigated.

4.1. Sleep staging network

Adversarial domain adaptation can be applied to any sleep staging network architecture, as it only requires taking features learned by this network and feeding these to a domain discriminator. Both recurrent and convolutional architectures were successfully used in combination with adversarial domain adaptation [26].

In order to validate the proposed methods, a state-of-the-art sleep staging network was used to extract relevant features from the EEG signals. The SeqSleepNet [6] architecture with a sequence length of $M = 10$ was selected as the baseline feature extractor. This network follows a many-to-many classification scheme, taking a sequence of multiple 30-second segments as input and predicting the corresponding sequence of sleep stages.

SeqSleepNet requires being fed with time-frequency images rather than raw EEG signals. As a pre-processing step, all the EEG signals were first bandpass filtered between 0.3 and 40 Hz and resampled to 100 Hz. Then, the logarithmically scaled spectrogram of every recording was computed and normalized to unit standard deviation and zero mean.

The network architecture consists of a first block of layers operating on an epoch level and a second block of layers operating on a sequence level. Computations at the epoch level are performed by a filterbank layer, a bidirectional RNN (biRNN) layer, and an attention layer. This block of three layers outputs one feature vector for every 30-second epoch. Then, the feature vectors of the whole sequence of epochs are combined and presented to a sequence-level biRNN layer. This layer transforms the given sequence of input feature vectors into a sequence of output feature vectors. Thus, M input vectors get transformed into M output vectors. These are then classified into M sleep stages by a series of M fully connected layers with softmax activation. SeqSleepNet is trained end-to-end, by minimizing the average cross-entropy loss over the M segments. The parametrization and training settings in this study were the same as in the original paper, using L2-regularization, the Adam optimizer, and a learning rate of $1e-4$. For

more details, see [6].

4.2. Experimental setup

4.2.1. Training parameters The base network was pre-trained for 10 training epochs on all 200 C4-A1 recordings of the MASS dataset, except for 10 recordings used as a validation set to retain the best model. Then, for every domain adaptation experiment, cross-validation was performed on the target dataset to obtain average performance values. For every fold of a cross-validation experiment, the target dataset was divided into a training set, a validation set, and a test set. The training set was used to train the adaptation method for 20 training epochs, and the validation set was used to evaluate the model after every 100 training steps and retain the best-performing one. The best model was then evaluated on the test set. The test performances were averaged over all the cross-validation folds. For the Surrey - cEEGrid dataset consisting of 12 recordings, 12-fold cross-validation was performed. For the Dreem - Headband dataset, 12-fold cross-validation was performed on the 25 recordings. The Leuven - CBTE dataset of 46 recordings was split through 23-fold cross-validation.

4.2.2. Minibatch construction During training, minibatches were constructed with

(i) labeled data from the source domain,
(ii) unlabeled data from the target domain,
(iii) only in the semi-supervised experiments: labeled data from the target domain.

In every minibatch, the number of labeled samples from the target domain was fixed to 8 in the semi-supervised experiments. The number of unlabeled samples of the target domain was such that one training epoch would correspond to one pass through both the labeled and unlabeled part of the dataset. The amount of data from the source domain was balanced with the amount of target data.

4.2.3. Performance metrics After applying the adversarial domain adaptation training procedure, the final model was evaluated on an independent test set of the target domain as $p_T = C_T(F(x_T))$. This result was compared to the ground truth target labels y_T using the classification accuracy (acc), Cohen's kappa coefficient (κ), and the weighted F1-score (wF1). κ measures the inter-rater agreement between the scorer and the model. wF1 is the mean of the per-class F1-scores, weighted by each class's number of true instances.

4.3. Influence of target (pseudo-)labels

In the first set of experiments, we aimed to quantify the effect of adversarial domain adaptation on the wearable EEG target datasets and determine the impact of pseudo-labels and real target labels on the performance. Various baseline methods were implemented in order to evaluate the three adversarial domain adaptation methods

defined in section 2. The lower limit baseline was defined as the performance obtained by directly applying the network to the target domain after pre-training on the MASS C4-A1 dataset ('direct transfer'). The model was also trained on the complete labeled target dataset, with supervised training from scratch and supervised fine-tuning of the pre-trained network. The latter method represents the upper limit to the performance.

