A Deep Learning Model for Automatic Sleep Scoring using Multimodality Time Series

Rui Yan
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
ruiyanmodel@foxmail.com

Fan Li
School of Biomedical Engineering
Dalian University of Technology
Dalian, China
lifandlpu@foxmail.com

DongDong Zhou
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
dongdongzhou1017@foxmail.com

Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
tapani.e.ristaniemi@jyu.fi

Fengyu Cong
School of Biomedical Engineering
Dalian University of Technology
Dalian, China
fengyu.cong@foxmail.com

***Abstract*—Sleep scoring is a fundamental but time-consuming process in any sleep laboratory. Automatic sleep scoring is crucial and urgent to help address the increasing unmet need for sleep research. Therefore, this paper aims to develop an end-to-end deep learning architecture using raw polysomnographic recordings to automate sleep scoring. The proposed model adopts two-dimensional convolutional neural networks (2D-CNN) to automatically learn features from multi-modality signals, together with a "squeeze and excitation" block for recalibrating channel-wise feature responses. The learnt representations are finally fed to a softmax classifier to generate predictions for each sleep stage. The model performance is evaluated on two public sleep datasets (SHHS and Sleep-EDF) with different available channels. The results have shown that our model achieves an overall accuracy of 85.2% on the SHHS dataset and an accuracy of 85% on the Sleep-EDF dataset. We have also demonstrated that the proposed architecture not only is able to handle various numbers of input channels and several signal modalities from different datasets but also exhibits short runtimes and low computational cost.**

***Keywords*—polysomnography, automatic sleep scoring, multi-modality analysis, deep learning, transfer learning**

I. Introduction

Adequate and high-quality sleep is vital to our physical and mental well-being. Nowadays, and likely because of our ephemeral lifestyle in modern society, complaints about sleep disorders increase dramatically among people. An effective way to diagnose sleep disorders and monitor sleep quality is overnight polysomnography (PSG), which is carried out in a specialized hospital-based sleep laboratory. A PSG test simultaneously records dozens of sleep signals including electroencephalograms (EEG), electrooculogram (EOG), electromyograms (EMG), electrocardiogram (ECG), pulse oximetry, airflow, respiratory effort etc. The standard method of analyzing PSG recordings is based on the criteria proposed by Rechtschaffen and Kales (R&K)[1] and the recently updated American Academy of Sleep Medicine (AASM) standards[2].

Based on the amplitude and frequency characteristics of sleep signals, the R&K rules divide sleep into five distinct stages: non-rapid eye movement (NREM) stages 1, 2, 3 and 4 and rapid eye movement stage (stage R), but the recently updated AASM standard merges stages 3 and 4 into N3 due to their prevalent low-frequency oscillations. The process of assigning a sleep stage to each sleep segment is called sleep scoring, which is a fundamental step in sleep research. However, the process of sleep scoring is labor-intensive, as studies have revealed that the annotation of an 8-h recording requires approximately 2-4 hours[3]. With the development of computerized methods, there is a growing interest in automatic scoring of PSG recordings.

Numerous attempts so far have been made to automate sleep scoring[4]. Conventional machine-learning methods mainly consist of two main components: feature extraction and classification. For the step of feature extraction, diverse features, such as statistic features, frequency features and nonlinear features, are extracted from the transformed or decomposed signals of EEG, EOG and/or EMG[5]. For classification, support vector machine, random forest, K-nearest neighbor classifier, Naive Bayes, artificial neural network etc. have been employed in the existing literature[6], [7]. In these studies, the agreement between automatic methods and human experts ranged from 0.8 to 0.9 and that value highly relied on the validity of employed features.

Most recently, in the field of automatic sleep scoring, there have sprung up many algorithms that adopted deep learning networks since it did not require explicit feature extraction and was especially suitable for big data approach[8]. Convolutional neural network (CNN) had been used on raw EEG signals to extract features automatically[9], which offered competitive scoring performance on a large multi-center sleep dataset. PSG signals were also transformed into time-frequency images using short-time Fourier transform[10] or wavelet transformation[11], given the superiority of CNNs in image processing. Moreover, deep learning algorithms had introduced some novel classification schemes to mimic the way sleep experts performed manual sleep scoring, such as one-input-multi-output scheme[12] and sequence-to-sequence model[13]. The novel classification schemes were impossible for conventional machine learning paradigms. Attempts on deep learning had yielded exciting results, although training models from scratch required a huge amount of training data and computational resources[14].

