

Multi-Channel Multi-Domain based Knowledge Distillation Algorithm for Sleep Staging with Single-Channel EEG

Chao Zhang[✉], Yiqiao Liao, Siqi Han, Milin Zhang[✉], *Senior Member, IEEE*,
Zhihua Wang[✉], *Fellow, IEEE*, Xiang Xie[✉]

Abstract—This paper proposed a Multi-Channel Multi-Domain (MCMD) based knowledge distillation algorithm for sleep staging using single-channel EEG. Both knowledge from different domains and different channels are learnt in the proposed algorithm, simultaneously. A multi-channel pre-training and single-channel fine-tuning scheme is used in the proposed work. The knowledge from different channels in the source domain is transferred to the single-channel model in the target domain. A pre-trained teacher-student model scheme is used to distill knowledge from the multi-channel teacher model to the single-channel student model combining with output transfer and intermediate feature transfer in the target domain. The proposed algorithm achieves a state-of-the-art single-channel sleep staging accuracy of 86.5%, with only 0.6% deterioration from the state-of-the-art multi-channel model. There is an improvement of 2% compared to the baseline model. The experimental results show that knowledge from multiple domains (different datasets) and multiple channels (e.g. EMG, EOG) could be transferred to single-channel sleep staging.

Index Terms—Sleep staging, Transfer learning, Knowledge distillation, Single-channel EEG, Brain-computer interface

I. INTRODUCTION

SLEEP staging is an essential technique for sleep-related disease diagnosis and treatment. According to the sleep staging definition by the American Academy of Sleep Medicine (AASM), there are 5 sleep stages: Wake, Non-Rapid Eye Movement 1 (N1), N2, N3, and Rapid Eye Movement (REM). The golden standard for sleep staging is manual labelling on Polysomnography (PSG) signals by doctors. The PSG signals consist of Electroencephalography (EEG), Electrooculography (EOG), Electromyography (EMG), and Electrocardiogram (ECG). A considerable amount of literature

[1–4] has been published on automatic sleep staging using PSG signals with feasible accuracy. Seqsleepnet [1] achieved a state-of-the-art accuracy based on multi-channel signals for sleep staging using a sequence-to-sequence hierarchical recurrent neural network (RNN). However, the acquisition of PSG signals involves bulky equipment, which limits the potential scenarios that can be applied.

In recent years, various wearable sleep monitoring devices have been released using single-channel EEG signal for sleep status analysis. Different single-channel EEG signal based sleep staging algorithms have been reported in literature [5–16]. DeepSleepNet [5] combines the time-invariant features from Convolutional Neural Network (CNN) and temporal features from Bidirectional Long Short-Term Memory (Bi-LSTM) for single-channel sleep staging. However, the insufficiency of the single-channel data for training is a big issue in improving the accuracy. The widely applied deep learning models are easy to be over-fitting while the dataset is too small. As the amount of dataset is fixed, the practical method to increase the available training data is to either introduce information from other domains into the training set or take the typical ignored information into account.

In order to introduce information from other domains into the training set, one choice is to get more knowledge from a bigger dataset. [17] proposed a Conditional Wasserstein Generative Adversarial Network (GAN) framework to generate EEG data for data augmentation, but the GAN failed to create knowledge that does not exist in the dataset. Recently, pre-trained representation model such as Bidirectional Encoder Representations from Transformers (BERT) [18] achieved state-of-the-art performance for natural language processing tasks. Following this trend, [6] proposed a transfer learning approach to transfer knowledge from a large dataset to a small cohort. The model was pre-trained in the source domain, and then was fine-tuned in the target domain. It achieved the state-of-the-art accuracy in single-channel sleep staging. However, the cross-channel knowledge transfer is ignored.

Another promising solution is to utilize knowledge from the ignored channels. The concept of knowledge distillation [19] was proposed for model compression. The teacher model, M_T , usually features higher accuracy but with higher complexity, while the student model, M_S , features lower accuracy with a

Chao Zhang and Yiqiao Liao contributed equally to this work.

