A Knowledge Distillation Framework for Enhancing Ear-EEG Based Sleep Staging with Scalp-EEG Data

Mithunjha Anandakumar*,1,2, Jathurshan Pradeepkumar*,1,2, Simon L. Kappel3, Chamira U. S. Edussooriya1,4, Anjula C. De Silva1

1Dept. of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka
2Faculty of Arts and Sciences, Harvard University, USA
3Dept. of Electrical and Computer Engineering, Aarhus University, Denmark
4Dept. of Electrical and Computer Engineering, Florida International University, USA

Abstract—Sleep plays a crucial role in the well-being of human lives. Traditional sleep studies using Polysomnography are associated with discomfort and often lower sleep quality caused by the acquisition setup. Previous works have focused on developing less obtrusive methods to conduct high-quality sleep studies, and ear-EEG is among popular alternatives. However, the performance of sleep staging based on ear-EEG is still inferior to scalp-EEG based sleep staging. In order to address the performance gap between scalp-EEG and ear-EEG based sleep staging, we propose a cross-modal knowledge distillation strategy†, which is a domain adaptation approach. We employ model architectures from the transformer and convolutional neural network families to demonstrate the model-agnostic nature of the method. Our experiments and analysis validate the effectiveness of the proposed approach with existing architectures, where it enhances the accuracy of the ear-EEG based sleep staging by 3.46% and Cohen’s kappa coefficient by a margin of 0.038. Furthermore, our findings indicate that our approach is not limited to a specific model architecture and can be applied to a wide range of deep learning models.

Index Terms—Sleep staging, cross-modal knowledge distillation, ear-EEG, domain adaptation.

I. INTRODUCTION

Humans spend one-third of their lifetime in sleep, and decades of research have underlined the negative effects of poor sleep on mental and physical well-being. Polysomnography (PSG) is a comprehensive and the gold-standard study currently used in clinics to assess the quality of sleep and to diagnose sleep related disorders. In PSG, several physiological signals such as; electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), electrooculogram (EOG), and SpO2 are recorded followed by manual annotation of sleep stages by sleep experts based on standard guidelines such as Rechtschaffen and Kales (R&K) [1] or American Academy of Sleep Medicine (AASM) [2], to aid sleep stage detection and sleep disorder diagnosis.

PSG enables high-quality sleep assessment however, the signal acquisition setup is obstructive to sleep since the patient has to wear multiple sensors and electrodes. The quality of patients’ sleep is compromised due to the discomfort caused by these sensor systems and the unfamiliarity of the hospital environment, thus affecting the diagnostic quality of the study, and leading to treatment errors. Additionally, PSG is a complex and expensive setup that requires expert assistance and specialized laboratories. Therefore, sleep studies are generally conducted at a sleep clinic, which limits PSG from being used in home-based and long-term settings.

In order to overcome the aforementioned limitations in sleep studies, there have been several efforts to build a simple and comfortable system for high-quality home-based sleep monitoring. Previous works have focused on utilizing either one or combinations of physiological signals such as ear-EEG [3]–[8], ECG [9], respiratory signal [10], EOG [10], heart rate [11], photoplethysmography [12], and actigraphy [11], [13] for sleep monitoring.

Ear-EEG based sleep monitoring has advantages in terms of enhanced comfort and portability when compared to a full PSG montage. This enables long-term sleep monitoring in home settings. Ear-EEG is a potential alternative for PSG, since there is a good correspondence between hypnograms measured with ear-EEG and scalp-EEG [14], [15]. Several studies have strongly validated that the ear-EEG recordings contain the necessary information to classify sleep stages with an acceptable level of accuracy [3]–[8]. Manually annotating sleep stages on the recordings is a labor-intensive and time-consuming process, hence past research has focused on automating the process [16]–[19]. Similar techniques have been explored to classify sleep stages from ear-EEG [5], [6]. However, the performance of sleep staging based on ear-EEG is still inferior to scalp-EEG based sleep staging even though there is high mutual information between ear-EEG and scalp-EEG. The performance gap could be attributed to the discrepancy in signal-to-noise ratio due to spatial differences in the signal acquisition site and poorer spatial resolution of ear-EEG [15]. The unavailability of large publicly available ear-EEG datasets makes this task more challenging. In our study, we focus on improving the sleep staging performance of existing deep-learning based algorithms on ear-EEG by leveraging effective training strategies and knowledge from scalp-EEG recordings.

