JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Yan, Rui; Li, Fan; Zhou, Dong Dong; Ristaniemi, Tapani; Cong, Fengyu

Title: Automatic sleep scoring : a deep learning architecture for multi-modality time series

Year: 2021

Version: Accepted version (Final draft)

Copyright: © 2020 Elsevier B.V.

Rights: CC BY-NC-ND 4.0

Rights url: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the original version:

Yan, R., Li, F., Zhou, D.D., Ristaniemi, T., & Cong, F. (2021). Automatic sleep scoring : a deep learning architecture for multi-modality time series. Journal of Neuroscience Methods, 348, Article 108971. https://doi.org/10.1016/j.jneumeth.2020.108971

Journal Pre-proof

Automatic sleep scoring: A deep learning architecture for multi-modality time series

Rui Yan (Conceptualization) (Methodology) (Software) (Writing - original draft), Fan Li (Validation) (Investigation), Dong Dong Zhou (Validation), Tapani Ristaniemi (Supervision), Fengyu Cong (Supervision) (Funding acquisition)

PII: S0165-0270(20)30394-0

DOI: https://doi.org/10.1016/j.jneumeth.2020.108971

Reference: NSM 108971

To appear in: *Journal of Neuroscience Methods*

Received Date: 9 August 2020

Accepted Date: 10 October 2020

Please cite this article as: { doi: https://doi.org/

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier.

Journal Pre-proof

Automatic Sleep Scoring: A Deep Learning Architecture for Multi-modality Time Series

Rui Yan[a,b], Fan Li [a], DongDong Zhou[a,b], Tapani Ristaniemi[b], Fengyu Cong[a,b,c,d*]

[a] *School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China*

[b] *Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland*

[c] *School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, 116024, Dalian, China*

[d] *Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province. Dalian University of Technology, 116024, Dalian, China*

Highlights:

- A deep learning architecture is proposed to automate sleep scoring using multi-modality PSG signals.
- A linear activation function is adopted in the first CNN layer to accommodate different numbers of input channels, which helps to addresschannel mismatches.
- One LSTM module and two CNN moduleswith different kernels sizesare employedtocapture information across temporal and spatial scales.
- The proposed model achieves good performanceon three disparate datasets withdifferent subject attributions, thereby demonstrating model generalizabilityon different disease populations.
- Model transferability is demonstrated across three datasetswith different input channels and signal modalities.

Abstract

Background: Sleep scoring is an essential but time-consuming process, and therefore automatic sleep scoring is crucial and urgent to help address the growing unmet needs for sleep research. This paper aims to develop a versatile deep-learning architecture to automate sleep scoring using raw polysomnography recordings.
Method: The model adopts a linear function to address different numbers of inputs, thereby extending model applications. Two-dimensional convolution neural networks are used to learn features from multi-modality polysomnographic signals, a "squeeze and excitation" block to recalibrate channel-wise features, together with a long short-term memory module to exploit long-range contextual relation. The learnt features are finally fed to the decision layer to generate predictions for sleep stages.
Result: Model performance is evaluated on three public datasets. For all tasks with different available channels, our model achieves outstanding performance not only on healthy subjects but even on patients with sleep disorders (SHHS: Acc-0.87, K-0.81; ISRUC: Acc-0.86, K-0.82; Sleep-EDF: Acc-0.86, K-0.81). The highest classification accuracy is achieved by a fusion of multiple polysomnographic signals.
Comparison: Compared to state-of-the-art methods that use the same dataset, the proposed model achieves a comparable or better performance, and exhibits low computational cost.
Conclusions: The model demonstrates its transferability among different datasets, without changing model architecture or hyper-parameters across tasks. Good model transferability promotes the application of transfer learning on small group studies with mismatched channels. Due to demonstrated availability and versatility, the proposed method can be integrated with diverse polysomnography systems, thereby facilitating sleep monitoring in clinical or routine care.
Keywords: polysomnography; automatic sleep scoring; multi-modality analysis; deep learning