In practice, however, some sleep studies may only focus on a small cohort, in which case the network's performance would decline significantly. Besides, different in monitor devices and specific experimental motivations cause channel mismatch, limiting model application across tasks[15]. To solve the above problems, this work proposes a deep learning

This work is supported by the scholarship from China Scholarship Council (Nos. 201606060227).

approach that consists of a very low number of layers, thus resulting in low computation cost compared to other deep learning approaches. Moreover, the proposed approach constructs an end-to-end structure without computing spectrograms or hand-crafted features. One of the most key contributions of this study is that the proposed model can handle various numbers of input channels and several signal modalities from different datasets without changing the model structure and hyperparameters to accommodate channel mismatch.

The article is organized as follows: Section 2 details the experimental data and the proposed deep learning architecture. Section 3 demonstrates the performance of the proposed model. Section 4 discusses the results and limitations of this study. Finally, section 5 gives conclusions.

II. Materials and Methods

A. Sleep Datasets

This study employed two common datasets to evaluate the proposed deep-learning architecture. The first one was from the Sleep Heart Health Study (SHHS)[16], in which only the first round (SHHS-1) was selected in this study. The SHHS dataset was a multi-center cohort study to investigate whether sleep-disordered breathing was associated with a higher risk of various cardiovascular diseases. Subjects employed in the present study were selected by restricting the Respiratory Disturbance Index 3 Percent (RDI3P) < 5 to have near-normal characteristics. In addition, the selected subjects did not use beta-blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease, or history of stroke.

The second one was the Sleep-EDF dataset[17], [18], of which the Sleep Cassette (SC) subset was adopted. It consisted of 20 subjects aged 25-34 years. Each subject had two PSG recordings from two consecutive day-night periods, except for one subject (subject 13) who had only the first-night data. PSG recordings from the second night were employed in this study, and thus 19 recordings were included. TABLE I summarized the characteristics of employed recordings.

Each recording was scored by an experienced research assistant or sleep technologist according to R&K rules. Sleep recordings were segmented into 30-second per epoch and labelled as wakefulness (W), non-rapid eye movement stage (NREM, containing N1, N2, N3 and N4) and rapid eye movement stage (R). According to the recently updated AASM standard, NREM stages 3 and 4 were merged into N3 in the present article. As signals sampled at different rates, we up-sampled those with sampling rates lower than 125 Hz to accommodate data from different datasets. In order to remove noise and artefacts, a simple filtering process, including notch filters, low-pass filters and high-pass filters, was performed. The filtered frequency bands of each signal were summarized in TABLE II. In addition, the long awake period before and after sleep was trimmed so that the number of awake epochs was not dominant in sleep cycles. Except that, the whole sleep recording was fully included in the analysis without discarding any recorded segments, thereby to have a near-clinical situation. To eliminate individual differences, sleep signals were normalized by mapping its mean to 0 and its deviation to 1. Then, signals were divided into 30-second per epoch, and each epoch corresponds to one sleep stage.

TABLE I. THE DESCRIPTION OF SUBJECTS FROM TWO DATASETS.

Parameters		SHHS	Sleep-EDF
Subjects		100	19
Age		46.86 (4.22)	28.74(2.99)
Power Frequency		60Hz	50Hz
Employed Channels		C3, C4, EOGR, EOGL, EMG, ECG	Fpz-Cz, Pz-Oz, EOG (horizontal)
Amplitude	***EEG***	[-26.0, 20.7]	[-208.2, 204.5]
	EOG	[-17.3, 17.7]	[-478.5, 449.6]
	EMG	[-22.3, 22.0]	--
	ECG	[-39.0, 44.2]	--
Sampling Freq.	***EEG***	125Hz	100Hz
	EOG	50Hz	100Hz
	EMG	125Hz	1Hz
	ECG	125Hz	--

B. Model Architecture

The proposed method expands raw EEG signals to multi-modality PSG signals consisting of EEG, EOG, EMG and ECG. The idea is to mimic the way sleep experts perform manual sleep scoring. When sleep experts label a 30-second PSG epoch, they visually inspect amplitude and frequency characteristics of EEG signals and sleep-related events such as spindles and K-complexes [1]. They also check eye movements and muscle activity levels as a reference for labelling some stages, such as stage R[1], [2]. Recent studies have revealed that analysis of heart-rate variations enables us to track the transition from wakefulness to sleep[19]. Similarly, the proposed model jointly processes multi-modality signals, thereby providing a comprehensive analysis.