†These authors contributed equally to the work.
†https://github.com/Mithunjha/EarEEG_KnowledgeDistillation
In our work, we employ cross-modal knowledge distillation, a domain adaptation technique to enhance the performance of ear-EEG based sleep staging. Knowledge distillation [20] has been explored in the scalp-EEG domain for various applications such as emotion recognition [21], improving ECG based sleep staging [22], etc. Since there is high mutual information between ear-EEG and scalp-EEG [15], we hypothesize that the performance of ear-EEG based sleep staging can be improved by forcing the model to learn a similar feature representation from scalp-EEG data, such that the common features in both domains can be extracted. This should enable the model to better extract important features contained in the ear-EEG data for better sleep staging. To the best of our knowledge, this is the first of such works focusing on enhancing ear-EEG based sleep staging by employing knowledge distillation strategies and opens up an unexplored domain of techniques to enhance ear-EEG based sleep staging. We employ model architectures from the transformer and convolutional neural network families, in particular cross-modal transformer [16] and Usleep [17]. Our experiments confirm the effectiveness of the proposed approach with the selected architectures. The results demonstrate an enhancement in both the accuracy of ear-EEG based sleep staging, with an improvement of 3.46%, and the Cohen’s kappa coefficient, which is increased by 0.038. Moreover, our findings indicate that our approach is not restricted to a particular model architecture, but rather can be successfully utilized across a diverse spectrum of deep learning models.

II. METHODOLOGY

A. Problem Definition

In this paper, we address the problem of cross-domain sleep stage classification to enhance ear-EEG based sleep staging. The dataset used in the study consists of simultaneously collected scalp-EEG and ear-EEG recordings [5]. Our training set, with the size of $N$, consists of labeled and paired 30 s epochs of scalp-EEG and ear-EEG signals $\{x_i^s, x_i^t, y_i\}_{i=1}^N$, where $(x_i^s, x_i^t, y_i) \in \mathcal{X}^s \times \mathcal{X}^t \times \mathcal{Y}$. Here, $\mathcal{X}^s, \mathcal{X}^t \in \mathbb{R}^{T \times C}$ are the input space of the recorded signals, where $\mathcal{X}^s$ is the source domain of scalp-EEG signals and $\mathcal{X}^t$ is the target domain of ear-EEG signals. $T$ and $C$ represent the time steps in a 30 s epoch and the recorded number of channels, respectively. $\mathcal{Y} \in \mathbb{R}^{K}$ denotes the output space of sleep stages, where $K$ represents the number of sleep stages considered. Based on AASM [2], we consider $K = 5$, such that $\mathcal{Y} \in \{\text{WAKE}, \text{N1}, \text{N2}, \text{N3}, \text{REM}\}$. Under a cross-domain setting, our goal is to learn a function $f_0 : \mathcal{X}^s \rightarrow \mathcal{Y}$ by utilizing the knowledge in the $\mathcal{X}^s$ and known $\mathcal{Y}$. This is achieved by minimizing the error $E = \mathbb{E}_{x_t^s, x_t^t, y_t} [L(f_0(x_t^s), x_t, y_t)]$ on the given training dataset. Here $L$ denotes a loss function.

B. Knowledge Distillation Framework

Knowledge distillation [20] is a domain adaptation technique used to transfer knowledge from a teacher model to a student model. In knowledge distillation, the student model learns to mimic the teacher model by leveraging the embedded knowledge to achieve similar or higher performance. In our study, we employ cross-modal knowledge distillation. During training, the teacher model is only exposed to the source domain of scalp-EEG signals. Then, the embedded knowledge is distilled to the student model, which is trained in the target domain of ear-EEG signals.