1. Introduction

Sleep is a vital physiological process as it covers approximately one-third of the human lifespan. Adequate and high-quality sleep is essential for physical restoration[1], memory processing[2] and metabolism[3]. Nowadays, probably due to our hectic lifestyle in modern society, complaints about sleep problems increase dramatically among people. An effective way to monitor sleep quality and diagnose sleep problems is overnight polysomnographic (PSG) test. The PSG test simultaneously records dozens of sleep signals, including electroencephalograms (EEG), electrooculogram (EOG), electromyograms (EMG),

Journal Pre-proof

electrocardiogram (ECG), airflow and respiratory effort. These recorded signals are generally analyzed by sleep experts based on the R&K rules[4] and recently updated American Academy of Sleep Medicine (AASM) standard [5].

Based on the amplitude and frequency characteristics of PSG 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). The most recent AASM standard merges NREM stages 3 and 4 into N3 due to their prevalent low-frequency oscillations in EEG signals. Assigning a sleep stage to each sleep segment, called sleep scoring, is a very important step in any sleep research. However, the manual sleep scoring is labor-intensive and subjective. Previous studies have reported that the annotation of an 8-h recording requires approximately 2-4 hours[6], and the inter-scorer reliability of sleep scorings is about 0.8[7]. Therefore, automatic scoring is deemed as a promising approach due to its cost efficiency and high precision.

Numerous attempts[8] so far have been made in the field of automatic sleep scoring. Scoring methods based on conventional machine-learning were prevalent, which usually included two main components: feature extraction and classification. There were wide varieties of techniques for feature extraction, including but not limited to statistic methods[9], Fourier transforms[10], wavelet analysis[11] and Hilbert transform[12]. These techniques were responsible for describing sleep signals from multiple aspects. In order to obtain an evaluation of sleep stages, these extracted features were then fed to a classifier[13], such as support vector machine[14], random forest[15], K-nearest neighbor classifier [16], Naive Bayes[10], artificial neural network[17]. These studies' accuracy ranged from 0.8 to 0.9 and highly depended on the validity of employed features.

Recently, approaches based on deep learning have sprung up since it avoided explicit feature extractions commonly seen in conventional machine-learning methods, and were especially suitable for big data approaches [18]. Mousavi et al. [19] proposed a convolutional neural network (CNN) to automate sleep scoring using EEG time series, which achieved competitive performance in the classification of 2 to 6 classes of sleep stages. Instead of raw signal inputs, time-frequency images, generated by the short-time Fourier transform [20] or the wavelet transform [21], were also explored in several studies. Zhang et al. [22] even compared these two different input representations and concluded that the network performance using the spectrogram as inputs was superior to that using time series as inputs, which was attributed to the compact information and less artifact in the spectrogram. Although CNN gave the most convincing performance in some fields, for example, computer vision and image recognition, it still suffered from some problems, such as the selection of hyper-parameters, feature redundancy, and vanishing gradients[23], which challenged the construction of deep convolutional networks.

Recurrent neural networks (RNN) were also important in deep learning networks because of their good performance in capturing temporal correlations of inputs[24]. One of the most popular was the long short-term memory network (LSTM) that solved the problem of vanishing gradients and long-term dependence in traditional RNN. The LSTM module had made great progress in the application of natural language processing[25]. In the field of automatic sleep scoring, some studies had revealed that the application of LSTM module helped to capture the inter-segment temporal contexts, thereby improving scoring accuracy [26]. However, the LSTM module required to calculate a lot of parameters and was prone to overfitting. In practical applications, the LSTM module usually relied on CNN modules[27] or conventional techniques of feature extraction[28] to compress the inputs, thereby saving computational cost.

Moreover, studies based on deep learning had introduced some novel classification schemes to mimic the way sleep experts performed in manual sleep scoring, such as one-input to multi-output schemes[29] and sequence-to-sequence models[30]. These novel schemes explicitly utilized the dependence of consecutive segments, which were impossible for conventional machine learning paradigms. According to their experiment results, the long-term dependence between segments led to significant performance improvement. In short, attempts on deep learning had yielded exciting results, although training models from scratch required a huge amount of training data and computational resources[31].