These experiments were performed for all three target domains: the Surrey - cEEGrid dataset, the Headband data of the Dreem dataset, and the crosshead behind-the-ear modality of the Leuven dataset. The Surrey - cEEGrid dataset and Dreem - Headband dataset both allow to investigate transfer learning on real wearable data acquired at different sleep laboratories from the source data of the MASS dataset. While the cEEGrid data are recorded from behind the ear, the Headband data are recorded with forehead electrodes. The Leuven-CBTE-46 target domain allows to analyze transfer learning on a new EEG modality acquired at another sleep laboratory and from a widely different population. The diseased and elderly patient population of this dataset reflects the real use case for wearable sleep monitoring and allows us to validate our sleep staging methods on the target population with suspected sleep-wake disturbances.

4.4. Influence of amount of domain mismatch

When a target and source domain are more similar, the domain mismatch is smaller, and the domain adaptation problem gets easier. The same holds for the difference in the source and target task in transfer learning. In each of the two following experiments, we selected a different source dataset with a closer resemblance to a target dataset. We tested whether such a better-matched source domain improved the performance after transfer learning to the target domain. Two different sources of mismatch were investigated in this way. To ensure a fair comparison, the source domain was only changed in the adversarial domain adaptation training, so pre-training was performed on the original source domain (C4-A1 of MASS) in all experiments.

4.4.1. Influence of recording setup The first experiment tested the influence of the mismatch between the recording setup of the source and target domain. The F7-F8 forehead derivation of the Dreem dataset greatly differs from the C4-A1 derivation of the MASS PSG dataset. However, the MASS dataset's EOG signal (EOG left-right) should resemble the F7-F8 derivation much more. For that reason, we tested adversarial domain adaptation with MASS's EOG derivation as a source domain and the Dreem dataset's F7-F8 derivation as a target domain. The obtained performance was compared to the performance obtained with MASS's C4-A1 channel as a source domain.

4.4.2. Influence of other factors The second experiment tested the impact of the mismatch between the population, recording equipment, and scorer of the source data and target data. The Leuven PSG dataset was specifically selected as an almost perfect match for the Leuven-CBTE-46 dataset with regard to all three factors. Indeed, both

datasets are recorded from elderly suspected sleep apnea patients in the same sleep laboratory and hence with the same equipment, and even scored by the same scorer. The sleep staging model was thus adapted to the Leuven-CBTE-46 dataset (right ear - A1 derivation) with two source domains: the highly matched Leuven PSG dataset and the general MASS PSG dataset. The source EEG derivation was C4-A1 in both cases.

4.5. Personalization

Adversarial domain adaptation could also provide an elegant way to personalize a sleep staging network to an individual subject. A last set of experiments was conducted to determine the effect of unsupervised personalization, using adversarial domain adaptation with pseudo-labels. Since only one recording was available per subject, these experiments required some adjustments of the experimental setup defined for the other experiments.

Regarding the pre-training, it made sense to start the personalization process for each subject with a network adapted to the relevant EEG modality and dataset. The feature extractor and classifier weights were thus initialized with network weights pre-trained with supervised fine-tuning on the relevant dataset and EEG derivation (Surrey - cEEGrid, Dreem - Headband, or Leuven-CBTE-46). Following the cross-validation scheme from section 4.2, in every fold, the base network was first fine-tuned to other recordings of the dataset (the training set of the fold). Then, adversarial domain adaptation was separately performed on each recording in the fold's test set. The new 'source dataset' in this personalization process consisted of the training set of the current fold, and the 'target dataset' was one of the test set recordings. To avoid confusion with the source and target datasets as defined for the other experiments, these are further referred to as the training set and test recording in the context of personalization.

Pre-training was achieved with transfer learning to the other recordings of the target dataset, and was thus performed with the same experimental setup as the transfer learning experiments (see section 4.2). In the personalization step, the network was trained for 10 training epochs instead of 20. Instead of selecting the best-performing model using a validation set, we retained the model with the lowest pseudo-label loss $\mathcal{L}_{ps}$ for evaluation. This change in the experimental setup was necessary because we could not use an independent labeled validation set as the target domain was just one recording. Minibatches consisted of 32 samples of the training set and the test recording. After performing the cross-validation as described, results were averaged over all the test recordings.

The personalization experiments were carried out on all three target datasets: the Surrey - cEEGrid dataset, the Dreem - Headband dataset, and the Leuven-CBTE-46 dataset. Adversarial domain adaptation for personalization was performed with the pseudo-labels method and compared with two baselines. The first baseline was the performance of the pre-trained baseline network, in this case, trained with supervised fine-tuning to the relevant target dataset. The second baseline was the result obtained

with a simple adaptation of the normalization statistics in the batch norm layers of the sleep staging network. This commonly known domain adaptation technique [28] was recently demonstrated in personalized sleep staging [40]. The method adjusts normalization statistics to a target domain without training any network weights. By comparing against this baseline, we investigated whether adversarial domain adaptation accomplishes more than only adapting these normalization statistics.