There are different types of knowledge distillation approaches [23] such as response-based, feature-based, and relation-based knowledge distillation. We focused on improving ear-EEG based sleep staging by forcing the model to learn a similar feature representation to scalp-EEG, such that the common features in both domains could be extracted. This enabled the model to extract important features buried in the ear-EEG signals for better sleep staging. To test and validate our approach, feature-based knowledge distillation [23] was employed under two different strategies; 1) offline, and 2) online knowledge distillation.

1) Offline Knowledge Distillation: In offline feature-based knowledge distillation as shown in Fig. 1, the teacher model $f_0^T$ is initially trained in the $\mathcal{X}^s$ domain in a supervised setting. In the offline knowledge distillation settings, the weights of $f_0^T$ are fixed, and the student model $f_0^S$ is randomly initialized and trained in the $\mathcal{X}^t$ domain by minimizing $L_{kd}^S$. 

![Fig. 1. The overview of the offline cross-modal knowledge distillation framework. The teacher model is initially trained on Scalp EEG recordings. During distillation, the teacher model is fixed and the student model is trained.](image)

![Fig. 2. The overview of the online cross-modal knowledge distillation framework. In contrast to the offline knowledge distillation, both teacher and student models are trained during distillation.](image)
which is defined as:

\[
L_{kd}(f_S^T(x^t_i), f_T^S(x^s_i), y_i) = L_{ce}(f_S^T(x^t_i), y_i) + L_{mse}(f_S^T(x^t_i), f_T^S(x^s_i)).
\]  

(1)

Here \(L_{ce}\) represents categorical cross-entropy loss, \(L_{mse}\) is employed between the intermediate features of \(f_S^T\) and \(f_T^S\) to enable \(f_S^T\) to learn a similar feature representation as \(f_T^S\) and is defined as

\[
L_{mse}(f_S^T(x^t_i), f_T^S(x^s_i)) = \frac{1}{N} \sum_{j=1}^{N} ||f_S^T(x^t_i) - f_T^S(x^s_i)||^2_2.
\]  

(2)

For clarity, \(f_S(x_i)\) is also used to denote intermediate feature maps.

2) **Online Knowledge Distillation:** In contrast to offline knowledge distillation, the weights of both \(f_S^T\) and \(f_T^S\) are randomly initialized and updated simultaneously in the online knowledge distillation setting as shown in Fig. 2. \(f_T^S\) is trained to minimize the loss mentioned in (1), and \(f_S^T\) is trained to minimize \(L_{ce}(f_S^T(x^t_i), y_i)\), which is a categorical cross-entropy loss. The goal is to improve ear-EEG based sleep staging by enabling \(f_S^T\) and \(f_T^S\) to jointly learn a common feature space between \(X^s\) and \(X^t\).

**C. Transfer Learning**

Transfer learning is a deep learning procedure used for storing knowledge gained on solving a particular task and reuse them as an initialization point for a similar new task. This method has been widely used in sleep stage classification domain [19], [24], where a model is pre-trained on larger datasets and then fine-tuned towards smaller datasets. In our work, we employed transfer learning to improve ear-EEG based sleep staging and compared its performance with the knowledge distillation approaches. Here, we hypothesized that the model would learn the characteristics of scalp EEG signals during pre-training and then adapt the weights toward ear-EEG during fine-tuning.

**III. EXPERIMENTS**

**A. Dataset**

We used the dataset from [5], consisting of one whole night of sleep recording from 9 healthy subjects (age: 26-44; 3 females and 6 males). The dataset contains simultaneously recorded PSG and ear-EEG. The PSG consists of 8 scalp-EEG channels (O1, O2, C3, C4, A1, A2, F3, and F4 according to the international 10-20 system), 2 EOG channels, and a chin EMG channel. Note that the EOG and EMG channels were not considered in our study. The ear-EEG consists of 12 channels, with 6 channels from each ear: ELA, ELB, ELE, ELI, ELG, ELK, ERA, ERB, ERE, ERI, ERG, and ERK according to [5]. All the EEGs were sampled at 200 Hz. The dataset was manually scored based on the international AASM [2] guideline.