However, in terms of conventional machine-learning methods in automatic sleep scoring, their classification performances highly rely on extracted features. The elaborate features may underperform in other datasets, thus limiting model generalizability. For automatic sleep scoring methods based on deep learning, most models are designed for specific datasets and certain input signals, which require task-specific modification when their models are used in different tasks. Moreover, that modification is difficult and even inefficient, especially for sleep studies focused on a small group because of insufficient training data. In practical applications, differences of monitor devices and experimental objectives

Table 1 . Subject characteristics.

Para.		SHHS	ISRUC	Sleep-EDF
Subjects		100	99	19
Attribute		Near-health	Sleep disturbance	Health
Age		46.86 ±4.22	51±16	28.74±2.99
Criterion		R&K	AASM	R&K
Power Frequency		60Hz	50Hz	50Hz
Employed Channels		C3, C4, EOGR, EOGL, EMG, ECG	F3, C3, O1, F4, C4, O2, ROC, LOC, EMG, ECG	Fpz-Cz, Pz-Oz, EOG (horizontal)
Amplitude	**EEG**	[-26.0, 20.7]	[-149.4, 151.8]	[-208.2, 204.5]
	EOG	[-17.3, 17.7]	[-138.0, 146.1]	[-478.5, 449.6]
	EMG	[-22.3, 22.0]	[-524.7, 518.8]	--
	ECG	[-39.0, 44.2]	[-145.3, 121.0]	--
Sampling	**EEG**	125Hz	200Hz	100Hz
	EOG	50Hz	200Hz	100Hz
	EMG	125Hz	200Hz	1Hz
	ECG	125Hz	200Hz	--

Note: Unless specifically indicated, the above EEG channels were referred to the left or the right mastoids (M1 or M2) according to the 10–20 international electrode placement system.

∗ Corresponding author.
E-mail addresses: cong@dlut.edu.cn (Fengyu Cong), ruiyanmodel@foxmail.com (Rui Yan)

Journal Pre-proof

induce channel mismatch, which challenges the application of transfer learning[32]. To tackle the above problems, this work proposes a simple but versatile deep learning architecture that does not require task-specific modifications to the model architecture or hyper-parameters. The proposed architecture employs very few numbers of layers, thus resulting in low computation cost compared to other deep learning approaches. The main contributions of this work are presented as follows.

a) A deep learning architecture is proposed to automate sleep scoring using multi-modality PSG signals.
b) A linear activation function is adopted in the first CNN layer to accommodate different numbers of input channels, which helps to address channel mismatches.
c) One LSTM module and two CNN modules with different kernels sizes are employed to capture information across temporal and spatial scales.
d) The proposed model achieves good performance on three disparate datasets with different subject attributions, thereby demonstrating model generalizability on different disease populations.
e) Model transferability is demonstrated across three datasets with different input channels and signal modalities.

The article is organized as follows: Section 2 presents details of 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.

2. Methodology

2.1 Data description

This study adopted three public datasets to evaluate model performance. The first one was from the Sleep Heart Health Study (SHHS)[33], in which only the first round (SHHS-1) was selected in this study. The SHHS dataset recruited thousands of participants from nine existing epidemiological studies to investigate the relationship between sleep-disordered breathing and various cardiovascular diseases. A total of 100 subjects were selected out by restricting the respiratory disturbance index (RDI3P) < 5 to have near-normal characteristics. Besides, the selected subjects did not use beta-blockers, alpha-blockers, inhibitors, and did not suffer documented hypertension, heart disease and stroke.

The second one was ISRUC-Sleep dataset[34], of which subgroup 1 was chosen in the present article. This subgroup included 100 PSG recordings from healthy subjects, patients with sleep disorders and patients under the effect of sleep medication. Subject 8 was excluded due to the lack of required channels, and therefore only 99 subjects were analyzed in the following experiments. Each recording was visually labelled by two sleep experts according to the AASM standard[5]. To improve signal quality, dataset providers had filtered all signals by a 50Hz notch filter. In addition, the signals of EEG and EOG were filtered between 0.3Hz and 35Hz, and EMG signals were filtered between 10Hz and 70Hz.

