Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance

Qi Xu[a,b,1,*], Dongdong Zhou[c,d,1], Jian Wang[c,d], Jiangrong Shen[e], Lauri Kettunen[d], Fengyu Cong[a,c,d]

[a]*School of Artificial Intelligence, Dalian University of Technology, Dalian, China*
[b]*Guangdong Laboratory of Artificial Intelligence and Digital Economy, Shenzhen, China*
[c]*School of Biomedical Engineering, Dalian University of Technology, Dalian, China*
[d]*Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland*
[e]*College of Computer Science and Technology, Zhejiang University, Hangzhou, China*
Email: xuqi@dlut.edu.cn, dongdong.w.zhou@student.jyu.fi, wangjian009@mail.dlut.edu.cn
jrshen@zju.edu.cn, lauri.y.o.kettunen@jyu.fi, cong@dlut.edu.cn

***Abstract*—Accurate sleep stage classification is vital to assess sleep quality and diagnose sleep disorders. Numerous deep learning based models have been designed for accomplishing this labor automatically. However, the class imbalance problem existing in polysomnography (PSG) datasets has been barely investigated in previous studies, which is one of the most challenging obstacles for the real-world sleep staging application. To address this issue, this paper proposes novel methods with signal-driven and image-driven ways of noise addition to balance the imbalanced relationship in the training dataset samples. We evaluate the effectiveness of the proposed methods which are integrated into a convolutional neural network (CNN) based model. Experimental results evaluated on Sleep-EDF-V1, Sleep-EDF and CCSHS databases demonstrate that the proposed balancing approaches with specific tensity Gaussian white noise could enhance the overall or stage N1 recognition to some degree, especially the combination of two types of Data augmentation (DA) strategies shows the superiority of overall accuracy improvement.**

***Index Terms*—Sleep stage classification, Class imbalance problem, Data augmentation, Time-frequency image**

I. Introduction

Sleep is one of the most important human activities, which makes great contributions to one's mental and physical health and recovery [1], [2]. However, millions of people around the world suffer from different degrees and types of sleep-related issues [3]. It is a time-consuming and labor-intensive procedure to diagnose and treat them, thereinto correct sleep stage classification is an essential step. Clinically, whole-night sleep PSG data, including electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG), electrooculogram (EOG), etc, are divided into 30s epochs with labels of Wake (W), Rapid Eye movement (REM), Non-REM1 (N1), Non-REM2 (N2) and Non-REM3 (N3) by hands [4]. Although large amounts of deep learning methods have been proposed to handle this task automatically [5]–[12], it seems that there is still a gap from real-world implementation, one of possibilities is that the class imbalance problem (CIP) of PSG datasets which has not been paid enough attention and solved well.

In simple terms, the CIP in sleep scoring refers to the duration of each sleep stage is not equal because of the special sleep structure. For instance, stage W and N2 occupy the dominant proportion of samples (more than 60%). By contrast, the N1 stage usually accounts for 2%-5% of overnight sleep time. It is not fair for minority classes when training a deep neural network model with the imbalanced dataset. In such a way, the major categories contribute the leading weight updating, while the contribution of minority ones is biased during the back-propagation. Whether the overall accuracy or the recognition rate is limited by the CIP, which is worth further exploration. As the representation of minority groups, N1 stage suffers from heavy discrimination with the highest misclassification rate. Only a few of works have focused on the solutions for CIP in the sleep scoring. Supratak *et al*. [6] duplicated the minority sleep stages in the training set in which each sleep stage is equally shown. Similarly, Dong *et al*. [13] used oversampling to generate new samples to keep the same percentage of all sleep stages. However, if increasing the number of minority classes in a mechanical way to reach a state that all sleep stags have an equal number of samples, the initial sleep structure was totally destroyed. Fan *et al*. [14] applied five DA methods to assess the enhancement of overall accuracy and N1 classification rate, although the overall performance was improved, the N1 accuracy showed a slight drop sadly.