Chao Zhang and Milin Zhang are with the Department of Electronic Engineering, Institute for Precision Medicine, and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China, 100084. Corresponding author e-mail: zhangmilin@tsinghua.edu.cn.

Yiqiao Liao, Zhihua Wang and Xiang Xie are with the School of Integrated Circuits, Tsinghua University, Beijing, China, 100084.

Siqi Han is with the School of Modern Post (School of Automation), Beijing University of Posts and Telecommunications, Beijing, China, 100876.

This work is supported in part by the National Key Research and Development Program of China (No.2018YFB220200*), in part by the Natural Science Foundation of China through grant 92164202, in part by the Beijing Innovation Center for Future Chip, in part by the Beijing National Research Center for Information Science and Technology.

Digital Object Identifier

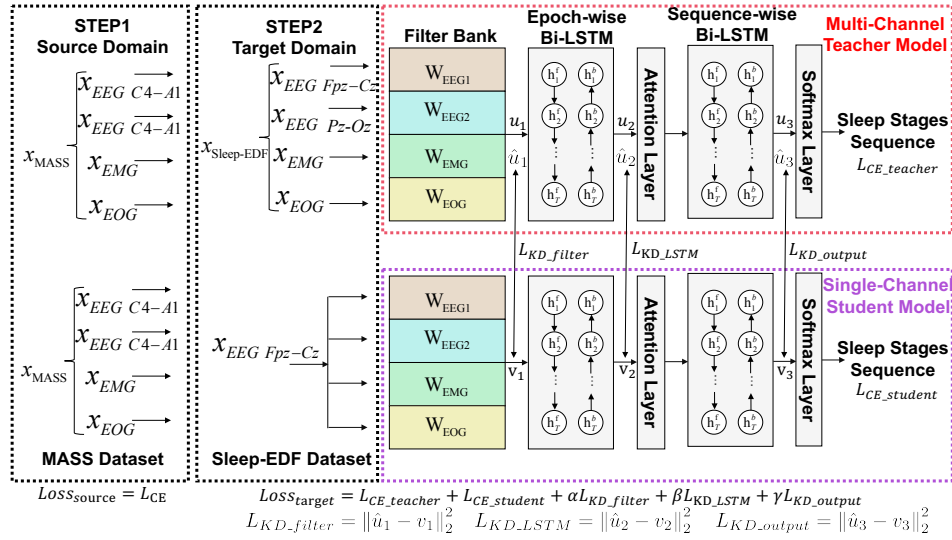


Fig. 1. The overall architecture of the proposed MCMD algorithm. It consists of two steps: 1) Source Domain Pre-training, 2) Target Domain Knowledge Distillation. In the step 1, four-channel model M_0 is trained using the three-channel MASS dataset. In the step 2, M_0 is used to initialize the teacher model M_T and the student model M_S . M_T takes multi-channel signals $x_{sleep-edf}$ as its input, while M_S only takes $x_{EEGFpz-Cz}$ as its input. The teacher and student models are trained simultaneously in the Sleep-EDF dataset while knowledge distillation is utilized between the filterbank, LSTM, and output of these two models.

light-weight architecture. M_S learns from M_T with satisfactory accuracy and appropriate complexity. Our previous work [7] proposed a competition and cooperation based knowledge distillation model, where M_T is a multi-channel model, and M_S is a single-channel model. It enhances the performance of single-channel EEG with knowledge transfer between channels. Yet, the knowledge in other domains is ignored.