The third dataset was the Sleep-EDF dataset[35], [36], in which the sleep cassette (SC) subset was adopted. It consisted of 20 healthy subjects whose age ranged from 25 years old to 34 years old. Each subject had 2 PSG recordings about 20 hours each, except for subject 13 who had only the first-night recording. The recorded two PSG recordings for each subject were from two consecutive day-night periods at subjects' home. To avoid "the first night effect", PSG recordings from the second night were employed in the present study, and thus a total of 19 recordings were analyzed. Table 1 summarized the characteristics of employed recordings, where the age was shown as mean age ± standard deviation.

To accommodate data from different datasets, all signals were sampled or resampled to 125Hz. In order to remove noise and artefacts, all signals were filtered by a notch filter, a high-pass filter and a low-pass filter. The effective frequency band of EEG and ECG signals was limited to 0.5Hz-30Hz, 0.5Hz-10Hz for EOG signals, and only information above 10Hz was retained for EMG signals. In addition, for recordings in the Sleep-EDF dataset, the long awake periods before and after sleep were trimmed to restrict our analysis to the nocturnal sleep. In order to minimize the variability between recordings, each signal was normalized by mapping its mean to 0 and its deviation to 1. Table 1 displayed the amplitude ranges of signals after preprocessing. Afterwards, all the signals were divided into 30-second segments, each segment corresponding to a single sleep stage. For PSG recordings scored using R&K rules, NREM stages 3 and 4 were merged into N3 in the present article according to the recently updated AASM standard.

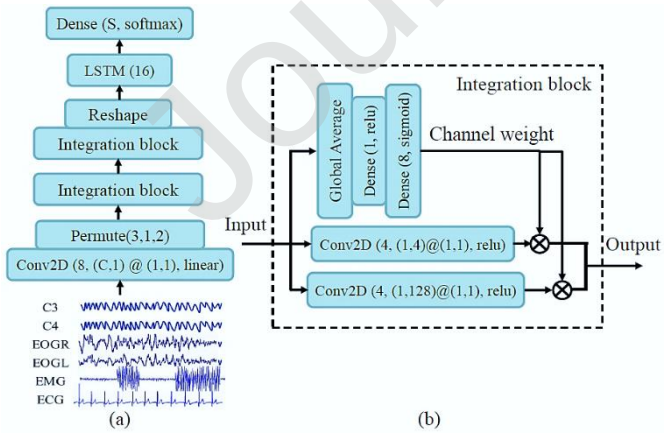

Figure 1. Overview of the proposed architecture.

2.2 Model Architecture

The proposed deep-learning architecture extends the input from EEG signals to a fusion of multiple PSG signals. The idea imitates the way sleep experts perform manual sleep scoring. Besides the characters of EEG signals, sleep experts also check eye movements and muscle activities as a reference when they label a 30-second PSG segment[4], [5]. For example, Stage R is characterized by low-amplitude and mixed-frequency EEG activities, rapid eye movements and the lowest EMG activity

Journal Pre-proof

level; Stage N3 is marked by high-amplitude slow waves and rare eye movement. Recent studies have revealed that analysis of cardiac electrophysiological activity enables us to track the transition from wakefulness to sleep[37]. Hence, there is every reason to believe that the joint processing of multiple PSG signals is conducive to an accurate scoring performance.

Figure 1 shows a schematic diagram of the proposed model to provide an intuitive manner to capture the model structure. Detailed model parameters and layer outputs are provided in Table 2. As can be seen from Figure 1, the proposed architecture comprises several CNN modules and one LSTM module to extract spatial and temporal features from raw PSG signals. The size of the input data is $\mathrm{T} \times \mathrm{C} \times 1$ where T is the number of time points and C is the number of input channels. Since the sampling rate of employed signals is set to 125Hz and each sample lasts 30 seconds, hence, $\mathrm{T} = 3750$ in the present article. The proposed architecture does not restrict the number of input channels C which may be diverse in different datasets.