To remedy the CIP in the sleep scoring task with deep learning based models, we aim to balance the dataset samples by only increasing the number of N1 stage in the original training set with Gaussian white noise addition, this way could retain the original sleep architecture as much as possible. Additionally, we further investigate two categories of Gaussian white noise addition to the EEG signal. One is to add Gaussian white noise to the raw EEG signals (signal-driven) and then transform the noisy EEG signals to time-frequency images. Another one is to convert the EEG signals to time-frequency images then add the noise to the images (image-driven). These two balancing methods are embedded into a CNN based model to show the effectiveness of the relatively balanced state

[1] Equal contribution to this work, * Corresponding author: xuqi@dlut.edu.cn

TABLE I
THE SCHEME OF SIGNAL-DRIVEN AND IMAGE-DRIVEN APPROACHES

Intensity	signal-driven	image-driven
low	10 dB	mean = 0, variance = 0.05
moderate	5 dB	mean = 0, variance = 0.1
high	1 dB	mean = 0, variance = 0.2

between the imbalanced training data and model. Both signal-driven and image-driven balancing methods could improve overall accuracy or N1 accuracy to varying degrees.

The rest of this paper is organized as follows: The Sec.II describes the class imbalance problem and defines the class imbalance factor of PSG datasets. We present the experiments and experimental results in Sec. III and IV, respectively. The final conclusion and discussion are included in Sec.V.

II. CLASS IMBALANCE PROBLEM

Class imbalance is a common yet easily overlooked issue in the sleep stage classification task, the class distribution of the PSG dataset not only depends on the physical or mental conditions but also depends on the ages and genders. When the number of each category is severely unequal, we can say the dataset suffers from the CIP. Here, we define a class imbalance factor (CIF) to quantify the degree of CIP as follow:

$$CIF = \frac{N}{2 \cdot c \cdot \min\{N_i\}} \quad i \in \{1, 2, \ldots, c\} \tag{1}$$

Where the N is the total samples, c refers to the number of sleep stages, and N_i represents the number of each stage. If the $CIF = 0.5$, it means the dataset is balanced. If the $CIF > 0.5$ in eq. (1), that dataset could be regarded as an imbalanced one. Furthermore, the larger CIF means that the PSG dataset is more imbalanced. The CIP mainly affects the training procedure of the deep model which leads to erroneous results in pattern classification tasks. For example, one of the most popular used training rules in deep learning is the back-propagation (BP) algorithm, in which the major classes are responsible for prime parts of weight update. As a consequence, the minority categories become the biased ones with relatively lower recognition rate.

The straightforward way is to increase the number of minority classes to keep equivalent with others [6]. However, the original sleep architecture is broken completely in such a way. Therefore, we only generate new epochs for the N1 stage in training set with the noise addition to maintain the intact sleep structure as far as possible. In this study, we adopt the scheme with a time-frequency image input, which is generally considered as a higher-level representation of the raw signal and can get a faster training speed [11], [15]. Furthermore, we also investigate whether the sequence order of Gaussian white noise addition (i.e., before and after the time-frequency transform) would affect the final result. To be specific, the same type of noise (Gaussian white noise)

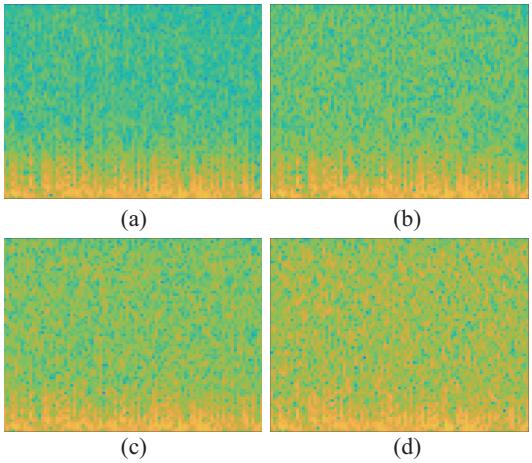

Fig. 1. (a) is the time-frequency image of raw EEG signal (N1 stage, Sleep-EDF-V1), the x-axis represents the time, the y-axis denotes the frequency. Subfigures (b), (c) and (d) illustrate the time-frequency images of raw EEG signal with 10, 5 and 1 dB Gaussian white noise addition respectively.

with three intensities is designed for comparison. The first method is to add the Gaussian white noise with 10, 5 and 1 dB (low, moderate and high intensities) to the raw EEG signal, respectively, then the noisy EEG signals are converted to time-frequency image using the short-time Fourier transform (STFT), it is a signal-driven approach to conduct the noise addition. As a comparison, the second scheme, the image-driven way, adds the similar intensities of Gaussian white noise (the mean (M) is 0, the variances (V) are 0.05, 0.1, 0.2 respectively) to the time-frequency image rather than the raw EEG signal. The scheme of two Gaussian white noise addition methods is demonstrated in Table I.