This paper proposed a Multi-Channel Multi-Domain (MCMD) based knowledge distillation algorithm for single-channel EEG based sleep staging. The domains consist of the MASS dataset [20] as source domain and the Sleep-EDF dataset [21, 22] as target domain. The channels consist of EEG, EMG and EOG channels. The proposed algorithm combines knowledge transfer in four different scenarios: Same-Domain Same-Channel (SDSC), Same-Domain Cross-Channel (SDCC), Cross-Domain Same-Channel (CDSC) and Cross-Domain Cross-Channel (CDCC). It consists of two steps: 1) Source domain pre-training. 2) Target domain knowledge distillation. In the first step, the pre-trained model from the source domain was used for initializing the teacher and the student models. Both CDSC and CDCC transfer was employed for those models. In the second step, knowledge distillation was applied from the multi-channel teacher model to the single-channel student model with the combination of output transfer, feature transfer, and filterbank transfer. SDSC and SDCC knowledge transfer was employed in this step. The proposed algorithm achieves an accuracy of 86.5%, which is higher than the previous single-channel sleep staging works reported on literature. There is only a 0.4% deterioration from our multi-channel teacher model, as well as a 2% improvement compared with the baseline Seqsleepnet [1]. The experimental results show the effectiveness of the four knowledge transfer

scenarios. Knowledge from different channels, or even different domains could be used in single-channel sleep staging.

The rest of this paper is organized as follows: Section II introduces the proposed MCMD knowledge distillation algorithm with the experimental results shown in Section III, while Section IV concludes our work.

II. MCMD BASED KNOWLEDGE DISTILLATION

Fig.1 illustrates the architecture of the proposed MCMD knowledge distillation algorithm. The training of the proposed algorithm consists of two steps: 1) Source domain pre-training. 2) Target domain knowledge distillation. The proposed algorithm takes four-channel signals from the source domain (MASS dataset) and the target domain (Sleep-EDF dataset) for training. It employs a multi-channel teacher model and a single-channel student model for sleep staging in the target domain.

The Seqsleepnet model is chosen for M_T and M_S . The input of the model is time-frequency representations of 30s PSG epochs. A short-time Fourier transform is applied to transform the 30s PSG raw data into power spectra with a number of frequency bins, F , of 129, a number of time indices, T , of 29, and a number of channels, C , of 4. The structure of the model consists of filterbank layers, epoch-wise Bi-LSTM with a input length of the frame number in one epoch, attention layers, sequence-wise Bi-LSTM with a input length of the epoch number in one sequence, and Softmax layers.

A. Source Domain Pre-Training

The source domain is MASS dataset consisting of EEG C4-A1, EMG, and EOG. The target domain is Sleep-EDF consisting of EEG Fpz-Cz, EEG Pz-Oz, EMG, and EOG. The EEG channel C4-A1 is duplicated for the corresponding of the

EEG Pz-Oz channel in the target domain, which eliminates the channel number mismatch and increases the number of cross domain knowledge transfer paths. The Seqsleepnet model M_0 is trained with sequence Cross-Entropy loss function $H(y, p)$, which is defined as:

$$H(y, p) = -\frac{1}{N_b} \frac{1}{L} \sum_{k=1}^{N_b} \sum_{j=1}^L \sum_{i=1}^5 y_{ji}^k \log(p_{ji}^k) \quad (1)$$

where y_{ji}^k denotes the i^{th} sleep stage in the one-hot ground truth label of the j^{th} sample in the k^{th} sequence of the corresponding batch. L is the length of a sequence, which is set as 30. N_b is the size of one batch. p is the prediction outputs after Softmax activation which can be calculated as:

$$p = \text{softmax}(u(x_{C4-A1}, x_{C4-A1}, x_{EMG}, x_{EOG})) \quad (2)$$

where $u(*)$ denotes the last hidden layer output from M_0 . The best Seqsleepnet model M_0 for the four channels with parameters W_0 can be found as:

$$W_0 = \underset{W_0}{\text{argmin}}(H(y, p(x_{C4-A1}, x_{C4-A1}, x_{EMG}, x_{EOG}))) \quad (3)$$

M_T and M_S will be both initialized from M_0 . Thus, multi-channel knowledge is transferred from the source domain to the target domain. M_T would be used for all the signals from Sleep-EDF dataset including EEG Fpz-Cz, EEG Pz-Oz, EMG and EOG. However, M_S is a single channel model only using EEG Fpz-Cz. Although the number of the input channels is different for M_T and M_S , they share the same network structure and initialization model M_0 . M_T and M_S are then fine-tuned in the target domain for knowledge transfer. As M_S requires input from four channels, the single-channel input EEG Fpz-Cz is duplicated for three times.