The proposed model, shown in Fig. 1, utilizes CNNs to extract features from raw PSG signals. The size of input data is $N \times 3750 \times C \times 1$ where N is the number of samples and C is the number of input channels. The first convolutional layer filters the input data using 8 kernels of size $C \times 1$ with a stride size of 1 point. The activation function of the first layer is a time-independent linear operation that projects diverse inputs into an optimal virtual space by adjusting weights and biases during model training. The dimension of virtual space is determined by the kernels of the first convolutional layer. To reduce the change of model parameters in transfer learning, we fix the dimension of virtual space. A permutation layer[20] is followed to hold channel information of virtual space and to transfer subsequent operations to the time domain.

The third layer is an integration block with three key components: a "squeeze and excitation" block to estimate channel weight[21], a convolutional layer with a smaller size to capture local features and a convolutional layer with a larger size for capturing the big context. Given the local receptive field of convolution operations, a global average pooling is used to squeeze global information which is then excited to generate channel-wise statistics[21]. We employ two CNNs

TABLE II. FILTERED FREQUENCY BANDS FOR EACH SIGNAL.

Signals	Frequency Band
EEG	0.5Hz-30Hz
EOG	0.5Hz-10Hz
EMG	10Hz-fs/2[a]
ECG	0.5Hz- 30Hz

[a] fs denotes sampling frequency.

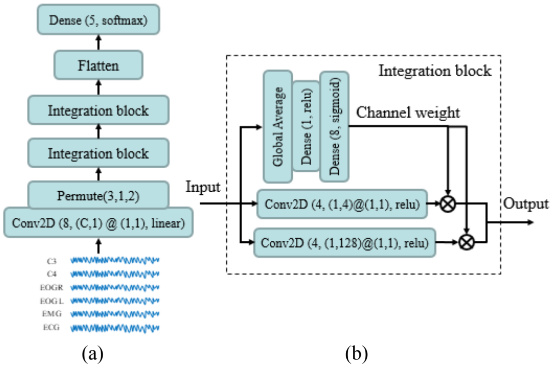

Fig. 1. Overview of the proposed architecture.

with small and large filter sizes to extract nonlinear features from the input data. Previous research[22] found that smaller filters are better to capture local contexts (e.g., when certain EEG patterns appear), while larger filters are better to capture big contexts. The outputs of these two CNNs are weighted and then concatenated into the final output of integration block. Two integration modules are adopted, each followed by a max-pooling layer with a size of (1,4), a dropout layer with a drop rate of 0.15 and a batch-normalization layer.

There is a dropout layer with a drop rate of 0.5 before the decision layer. The decision layer is a fully-connected layer with 5 units, which is activated by the softmax function. The output of the proposed architecture is a probability matrix with size N × 5, where N is the number of samples and 5 is the number of sleep stages. The stage prediction for each sample corresponds to the stage with the maximum probability in the probability matrix.

C. Model training

Only PSG recordings from the SHHS dataset were used to determine the structure and hyperparameters of the proposed model. The whole dataset was split into train, validation and test sets. We used PSG recordings from 20 subjects as the final test set, and recordings from 80 subjects for training and validation. It should be noted that only the data from the training set and validation set was used in the process of parameters selection and model training, and thereby the test set was completely independent. In order to find the best hyperparameters for the proposed architecture, we performed a random search using a Python package named hyperopt[23]. For each set of hyperparameters, we used 5-fold cross-validation to train and evaluate the classifier (64 subjects for training and 16 subjects for evaluation). TABLE III summarized the distribution of each hyperparameter value. The parameter set leading to the highest accuracy and the least variability was adopted. Finally, the optimal model was achieved by using Adam optimizer with a learning rate of 0.002 and a batch size of 256. The network was trained by minimizing categorical cross-entropy. The code was written in Keras[24] with a Tensorflow backend[25].

TABLE III. DISTRIBUTION OF HYPERPARAMETERS.