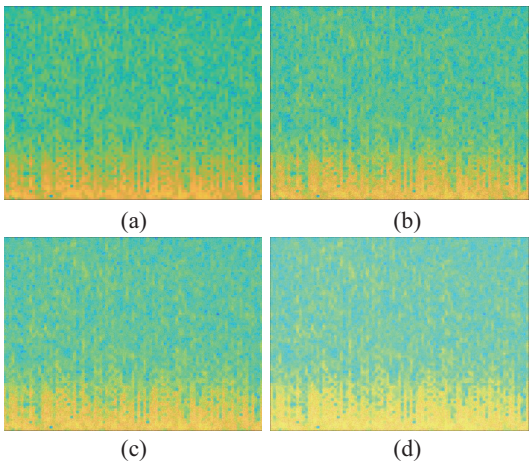

Fig. 2. (a) is the time-frequency image of raw EEG signal (N1 stage, Sleep-EDF-V1) which is the same as Fig. 1. The x-axis represents the time, the y-axis denotes the frequency. (a). Subfigures (b), (c) and (d) present the time-frequency images with three intensities of Gaussian white noise addition (variances are respectively 0.05, 0.1 and 0.2).

When attaining the optimal intensity of two balancing methods, the efficiency of the combination of two proposed

methods is further tested. We visualize the time-frequency images of two noise addition methods in Fig. 1 and Fig. 2.

III. Experiments

A. Experimental Datasets

1) Sleep-EDF-V1: The Sleep-EDF-V1 dataset has two subsets: sleep-cassette (SC) and sleep-telemetry (ST). In this study, we choose the 20 individuals with 39 overnight PSG recordings from the SC cohort, the age ranges from 25 to 34 years. As the suggestion of the American Academy of Sleep Medicine (AASM) manual, the frontal lobes Fpz-Cz channel EEG with a sampling rate of 100 Hz is adopted. More details are described in [16], [17]. The whole PSG recording was labeled with different sleep stages (i.e., W, N1-N4 and REM) based on the Rechtschaffen and Kales (R&K) [18], we merged the stages N3 and N4 into stages N3 for being consistent with the latest AASM standard.

2) Sleep-EDF: The Sleep-EDF dataset is the expanded version, including 78 subjects whose age stretches to 101 years. It has a higher proportion of N1 stages with the increase of age. In order to mitigate the negative impact of the long W stage period on overall accuracy (e.g., stage W has the highest classification accuracy), 30 minutes of W stages before and after regular sleep stages are employed for both versions of Sleep-EDF datasets.

3) CCSHS: The last PSG dataset used in this study is the Cleveland Children's Sleep and Health Study (CCSHS), which includes 515 children aged from 16-19 years. Due the absence of the FPz-Cz, we employ the the C4/A1 (sampled at 128 Hz) channel EEG instead. The main description can be found in [19], [20]. Here, we implement the many-to-one scheme which treats the combination of one 30 s epoch and its neighboring epochs as the contextual input (i.e., 90 s epoch). In Table II, we conclude the number of each sleep stage, the CIF is respectively 6.3%, 6.6% and 2.8% for the Sleep-EDF-V1, Sleep-EDF and CCSHS datasets. Although the sleep stage with the minimum number of Sleep-EDF is different from the other two datasets, we adopt the proposed balancing method only to increase the samples of stage N1 on all experimental datasets.

