Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

Personalized automatic sleep staging with single-night data: a pilot study with Kullback–Leibler divergence regularization

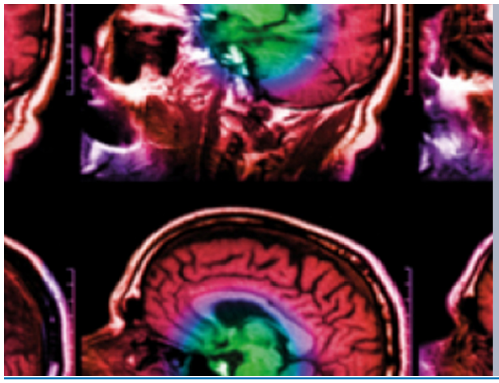
To cite this article: Huy Phan *et al* 2020 *Physiol. Meas.* **41** 064004

View the article online for updates and enhancements.

Recent citations

- Intelligent fault diagnosis of rotating components in the absence of fault data: a transfer-based approach
Minqiang Deng *et al*

- Editorial on Remote Health Monitoring: from chronic diseases to pandemics
Joachim A Behar *et al*



IPEM | IOP
Series in Physics and Engineering in Medicine and Biology

Your publishing choice in medical physics, biomedical engineering and related subjects.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 134.58.253.30 on 12/04/2021 at 12:36

IOP Publishing Physiol. Meas. 41 (2020) 064004 https://doi.org/10.1088/1361-6579/ab921e

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

CrossMark

OPEN ACCESS

RECEIVED
9 January 2020

REVISED
29 April 2020

ACCEPTED FOR PUBLICATION
11 May 2020

PUBLISHED
30 June 2020

Original Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Personalized automatic sleep staging with single-night data: a pilot study with Kullback–Leibler divergence regularization

Huy Phan[1], **Kaare Mikkelsen**[2], **Oliver Y Chén**[3], **Philipp Koch**[4], **Alfred Mertins**[4], **Preben Kidmose**[2] and **Maarten De Vos**[3,5]

[1] School of Electronic Engineering and Computer Science, Queen Mary, University of London, United Kingdom
[2] Department of Engineering, Aarhus University, Denmark
[3] Institute of Biomedical Engineering, University of Oxford, United Kingdom
[4] Institute for Signal Processing, University of Lübeck, Germany
[5] Department of Electrical Engineering, KU Leuven, Belgium

E-mail: h.phan@qmul.ac.uk

Keywords: automatic sleep staging, personalization, transfer learning, single-night data, KL-divergence regularization

Abstract

Objective: Brain waves vary between people. This work aims to improve automatic sleep staging for longitudinal sleep monitoring via personalization of algorithms based on individual characteristics extracted from sleep data recorded during the first night. *Approach*: As data from a single night are very small, thereby making model training difficult, we propose a Kullback–Leibler (KL) divergence regularized transfer learning approach to address this problem. We employ the pretrained SeqSleepNet (i.e. the subject independent model) as a starting point and finetune it with the single-night personalization data to derive the personalized model. This is done by adding the KL divergence between the output of the subject independent model and it of the personalized model to the loss function during finetuning. In effect, KL-divergence regularization prevents the personalized model from overfitting to the single-night data and straying too far away from the subject independent model. *Main results*: Experimental results on the Sleep-EDF Expanded database consisting of 75 subjects show that sleep staging personalization with single-night data is possible with help of the proposed KL-divergence regularization. On average, we achieve a personalized sleep staging accuracy of 79.6%, a Cohen's kappa of 0.706, a macro F1-score of 73.0%, a sensitivity of 71.8%, and a specificity of 94.2%. *Significance*: We find both that the approach is robust against overfitting and that it improves the accuracy by 4.5 percentage points compared to the baseline method without personalization and 2.2 percentage points compared to it with personalization but without regularization.

1. Introduction

The increasing awareness of the important role of sleep in protecting our mental and physical health (Siegel 2005) has been translated in an increasing demand in personal sleep monitoring tools. For such purpose, automating sleep scoring is vital and indispensable since manual scoring is simply too expensive, time-consuming, and labor-intensive (Iber *et al* 2007, Hobson 1969). The advance of machine learning, deep learning in particular, coupled with the availability of large sleep databases (National Sleep Research Resource 2019, O'Reilly *et al* 2014, Stephansen *et al* 2018) has stimulated a new wave of interest in developing automatic sleep staging methods. In fact, machine performance in sleep staging has progressed significantly, being on par with manual scoring made by sleep experts, thanks to recent methods based on deep learning (Phan *et al* 2019c, Supratak *et al* 2017, Stephansen *et al* 2018).

The above-mentioned state-of-the-art performance is only possible using supervised learning. That is, we need data to be recorded and manually labeled from a cohort of subjects, followed by model training based on the labeled data. In fact, the recent expert-level sleep staging performance is only obtainable with a large cohort (i.e. hundreds or thousands of subjects) (Phan *et al* 2019c, Stephansen *et al* 2018). Collecting

© 2020 Institute of Physics and Engineering in Medicine

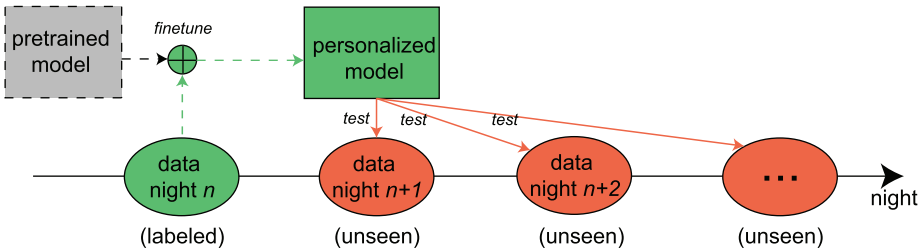


Figure 1. Personalization with single-night data: a pretrained model is finetuned with the labeled data of night n of an individual to yield the personalized model which is tested on the same individual's unseen data of nights $n+1$, $n+2$, ...