5. Results

5.1. Influence of target (pseudo-)labels

Figure 2 shows the influence of adversarial domain adaptation on the sleep staging performance in the target domain. The impact of using (pseudo-)labels of the target domain is also shown. The accuracy is plotted for all three proposed adversarial domain adaptation methods: the basic method (ADA-0), the pseudo-labels method (ADA-ps), and the method with real labels of two recordings (ADA-2). The performance of these adversarial domain adaptation methods is compared to the three baselines: direct transfer of the pre-trained base network (DT), supervised training from scratch (FS), and fine-tuning on the target dataset (FT). Results are ranked from least to most relevant training data for training the model. Table 2 shows the complete results, with all three performance metrics. The performance metrics are represented as their mean and standard error over all cross-validation folds.

The results clearly show that all three adversarial domain adaptation methods achieve performances between the direct transfer and training from scratch baselines. The supervised fine-tuning baseline performs better than training from scratch. Overall, out of the three adversarial domain adaptation methods, the version with true target labels of two recordings (ADA-2) performs the best, followed by the pseudo-label version (ADA-ps), and lastly the basic method (ADA-0).

The hyperparameters λ_{adv} and λ_{ps} define the relative weight of the adversarial domain loss function and the pseudo-label loss function during training. In this study, $\lambda_{adv} = 0.01$ and $\lambda_{ps} = 0.01$ were chosen. Figure 3 shows a sensitivity analysis to these hyperparameters on the Surrey - cEEGrid dataset. As can be seen, the hyperparameter values are optimal for this dataset. However, the same values were applied for all three datasets, and the trends and results were consistent across the three datasets without overfitting to these hyperparameters.

5.2. Influence of domain mismatch

Figure 4 shows the effect of different sources of domain mismatch on the adversarial domain adaptation results. The figure shows the mean and standard error of the sleep staging accuracy for different training methods: adversarial domain adaptation with no target labels (ADA-0), with pseudo-labels (ADA-ps) and with labels of two recordings (ADA-2).