4) Data preprocessing: In this work, we adopt the STFT with a window size of two seconds and 50% overlap to convert the EEG signal to the image. Firstly, the EEG signal (with/without Gaussian white noise addition) is filtered by a notch filter, a high-pass filter and a low-pass filter in sequence. Hamming window and 256 points Fast Fourier Transform (FFT) [21] are further conducted to obtain the time-frequency image (efficient frequency band: 0.5-30 Hz).

B. Experimental setting

The whole dataset is divided into the training and test sets randomly based on the ratio of 4 to 1 (i.e., 80% subjects as the training set, 20% subjects as the test set). We use the Adam optimizer to train the model within 30 iterations, the model with the best performance in the test set is saved in all epochs. In addition, the learning rate would drop to half value when the test accuracy shows no enhancement within three epochs. The categorical cross-entropy is chosen as the model loss function. To find out a proper batch size, we assess four batch sizes (32, 64, 128 and 256), the batch size of 64 achieves the best performance. In our cases, a workstation with two Inter Xeon E5-2640 V4 CPUs and four Nvidia Tesla P100 GPUs with 16 GB memory is applied to conduct all experiments.

TABLE II
The data distribution of the experimental datasets

Stage	Sleep-EDF-V1	Sleep-EDF	CCSHS
W	10197 (23.1%)	69518(34.9%)	211030 (30.6%)
N1	2804 (6.3%)	21522 (10.8%)	19211 (2.8%)
N2	17799 (40.3%)	69132 (34.7%)	249681 (36.2%)
N3	5703 (13.0%)	13039 (6.6%)	110188 (16.0%)
REM	7717 (17.5%)	25835 (13.0%)	100252 (14.5%)

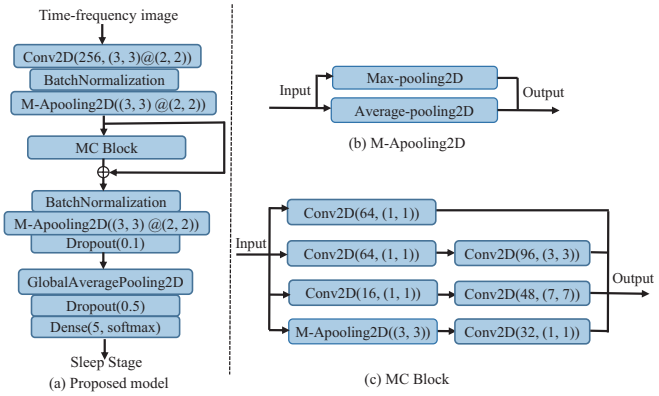

Fig. 3. The overall construct of the evaluation model.

C. The evaluation model

We construct a convolutional neural network based model to assess the efficiency of the proposed balancing method, it is treated as the baseline model (shown as Fig.3). The baseline model is mainly composed of a two-dimensional convolutional (Conv2D) layer, a multi-convolution (MC) block, two Max-Apooling2D layers and several BatchNormalization and dropout layers. The MC block, containing three filter sizes (1×1, 3×3, 7×7), is inspired by the inception module [24] to obtain the multi-scale feature representations. Similarly, we concatenate the outputs of Max-pooling2D and Average-pooling2D layers to rebuild as the Max-Apooling layer. The dropout layer aims to prevent the overfitting problem with a drop rate of 0.1 and 0.5. In addition, the Global Average Pooling (GAP) layer is used to replace the fully connected layer, which is considered more robust spatial translations of the input without parameter optimization [25]. The final dense layer employing the softmax as the activation function is implemented for predicting the sleep stage. We also apply the shortcut connection strategy to combine the input of the

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT INTENSITIES OF THE GAUSSIAN NOISE ADDITION (SIGNAL-DRIVEN AND IMAGE-DRIVEN WAYS) IN THIS WORK