Hyperparameter		Distribution
First CNN	***Filters***	[4, 6, 8, 16, 32]
	Strides	[1, 2, 3, 5, 7]
Integration Block	***Filters***	[4, 8, 16, 32, 64, 128]
	Smaller Kernel size	[2, 4, 8, 16, 32]
	Bigger Kernel size	[32, 64, 128, 256, 512]
	Strides	[1, 2, 3, 5, 7]
	Activation	{'relu',' tanh'}
Pooling Size		[2, 3, 4, 5]
Dropout Rate		[0.05, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5]
Learning Rate		[0.001, 0.002, 0.003, 0.004, 0.005, 0.01]
Optimizer		{'Adam', 'SGD'}
Batch Size		[64, 128, 256, 512]

D. Transfer learning on small datasets

Usually, training of deep learning networks required large amounts of data, which was expensive and difficult for many sleep studies. Thus, model transferability became crucial because it made deep learning research on a small cohort a reality. To demonstrate the transferability of the proposed model, we evaluated model performance with the SHHS dataset as the source domain and the Sleep-EDF dataset as the target domain. The proposed model was firstly trained using data from six channels of 80 subjects from the SHHS dataset and then transferred the model to three channels of data from the Sleep-EDF dataset. Note that we adopted not only different signal modalities but also different numbers of channels for the source and target domains on purpose because we wanted to enforce more channel mismatches.

To evaluate the efficiency of transfer learning in the target domain, a leave-one-out cross-validation was conducted. It means that for each iteration, there are 18 recordings for fine-tuning the entire pre-trained network and one independent recording for testing. It's worth mentioning that fine-tuning does not change the model structure and hyperparameters, but only adjusts weights and biases of notes. That iteration repeats 19 times. The aggregated performance of 19 recordings will be reported in the next section.

III. PERFORMANCE ASSESSMENT

Model performance was evaluated by accuracy, sensitivity, precision, Cohen's kappa and F1 score. The detailed definition can be found in previous studies[7], [26], [27].

A. Performance on the SHHS dataset

To illustrate model performance, TABLE IV showed the confusion matrix obtained by test subjects from the SHHS dataset, where we can verify the distribution of samples that were correctly or incorrectly classified. As can be seen from Table IV, the overall classification accuracy was 85.2%, which exceeded the accepted benchmark $Acc = 80\%$ among trained human scorers[28]. The most correctly classified stage was wakefulness with a precision of 92.6%. It was followed by N3 (88.0%), R (86.2%) and N2 (85.1%). Stage N1 was the hardest to classify with 35.2% of samples correctly assigned. 34% of samples were misclassified as N2, 19% as R and 12% as W. That result was consistent with previous results. Stage N1 was considered as a transition state between wakefulness and "real" sleep, thereby including information from two or three sleep stages. As a result, the scoring of N1 was quite obscure, even for sleep scoring experts[29]. Closer inspection of TABLE IV showed that most misclassifications occurred in contiguous stages in the sleep cycle. For example, N3 was most often misclassified as N2, and rarely as N1. This error was mainly due to similar electrophysiological characteristics between adjacent stages, rather than defects of model design.

TABLE IV. CONFUSION MATRIX OF TEST RECORDINGS FROM THE SHHS DATASET.

	Technologists' score stage						Pre.
	Stage	***W***	***N1***	***N2***	***N3***	***R***	
Proposed	***W***	3369	68	119	3	80	92.6%
	N1	125	187	92	0	128	35.2%
	N2	139	184	7986	804	275	85.1%
	N3	0	0	325	2392	0	88.0%
	R	34	106	467	2	3018	83.2%
Sen.		91.9%	34.3%	88.8%	74.7%	86.2%	
Acc.							85.2%

B. Transfer learning

Model generalizability was tested on the Sleep-EDF dataset. As shown in TABLE I, the available channel, amplitude distribution and acquisition environment were significantly different between the two datasets, which may limit the use of models proposed by some studies. TABLE V displayed the performance of the transfer learning scenario (indicated by Sleep-EDF^b) compared to the model trained from scratch using only the Sleep-EDF data (shown by Sleep-EDF^a). The results in TABLE V showed that the proposed model outperformed all other state-of-the-art results on the Sleep-EDF dataset, whether training from scratch or using transfer learning. Moreover, despite the serious channel mismatch, transferring the knowledge of the source domain (the SHHS dataset) to the target domain (the Sleep-EDF dataset) brought up the accuracy compared to a network trained from scratch using only data from the target domain.

TABLE V also listed the model architecture, its approach, input channels, input types, the number of subjects and other parameters for comparison. As can be seen from TABLE V, our method achieved a comparable or better performance compared to state-of-the-art methods that used the same dataset but more complex model structure. For example, the deep-learning architecture proposed in Sors et al.'s study [30] had up to 10^6 parameters, while the proposed architecture did not exhibit more than 10^4 parameters. Note that this was at least two orders of magnitude lower than the architecture proposed by Sors et al. Moreover, to the best of our knowledge, the proposed model was the first one that can handle different numbers of input channels without changing model structure and hyperparameters. The compact and versatile structure was conducive to clinical applications.