There are two types of transfer in the first step: 1) Cross-Domain Same-Channel (CDSC) and 2) Cross-Domain Cross-Channel (CDCC). The knowledge from the EEG C4-A1 is transferred to similar channels such as EEG Fpz-Cz and EEG Pz-Oz from other domain by applying the CDSC transfer. In particular, with the CDCC transfer, the EMG and EOG channels from other domain would also be transferred to the EEG channels.

B. Target Domain Knowledge Distillation

The multi-channel knowledge in the target domain is even more crucial for the single-channel sleep staging as there is no domain shift. There is a knowledge distillation between M_T and M_S after the pre-training. The M_T model can learn from the multi-channel signals and teach the M_S model. The loss function for this knowledge distillation is defined as:

$$\begin{aligned} Loss_{target} = & L_{CE_teacher} + L_{CE_student} + \alpha L_{KD_filter} \\ & + \beta L_{KD_LSTM} + \gamma L_{KD_output} \end{aligned} \quad (4)$$

where $L_{CE_teacher}$ and $L_{CE_student}$ are sequence cross-entropy loss functions for M_T and M_S , respectively. L_{KD_filter} , L_{KD_LSTM} and L_{KD_output} are knowledge distillation losses of the filterbank, the epoch-wise LSTM, and

the hidden layer output between the teacher and the student, respectively. The hyper-parameters α , β and γ are all set as 1500. The loss of the knowledge distillation is defined as:

$$L_{KD_filter} = \|\hat{u}_1 - v_1\|_2^2 \quad (5)$$

$$L_{KD_LSTM} = \|\hat{u}_2 - v_2\|_2^2 \quad (6)$$

$$L_{KD_output} = \|\hat{u}_3 - v_3\|_2^2 \quad (7)$$

where \hat{u} and v are output features after the filterbank, the epoch-wise LSTM and the final hidden layer for M_T and M_S , respectively.

Simultaneous training is applied to the proposed teacher-student system. A more robust model can be expected since M_S could learn from the dynamic paths for training instead of a static pre-trained M_T with fixed parameters. However, the knowledge distillation loss increases the similarity between M_T and M_S , which may result in a reduction in the accuracy of M_T and M_S simultaneously. A gradient block is applied to the knowledge distillation loss to stop M_S decreasing the performance of M_T during the simultaneous training.

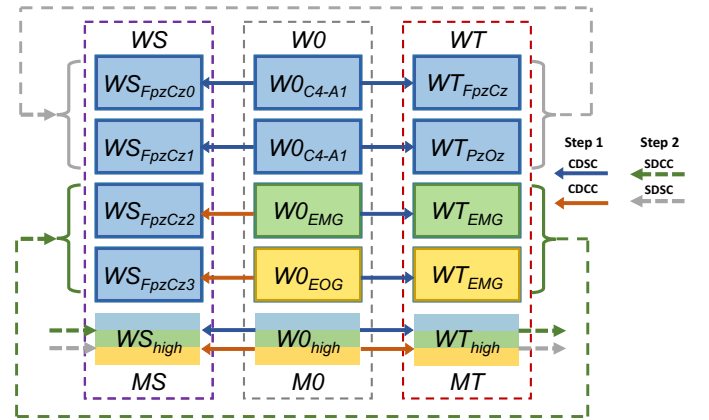


Fig. 2. Knowledge Transfer Paths in the proposed work. There are four types of knowledge transfer methods: CDSC, CDCC, SDCC, and SDSC.