and manually scoring a large amount of sleep data is a vast burden, particularly for wearable EEG devices like in-ear EEG (Mikkelsen *et al* 2019b) and around-the-ear EEG (Mikkelsen *et al* 2019a, Sterr *et al* 2018), in which case the workload is increased by the need for an added PSG for reference. Including and utilizing available sleep data for training a sleep staging algorithm in novel settings is not easy, due to channel mismatch caused by differences in channel layouts, electrode placements, recording devices and software, preprocessing procedure, normalization parameters, clinical cohort characteristics, etc. (Phan *et al* 2019c). The work in Phan *et al* (2019d) and Phan *et al* (2019c) proposed a transfer learning approach to circumvent the above-mentioned channel mismatch and enable knowledge transfer from a large dataset to a small cohort, making a deep learning model for a different, specific setting with a small amount of data possible. However, such approach still requires data from dozens of subjects to succeed. Although collecting and labeling this relatively small amount of sleep data is not technically challenging, here we want to push this data constraint further and ask whether it is possible to adapt a pretrained model with single-night data of a particular subject, i.e. personalization, even without knowing in which setting the data is recorded. By personalization, we mean that the parameters of the pretrained model are adapted to an individual's data to convert into a personalized model which is later tested on the same individual's future data. The procedure is illustrated in figure 1. If personalization with single-night data in an unknwon setting is possible, it would be therefore possible for one to build a model for personalized sleep monitoring using his/her own minimal data recorded with a private device. It is equally important and necessary when privacy and security become serious concerns (Agarwal *et al* 2019, Martinovic *et al* 2012, Bonaci *et al* 2015), and thus, owning EEG data from others to form a cohort for transfer learning (Phan *et al* 2019c) would be more and more difficult. An additional and very important benefit of personalization is that it has previously been shown that automatic sleep scoring becomes more accurate when the classifier can focus on the peculiarities of the individual (see Mikkelsen *et al* (2017) and Mikkelsen *et al* (2019b)). This is especially the case when using non-standard EEG montages, for instance in in-ear EEG and around-the-ear EEG. It should be noted that this personalization problem is different from that in Mikkelsen and De Vos (2018) in which a cohort of subjects is known and a model is trained on the cohort before personalizing for a subject in the same cohort. Here, we assume there exists no prior information about the cohort or recording settings but only data from a single night of a target subject is available.

Building a deep-learning model using single-night data is challenging. First, the model can easily overfit the data regardless of whether we train a model from scratch or finetune a pretrained model (Phan *et al* 2019c). Second, different subjects are expected to have varying convergence/overfitting rate when training/finetuning the personalized model. Therefore, we do not know when the model will start overfitting, as we do not have validation data at hand for model selection as in the case of a cohort (Phan *et al* 2019c, Phan *et al* 2019d). Third, regular data normalization cannot be done as a cohort's statistics are unknown. In this work, we aim to tackle this 'personalization with single-night data' challenge using a novel approach based on transfer learning. To that end, we employ the pretrained SeqSleepNet (Phan *et al* 2019a) (i.e. the subject independent (SI) model), and finetune it with single-night data from an individual of an unknown cohort to accomplish personalization. Note that, the source-domain cohort which was used for pretraining the model is also assumedly unknown. To deal with the overfitting problem, Kullback–Leibler (KL)-divergence between the output of the SI model and the personalized model is used to regularize the personalized model. The KL-divergence regularization, in effect, prevents the personalized model from drifting too far away from the SI model. Once the problem of overfitting has been addressed, model selection is no longer an issue as we can keep finetuning the SI model as long as we need. Experiments on 75 subjects of the Sleep-EDF Expanded database (Kemp *et al* 2000, Goldberger *et al* 2000) show that KL-divergence regularized personalization with single-night data is robust against overfitting and achieves an average sleep staging accuracy of 79.6%, improving 4.5 and 2.2 percentage points over non-personalization and personalization without KL-divergence regularization, respectively.

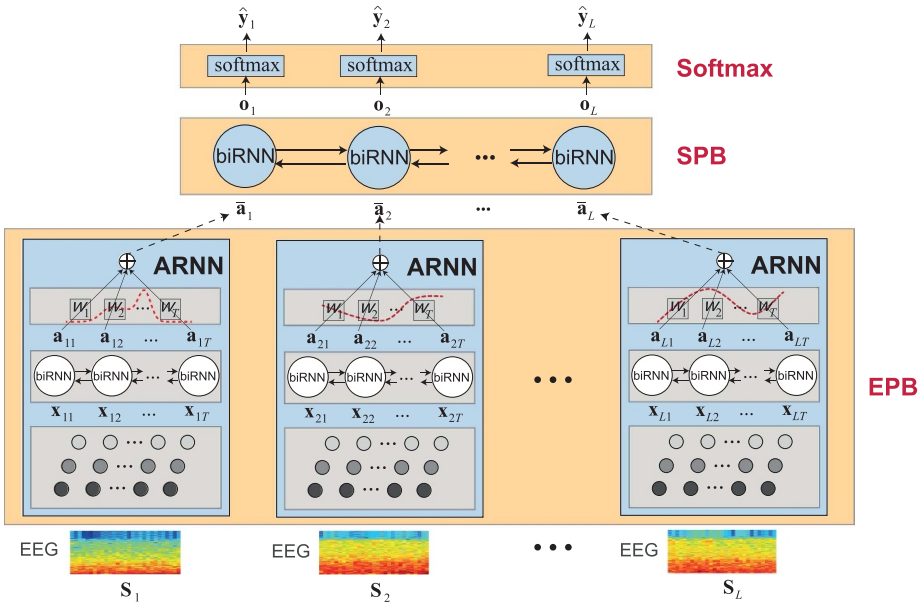


Figure 2. Illustration of SeqSleepNet. The model consists of three components: epoch processing block (EPB), sequence processing block (SPB), and Softmax. © 2019 IEEE. Reprinted, with permission, from Phan *et al* 2019a.