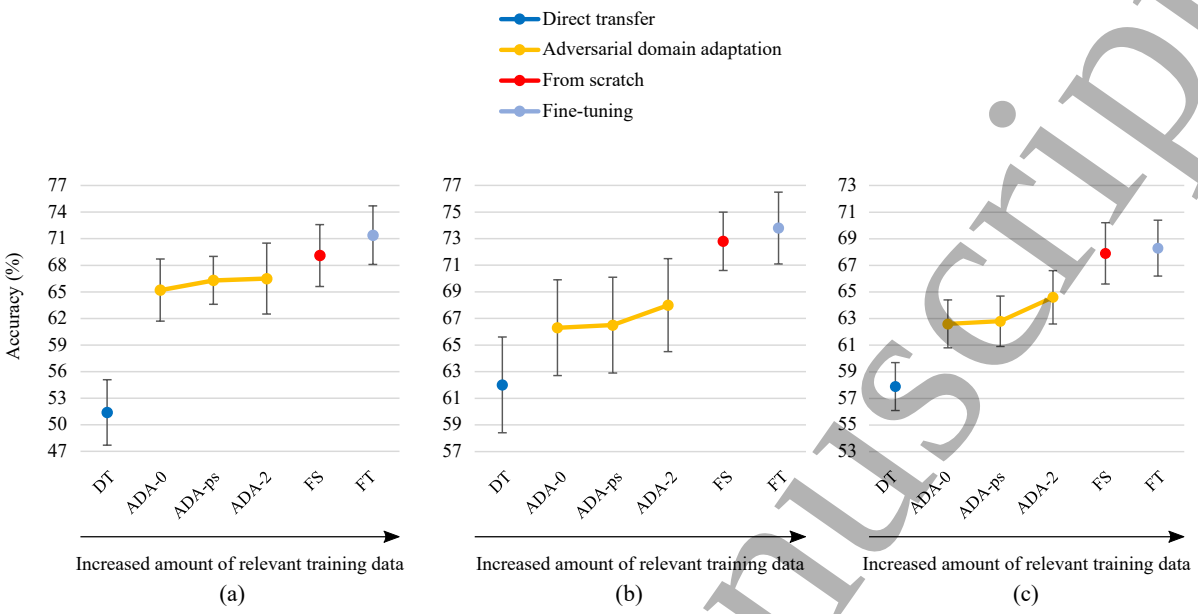

Figure 2. The accuracy of the three adversarial domain adaptation techniques on the three target datasets. Direct transfer (DT), fully supervised training (FS), and supervised transfer learning (TL) are shown as baselines. ADA designates adversarial domain adaptation. ADA-0 is the basic ADA, ADA-ps is the pseudo-label version, and ADA-2 is the ADA with labels of two target recordings. From left to right, there is an increased amount of relevant training data. The error bars indicate the standard error over the cross-validation folds. (a) The Surrey - cEEGrid dataset. (b) The Dreem - Headband dataset. (c) The Leuven - crosshead behind-the-ear dataset.

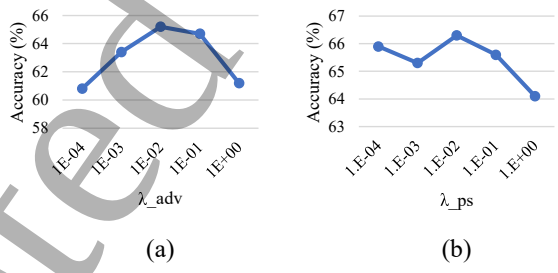

Figure 3. Sensitivity analysis to the hyperparameters λ_{adv} and λ_{ps} in Surrey - cEEGrid dataset. (a) λ_{adv} in the basic adversarial domain adaptation experiment ($\lambda^*_{adv} = 0.01$), (b) λ_{ps} in the pseudo-label experiment ($\lambda^*_{ps} = 0.01$).

Figure 4(a) displays the influence of the mismatch between recording setups on the classification performance on the Dreem - Headband dataset. There is a clear difference in performance between the scenario with the EOG channel and the C4-A1 channel as the source domain. With the EOG channel as a source domain, the mean accuracy is higher for all three adversarial domain adaptation methods. Figure 4(b) shows the impact of the mismatch between the population, scorer, and equipment on the classification performance in the Leuven-CBTE-46 dataset. The sleep staging accuracy on this target dataset is higher with the Leuven PSG dataset as a source domain than

Table 2. The results of the three adversarial domain adaptation techniques: the basic version using no labels of the target domain, the pseudo-labels version and the version with target domain labels of two recordings. The compared baselines are direct transfer with no retraining, fully supervised training from scratch on the target dataset, and supervised fine-tuning on the target dataset. The sleep staging performance is reported for the three target domains, as mean ± standard error over all the cross-validation folds. Metrics are the accuracy (acc), Cohen's kappa (κ) and weighted F1-score (wF1).

Surrey - cEEGrid				
Method		Acc	κ	wF1
Direct transfer		51.4±3.7	0.375±0.032	50.4±3.1
ADA	Basic	65.2±3.5	0.512±0.046	62.4±3.3
ADA	Pseudo-labels	66.3±2.7	0.517±0.039	63.2±2.8
ADA	Target labels	66.5±4.0	0.527±0.055	64.3±4.0
From scratch		69.1±3.5	0.575±0.041	66.3±3.6
fine-tuning		71.4±3.3	0.597±0.046	70.5±3.2
Dreem - Headband				
Method		Acc	κ	wF1
Direct transfer		62.0±3.6	0.460±0.045	61.4±3.6
ADA	Basic	66.3±3.6	0.512±0.049	66.0±3.6
ADA	Pseudo-labels	66.5±3.6	0.517±0.047	66.3±3.5
ADA	Target labels	68.0±3.5	0.537±0.048	67.8±3.5
From scratch		72.8±2.2	0.599±0.030	71.1±2.1
Finetuning		73.8±2.7	0.619±0.035	73.5±2.6
Leuven - crosshead behind-the-ear				
Method		Acc	κ	wF1
Direct transfer		57.9±1.8	0.440±0.024	58.4±1.8
ADA	Basic	62.6±1.8	0.482±0.024	63.1±1.8
ADA	Pseudo-labels	62.8±1.9	0.485±0.025	63.2±2.0
ADA	Target labels	64.6±2.0	0.506±0.028	64.3±2.2
From scratch		67.9±2.3	0.538±0.032	65.2±2.4
fine-tuning		68.3±2.1	0.550±0.030	66.9±2.2

with the MASS dataset as a source domain. Again, this is true for all three adversarial domain adaptation methods (see the green line compared to the yellow line in figure 4(b)).