IV. DISCUSSION

In this work, we presented a convolutional network to automatically classify sleep stages which would help alleviate the burden of practitioners. Most automatic methods reported so far were based on hand-crafted features or designed for certain datasets. Thus, it was hard to find a method that generalized correctly to other datasets, especially when the channel did not match. To solve this problem, we proposed a compact end-to-end recognition structure that can handle various numbers of input channels and several signal modalities without changing any layer or hyperparameter values. Experimental results have demonstrated that the proposed model exhibited strong classification performance and low computational cost on both datasets compared with state-of-the-art results. More importantly, the proposed model showed potential transferability on data with different channels. As shown in TABLE V, transfer learning brought a slight accuracy improvement compared to a network trained from scratch using only data from the target domain, and the authors believed that using a huge and high-quality source dataset contributes to performance improvement of transfer learning.

Few studies tested their models on recordings collected from different record environments and hardware platforms. Zhang et al. [28] did so, demonstrating generalizability by testing model on two novel datasets without using transfer learning. In this article, it was impossible to directly test model performance on the Sleep-EDF dataset due to different channel numbers and signal modalities. Phan et al.[15] evaluated different fine-tuning strategies of transfer learning using the SeqSleepNet+ model and the DeepSleepNet+ architecture. However, they only used two signal modalities (EEG and EOG) and the same number of input channels. The proposed model can simultaneously handle four commonly used signal modalities without limiting the number of input channels. In addition, we adopted a "squeeze and excitation" block[21] to adaptively recalibrate channel-wise feature responses, thereby making full use of channel information.

TABLE V. PERFORMANCE COMPARISON.

Ref.	Dataset	Subject	Input Channel	Input Type	Deep Learning Architecture		Approach	Result			
					Structure	***Layer***		***Accuracy***	***K***	***Macro F1***	***Micro F1***
Ref[31]	SHHS	1000	EEG: C3, C4 EMG 2 EOGs	Time series	1DCNN	37 CNN	One-to-one	0.78	0.83	0.76	--
Ref [30]	SHHS-1	5728	C4-A1	Time series	1DCNN	12 CNN	Many-to-one	0.87	0.81	0.78	0.87
Ref [32]	Sleep-EDF	20	Fpz-Cz	Time series	CNN+LSTM	--	Many-to-one	0.84	0.78	0.78	--
Ref [12]	Sleep-EDF	20	EEG EOG	Spectrogram	2DCNN	2 CNN	One-to-many	0.82	0.75	0.75	--
Proposed	SHHS	100	EEG: C3, C4 2 EOGs EMG ECG	Time series	2DCNN	5 CNN	One-to-one	0.85	0.79	0.76	0.85
	Sleep-EDF [a]	19	EEG: FpzCz, PzOz 1 EOG					0.84	0.77	0.78	0.84
	Sleep-EDF [b]	19	EEG: FpzCz, PzOz 1 EOG					0.85	0.79	0.80	0.85

LSTM: Long short-term memory unit which is an artificial recurrent neural network (RNN) architecture.

Sleep-EDF [a]: The model was trained from scratch.

Sleep-EDF [b]: The model was fine-tuned for the Sleep-EDF dataset.

Even though our results are encouraging, the proposed model is still subject to several limitations. Firstly, our model requires to be trained with a sufficient amount of sleep data to improve generalization. Secondly, it is worth to explore the addition of different deep learning modules, such as a combination of CNN and LSTM, since studies have found that the use of temporal context can significantly improve model performance. Thirdly, more fine-tuning strategies will be explored in our future work. For example, training only the first and decision layers can help speed up the use of transfer learning in the target domain.

V. CONCLUSION

This paper proposed an automatic sleep scoring model based on raw PSG recordings. The deep-learning network was composed of two parallel convolution layers with different filter sizes for capturing both fine and coarse temporal features, and a "squeeze and excitations" block to recalibrate channel-wise feature responses. Experiments on two common sleep datasets showed that the model achieved comparable performance and low computational cost compared to state-of-the-art methods. In addition, our results proved that the proposed model was able to handle various numbers of input channels and several signal modalities from different datasets without changing model architecture and hyperparameters. The versatile model can be integrated with diverse sleep monitoring devices, thereby facilitating sleep research in clinical or routine care.