The first convolution layer filters the input data using 8 kernels of size $\mathrm{C} \times 1$ with the stride size of 1 point. The activation function of the first layer is a time-independent linear operation. Similar linear functions have been used as spatial filters in the study of Chambon et al. [38]. Here, we use the linear function to accommodate mismatched input channels, and thus subsequent model parameters can be free from the influence of varying numbers of input channels. The outputs of the first layer are a set of linear combinations of input signals. The optimal combinations can be achieved by adjusting weights and biases of kernels during model training. This operation can be considered as a projection that maps diverse inputs into the optimal virtual space, thereby compensating channel mismatch. In addition, in order to prevent the model from overfitting[39], we apply a L2 weight regularization with a value of 0.01 in the first convolution layer. A permutation layer[38] is followed to hold channel information of the 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 weights, a convolution layer with a smaller kernel size to capture local features and a convolution layer with a larger kernel size for capturing the big context. In view of the local receptive field of convolution operations[40], global information is required to evaluate channel weights, which is achieved by a global average pooling. Two fully-connected layers followed the global average pooling are to excite the nonlinearity of among weights[40]. We employ two CNNs with small and large filter sizes to extract nonlinear features from its input. The previous study[39] has found that smaller kernels are better to capture local contexts (i.e., when certain of EEG patterns appear), while larger kernels are conducive to capturing big contexts. The outputs of the two CNN modules are weighted by channel-wise statistics and then concatenated into the final output of the integration block. Two integration modules are adopted, each followed by a max-pooling layer with a size of (1, 16), a dropout layer with a drop rate of 0.15 and a batch-normalization layer. Here, the large pooling size is to compress temporal information, thereby reducing model parameters and memory requirements. The layers of dropout and batch-normalization help to control overfitting.

The long short-term memory (LSTM) module is arranged before the decision layer to dig up long-range contextual information. A typical LSTM unit has a memory cell and three gates, namely an input gate, an output gate and a forget gate, to regulate the retention or discard of information flow. The unique mechanism allows LSTM units to selectively remember the previous information, thereby facilitating the current decision. Previous studies[39] have revealed that the context information helps to capture the transition rules among sleep stages. The conclusion is consistent with manual scoring rules[4], [5]. The transition rules allow sleep experts to predict possible sleep stages for the current segment based on a sequence of PSG segments. These transition rules are especially helpful for decision-making when signal characters of the current segment are ambiguous.

The final layer is the decision layer, which is a fully-connected layer activated by the softmax function. The number of units is equal to the number of classes. In the present article, we split sleep segments into five sleep stages, namely W, N1, N2, N3 and R, and therefore $\mathrm{S} = 5$. The output of the decision layer is a probability matrix with a size of $\mathrm{N} \times \mathrm{S}$, where N is the number of samples (or sleep segments) and S is the number of sleep stages. The stage prediction for each sample corresponds to the stage with the maximum probability.

Table 2. Architecture detail

Layer	Type	Units	Size	Stride	Activation	Output size
Input						(3750, C, 1)
1	Conv2D	8	(1, C)	(1, 1)	linear	(3750, 1, 8)
2	Permute					(8, 3750, 1)
3	Integration block					(8, 3750, 8)
4	Max-pooling		(1, 16)			(8, 234, 8)
5	Dropout (0.15)					(8, 234, 8)
6	Batch normalization					(8, 234, 8)
7	Integration block					(8, 234, 8)
8	Max-pooling		(1, 16)			(8, 14, 8)
9	Dropout (0.15)					(8, 14, 8)
10	Batch normalization					(8, 14, 8)
11	Permute					(14, 8, 8)
12	Reshape					(14, 64)
13	LSTM	16			tanh	16
14	Dense	5			softmax	5

2.3 Hyper-parameter optimization

The selection of hyper-parameters was carried out on only the SHHS dataset via 5-fold cross-validation. The whole dataset was split into five subsets, each with 20 subjects. For a given hyper-parameter set, the proposed model was trained on data from 4 subsets and tested on data from the remaining subset. In addition, we used 20% of training data for model validation. This process was repeated 5 times, with each subset being used as test data once. The final performance on this hyper-parameter set was determined by the aggregated test performance across all five folds. It should be noted that once the optimal hyper-parameter set determined, it would be used in all experiments.