**B. Data Preprocessing**

All the channels (scalp and ear-EEG) were bandpass filtered between 0.2—42 Hz. Noisy ear-EEG channels were removed based on the mean power. Initially, all ear derivations were calculated, and the mean power (\(P_{\text{ij}}\)) for derivation consisting of channel \(i\) and \(j\) within the frequency range of 10—35 Hz was calculated [5]. A channel \(i\) was rejected if \(m_i = \text{median}(P_{\text{ij}})\) for derivation consisting of channel \(i\) and \(j\) within the frequency range of 10—35 Hz was calculated [5]. A channel \(i\) was rejected if \(m_i = \text{median}(P_{\text{ij}})\) for derivation consisting of channel \(i\) and \(j\) within the frequency range of 10—35 Hz was calculated [5]. A channel \(i\) was rejected if \(m_i = \text{median}(P_{\text{ij}})\). Altogether 15 channels were rejected from the dataset. Recordings of subject 5 were removed from the dataset as both the ear canal channels (ERA and ERB) from the right ear were rejected, thus average electric potential difference cannot be calculated. Table I shows the number of rejected and usable ear-EEG channels that remained after the channel rejection. As inputs to the network, derivations of scalp-EEG and ear-EEG were extracted as follows:

**Scalp-EEG:** Three scalp-EEG derivations were extracted from 8 scalp-EEG channels: C3-O1, C4-O2, and A1-A2. These three scalp-EEG derivations were chosen, because of their spatial correspondence with ear-EEG derivations.

**Ear-EEG:** Three ear-EEG derivations were extracted from the 12 ear-EEG channels [5] as shown below, where (\(\bar{7}\)) denotes the mean.

\[
L_1 = \text{ELA}, \text{ELB}, \text{ELE}, \text{ELI}, \text{ELG}, \text{ELK}
\]  

(3)

\[
R_1 = \text{ERA}, \text{ERB}, \text{ERE}, \text{ERI}, \text{ERG}, \text{ERK}
\]  

(4)

\[
L - R = L_1 - R_1
\]  

(5)

\[
LE = \text{ELA}, \text{ELB} - \text{ELE}, \text{ELI}, \text{ELG}, \text{ELK}
\]  

(6)

\[
RE = \text{ERA}, \text{ERB} - \text{ERE}, \text{ERI}, \text{ERG}, \text{ERK}
\]  

(7)

Here, \(L - R\) derivation gives the average electric potential difference between the left and right ears. \(LE\) and \(RE\) derivations are the average potential difference between the concha and ear canal in the left and right ears, respectively.

**C. Training and Experimental Setup**

The model under all training settings were trained using the Adam optimizer [25] with learning rate (\(lr\), \(beta_{1}\) and \(beta_{2}\) set to, \(10^{-3}\), 0.9 and 0.999, respectively for 100 epochs. The batch size was experimentally chosen as 32. The dataset was partitioned in a leave-one-subject-out (LOSO) fashion, as this is the most probable real-application scenario. Hence, at a given iteration, the model was trained on seven subjects and tested on one, which the model has not seen during training. The model was implemented in PyTorch environment.
### TABLE II
Overall performance of supervised learning, transfer learning (TL), and knowledge distillation (KD).

| Model Architecture | Modality  | Method     | ACC   | κ  |
|--------------------|-----------|------------|-------|----|
| Scalp-EEG [17]     | Supervised| Supp.      | 73.19 | 0.631 |
| Ear-EEG            | Supervised| TL         | 64.58 | 0.524 |
|                    |           | KD (Offline)| 66.25 | 0.544 |
|                    |           | KD (Online) | 68.04 | 0.562 |
| Epoch              | Supervised| Supp.      | 78.11 | 0.670 |
| Cross-Modal        | Supervised| TL         | 63.84 | 0.481 |
| Transformer [16]   |           | KD (Offline)| 69.61 | 0.551 |
|                    |           | KD (Online) | 69.19 | 0.551 |