ACKNOWLEDGEMENT

The authors would like to thank the Sleep Heart Health Study(SHHS)[16] and Physionet[17], [18] for providing the polysomnographic data.

REFERENCES

[1] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," Washingt. DC US Natl. Inst. Heal. Publ., 1968.

[2] R. B. Berry et al., "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events," J. Clin. Sleep Med., vol. 8, no. 5, pp. 597–619, 2012.

[3] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," J. Neurosci. Methods, vol. 271, pp. 107–118, 2016.

[4] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," Comput. Methods Programs Biomed., vol. 176, pp. 81–91, 2019.

[5] K. Šušmáková and A. Krakovská, "Discrimination ability of individual measures used in sleep stages classification," Artif. Intell. Med., vol. 44, no. 3, pp. 261–277, 2008.

[6] S. Özşen, "Classification of sleep stages using class-dependent sequential feature selection and artificial neural network," Neural Comput. Appl., vol. 23, no. 5, pp. 1239–1250, 2013.

[7] R. Yan et al., "Multi-modality of polysomnography signals' fusion for automatic sleep scoring," Biomed. Signal Process. Control, vol. 49, pp. 14–23, 2019.

[8] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, "DOSED: a deep learning approach to detect multiple sleep micro-events in EEG signal," J. Neurosci. Methods, vol. 321, pp. 64–78, 2019.

[9] Z. Mousavi, T. Yousefi Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi, "Deep convolutional neural network for classification of sleep stages from single-channel EEG signals," J. Neurosci. Methods, vol. 324, pp. 108312, 2019.

[10] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," Sleep, vol. 41, no. 5, pp. zsy041, 2018.

[11] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," arXiv Prepr. arXiv1610.01683, 2016.

[12] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 1285–1296, 2019.

[13] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, 2019.

[14] A. Malafeev et al., "Automatic human sleep stage scoring using deep neural networks," Front. Neurosci., vol. 12, pp. 781, 2018.

[15] H. Phan et al., "Towards more accurate automatic sleep staging via deep transfer learning," arXiv Prepr. arXiv1907.13177, 2019.

[16] D. A. Dean et al., "Scaling up scientific discovery in sleep medicine: The National Sleep Research Resource," Sleep, vol. 39, no. 5, pp. 1151–1164, 2016.

[17] 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.

[18] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," IEEE Trans. Biomed. Eng., vol. 47, no. 9, pp. 1185–1194, 2000.

[19] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," Physiol. Meas., vol. 36, no. 10, pp. 2027–2040, 2015.

[20] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 4, pp. 758–769, 2018.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. IEEE Conf. Comput. Vis. pattern Recognit., pp. 7132–7141, 2018.

[22] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, 2017.

[23] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," 30th Int. Conf. Mach. Learn. ICML 2013, no. PART 1, pp. 115–123, 2013.

[24] F. Chollet, "Keras: Deep learning library for theano and tensorflow," URL: https//keras. io/k, vol. 7, no. 8, pp. T1, 2015.

[25] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, pp. 265–283, 2016.

[26] E. Fernandez-Blanco, D. Rivero, and A. Pazos, "Convolutional neural networks for sleep stage scoring on a two-channel EEG signal," Soft Comput., vol. 24, no. 6, pp. 4067–4079, 2020.

[27] R. Yan, F. Li, X. Wang, T. Ristaniemi, and F. Cong, "An automatic sleep scoring toolbox: multi-modality of polysomnography signals' processing," ICETE 2019 - Proc. 16th Int. Jt. Conf. E-bus. Telecommun., vol. 1, pp. 301–309, 2019.

[28] L. Zhang, D. Fabbri, R. Upender, and D. Kent, "Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks," Sleep, vol. 42, no. 11, pp. 1–10, 2019.

[29] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," Artif. Intell. Med., vol. 53, no. 1, pp. 25–33, 2011.

[30] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J. F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," Biomed. Signal Process. Control, vol. 42, pp. 107–114, 2018.

[31] I. Fernández-Varela, E. Hernández-Pereira, and V. Moret-Bonillo, "A convolutional network for the classification of sleep stages," Multidiscip. Digit. Publ. Inst. Proc., vol. 2, no. 18, pp. 1174, 2018.

[32] S. Back, S. Lee, H. Seo, D. Park, T. Kim, and K. Lee, "Intra- and Inter-epoch Temporal Context Network (IITNet) for automatic sleep stage scoring," arXiv Prepr. arXiv1902.06562, 2019.