Journal Pre-proof

Therefore, there was no task-specific modification to model structure and hyper-parameters, except for the kernel size of the first convolution layer that was determined by the number of input channels.

In order to find the best hyper-parameters for the proposed architecture, we performed a random search using a Python package named hyperopt[41]. The number of iteration was set to 50. The search space of hyper-parameters was summarized in Table 3. The parameter set leading to the highest accuracy and the lowest variability was adopted as the optimal parameters. If two sets of parameters gave a similar performance, the one with lower computational costs would be selected. 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 using the Keras package[42] with the Tensorflow backend[43].

3. Performance assessment

Model performance is evaluated by accuracy, precision, recall, F1 score and Cohen's kappa.

Accuracy (Acc.) measures the proportion of samples that the model correctly predicted.

Precision (P) is the fraction between true positives and the predicted positives.

Recall (R), also named sensitivity, calculates the percentage of actual positives that the model correctly identified.

F1 score (F1) represents the harmonic mean between precision and sensitivity.

Kappa (K) is an agreement measure between the proposed model and a human expert, which takes into account the chances of random agreement. A Large value indicates a high agreement between two classification results, and the perfect agreement gets a value of 1.

3.1 Classification performance

In order to illustrate model performance, Table 4 showed the aggregated confusion matrix from 5-fold cross-validation on the SHHS dataset. The confusion matrix clarified the distribution of samples that were correctly or incorrectly classified. From Table 4, we can see that the total classification accuracy was 0.87, which exceeded the accepted benchmark $Acc = 80\%$ among trained human scorers[7]. The best classification was wakefulness with the precision of 0.93. It followed by N2 (0.87), N3 (0.87) and R (0.83). Stage N1 was the hardest class to classify, with 31% of samples correctly assigned. There were 25% of N1 samples misclassified as R, 25% as N2 and 19% as W. The low precision of N1 stage was common in studies. Stage N1 was considered 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[44]. Closer inspection of Table 4 showed that most misclassifications occurred in contiguous stages in the sleep cycle. For example, N3 was often misclassified as N2, and rarely misclassified as N1. These misclassifications were mainly due to similar or mixed electrophysiological characteristics between adjacent stages, rather than the defect of model design.

Table 4. Confusion matrix for test recordings from the SHHS dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	18925	575	456	9	281	0.93	0.93	0.93
	N1	330	937	493	0	224	0.47	0.31	0.37
	N2	712	740	38442	3599	882	0.87	0.90	0.88
	N3	21	0	1512	10662	1	0.87	0.75	0.81
	R	433	762	1947	2	15569	0.83	0.92	0.87
Accuracy									0.87
Kappa									0.81

Table 5. Confusion matrix for test recordings from the Sleep-EDF dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	2347	81	22	3	8	0.95	0.84	0.89
	N1	232	901	333	6	107	0.57	0.58	0.57
	N2	34	227	7430	234	95	0.93	0.86	0.89
	N3	5	8	392	2476	1	0.86	0.91	0.88
	R	166	347	422	3	3750	0.80	0.95	0.87
Accuracy									0.86
Kappa									0.81

Table 6. Confusion matrix for test recordings from the ISRUC dataset.