Fig.2 illustrates the knowledge transfer paths in the proposed work. There are four types of knowledge transfer methods: CDSC, CDCC, SDCC, and SDSC. The model is divided into two parts: the low-level filterbank and the high-level network. The filterbank consists of fully connected layers. The network consists of a hierarchical Bi-LSTM, an attention layer and a fully connected layer. In the low-level filterbank, each weight is responsible for one channel independently. The weights in the high-level network deal with all the four channels simultaneously. In the first step, there are CDSC and CDCC knowledge transfer between M_0 and M_T , M_S with the pre-training and fine-tuning scheme. The knowledge could directly be transferred between the models through the different weights. For CDSC, the knowledge is transferred between similar channels from different domains. For instance, the EEG C4-A1 of the MASS dataset is transferred to the EEG Fpz-Cz of the Sleep-EDF dataset. Between these two

domains, there is a significant difference due to the usage of different acquisition devices. In addition, the locations of the EEG channels are also different. However, as they share a low level representation of EEG, a simple duplication of the weight in M_0 still provides a proper initialization for the training, which prevents the model from over-fitting with the increase of the amount of training data. For CDCC, the knowledge origins from different channels of different domains. For example, the EMG and EOG of the source domain are transferred to the EEG channels in the target domain. EMG and EOG are meaningful for sleep staging since the EMG signal is usually used in the classification among the REM and NREM stages, and the EOG helps determine when sleep occurs as well as whether the subject is in the REM sleep stage. The weights of the two channels, $W_{0_{EMG}}$ and $W_{0_{EOG}}$, contain information that is useful in the determination of those sleep stages. Thus the introduction of the EMG and EOG channels is helpful for the training of M_S and M_T .

In the second step, there are SDCC and SDSC transfer between M_T and M_S . The knowledge distillation loss forces the intermediate layers of these two networks to output the same results. The knowledge distillation loss also causes an indirect knowledge transfer between the weights. There is knowledge transfer from the EMG and EOG channels to the EEG channel in SDCC transfer, as well as from the two EEG channels to the single-channel EEG Fpz-Cz with SDSC transfer. Since M_T is trained based on multi-channel signals, it converges to a smoother feature space. Even if the weights responsible for the single-channel EEG processing are taken out separately, it would be better than the models trained using the single-channel EEG. Therefore, in SDSC, besides EEG Pz-Oz, EEG Fpz-Cz in M_T will also be transferred to the same channel in M_S . There are three different knowledge distillation losses monitored in this step. Firstly, L_{KD_output} pushes the output of the final hidden layer from the two models be similar to each other. The output is a posterior probability distribution for sleep staging. As not disturbed by the number of channels, it is the most effective knowledge distillation approach. Secondly, L_{KD_filter} compels these two models to output equal time-invariant features extracted by the filterbank, which is equivalent to the translation of different channel features. Thirdly, L_{KD_LSTM} makes the epoch-wise LSTM of the outputs from M_T and M_S share the same temporary feature. It is a supplement for time-invariant feature translation. In the combination of the filterbank transfer, the LSTM feature transfer, and the output transfer, the knowledge distillation loss enforces the two models similar not only in output, but also in the features of the intermediate layers with different inputs and weights. In this case, M_S learns the relationship between the different channels, although the input signal is only EEG Fpz-Cz, the intermediate features are similar to M_T with a four-channel input.

III. EXPERIMENTAL RESULTS

The proposed work applies pre-training in the source domain using the MASS dataset. The pre-training was completed

under a 20-fold cross validation protocol for 100 epochs. Then the model with the highest accuracy was retained. The knowledge distillation is applied in the target domain using the Sleep-EDF dataset. The Sleep-EDF dataset is used for testing by a leave-one-subject-out 20-fold cross-validation. The MASS dataset contains 200 participants with overnight EEG records and corresponding sleep stages. The dataset contains one scalp-EEG signals from C4-A1 channel, a submental chin EMG, and a horizontal EOG. All the three channels were used for the pre-training of the four-channel model M_0 . The Sleep-EDF dataset contains 20 participants with two scalp-EEG signals from Fpz-Cz and Pz-Oz channels, an EMG, and an EOG. For the multi-channel teacher model, signals from all four channels are employed, while only the EEG Fpz-Cz channel is employed for the single-channel student models. The sleep stages consist of Wake, N1, N2, N3 and REM.