	Sleep-EDF-V1			Sleep-EDF			CCSHS		
	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)	ACC (%)	K (%)	RE_N1 (%)
Without DA	86.3	81.1	38.9	84.5	79.0	24.6	87.0	82.0	22.9
DA (signal-driven, 10 dB)	85.4	80.0	35.2	84.5	79.0	26.1	87.3	82.4	25.7
DA (signal-driven, 5 dB)	86.8	81.1	42.7	84.3	78.6	18.7	87.2	82.3	24.1
DA (signal-driven, 1 dB)	87.1	82.3	34.8	84.7	79.3	24.0	87.5	82.5	27.3
DA (image-driven, V = 0.05)	87.0	82.1	30.8	84.6	79.0	15.6	87.1	82.1	21.9
DA (image-driven, V = 0.1)	87.0	82.2	34.5	84.6	79.1	21.1	87.3	82.3	22.2
DA (image-driven, V = 0.2)	86.1	80.7	30.3	84.6	79.1	25.9	87.3	82.3	23.0
DA (Combination, 1 dB & V = 0.1)	87.2	82.4	28.5	84.9	79.4	19.1	87.9	82.9	20.7

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHODS AND PREVIOUS METHODS ON THE CCSHS DATASET

Study	Method	Input channel	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Nakamura *et al.* [22]	HMM	C4/A1 + C3/A2	Spectrogram	515	-	73.0	-
Li *et al.* [23]	Random Forest	C4/A1	Features	116	86.0	80.5	7.3
Baseline	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.0**	**82.0**	**22.9**
DA (signal-driven, 1 dB)	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.5**	**82.5**	**27.3**
DA (image-driven, V = 0.1)	**CNN**	**C4/A1**	**Time-frequency image**	**515**	**87.3**	**82.3**	**23.0**

MC block with features learned from the MC block, in which 240 filters with size of 1×1 are used to unify the dimension.

IV. EXPERIMENTAL RESULTS

A. Overall performance

We employ the overall accuracy (ACC), Cohen's kappa coefficient (K) and class-wise recall of N1 (RE_N1) to assess the performance. The RE, ACC and K are defined as follows:

$$RE = \frac{TP}{TP + FN}. \tag{2}$$

$$ACC = \frac{\sum_{i=1}^{n} x_{ii}}{N} \tag{3}$$

$$K = \frac{\frac{\sum_{i=1}^{n} x_{ii}}{N} - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}{1 - \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} x_{ji}\right)}{N^2}}. \tag{4}$$

where TP and FN denote the true positives and false negatives respectively, N is the total number of all sleep stages, x_{ii} represents the diagonal value of the confusion matrix, n refers to the number of classes.

Table III illustrates the results of the baseline model without DA methods and different intensities Gaussian white noise addition using two balancing methods. We can see that the 1 dB Gaussian white addition could obtain the most significant improvement of ACC and K for the signal-driven DA on three datasets (Sleep-EDF-V1: ACC-0.8%, K-1.2%; Sleep-EDF: ACC-0.2%, K-0.3%; CCSHS: ACC-0.5%, K-0.5%). In terms of the RE of N1 stage, the 5, 10 and 1 dB achieve 3.8%, 1.5% and 4.4% enhancement from 38.9%, 24.6% and 22.9% on the Sleep-EDF-V1, Sleep-EDF and CCSHS datasets, respectively. Regarding the image-driven DA, the low and moderate intensities (V = 0.05 and 0.1) have the same ACC improvement on Sleep-EDF-V1 and Sleep-EDF datasets. Nevertheless, only the heavy intensity (V = 0.2) gains a gentle enhancement (1.3% and 0.1%) of RE_N1 on the Sleep-EDF and CCSHS databases. It is pleasant that the combination of two intensities (1 dB and V = 0.1) realise the most considerable ACC and K improvement (ACC-0.9%, K-1.3%; ACC-0.4%, K-0.4%; ACC-0.9%, K-0.9%) on the experimental datasets, but an unfavorable decrease in the RE_N1 on three datasets. In addition, two balancing approaches fail to show remarkable distinctions concerning the accuracy improvement with the experimental datasets.

B. Performance comparison

In order to further validate the efficiency of proposed methods, we also compare the overall and N1 accuracies with other works on the same dataset in Tables IV and V. It can be observed in Table IV that the proposed methods can outperform [22], [23] on the CCSHS dataset. Similarly, the baseline model shows better overall accuracy than the performance of [14], [15], [21] on the Sleep-EDF-V1 and Sleep-EDF datasets. Moreover, the performance (i.e., accuracies of all stages and N1) obtain further enhancement with proposed balancing methods.