### D. Architecture details

In our study, we employed two distinct deep learning-based sleep stage classifiers, belonging to two different model classes, namely transformers and convolutional neural networks (CNNs). Specifically, we utilized cross-modal transformers [16], within the transformer category. For our analysis, we opted for epoch cross-modal transformers. We extended the model by incorporating 3 EEG inputs, where each EEG signal is processed by an intra-modal attention block, followed by a cross-modal attention network and feedforward networks. The resulting learned representations from each EEG are concatenated and utilized for classification. On the other hand, we utilized Usleep [17] which is a CNN-based architecture particularly designed for time-series segmentation tasks such as sleep staging. Usleep consists of three components: encoder, decoder, and segment classifier. The encoder encodes the input signal into a dense feature representation, while the decoder maps the learned feature representation back to the input feature space. Finally, the segment classifier outputs the sleep stages. For more details of the cross-modal transformer and Usleep, the reader is referred to [16] and [17], respectively. In both offline and online knowledge distillation approach, the student and teacher models are clones of the same architecture - either USleep or epoch cross-modal transformer.

### IV. RESULTS AND DISCUSSION

We compared both our knowledge distillation and transfer learning strategies with the two existing supervised learning architectures used in the domain of sleep stage classification. Accuracy (ACC) and Cohen’s kappa coefficient (κ) were considered as performance metrics in LOSO cross-validation method. To further assess the effectiveness of our proposed method, we utilized T-distributed stochastic neighbor embedding (T-SNE) to analyze the learned representations.

#### A. Sleep Staging Performance

Overall performance and comparison of the explored training strategies are given in Table II. The results clearly state the effectiveness of both knowledge distillation training strategies in improving ear-EEG based sleep staging. For U-Sleep [17], the accuracy of the sleep staging was increased by 3.46% and in Cohen’s kappa coefficient by a margin of 0.038, when compared to the supervised training. A similar performance increase was observed with epoch cross-modal transformers [16], which showed that the proposed knowledge distillation strategies are agnostic to model architectures. The Cohen’s kappa coefficients obtained for supervised learning were low, when compared to more recent ear-EEG based sleep studies [8], but higher than the results presented in [5], where the dataset we used was originally presented. This lower Cohen’s kappa coefficients could be attributed to a combination of several factors, including a lower sample size when compared to more recent studies [8] and advanced equipment used to record ear-EEG in more recent studies [8], [26].

Both online and offline knowledge distillation methods yielded similar scores. Considering the similar scores, offline knowledge distillation is preferred, because online knowledge distillation requires large computational resources as shown in Table IV, since both teacher and student models were trained simultaneously. On the other hand, offline knowledge distillation requires the teacher model to be pre-trained beforehand, thus during the knowledge distillation process, computational resources are only needed to train a single model (the student). This makes offline knowledge distillation more computationally efficient compared to the online approach. According to Table II the proposed knowledge distillation strategies outperformed transfer learning in both model architectures. This indicates that the pre-trained model on the scalp-EEG domain was a poor initialization for the ear-EEG domain.

Class-wise performances using macro averaged F1-score (MF1) of the training strategies are given in Table III. In most

### TABLE III
Class-wise MF1 performance of supervised learning, transfer learning (TL), and knowledge distillation (KD).

| Model       | Per-class Performance |
|-------------|-----------------------|
| Method      | N1 | N2 | N3 | REM |
| Scalp-EEG   | 82.1| 6.0 | 88.7 | 79.3 | 74.1 |
| Ear-EEG     | 62.2| 11.5 | 72.6 | 74.0 | 60.0 |
| TL          | 65.1| 6.3 | 75.3 | 74.6 | 57.3 |
| KD (Offline)| 62.2| 10.3 | 74.7 | 76.8 | 60.8 |
| KD (Online) | 64.3| 11.0 | 77.5 | 76.5 | 61.0 |

* denotes supervised learning. ↑ and ↓ indicates performance increase and decrease compared to Ear-EEG.*
sleep stages, the knowledge distillation methods improved the prediction, but it under performed predicting the N1 sleep stage, when compared to supervised learning. Generally, the Cohen’s kappa for classification of the N1 sleep stage was low for all classification methods given in Table III. A low Cohen’s kappa for classification of N1 is a common trend for automatic sleep staging algorithms [16], [17] and therefore not a concern.