2. Material

We used the Sleep-EDF Expanded database (Sleep Cassette subset, version 2018) (Kemp *et al* 2000, Goldberger *et al* 2000) in this study. This database consists of 78 healthy Caucasian subjects aged 25-101. The database is particularly suitable for this study as there are 75 out of 78 subjects with two consecutive day-night PSG recordings collected for each. Three subjects (subjects 13, 36, and 52) whose one recording was lost due to device failure were excluded from the personalization experiments. Manual scoring was done by sleep experts according to the R&K standard (Hobson 1969) and each 30-second PSG epoch was labeled as one of the eight categories W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN. We merged N3 and N4 into a single stage N3 and excluded MOVEMENT and UNKNOWN categories as done in previous experiments in earlier versions of the database (Imtiaz and Rodriguez-Villegas 2015, Tsinalis *et al* 2016a, Tsinalis *et al* 2016b, Supratak *et al* 2017, Phan *et al* 2019b). We used the Fpz-Cz EEG channel sampled at 100 Hz in this study. As different portions of this database have been used in the literature, it should be stressed that we only made use of the *in-bed* parts (from *lights off* time to *lights on* time) recommended by Imtiaz and Rodriguez-Villegas (2014), Imtiaz and Rodriguez-Villegas (2015) which was adopted in many existing works (Tsinalis *et al* 2016a, Tsinalis *et al* 2016b, Phan *et al* 2019b, Phan *et al* 2019c, Phan *et al* 2018a, Phan *et al* 2018b, Mikkelsen and De Vos 2018, Andreotti *et al* 2018).

3. Methods

3.1. Sequence-to-sequence sleep staging with SeqSleepNet

SeqSleepNet, recently proposed in Phan *et al* (2019a), has demonstrated state-of-the-art performance on several sleep databases (Phan *et al* 2019a, Phan *et al* 2019d) and its suitability for transfer learning tasks (Phan *et al* 2019d, Phan *et al* 2019c). We employ it in this work to study sleep-staging personalization. As a sequence-to-sequence sleep-staging model, SeqSleepNet learns to maximize the joint conditional probability $p(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L \,|\, \mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$. In other words, it receives a sequence of L consecutive epochs $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$ and classifies them at once into a sequence of corresponding sleep stages $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L)$, where $\mathbf{y}_l$, for $1 \leq 1 \leq L$, is a one-hot encoding vector.

To be fed into the network, the EEG signal of a 30-second epoch is transformed into a time-frequency image $\mathbf{S} \in \mathbb{R}^{F \times T}$ obtained via short-time Fourier transform (STFT), where F is the number of frequency bins and T is the number of time instances (cf. section 4). The network is composed of three main components: epoch processing block (EPB), sequence processing block (SPB), and Softmax, as illustrated in figure 2.

EPB. EPB is an attention-based RNN (ARNN) (Phan *et al* 2018b) that is shared by all epochs in the input sequence for short-term (i.e. intra-epoch) sequential modeling. The ARNN subnetwork consists of a *filterbank* layer (Phan *et al* 2018a), a bidirectional RNN realized by long short-term memory (LSTM) cells (Hochreiter and Schmidhuber 1997) with recurrent batch normalization (Cooijmans *et al* 2016), and an

attention layer (Luong *et al* 2015). The trainable filterbank layer with M filters is designed to smooth and reduce the frequency dimension of each epoch $\mathbf{S}$ from F to M, where $M < F$ (Phan *et al* 2018a). The resulting image is then treated as a sequence of T local feature vectors $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ (corresponding to T spectral columns) which is encoded by the bidirectional RNN into a sequence of output vectors $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T)$. The attention layer (Luong *et al* 2015) is trained to produce attention weights $(w_1, w_2, \ldots, w_T)$ and to combine the output vectors into a single feature vector $\bar{\mathbf{a}} = \sum_{t=1}^{T} w_t \mathbf{a}_t$ to represent the epoch $\mathbf{S}$.

SPB. SPB is a bidirectional RNN for long-term (i.e. inter-epoch) sequential modeling. Similar to the RNN in EPB, this RNN is also realized by LSTM cells (Hochreiter and Schmidhuber 1997) with recurrent batch normalization (Cooijmans *et al* 2016). After EPB, the input sequence $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$ is converted into a sequence of feature vectors $(\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \ldots, \bar{\mathbf{a}}_T)$. In turn, the SPB's bidirectional RNN iterates over the sequence of induced feature vectors and encodes it into a sequence of output vectors $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)$.

Softmax. Given the sequence of output vectors $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)$, classification takes place at the Softmax component to produce the sequence of posterior probabilities $(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_L)$, where $\hat{\mathbf{y}}_l$ coresponds to the epoch at index l, $1 \leq l \leq L$, of the input sequence. Similar to the SeqSleepNet+ variant in Phan *et al* (2019c), the softmax layer is shared by all epochs.

The network is trained end-to-end to minimize the sequence classification loss over all N training sequences in the training data:

$$\begin{aligned} E(\Theta) &= -\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\mathbf{y}_{nl}\log\left(\hat{\boldsymbol{y}}_{nl}(\Theta)\right) + \frac{\lambda}{2}\|\Theta\|_2^2 \\ &= -\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}\mathbb{I}(y_{nl}=c)\log P_\Theta(\hat{y}_{nl}=c) + \frac{\lambda}{2}\|\Theta\|_2^2, \end{aligned} \tag{1}$$

where $\mathcal{C} = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ is the set of all possible sleep stages. In (1), $\mathbb{I}(\cdot)$ is the indicator function, y_{nl} and $\hat{y}_{nl}$ denotes the ground-truth and output discrete labels of the lth epoch in the nth sequence, respectively. Θ denotes the trainable parameters of the network and λ is the coefficient of the ℓ_2-norm regularization term.

3.2. KL-divergence regularization for personalization

Given the small amount of data (from one night), it is not feasible to train a deep learning model like SeqSleepNet from scratch. As mentioned before, we, therefore, pursue a transfer learning approach similar to Phan *et al* (2019c), Phan *et al* (2019d) for personalization. We use the SeqSleepNet model from Phan *et al* (2019c), which was pretrained using the C4-A1 EEG data from 200 subjects (686,610 epochs in total) of the Montreal Archive of Sleep Studies (MASS) database (O'Reilly *et al* 2014) (i.e. the source database), as the subject independent (SI) model denoted by Θ. We would like to remind the readers that the MASS cohort is assumedly unknown here. The SI model Θ then serves as the starting point and is finetuned using data from a single night obtained from a target subject to derive the personalized model, denoted by Θ^p, as illustrated in figure 3. Note that channel mismatch is expected between the source-domain MASS database and the target subject's personalization data, and finetuning is supposed to address both channel mismatch and personalization. We investigate four finetuning strategies {*All*, *EPB+Softmax*, *SPB+Softmax*, *Softmax*} similar to those in (Phan *et al* 2019c, Phan *et al* 2019d). When components of the pretrained network (i.e. the entire network, EPB+Softmax, SPB+Softmax, or Softmax depending on the finetuning strategies) are finetuned, their weights are adapted with the personalization data while the rest remains fixed.