TABLE V
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHODS AND PREVIOUS METHODS ON THE SLEEP-EDF-V1 AND SLEEP-EDF DATASETS

Study	Database	Method	Input type	Subjects	$ACC(\%)$	$K(\%)$	RE_N1
Ref. [14]	Sleep-EDF-V1	Deep CNN	Time series	20	74.8	66.0	-
Ref. [21]	Sleep-EDF-V1	1-max CNN	Time-frequency image	20	82.6	76	29.9
Ref. [15]	Sleep-EDF-V1	CNN	Spectrogram	20	86.1	81.0	-
Baseline	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**86.3**	**81.1**	**38.9**
DA (signal-driven, 5 dB)	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**86.8**	**81.1**	**42.7**
DA (image-driven, V = 0.1)	**Sleep-EDF-V1**	**CNN**	**Time-frequency image**	**20**	**87.0**	**82.2**	**34.5**
Ref. [26]	Sleep-EDF	CNN + LSTM	Time series	78	80.0	73	-
Ref. [27]	Sleep-EDF	CNN + LSTM	Time series	78	83.1	77	-
Ref. [11]	Sleep-EDF	RNN	Time series	78	84.0	77.8	-
Baseline	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.5**	**79.0**	**24.6**
DA (signal-driven, 10 dB)	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.5**	**79.0**	**26.1**
DA (image-driven, V = 0.2)	**Sleep-EDF**	**CNN**	**Time-frequency image**	**78**	**84.6**	**79.1**	**25.9**

V. CONCLUSION AND DISCUSSION

The inherent CIP existing in the PSG datasets has hindered the real-world application of automatic sleep scoring models greatly. In this paper, we try to explore the solutions for the CIP in the sleep stage classification procedures. We first define the CIF to quantify the imbalance degree in three common PSG datasets. Two balancing methods are further introduced to mitigate the undesirable effect from the types of noise addition. The first one is to add different intensities of Gaussian white noise to the raw EEG signal, the noisy EEG signals are then converted to the time-frequency images. In this way, extra frequency components could be added to the time-frequency images, it is called the signal-driven way. Another noise addition way is to add the Gaussian white noise to the time-frequency image directly, it is more similar to the implementation in the computer vision field, we name it the image-driven method. Different from previous studies balancing the PSG datasets with equal proportion [6], [13], [14], we argue that it would break the original overnight sleep structure and hide the physiological mechanism related to sleep. By contrast, we only increase the number of the minority class (N1 stage in this study) that we intend to improve to keep consistent with the test set as much as possible. The proposed methods are validated on a CNN based model with three public PSG datasets.

According to the experimental results, although there is no fixed intensity Gaussian white noise suitable for the enhancement of ACC, K and the recognition of N1 stage on experimental PSG datasets, the overall and N1 stage classification rate could be improved with different intensities. In addition, two DA methods do not show significant differences regard to the improvement of model performance. It can be inferred that it should be tailored to adopt the different intensities and types of Gaussian white noise addition based on the practical results on different properties of PSG datasets. In future work, we will explore more data argumentation methods to deal with the CIP of PSG datasets. Except for balancing the samples, how to balance the deep network is another aspect that can be considered.

ACKNOWLEDGMENT

This work was supported by National Key R&D Program of China National (No.2021ZD0109803), Natural Science Foundation of China (No.91748105), National Foundation in China (No. JCKY2019110B009, 2020-JCJQ-JJ-252), the Fundamental Research Funds for the Central Universities [DUT20LAB303, DUT20LAB308, DUT21RC(3)091] in Dalian University of Technology in China, Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ, No. GML-KF-22-11), CAAI-Huawei Mindspore Open Fund (CAAIXSJLJJ-2021-003A) and the Scholarships from China Scholarship Council (No.201806060164, No.202006060226). This study is to memorize Prof. Tapani Ristaniemi from University of Jyväskylä. We also thank Prof. Hämäläinen Timo from University of Jyväskylä for his great help in this study.