TABLE I
COMPARISON BETWEEN TRANSFER METHODS

Output Transfer	Transfer Methods		ACC
	Filterbank Transfer	LSTM Transfer	
✓	✓	✓	86.52%
✓	×	×	86.44%
×	✓	×	86.31%
×	×	✓	86.26%
×	×	×	85.77%

Table I compares the performance between different transfer methods from the proposed work. It is noted that the proposed output transfer features the best, since the posterior probability distribution from the final hidden layer contains more information concerning sleep staging without being contaminated by other channels. In addition, with the combination of these three transfer losses, a better result is achieved. The output of the filterbank contains more useful knowledge than the LSTM, and this might because that the filterbank transfer enables one-to-one channel transfer by extracting features for different channels independently using different weights. In contrast, the LSTM transfer employs all the four channels simultaneously. The results also indicate that any transfer method is better than nothing.

Table II compares the proposed work with previous works. The MCMD model achieves an accuracy (ACC) of 86.5% and a Macro-F1 (MF1) of 80.9 in single-channel EEG sleep staging. With the combination of four types of transfer methods, our model achieves a comparable result with multi-channel model using only single-channel EEG. An improvement of 0.4% is achieved while comparing to the state-of-the-art single channel sleep staging method in literature [14] based on subject independent 20-fold cross validation. Compared with our previous work [7], there is a 2.8% improvement of ACC and a 5.4 improvement of MF1 from the introduction of multi-domain knowledge, which also indicates that it solves the problem of unbalanced sample distribution.

Table III demonstrates the Ablation experiment results of the MCMD algorithm. Different knowledge transfer methods and model capacities are applied. The baseline Seqsleepnet model achieves an ACC of 84.57%. According to [6], there is a 1.04%

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART

Methods	Dataset	Channels	Transfer Methods	Overall Metrics		
				ACC	MF1	kappa
This work	Sleep-EDF (MASS)	Single	CDSC CDCC SDSC SDCC	86.5	80.9	0.82
TNSRE18 [2]	Sleep-EDF	Single	None	81.4	72.2	-
TBE18 [3]	Sleep-EDF	Single	None	81.9	73.8	0.74
TNSRE17 [5]	Sleep-EDF	Single	None	82.0	76.9	0.76
EMBC18 [8]	Sleep-EDF	Single	None	82.5	72.0	0.76
EMBC18 [10]	Sleep-EDF	Single	None	82.6	74.2	0.76
ISCAS20 [7]	Sleep-EDF	Single	SDSC SDCC	83.7	75.7	0.78
TCASII21 [13]	Sleep-EDF	Single	None	83.8	75.3	0.78
TBE20 [6]	Sleep-EDF (MASS)	Single	CDSC	85.2	79.6	0.79
TNSRE21 [14]	Sleep-EDF	Single	None	86.1	79.2	0.81
TNSRE19 [1]	MASS	Three	None	87.1	81.5	0.833

TABLE III
ABLATION EXPERIMENTS IN MCMD ALGORITHM

Student Model				
Transfer Method	Source Domain	Target Domain	Capacity	ACC
Baseline-1C	No	Single-Channel	1C	84.57%
CDSC	Single-Channel	Single-Channel	1C	85.61%
CDSC+CDCC	Multi-Channel	Single-Channel	4C	85.77%
CDSC+CDCC +SDSC+SDCC	Multi-Channel	Multi-Channel	4C	86.52%
Teacher Model				
Transfer Method	Source Domain	Target Domain	Capacity	ACC
Baseline	No	Multi-Channel	4C	86.27%
CDSC+CDCC	Multi-Channel	Multi-Channel	4C	86.91%

improvement compared with the baseline by using single-channel pre-training from the source domain. The proposed multi-channel pre-training features better ACC by applying a combination of CDSC and CDCC. As only a small part of the knowledge transfer is taken across different channels and domains, the improvement is not high. With a teacher-student knowledge distillation scheme combined with all the four transfer scenarios, M_S achieves an ACC of $86.52 \pm 5.61\%$ for single-channel sleep staging.