The experiments in Phan *et al* (2019c) showed that sleep transfer learning required roughly at least ten subjects' data, leaving personalization with a single-night's data of the target subject exposed to the substantial risk of overfitting. Indeed, we did see empirically that the personalized model tends to overfit the personalization data very easily during experiments. Unfortunately, at present these seems to be no viable way to select the right model during finetuning before overfitting starts. One potential solution here is to leave out a portion of the one-night data for validation. However, since this leave-out validation data is distributed very similarly to the finetuning data, it can also be overfitted easily and thus cannot be used to identify overfitting. To remedy overfitting, we propose to regularize the sequential classification loss function in (1) with the KL divergence between the posterior probability outputs of the SI model Θ and the ones from the personalized model Θ^p, which constrains the personalized model from departing too far from the SI model (Yu *et al* 2013). Formally, given an input sequence $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L)$, KL divergence between the outputs of the two models reads

$$D_{\mathrm{KL}} = \frac{1}{L}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}} P_\Theta(\hat{y}_l = c)\log\left(\frac{P_\Theta(\hat{y}_l = c)}{P_{\Theta^p}(\hat{y}_l = c)}\right). \tag{2}$$

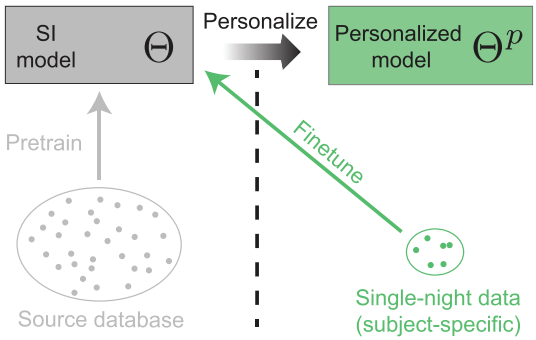


Figure 3. Illustration of sleep personalization using data from a single night. The subject independent (SI) model Θ is pretrained with a source-domain database (assumedly unknown) Subsequently, it finetuned using the single-night data of a target subject to derive the personalized model Θ^p.

The KL-divergence regularization is added into the sequential classification loss function in (1) to form the loss function for personalization:

$$E(\Theta^p) = -(1-\alpha)\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}\mathbb{I}(y_{nl}=c)\log P_{\Theta^p}(\hat{y}_{nl}=c) + \frac{\lambda}{2}\|\Theta^p\|_2^2 + \alpha\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}P_{\Theta}(\hat{y}_{nl}=c)\log\left(\frac{P_{\Theta}(\hat{y}_{nl}=c)}{P_{\Theta^p}(\hat{y}_{nl}=c)}\right), \quad (3)$$

where $\alpha \in [0, 1]$ is the KL-divergence regularization coefficient, regulating how far the personalized model Θ^p deviates from the SI model Θ. When $\alpha = 0$, the KL-divergence regularization is cancelled out and the personalization turns out to be the same as regular finetuning in Phan *et al* (2019c), Phan *et al* (2019d). In this case, the pretrained SI model is adapted solely on the personalization data. In contrast, when $\alpha = 1$, we trust the pretrained SI model completely and ignore all new information from the personalization data. Since the term $\alpha\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}P_{\Theta}(\hat{y}_{nl}=c)\log P_{\Theta}(\hat{y}_{nl}=c)$ in the KL-divergence regularization term in (3) does not depend on the personalized network Θ^p, the KL-divergence regularized loss function can be simplified as

$$E'(\Theta^p) = -(1-\alpha)\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}\mathbb{I}(y_{nl}=c)\log P_{\Theta^p}(\hat{y}_{nl}=c) + \frac{\lambda}{2}\|\Theta^p\|_2^2 - \alpha\frac{1}{L}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{c\in\mathcal{C}}P_{\Theta}(\hat{y}_{nl}=c)\log P_{\Theta^p}(\hat{y}_{nl}=c). \quad (4)$$

It turns out that the loss function for personalization in (4) consists of two components: (1) the cross-entropy between the output of the personalized model Θ^p and the ground-truth, and (2) the cross-entropy between the output of the personalized model Θ^p and the output of the pretrained SI model Θ. Thus, model personalization is equivalent to changing the target distribution from the unknown source-domain database (the MASS database used for pretraining) to a linear interpolation of the source-domain data distribution and the personalized data distribution (Yu *et al* 2013). This interpolation prevents the network from overfitting the personalization data.

4. Experimental setup

For each of the 75 subjects with two day-night recordings of the Sleep-EDF Expanded database, we conducted finetuning of the pretrained SeqSleepNet (Phan *et al* 2019c) using data from the first night and evaluating the personalized model on data from the second night. We experimented with different values for the KL-divergence regularization coefficient α in the set of {0, 0.2, 0.4, 0.6, 0.8} to investigate its influence. Note that, when $\alpha = 0$, we excluded the KL-divergence regularization completely. This case is used as the baseline for comparison with the proposed approach.

The EEG signal was divided into 30-second epochs. Each epoch was transformed into a log-magnitude time-frequency image using the following procedure: the signal was divided into two-second windows with

Table 1. Performance on regular (scratch) training setup via 10-fold cross validation.