REFERENCES

[1] Pierre Maquet, "The role of sleep in learning and memory," *Science.*, vol. 294, no. 5544, pp. 1048–1052, 2001.

[2] Raffaele Ferri, Mauro Manconi, Plazzi, et al., "A quantitative statistical analysis of the submentalis muscle emg amplitude during sleep in normal controls and patients with rem sleep behavior disorder," *J. Sleep Res.*, vol. 17, no. 1, pp. 89–100, 2008.

[3] Vijay Kumar Chattu, MD Manzar, Soosanna Kumary, et al., "The global problem of insufficient sleep and its serious public health implications," in *Healthcare*. Multidisciplinary Digital Publishing Institute, 2019, vol. 7.

[4] Conrad Iber, Sonia Ancoli-Israel, Andrew L Chesson, et al., *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, vol. 1, Westchester, IL, USA: Amer. Acad. Sleep Med., 2007.

[5] Rui Yan, Fan Li, Dongdong Zhou, et al., "Automatic sleep scoring: A deep learning architecture for multi-modality time series," *J. Neurosci. Methods.*, vol. 348, pp. 108971, 2021.

[6] Akara Supratak, Hao Dong, Chao Wu, et al., "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.

[7] Stanislas Chambon et al., "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.

[8] Dongdong Zhou, Jian Wang, Guoqiang Hu, et al., "Singlechannelnet: A model for automatic sleep stage classification with raw single-channel eeg," *Biomed. Signal Process. Control.*, vol. 75, pp. 103592, 2022.

[9] Huy Phan, Fernando Andreotti, Navin Cooray, et al., "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, 2018.

[10] Wei Qu, Zhiyong Wang, Hong Hong, et al., "A residual based attention model for eeg based sleep staging," *IEEE J. Biomed. Health. Inf.*, vol. 24, no. 10, pp. 2833–2843, 2020.

[11] Huy Phan, Oliver Y Chén, Minh C Tran, et al., "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[12] Rui Yan et al., "A deep learning model for automatic sleep scoring using multimodality time series," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 1090–1094.

[13] Hao Dong, Akara Supratak, Wei Pan, et al., "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, 2017.

[14] Jiahao Fan, Chenglu Sun, Chen Chen, et al., "Eeg data augmentation: towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, pp. 056017, 2020.

[15] Dongdong Zhou, Qi Xu, Jian Wang, et al., "Lightsleepnet: A lightweight deep model for rapid sleep stage classification with spectrograms," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2021, pp. 43–46.

[16] Ary L Goldberger, Luis AN Amaral, Leon Glass, 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.

[17] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, et al., "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.

[18] Allan Rechtschaffen, "A manual of standardized terminology and scoring system for sleep stages of human subjects," *Electroencephalogr. Clin. Neurophysiol.*, vol. 26, no. 6, pp. 644, 1969.

[19] Guo-Qiang Zhang, Licong Cui, Remo Mueller, et al., "The national sleep research resource: towards a sleep data commons," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[20] Carol L Rosen, Emma K. Larkin, H. Lester Kirchner, et al., "Prevalence and risk factors for sleep-disordered breathing in 8-to 11-year-old children: association with race and prematurity," *J. Pediatr.*, vol. 142, no. 4, pp. 383–389, 2003.

[21] Huy Phan, Fernando Andreotti, Navin Cooray, et al., "Dnn filter bank improves 1-max pooling cnn for single-channel eeg automatic sleep stage classification," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2018, pp. 453–456.

[22] Takashi Nakamura, Harry J Davies, and Danilo P Mandic, "Scalable automatic sleep staging in the era of big data," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2019, pp. 2265–2268.

[23] Xiaojin Li, Licong Cui, Shiqiang Tao, et al., "Hyclasss: a hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, 2017.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, et al., "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit (CVPR).* 2015, pp. 1–9, IEEE.

[25] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv Preprint. arXiv:1312.4400,*, 2013.

[26] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLOS ONE.*, vol. 14, no. 5, pp. 1–15, 2019.

[27] Akara Supratak and Yike Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg," in *Proc. IEEE Eng. Med. Biol. Soc (EMBC).*, 2020, pp. 641–644.