These results clearly indicate that the mismatch between the source and target dataset influences the obtained classification performance on the target dataset (for a further discussion, see section 6).

5.3. Personalization

Table 3 shows the effect of personalization on individual patients. The performance before personalization is shown (non-pers), as well as the performance after personalization with the batch norm method (BN pers) and with adversarial domain adaptation (ADA pers). For each metric and target dataset, the table reports the mean

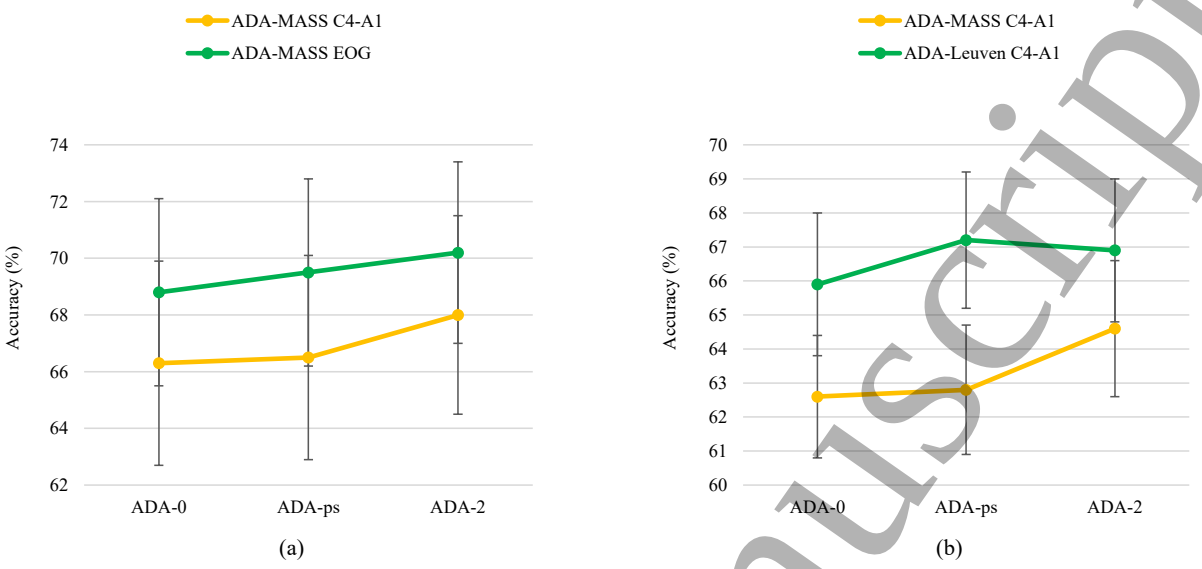

Figure 4. The influence of the domain mismatch between the source domain and target domain. ADA designates adversarial domain adaptation. ADA-0 is the basic ADA, ADA-ps is the pseudo-label version, and ADA-2 is the ADA with labels of two target recordings. The error bars indicate the standard error over the cross-validation folds. (a) For the Dreem - Headband dataset, the MASS dataset's EOG is compared to the MASS dataset's C4-A1 derivation as a source domain. (b) For the Leuven - crosshead behind-the-ear dataset, the Leuven PSG dataset is compared to the MASS dataset as a source domain.

and standard error over all the patients in this target dataset.

In almost all experiments, both personalization methods improve the performance compared to the baseline of fine-tuning to other patients of the same dataset. It is also observed that adversarial domain adaptation outperforms or achieves similar results to batch norm personalization in all datasets.

6. Discussion

This study proposes a unified adversarial domain adaptation framework for real-world use cases and evaluates its performance in the context of sleep staging on wearable EEG recordings. Experiments were performed to investigate the influence of key elements in this framework: the use of pseudo-labels and real labels from the target domain, the similarity of the source domain to the target domain, and the potential for personalization.

First, the adversarial domain adaptation strategy clearly achieves its purpose in real applications of wearable EEG recordings and unhealthy patient populations. The basic adversarial domain adaptation achieved relative accuracy improvements of 7% to 27% compared to the direct transfer scenario (see table 2). These gain margins are higher than those reported for the SeqSleepNet network by Yoo et al. [26], ranging from -1% to 13%. Our higher margins of improvement can be directly related to the increased

Table 3. Results of personalization on the three target domains. Mean ± standard error of the sleep staging performance over all the subjects after transfer learning to the dataset without personalization (Non-pers), batch normalization personalization (BN pers), and adversarial domain adaptation personalization (ADA pers). Accuracy (acc), Cohen's kappa (κ), and weighted F1-score (wF1) are shown. For each metric, the highest mean is shown in bold.