		Technologists' score stage					P	R	F1
	Stage	W	N1	N2	N3	R			
Proposed	W	22804	905	174	19	127	0.95	0.94	0.94
	N1	1091	7021	1708	13	631	0.67	0.68	0.67
	N2	166	1703	24175	2196	410	0.84	0.88	0.86
	N3	18	19	865	12637	7	0.93	0.84	0.89
	R	246	749	612	132	7723	0.82	0.87	0.84
Accuracy									0.86
Kappa									0.82

Table 3. Search set for hyper-parameters

Hyper-parameter	Distribution
Integration CNN Block	
Filters	[4, 6, 8, 16, 32]
Strides	[1, 2, 3, 5, 7]
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]
LSTM	
Unit	[6, 8, 16, 32, 64, 128]
Activation	{ 'relu', 'tanh' }
Pooling Size	[2, 3, 4,...,15, 16]
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]

Journal Pre-proof

To test the generalization capability, the proposed model was further evaluated on two independent datasets, the Sleep-EDF and the ISRUC dataset, in which subjects in the ISRUC study suffered from diverse sleep disorders. As shown in Table 1, the available channels, amplitude distributions and acquisition environments were significantly different among these three datasets. Besides, the model architecture and hyper-parameters were determined by recordings from the SHHS dataset, and they would remain unchanged in the classification of sleep segments from the other two datasets. In terms of the Sleep-EDF dataset, signals from three available channels (FpzCz, PzOz, EOG) were employed as model inputs, and a leave-one-out cross-validation was performed to evaluate model performance. For the ISRUC dataset, 10 available channels were adopted including six EEG, two EOG channels, one EMG channel and one ECG channel. Model performance was evaluated using 5-fold cross-validation to provide a generalized model evaluation. Table 5 and Table 6 presented the confusion matrix obtained on test recordings from the datasets of Sleep-EDF and ISRUC, respectively.

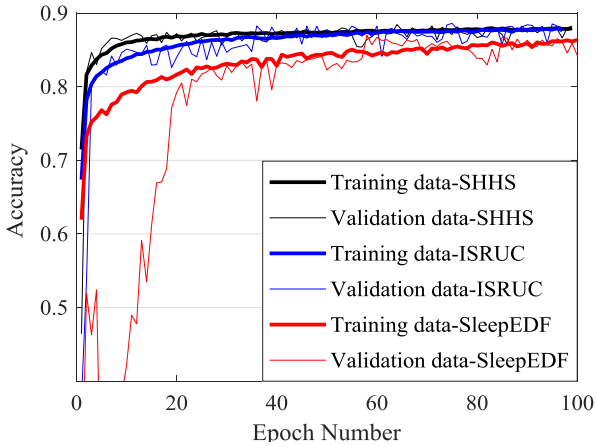

Figure 2. Train curve of three datasets

Comparing Table 4, Table 5 and Table 6, we can get that the proposed model gave outstanding performance on three disparate datasets, no matter healthy subjects from the Sleep-EDF dataset or patients with complex sleep disturbances from the ISRUC dataset. For recordings from the ISRUC dataset or Sleep-EDF dataset, the N1 stage got acceptable precision despite its small sample size, which further demonstrated that our method could tackle the problem of unbalanced classes.

Furthermore, we displayed the learning curves of the proposed model on three datasets. Figure 2 showed the changes of accuracy versus the number of iterations for training data and validation data. As it is seen, the network accuracy improved with increasing numbers of iteration (indicating by “Epoch Number” in Figure 2). Since model hyper-parameters were selected based on the data from the SHHS dataset, the convergence speed of the network was the fastest on the SHHS dataset, followed by the ISRUC dataset, and that on the Sleep-EDF dataset was the slowest. Nevertheless, using early stopping with a patience of 10 epochs to monitor the validation loss, the model training could complete within 100 iterations. Given the limited sample amount, the final accuracy on the Sleep-EDF dataset was inferior to those on the SHHS dataset and the ISRUC dataset. In order to improve classification accuracy of the Sleep-EDF dataset, we applied a fine-tuning strategy, which would be introduced in detailed in section 3.4.