B. Analysis on Learned Representations

In order to analyze and validate the learned representations by the student model under different training strategies, we extracted the features from intermediate layers and conducted T-SNE. T-SNE is a widely used technique for visualizing high-dimensional data in a low-dimensional space. The visualizations of their distributions are shown in Fig. 3. The plots clearly show two different clusters of scalp-EEG and ear-EEG features from the supervised trained models (Fig. 3(a)). Fig. 3(c) shows that the offline knowledge distillation strategy learns better features, which were similar to scalp-EEG, when compared to transfer learning (Fig. 3(b)). Similarly, online knowledge distillation (Fig. 3(d)) confirms our hypothesis on jointly learning a common feature space for better sleep staging.

The visualizations shows that the learned representations of ear-EEG data after knowledge distillation were more similar to the representations of scalp-EEG data than before knowledge distillation. This indicates that our method successfully transferred the knowledge from scalp-EEG to ear-EEG data, resulting in more informative and accurate representations of ear-EEG data. Also, the improved similarity of the learned representations of ear-EEG to those of scalp-EEG data suggests that our method can effectively bridge the gap between the limitations of ear-EEG data and the high accuracy of scalp-EEG data.

C. Limitations and Future Work

Although our proposed method for improving ear-EEG based sleep staging by transferring knowledge from scalp-EEG data has shown promising results, there are some limitations that must be acknowledged. Firstly, our method requires a paired scalp-EEG and ear-EEG dataset for effective domain adaptation during the training. This may limit its applicability in cases where such paired data are not readily available or are difficult to obtain. Secondly, the achievable improvement in ear-EEG based sleep staging may be dependent on the performance that can be achieved on scalp-EEG data. If the performance on scalp-EEG data is low, then the improvement in ear-EEG based sleep staging may be limited. Lastly, the performance of our method is also dependent on the signal quality of the ear-EEG data, which in turn is influenced by the hardware used for the experiments. In summary, while our proposed method offers a promising approach for improving ear-EEG based sleep staging, it is important to acknowledge its limitations. Future research should focus on addressing these limitations and further refining the method for optimal performance across different conditions.

V. Conclusion

In this paper, we present a domain adaption based training strategy to improve ear-EEG based sleep staging. We validated that the knowledge from scalp-EEG can be leveraged to enhance ear-EEG based sleep staging by model architectures from the transformer and convolutional neural network families. In particular, the proposed approach with existing architectures enhances the accuracy of the ear-EEG based sleep staging by 3.46% and Cohen’s kappa coefficient by a margin of 0.038. Furthermore, our findings indicate that our approach is not limited to a specific model architecture and can be applied to a wide range of deep learning models.

We believe that this work opens the door for future research on unpaired domain adaptation techniques to utilize largely available scalp-EEG datasets and to analyze largely unexplored deep learning strategies to improve and achieve clinical standards in ear-EEG based sleep studies. Thus, our work not only contributes to the field of sleep research but also has potential implications for clinical applications.

ACKNOWLEDGMENT

We would like to extend our gratitude to the Center for Ear-EEG, Aarhus University, Denmark for providing us with the ear-EEG dataset. We also would like to thank Dhinesh Suntharalingham and Vinith Kugathasan for their support.
REFERENCES

[1] A. Rechtschaffen, A. manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. Brain information service, Brain Research Institute. US Dept. of Health, Education, and Welfare, 1968.

[2] C. Iber et al., The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specification, 1st ed., American Academy of Sleep Medicine, 2007.

[3] T. Nakamura, Y. D. Akiyoshi, M. J. Morrell, and D. P. Mandic, “Hearables: Automatic overnight sleep monitoring with standardized in-ear EEG sensor,” IEEE Trans. Biomed. Eng., vol. 67, no. 1, pp. 203–212, 2019.

[4] T. Nakamura, V. Govendovsky, M. J. Morrell, and D. P. Mandic, “Automatic sleep monitoring using ear-EEG,” IEEE J. of Transl. Eng. Health Med, vol. 5, pp. 1–8, 2017.