Surrey - cEEGrid			
	Acc	κ	wF1
Non-pers	71.4±3.3	0.597±0.046	70.5±3.2
BN pers	72.5±3.4	0.613±0.047	72.4±2.9
ADA pers	**72.8±3.6**	**0.618±0.049**	**72.7±3.2**
Dreem - Headband			
	Acc	κ	wF1
Non-pers	73.6±2.7	0.625±0.030	73.8±2.6
BN pers	72.9±2.5	0.610±0.032	73.1±2.4
ADA pers	**74.5±2.7**	**0.638±0.030**	**74.6±2.6**
Leuven - crosshead behind-the-ear			
	Acc	κ	wF1
Non-pers	68.4±1.8	0.544±0.026	67.4±1.8
BN pers	**69.1±1.7**	0.551±0.026	**68.4±1.6**
ADA pers	69.0±1.8	**0.558±0.025**	68.2±1.7

domain mismatch in our experiments. In [26], the source and target domains consisted of similar, mostly healthy populations with comparable EEG channel configurations. In the present study, there was a large mismatch between the populations and channel configurations. This resulted in lower baseline accuracies for the direct transfer scenario and larger gain margins on average.

Second, this study generalizes adversarial domain adaptation to a semi-supervised framework for EEG-based classification and sleep staging applications. The addition of a separate target domain classifier trained with pseudo-labels or real labels from the target domain clearly has a positive impact on the classification performance in the target domain. Pseudo-labels resulted in rather small improvements, with relative increases in accuracy 0% to 2%. The improvement by using real target labels led to relative accuracy gains of 2% to 3%. We conclude from these results that the adversarial domain adaptation backbone is easily generalized to a semi-supervised framework. If a limited amount of target labels is available, it is beneficial to use them. If no labels of the target domain are available, pseudo-labels can be employed as an imperfect substitute. Our results support the idea that the classification tasks in the source dataset and target dataset differ to an extent, as this could explain the large gains obtained from training a separate target domain classifier with target labels.

The domain mismatch experiments were designed to investigate the influence of the similarity between the source domain (and task) and the target domain (and task) in the transfer learning process. In both experiments, during the adversarial domain adaptation training, the original source dataset was replaced with a new source dataset, which more closely matched the target dataset. The pre-training of the network was still performed on the original source dataset in both cases, so all adversarial domain adaptation experiments started from the same baseline network. The first domain mismatch experiment investigated the influence of the source domain channel (see figure 4(a)). Using the EOG channel of the MASS dataset instead of the C4-A1 channel, the accuracy was improved by a relative 3% to 5%. This clearly indicates that the match between a source channel and target channel has a large influence over the target performance. In practical applications with wearable EEG datasets, it makes sense to choose as the source domain channel the available PSG channel with the highest possible similarity to the new wearable derivation.

The second domain mismatch experiment tested the influence of other factors of the source dataset (see figure 4(b)). Being a highly controlled dataset of the same population, acquired in the same sleep laboratory, and scored by the same scorer, the Leuven PSG dataset is an almost perfect match for the Leuven-CBTE-46 dataset. The only substantial difference between this source and target dataset is the EEG channel, which is the standard C4-A1 in the source domain and the right ear - A1 channel in the target domain. The accuracy improved by 4% to 7% when using this matched source domain instead of the MASS source domain. Therefore, we conclude from this experiment that the match between the population, scorer, and recording equipment of the source and target dataset also greatly influences the target performance. In real clinical applications, it will often be unrealistic to find a source dataset with such a good correspondence on all these aspects. However, any parameter that influences the similarity of the source domain and target domain should be considered when choosing a source domain to achieve the best possible performance in the target domain. A second noteworthy observation with the Leuven PSG as a source dataset was the lack of improvement when using true labels instead of pseudo-labels (see ADA-ps and ADA-2 in figure 4(b)). As the population and scorer are the same in the Leuven-CBTE-46 and Leuven PSG datasets, this is a more 'pure' domain adaptation problem in which the sleep scoring task remains unchanged. This argument could explain why the target labels in this scenario did not add value. Again, this result supports the idea that the task mismatch between datasets is the reason why target labels are useful.