System	Data portion	Overall metrics				
		Acc.	κ	MF1	Sens.	Spec.
SeqSleepNet	*in-bed* only	79.1	0.708	74.6	74.2	94.2
DeepSleepNet (Supratak *et al* 2017)	*in-bed* only	78.5	0.702	75.3	75.0	94.1
SeqSleepNet	*in-bed* $\pm$ 30 min	82.6	0.760	76.4	76.3	95.4
SleepEEGNet (Mousavi *et al* 2019)	*in-bed* $\pm$ 30 min	80.0	0.730	73.6	–	–

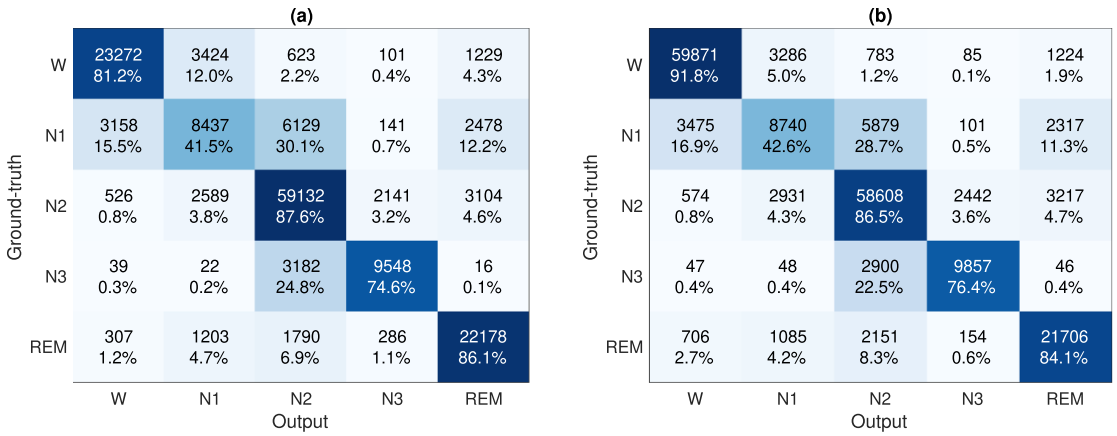


Figure 4. Confusion matrices obtained by SeqSleepNet. (a) *in-bed* data only, (b) *in-bed* data $\pm$ 30 min.

50% overlap, multiplied with a Hamming window, transformed to the frequency domain by means of a 256-point fast Fourier transform (FFT), and the amplitude spectrum was log-transformed. This resulted in an image of size $F \times T$ where $F = 129$ (the number of frequency bins) and $T = 29$ (the number of spectral columns).

5. Results

5.1. SeqSleepNet's performance on regular training setting

SeqSleepNet requires the data to be normalized to zero mean and unit standard deviation (Phan *et al* 2019a, Phan *et al* 2019c). Unfortunately, in our case neither the source-domain cohort (i.e. the MASS cohort) nor the target subject's cohort (i.e. the Sleep-EDF cohort) is known. We, therefore, cannot normalize the personalization data using the cohort's statistics. In addition, we found that model personalization is sensitive to differences in magnitude of data between two nights, and subject-specific data normalization resulted in poor performance in some subjects with such substaintial magnitude difference. To rule out this difference, we performed night-specific *normalization* in which data of one night recording was normalized by its mean and standard deviation.

The implementation was based on the Tensorflow framework (Abadi *et al* 2016). The pretrained SeqSleepNet was parametrized similarly to the one in Phan *et al* (2019c) and used a sequence length of $L = 20$. For personalization, the pretrained SeqSleepNet was finetuned on a single-night's finetuning data for 50 finetuning epochs and the performance was recorded every 5 (finetuning) epochs. Finetuning was performed using the Adam optimizer (Kingma and Ba 2015) with a learning rate of 10^{-4}.

SeqSleepNet has been reported to achieve state-of-the-art performance on the MASS database (O'Reilly *et al* 2014) (i.e. the source domain used for pretraining) and the earlier version of the Sleep-EDF Expanded database with 20 subjects (Kemp *et al* 2000, Goldberger *et al* 2000). It is worth assessing its performance on the experimental database on a regular (scratch) training setup. To this end, we conducted 10-fold cross-validation on all 78 subjects. At each iteration, seven subjects were left out for validation (i.e. model selection). During training, the network that achived the best overall accuracy on the validation subjects was retained for evaluation on the test subjects. The results of 10 cross-validation folds were pooled to calculate the overall metrics, including accuracy, macro F1-score (MF1) (Yang and Liu 1999), Cohen's kappa (κ) (McHugh 2012), sensitivity, and specificity. Beside SeqSleepNet, we also implemented the end-to-end variant of the popular DeepSleepNet (Supratak *et al* 2017, Phan *et al* 2019a) for comparison. In addition, we included results for another common usage of the database in which 30 minutes of data before and after

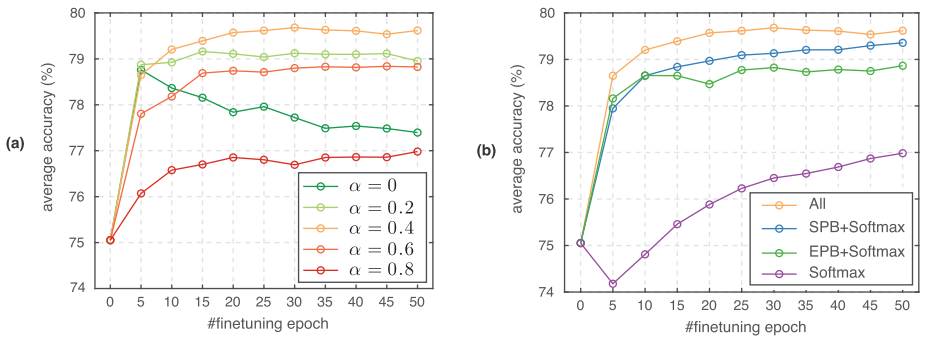


Figure 5. (a) Variation of average accuracy of 75 target subjects during finetuning (*All* strategy) with different values of α . (b) Variation of average accuracy of 75 target subjects during finetuning with different finetuning strategies (α was fixed to 0.4).

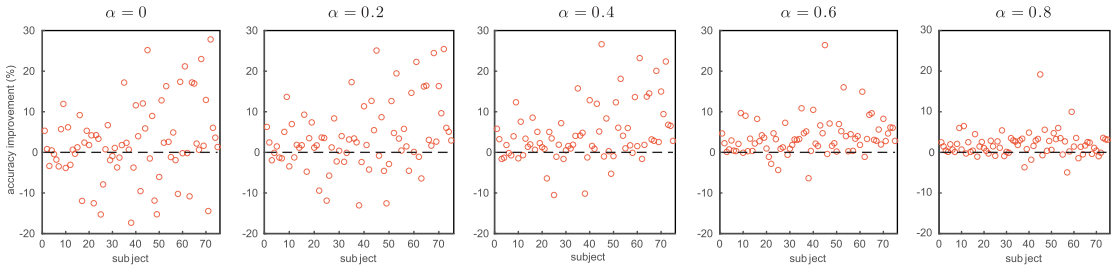


Figure 6. Individual accuracy improvements of 75 target subjects after 50 finetuning epochs when α takes different values in {0, 0.2, 0.4, 0.6, 0.8} (*All* finetuning strategy was employed).

in-bed parts are additionally included. The experimental results are shown in table 1 in which SeqSleepNet not only obtains better performance than the DeepSleepNet but also outperforms recent results in Mousavi *et al* (2019) on this latest version of the Sleep-EDF Expanded database. The accuracy of the sleep stages is also shown in the confusion matrices in figure 4.