[5] K. B. Mikkelsen, D. B. Villadsen, M. Otto, and P. Kidmose, “Automatic sleep staging using ear-EEG,” BioMed. Eng. OnLine, vol. 16, no. 111, pp. 1–15, 2017.

[6] K. B. Mikkelsen et al., “Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy,” J. Sleep Res., vol. 28, no. 2, p. e12786, 2019.

[7] Y. R. Tabar, K. B. Mikkelsen, M. L. Rank, M. C. Hemmsen, M. Otto, and P. Kidmose, “Ear-EEG for sleep assessment: a comparison with actigraphy and PSG,” Sleep Breath., vol. 25, no. 3, pp. 1693–1705, 2021.

[8] K. B. Mikkelsen et al., “Accurate whole-night sleep monitoring with dry-contact ear-EEG,” Sci. Rep., vol. 9, no. 16824, pp. 1–12, 2019.

[9] Y. Wei, X. Qi, H. Wang, Z. Liu, G. Wang, and X. Yan, “A multi-class automatic sleep staging method based on long short-term memory network using single-lead electrocardiogram signals,” IEEE Access, vol. 7, pp. 85,959–85,970, 2019.

[10] C. Sun, C. Chen, J. Fan, W. Li, Y. Zhang, and W. Chen, “A hierarchical sequential neural network with feature fusion for sleep staging based on EOG and RR signals,” J. Neural Eng., vol. 16, no. 6, p. 066020, 2019.

[11] X. Zhang, W. Kou, E. I. Chao, H. Gao, Y. Fan, and Y. Xu, “Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device,” Comput. Biol. Med., vol. 103, pp. 71–81, 2018.

[12] H. Korkalainen et al., “Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea,” Sleep, vol. 43, no. 11, p. zsaa098, 2020.

[13] J. Korkalainen, C. Sun, C. Chen, J. Fan, W. Li, Y. Zhang, and W. Chen, “Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea,” Sleep, vol. 43, no. 11, p. zsaa098, 2020.

[14] J. Pradeepkumar et al., “Towards interpretable sleep stage classification using cross-modal transformers,” arXiv:2208.06991, 2022.

[15] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jenum, and C. Igel, “U-sleep: resilient high-frequency sleep staging,” NPJ Digit. Med., vol. 4, no. 72, pp. 1–12, 2021.

[16] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, “Xsleepnet: Multi-view sequential model for automatic sleep staging,” IEEE Trans. Pattern Anal. Mach. Intell., 2021.

[17] H. Phan, K. B. Mikkelsen, O. Chen, P. Koch, A. Mertins, and M. De Vos, “Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification,” IEEE Trans. Biomed. Eng., vol. 69, pp. 2456–2467, 2022.

[18] H. Phan, O. Y. Chen, J. Dean et al., “Distilling the knowledge in a neural network,” arXiv:1503.02531, 2015.

[19] H. Phan, K. B. Mikkelsen, O. Chen, P. Koch, A. Mertins, and M. De Vos, “Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification,” IEEE Trans. Biomed. Eng., vol. 69, pp. 2456–2467, 2022.

[20] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, “Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification,” IEEE Trans. Biomed. Eng., vol. 69, pp. 2456–2467, 2022.

[21] S. Zhang, C. Tang, and C. Guan, “Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition,” Pattern Recognit., vol. 130, p. 108833, 2022.

[22] V. Joshi, S. Vijayarangan, P. SP, and M. Sivaprakasam, “A deep knowledge distillation framework for EEG assisted enhancement of single-lead ECG based sleep staging,” arXiv:2112.07252, 2021.

[23] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” Int. J. of Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

[24] H. Phan, O. Y. Chen, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, “Towards more accurate automatic sleep staging via deep transfer learning,” IEEE Trans. Biomed. Eng., vol. 68, no. 6, pp. 1787–1798, 2020.

[25] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980, 2017.

[26] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, “Dry-contact electrode ear-EEG,” IEEE Trans. on Biomed. Eng., vol. 66, no. 1, pp. 150–158, 2018.