Lastly, the personalization experiments compared a baseline model fine-tuned on the relevant dataset to a model personalized to a specific recording (see table 3). The adversarial personalization results were also compared to batch norm personalization as a simpler alternative method. Overall, both personalization strategies improved upon the non-personalized baseline. The adversarial strategy systematically improved the baseline by a relative 1% to 2%. For all three datasets, it outperformed or performed on par with the batch norm method, which improved the baseline in two out of three

datasets. Hence, this investigation indicates that adversarial domain adaptation can be applied for personalization. Further work is needed to determine to what extent it outperforms the simple batch norm method. The margin of improvement obtained with the adversarial domain adaptation strategy may well increase with increasing numbers of recordings of a subject. Indeed, more recordings would lead to a more accurate target distribution with sufficient representation for all sleep stages.

7. Conclusion

In this study, adversarial domain adaptation was applied to several real-world sleep staging problems. We can summarize our findings in a number of recommendations for adversarial domain adaptation in general, and some for sleep staging in particular.

First, adversarial domain adaptation is mostly implemented as an unsupervised method, but it can easily be generalized to a semi-supervised method. When some labels of the target domain are available, they should be used for improving the accuracy on the target domain. This is mostly beneficial when the source and target classification task may differ. In sleep staging tasks, the classification task may change with the study population and scorer, making semi-supervised learning with some target labels the superior option. Pseudo-labels can serve as an imperfect substitute when no true labels are available. When multiple labeled source domains are available to choose from, it is advisable to select the one that is the most similar to the target domain. For sleep staging, both the recording setup and the practical aspects of the sleep study such as the scorer, patient population, and recording equipment are of consequence. Lastly, we demonstrated that adversarial domain adaptation can elegantly achieve personalization of a model to a specific recording of an individual subject.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgement

This research was supported by the Research Foundation - Flanders (FWO) [grant number 1SC2921N]; by the 'Bijzonder Onderzoeksfonds KU Leuven (BOF)' ('Prevalence of Epilepsy and Sleep Disturbances in Alzheimer Disease - C24/18/097' and 'Starting Grant: Artificial Intelligence (AI)-enabled mining of big longitudinal datasets collected with wearable sensors'); and by the Flemish Government (AI Research Program). M.D.V and E.R.M.H. are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

References

[1] Nathaniel F. Watson, M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, James Gangwisch, Michael A. Grandner, Clete Kushida, Raman K. Malhotra, Jennifer L. Martin, Sanjay R. Patel, Stuart F. Quan, Esra Tasali, Michael Twery, Janet B. Croft, Elise Maher, Jerome A. Barrett, Sherene M. Thomas, and Jonathan L. Heald. Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society. *Journal of Clinical Sleep Medicine*, 11(6):591–592, 2015.

[2] Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M. Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital Medicine*, 3(1):1–15, dec 2020.

[3] Conrad Iber, Sonia Ancoli-Israel, A L Chesson, and Stuart Quan. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. *Westchester, IL: American Academy of Sleep Medicine*, 2007.

[4] Anthony Kales and Allan Rechtschaffen. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. United States Government Printing Office, Washington DC, 1968.

[5] Kaare B. Mikkelsen, James K. Ebajemito, Maria A. Bonmati-Carrion, Nayantara Santhi, Victoria L. Revell, Giuseppe Atzori, Ciro della Monica, Stefan Debener, Derk-Jan Dijk, Annette Sterr, and Maarten De Vos. Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. *Journal of Sleep Research*, 28(2), apr 2019.

[6] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, sep 2018.

[7] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, and Jimeng Sun. SLEEPNET: Automated Sleep Staging System via Deep Learning. jul 2017.

[8] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, 4(1):72, dec 2021.

[9] Huy Phan, Oliver Y. Chén, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, jul 2020.

[10] Orestis Tsinalis, Paul M. Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. oct 2016.

[11] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, mar 2017.

[12] Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, apr 2018.

[13] Huy Phan, Oliver Y Chén, Philipp Koch, Zongqing Lu, Ian Mcloughlin, Alfred Mertins, and Maarten De Vos. Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning. *IEEE Transactions on Biomedical Engineering*, 68(6):1787–1798, 2021.

[14] Antoine Guillot and Valentin Thorey. RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1441–1451, 2021.

[15] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, Preben Kidmose, and Maarten De Vos. Personalized automatic sleep staging with single-night data: a pilot study with Kullback–Leibler divergence regularization. *Physiological Measurement*, 41(6):064004, jun 2020.