5.2. Influence of KL-divergence regularization

It should be emphasized again that, different from the regular-setting experiment in Secion 5.1, only 75 subjects with two recordings were used for the personalization experiment and three subjects with one recording were excluded. The effect of KL-divergence regularization in avoiding overfitting for model personalization is exhibited in figure 5(a) when α takes different values in {0, 0.2, 0.4, 0.6, 0.8}. Without KL-divergence regularization (i.e. $\alpha = 0$), the average accuracy of the personalized models on 75 target subjects starts declining after five finetuning epochs during which the models most likely start overfitting the personalization data. The overfitting appears to get worse and worse with ongoing finetuning process as the average accuracy keeps decreasing. When being regularized with KL divergence (i.e. $\alpha > 0$), the pattern of the average accuracy curve is gradually reversed when α increases, exhibiting a negligible downward tendency when $\alpha = 0.2$, plateauing after 25 finetuning epochs with $\alpha = 0.4$, and trending upward with larger values for α.

The results in figure 5(a) also indicate that α plays the role of a trade-off parameter between the pretrained SI model and the purely personal model. When α is set small, we allow the personalized model to aggressively fit to the personalization data at the risk of severe overfitting. In contrast, when α is large, the personalized model is conservatively tied to the SI model and has less freedom to adapt to the personalization data. By doing so, it effectively avoids overfitting, although at cost of ignoring some individual-specific information. This argument is strengthened with the results in figure 6. In this figure, the individual accuracy improvements of 75 target subjects vary widely around the zero line when $\alpha = 0$ and become more and more concentrated towards zero with increasing values of α. For this dataset, it seems that a value around 0.4 is a reasonable choice for α.

Table 2 further provides a comparison of average performance obtained by personalization with different values of α and that before personalization. After personalization, the best performance is obtained with $\alpha = 0.4$, reaching an accuracy of 79.6% and improving over that of personalization without KL-divergence regularizaion and that of no-personalization by 2.2 and 4.5 percentage points absolute, respectively. Significant improvement on accuracy can also be seen from the confusion matrices in figure 7 for most of the

Table 2. Average sleep staging performance before and after personalization. Personalization without KL-divergence regularization corresponds to $\alpha = 0$ and personalization with KL-divergence regularization corresponds to $\alpha > 0$. *All* finetuning was employed and personalization was run for 50 finetuning epochs.

		Overall metrics				
		Acc.	κ	MF1	Sens.	Spec.
Before personalization		75.1 ± 11.2	0.648 ± 0.140	67.2 ± 11.4	69.7 ± 11.4	93.1 ± 2.8
After personalization	$\alpha = 0$	77.4 ± 10.0	0.677 ± 0.131	71.4 ± 9.7	69.6 ± 10.8	93.6 ± 2.6
	$\alpha = 0.2$	79.0 ± 8.4	0.697 ± 0.114	72.5 ± 8.9	71.2 ± 10.2	94.0 ± 2.3
	$\alpha = 0.4$	$\mathbf{79.6 \pm 8.4}$	$\mathbf{0.706 \pm 0.113}$	$\mathbf{73.0 \pm 8.8}$	$\mathbf{71.8 \pm 10.1}$	$\mathbf{94.2 \pm 2.2}$
	$\alpha = 0.6$	78.8 ± 10.0	0.697 ± 0.128	72.0 ± 10.0	71.6 ± 10.9	94.0 ± 2.5
	$\alpha = 0.8$	77.0 ± 10.9	0.672 ± 0.138	69.2 ± 12.0	70.2 ± 11.8	93.5 ± 2.7

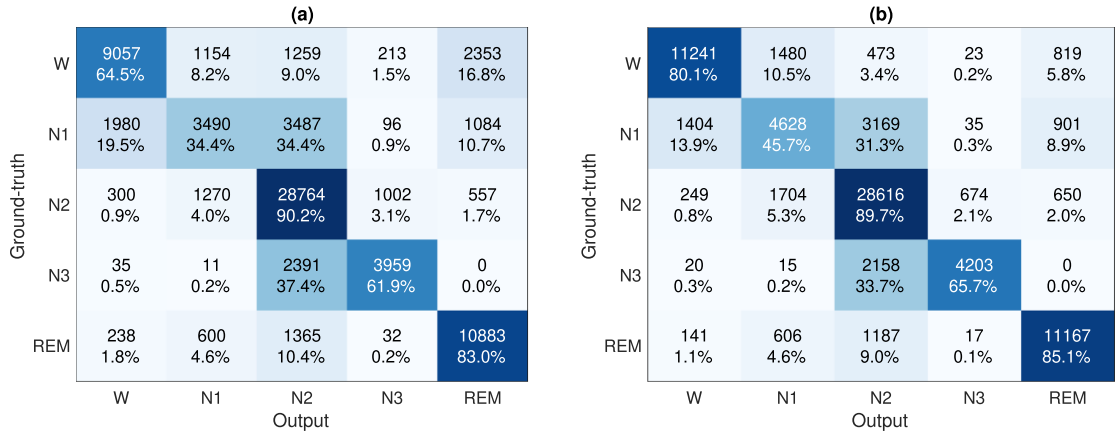


Figure 7. Confusion matrices obtained by SeqSleepNet before and after personalization. (a) Before personalization, (b) after personalization.

sleep stages. Furthermore, this accuracy level is similar to that of the model trained on the entire (known) cohort in table 1 even though only data from a single-night of the subjects was used and the cohort was unknown.