IV. CONCLUSION

This paper proposed a Multi-Channel Multi-Domain knowledge distillation algorithm for single-channel sleep staging. The proposed algorithm combines knowledge transfer in four different scenarios: Cross-Domain Same-Channel (CDSC), Cross-Domain Cross-Channel (CDCC), Same-Domain Cross-Channel (SDCC) and Same-Domain Same-Channel (SDSC), achieving a state-of-the-art single-channel sleep staging ACC of 86.5%, with only a 0.6% deterioration from the state-of-the-art multi-channel model. Experimental results show that the knowledge from multiple domains and multiple channels could be transferred to single-channel EEG sleep staging, bringing an accuracy improvement of 2%.

REFERENCES

- [1] H. Phan *et al.*, “Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.
- [2] S. Chambon *et al.*, “A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.
- [3] H. Phan *et al.*, “Joint classification and prediction cnn framework for automatic sleep stage classification,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.
- [4] J. Zhang and Y. Wu, “A new method for automatic sleep stage classification,” *IEEE transactions on biomedical circuits and systems*, vol. 11, no. 5, pp. 1097–1110, 2017.
- [5] A. Supratak, H. Dong, C. Wu, and Y. Guo, “Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
- [6] H. Phan *et al.*, “Towards more accurate automatic sleep staging via deep transfer learning,” *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 1787–1798, 2020.
- [7] Y. Liao, M. Zhang, Z. Wang, and X. Xie, “Design of a hybrid competition-cooperation teacher-students model for single channel based sleep staging,” in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.
- [8] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks,” in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 1452–1455.
- [9] Y. Liao, M. Zhang, Z. Wang, and X. Xie, “Tri-featurenet: An adversarial learning-based invariant feature extraction for sleep staging using single-channel eeg,” in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.
- [10] H. Phan *et al.*, “Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification,” in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 453–456.
- [11] S.-Y. Chang *et al.*, “An ultra-low-power dual-mode automatic sleep staging processor using neural-network-based decision tree,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 9, pp. 3504–3516, 2019.
- [12] Q. Cai *et al.*, “A graph-temporal fused dual-input convolutional neural network for detecting sleep stages from eeg signals,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 2, pp. 777–781, 2021.
- [13] Y. Liao *et al.*, “Lightsleepnet: Design of a personalized portable sleep staging system based on single-channel eeg,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2021.
- [14] L. Fiorillo, P. Favaro, and F. D. Faraci, “Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2076–2085, 2021.
- [15] A. R. Hassan and M. I. H. Bhuiyan, “An automated method for sleep staging from eeg signals using normal inverse gaussian parameters and adaptive boosting,” *Neurocomputing*, vol. 219, pp. 76–87, 2017.
- [16] A. R. Hassan and A. Subasi, “A decision support system for automated identification of sleep stages from single-channel eeg signals,” *Knowledge-Based Systems*, vol. 128, pp. 115–124, 2017.
- [17] Y. Luo and B.-L. Lu, “Eeg data augmentation for emotion recognition using a conditional wasserstein gan,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 2535–2538.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [19] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [20] C. O’reilly, N. Gosselin, J. Carrier, and T. Nielsen, “Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research,” *Journal of sleep research*, vol. 23, no. 6, pp. 628–635, 2014.
- [21] B. Kemp *et al.*, “Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg,” *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.
- [22] A. L. Goldberger *et al.*, “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals,” *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.