[16] Xiaolin Hong, Qingqing Zheng, Luyan Liu, Peiyin Chen, Kai Ma, Zhongke Gao, and Yefeng Zheng. Dynamic Joint Domain Adaptation Network for Motor Imagery Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:556–565, 2021.

[17] Guangcheng Bao, Ning Zhuang, Li Tong, Bin Yan, Jun Shu, Linyuan Wang, Ying Zeng, and Zhichong Shen. Two-Level Domain Adaptation Neural Network for EEG-Based Emotion Recognition. *Frontiers in Human Neuroscience*, 0:620, jan 2021.

[18] He Zhao, Qingqing Zheng, Kai Ma, Huiqi Li, and Yefeng Zheng. Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):535–545, feb 2021.

[19] Xin Chai, Qisong Wang, Yongping Zhao, Xin Liu, Ou Bai, and Yongqiang Li. Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Computers in Biology and Medicine*, 79:205–214, dec 2016.

[20] Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):344–353, jun 2020.

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Advances in Computer Vision and Pattern Recognition*, 17(9783319583464):189–209, may 2015.

[22] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep Domain Confusion: Maximizing for Domain Invariance. dec 2014.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, Michael I Jordan, and Jordan@berkeley Edu. Learning Transferable Features with Deep Adaptation Networks. In *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 97–105, 2015.

[24] Samaneh Nasiri and Gari D Clifford. Attentive Adversarial Network for Large-Scale Sleep Staging. Technical report, 2020.

[25] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Adversarial Domain Adaptation with Self-Training for EEG-based Sleep Stage Classification. jul 2021.

[26] Chaehwa Yoo, Hyang Woon Lee, and Jewon Kang. Transferring Structured Knowledge in Unsupervised Domain Adaptation of a Sleep Staging Network. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, aug 2021.

[27] Karl Weiss, Taghi M. Khoshgoftaar, and Ding Ding Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, dec 2016.

[28] Garrett Wilson and Diane J. Cook. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1–46, dec 2018.

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:2962–2971, feb 2017.

[31] Christian O'Reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, 23(6):628–635, dec 2014.

[32] Richard Berry, Rita Brooks, Charlene Gamaldo, Susan Harding, Robin Lloyd, Stuart Quan, Matthew Troester, and Brad Vaughn. AASM Scoring Manual Updates for 2017 (Version 2.4).

Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine, 13, 2017.

[33] Dorien Huysmans, Pascal Borzée, Dries Testelmans, Bertien Buyse, Tim Willemen, Sabine Van Huffel, and Carolina Varon. Evaluation of a Commercial Ballistocardiography Sensor for Sleep Apnea Screening and Sleep Monitoring. *Sensors 2019, Vol. 19, Page 2133*, 19(9):2133, may 2019.

[34] Annette Sterr, James K. Ebajemito, Kaare B. Mikkelsen, Maria A. Bonmati-Carrion, Nayantara Santhi, Ciro della Monica, Lucinda Grainger, Giuseppe Atzori, Victoria Revell, Stefan Debener, Derk-Jan Dijk, and Maarten De Vos. Sleep EEG Derived From Behind-the-Ear Electrodes (cEEGrid) Compared to Standard Polysomnography: A Proof of Concept Study. *Frontiers in Human Neuroscience*, 12:452, nov 2018.

[35] Stefan Debener, Falk Minow, Reiner Emkes, Katharina Gandras, and Maarten de Vos. How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, 49(11):1617–1621, nov 2012.

[36] Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific Reports*, 5(1):1–11, nov 2015.

[37] Pierrick J Arnal, Valentin Thorey, Eden Debellemaniere, Michael E Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, Mounir Chennaoui, and Fabien Sauvet. The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, 43(11):1–13, nov 2020.

[38] Célyne H. Bastien, Annie Vallières, and Charles M. Morin. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep medicine*, 2(4):297–307, 2001.

[39] Elisabeth R. M. Heremans, Huy Phan, Amir H. Ansari, Pascal Borzée, Bertien Buyse, Dries Testelmans, and Maarten De Vos. Feature matching as improved transfer learning technique for wearable EEG. dec 2021.

[40] Jiahao Fan, Hangyu Zhu, Jiang Xinyu, Long Meng, Chen Chen, Cong Fu, Yu Huan, Chenyun Dai, and Wei Chen. Unsupervised Domain Adaptation by Statistics Alignment for Deep Sleep Staging Networks. *TechRxiv*, 2021.