5.3. Influence of finetuning strategies

It was shown in Phan *et al* (2019c), Phan *et al* (2019d) that, in sleep transfer learning, it is important to finetune feature-learning parts of a pretrained network to overcome the channel mismatch between a source domain and a target domain. This principle also applies to personalization as shown in figure 5(b). Although finetuning the Softmax component alone improves the performance, the improvement is significantly lower than the ones obtained by other finetuning strategies in which the feature-learning components of the pretrained SeqSleepNet (i.e. EPB or SBP or both) and the Softmax component are collectively adapted. For instance, the *All* finetuning strategy produces an accuracy improvement of 4.6 percentage points which is more than twice as much as the 1.9 percentage points obtained using the *Softmax* finetuning strategy after 50 finetuning epochs.

5.4. To personalize or not personalize?

In sleep transfer learning, when there is a mismatch between the source domain (the MASS databased used for pretraining in our case) and the target domain (the personalization data in our case), it is vital to perform some form of finetuning. In case of personalization, besides possible discrepancies between the source-domain data and the personalization data (Phan *et al* 2019c), this data mismatch is further topped up with the target subject's peculiarities. On the contrary, when there is no data mismatch, finetuning could be omitted because no significant improvement is expected and one faces an increasing risk of overfitting. If there is a way to determine whether data distributions mismatch, one can decide to personalize the sleep staging model or not. Fortunately, we have access to the ground truth of a target subject's one-night data which can be utilized to assess the performance of the pretrained SI model. If the pretrained SI model performs well on this data, the personalization data distribution is very likely matched to the source-domain data distribution. Reversely, poor performance of the pretrained SI model on this personalization data is an indicator of data mismatch.

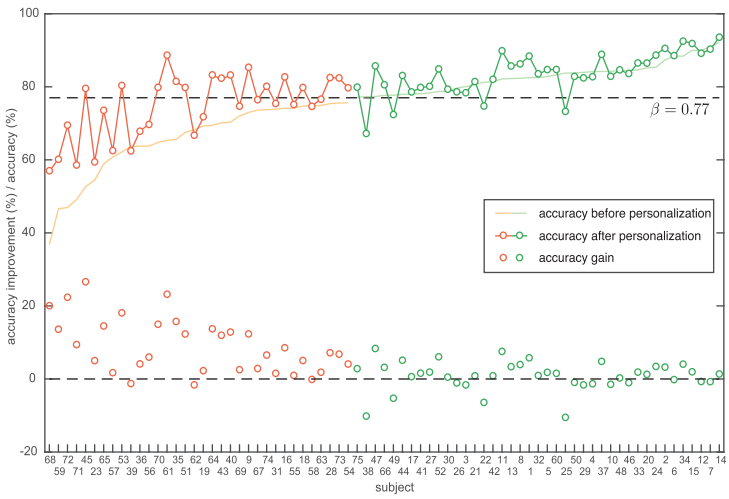


Figure 8. Individual accuracies before personalization, individual accuracies after personalization, and individual accuracy improvements of 75 target subjects (sorted by increasing accuracy before personalization). Group A (i.e. subjects with accuracy before personalization below β) is marked in orange and Group B (i.e. subjects with accuracy before personalization equal or greater than β) are marked in green. *All* finetuning was employed, α was fixed to 0.4, and personalization was run for 50 finetuning epochs.

In light of this observation, we applied a threshold β to the individual accuracy obtained on the data of the first night to group 75 target subjects into two groups: Group A, which consists of subjects with accuracy before personalization below β and Group B, which consists of subjects with accuracy before personalization equal or above β. Figure 8 shows the individual accuracies before and after personalization, and the individual accuracy improvements of the subjects in both groups with $\beta = 0.77$. As can be seen, most significant accuracy improvements correspond to the subjects in Group A while those improvements of the subjects in Group B are much more subtle. On average, personalization for Group A's subjects results in an average improvement of 9.0 percentage points, ten times larger than that for Group B's subjects which is 0.9 percentage points.

6. Discussion

The personalization results in figure 8 reveal uneven distribution of accuracy improvement across subjects. Those subjects on which the pretrained SI model performs poorly (i.e. severe data mismatch) benefit the most from personalization. However, only modest improvements were seen for those subjects on which the pretrained SI model performs well, despite the fact that there is a similar channel mismatch: the C4-A1 EEG channel was used for pretraining the SI model and the Fpz-Cz EEG channel was used for personalization data. We speculate that personalization will be crucial for all target subjects when a completely different channel layout is used, for example in-ear EEG (Mikkelsen *et al* 2019b) and around-the-ear EEG (Mikkelsen *et al* 2019a, Sterr *et al* 2018).

Setting a right value for the coefficient α was shown to play an important role in personalization performance. Although we have studied a common α for all target subjects and fixed its value during the personalization process, it is beneficial if α can be made adaptive. For example, for subjects with significant peculiarities (e.g. those subjects in Group A in section 5.4), one should start with a large α to impose strong personalization inititally and reduce it along the personalization process to gradually lower this risk. The amount of personalization data should also be taken into account when setting the value for the KL-divergence regularization coefficient α. As a matter of fact, using single-night data for personalization is convenient. However, when more data is available, improvement on personalization performance can be expected. In intuition, α should be proportional to the amount of personalization data, i.e. we should use a small α for small personalization data (we trust the SI model more) and a large α for large personalization data (we trust the personalization data more).

7. Conclusions

We introduced the problem of sleep-staging personalization with data from a single night and discussed its benefits and challenges in the context of personal sleep monitoring. We then tacked this problem using transfer learning. The subject independent (SI) model (i.e. the pretrained SeqSleepNet) was used as the

starting point and finetuned on a single-night's data of a target subject to accomplish personalization. KL-divergence between the personalized model's and the SI model's outputs is used to regularize the network's loss function during personalization. The regularization anchors the personalized model, effectively preventing it from overfitting to the personalization data. Experiments on data of 75 subjects from the Sleep-EDF Expanded database, we demonstrated that sleep personalization with a single-night's data is possible. Additionally, we showed that personalization with KL-divergence regularization is useful to prevent model overfitting and achieves more favorable results compared to the baseline model without personalization and the model with personalization but without regularization.

To sum up we demonstrated that automatic sleep staging with a single-night's data is possible with encouraging results. However, while the number of subjects, at 75, is decently high, the population could still be considered quite homogeneous, which could impact the results shown. In addition, the database used in this study was labeled according to the old R&K guidelines (Hobson 1969) rather than the new and more robust AASM ones (Iber *et al* 2007) This could introduce biases to the results. Larger databases with heterogeneous characteristics (e.g. different demographics, sleep diseases, and electrode placements, etc.) is desirable for furture work. Such databases should be labeled (or re-relabeld by an independent sleep technician) according to the AASM guidelines (Iber *et al* 2007).

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

ORCID iD

Preben Kidmose https://orcid.org/0000-0001-8628-8057

References

Abadi M *et al* 2016 Tensorflow: Large-scale machine learning on heterogeneous distributed systems *arXiv:1603.04467*

Agarwal A, Dowsley R, McKinney N D, Wu D, Lin C-T, De Cock M and Nascimento A C A 2019 Protecting privacy of users in brain-computer interface applications *IEEE Trans. Neural Syst. Rahabil. Eng.* **27** 1546–55

Andreotti F, Phan H, Cooray N, Lo C, Hu M T M and De Vos M 2018 Multichannel sleep stage classification and transfer learning using convolutional neural networks *Proc. EMBC* pp 171–4

Bonaci T, Calo R and Chizeck H J 2015 App stores for the brain: Privacy & security in brain-computer interfaces *IEEE Technol. Soc. Mag.* **34** 32–9

Cooijmans T, Ballas N, Laurent C, Gülçehre Ç and Courville A 2016 Recurrent batch normalization *arxiv:1603.09025*

Goldberger A L *et al* 2000 Physiobank, physiotoolkit and physionet: Components of a new research resource for complex physiologic signals *Circulation* **101** e215–e220

Hobson J A 1969 A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects *Electroencephalogr. Clin. Neurophysiol.* **26** 644

Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80

Iber C, Ancoli-Israel S, Chesson A L and Quan S F 2007 The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications *American Academy of Sleep Medicine*

Imtiaz S A and Rodriguez-Villegas E 2014 Recommendations for performance assessment of automatic sleep staging algorithms *Proc. EMBC* pp 5044–7

Imtiaz S A and Rodriguez-Villegas E 2015 An open-source toolbox for standardized use of PhysioNet Sleep EDF Expanded Database *Proc. EMBC* pp 6014–17

Kemp B, Zwinderman A H, Tuk B, Kamphuisen H A C and Oberye J J L 2000 Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG *IEEE Trans. Biomed. Eng.* **47** 1185–94

Kingma D P and Ba J L 2015 Adam: a method for stochastic optimization *Proc. ICLR* pp 1–13

Luong T, Pham H and Manning C D 2015 Effective approaches to attention-based neural machine translation *Proc. EMNLP* pp 1412–21

Martinovic I, Davies D, Frank M, Perito D, Ros T Song D 2012 On the feasibility of side-channel attacks with brain-computer interfaces *Proc. 21st USENIX Conf. Secur. Symp.* pp 143–58

McHugh M L 2012 Interrater reliability: the kappa statistic *Biochemia Medica*

Mikkelsen K B, Villadsen D B, Otto M and Kidmose P 2017 Automatic sleep staging using ear-EEG *BioMed. Eng. OnLine* **16** 111

Mikkelsen K and De Vos M 2018 Personalizing deep learning models for automatic sleep staging *arXiv Preprint arXiv:1801.02645*

Mikkelsen K B *et al* 2019a Machine learning derived sleep–wake staging from around the ear electroencephalogram outperforms manual scoring and actigraphy *J. Sleep Res.* **28** e12786

Mikkelsen K B, Tabar Y R, Kappel S L, Christensen C B, Toft H O, Hemmsen M C, Rank M L, Otto M and Kidmose P 2019b Accurate whole-night sleep monitoring with dry-contact ear-EEG *Sci. Rep.* **9** 1–12

Mousavi S, Afghah F and Acharya R 2019 SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach *PLoS One* **14** e0216456

NHLBI 2020 National sleep research resource: free research data and tools (http://sleepdata.org/) (Accessed: 1 Jan 2020)

O'Reilly C, Gosselin N, Carrier J and Nielsen T 2014 Montreal archive of sleep studies: An open-access resource for instrument benchmarking & exploratory research *J. Sleep Res.* **23** 628–35

Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2018a DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification *Proc. EMBC* pp 453–6
Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2018b Automatic sleep stage classification using single-channel EEG: learning sequential features with attention-based recurrent neural networks *Proc. EMBC* pp 1452–5
Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2019a SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging *IEEE Trans. Neural Syst. Rehabil. Eng. (TNSRE)* **27** 400–10
Phan H, Andreotti F, Cooray N, Chén O Y and De Vos M 2019b Joint classification and prediction CNN framework for automatic sleep stage classification *IEEE Trans. Biomed. Eng. (TBME)* **66** 1285–96
Phan H, Chén O Y, Koch P, Lu Z, McLoughlin I, Mertins A and De Vos M 2019c Towards more accurate automatic sleep staging via deep transfer learning arXiv:1907.13177
Phan H, Chén O Y, Koch P, Mertins A and De Vos M 2019d Deep transfer learning for single-channel automatic sleep staging with channel mismatch *Proc. 27th European Conf. Signal Process. (EUSIPCO)* pp 1–5
Siegel J M 2005 Clues to the functions of mammalian sleep *Nature* **437** 1264–71
Stephansen J B *et al* 2018 Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy *Nat. Commun.* **9** 5229
Sterr A *et al* 2018 Sleep EEG derived from behind-the-ear electrodes (ceegrid) compared to standard polysomnography: a proof of concept study *Front. Human Neurosci.* **12** 452
Supratak A, Dong H, Wu C and Guo Y 2017 DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG *IEEE Trans. Neural Syst. Rehabil. Eng.* **25** 1998–2008
Tsinalis O, Matthews P M, Guo Y and Zafeiriou S 2016a Automatic sleep stage scoring with single-channel EEG using convolutional neural networks *arXiv:1610.01683*
Tsinalis O, Matthews P M and Guo Y 2016b Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders *Ann. Biomed. Eng.* **44** 1587–97
Yang Y and Liu X 1999 A re-examination of text categorization methods *Proc. SIGIR* **99** 42–9
Yu D, Yao K, Su H, Li G and Seide F 2013 Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition *Proc. ICASSP* pp 7893–7