Journal Pre-proof

Journal Pre-proof

3.2 Model performance under different signals' fusions

In order to explore the effect of signal type on classification performance, we investigated different fusions of PSG signals. The analyzed signals included EEG, EOG, EMG and ECG. This experiment was performed on the ISRUC dataset due to its abundant channels. To provide an unbiased estimate of the model performance, we conducted a subject independent 5-fold cross-validation on 99 recordings. The results were shown in Figure 3, where the column represented the average accuracy of 5-fold cross-validation and the bar denoted the standard deviation. For the fusions of more than two signals, the signal name was abbreviated to its middle letter, such as C&E denoted the fusion of ECG signals and EEG signals, and M&O&E meant the fusion of signals of EMG, EOG and EEG.

As can be seen from Figure 3, abundant signals were conducive to improving accuracy and reducing uncertainty. Specifically, in the perspective of single-channel EEG inputs, time series from the C4 channel achieved the best performance with the mean accuracy of 0.78 and the standard deviation of 0.004. Time series from the O2 channel performed the worst, which may be attributed to the poor signal quality caused by the uncomfortable electrode locations. Adding EEG channels or other PSG modalities enhanced model performance, but up to a certain extent. The fusion of EEG, EOG and EMG signals produced the best performance in this experiment, with the average accuracy of 0.87 and the standard deviation of 0.002. In terms of signal types, the performance of EMG signals and EOG signals was superior to ECG signals, likely due to the morphological difference of ECG signals.

3.3 Performance comparison

The performance of the proposed model was compared with recent studies that used the same datasets. Table 7 showed model performance, together with model architectures, their approaches, input channels, input types, subject numbers and other parameters for comparison. What stood out in Table 7 was that our method achieved a comparable or better performance compared to the state-of-the-art methods that used the same dataset but more complex model structure.

More specifically, for studies on the Sleep-EDF dataset, our model achieved an accuracy of 0.86 and a kappa value of 0.81, which exceeded 2% on accuracy and 3% on kappa value compared to the "many to one" classification scheme proposed by Back et al.[45]. For studies on the ISRUC dataset, there was

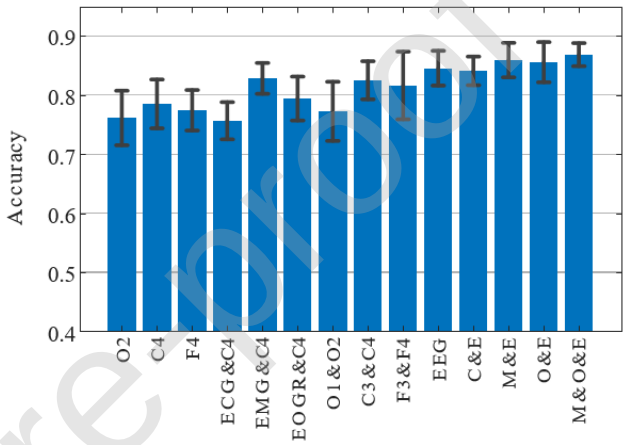

Figure 3. The classification accuracy for different signal fusions

Table 7. Performance comparison

Ref.	Dataset	Subjects	Input Channel	Input Type	Architecture		Approach	Result	
					Structure	*Layers*		*Acc.*	*Kappa*
Ref[48]	SHHS	1000	EEG: C3, C4 EOG: ROC, LOC EMG	Time series	1DCNN	37 CNN	One-to-one	0.78	0.83
Ref [47]	SHHS-1	5728	C4-A1	Time series	1DCNN	12 CNN	Many-to-one	0.87	0.81
Ref [45]	Sleep-EDF	20	Fpz-Cz	Time series	CNN+LSTM	--	Many-to-one	0.84	0.78
Ref [29]	Sleep-EDF	20	EEG EOG	Spectrogram	2DCNN	2 CNN	One-to-many	0.82	0.75
Ref[46]	ISRUC	40	EEG: F3, C3, O1, F4, C4, O2 EOG: ROC, LOC EMG	Features	Random forest	--	--	0.82	--
Proposed	SHHS	100	EEG: C3, C4 EOG: ROC, LOC EMG ECG	Time series	2DCNN+LSTM	5 CNN	One-to-one	0.87	0.81
	ISRUC	99	EEG: F3, C3, O1, F4, C4, O2 EOG: ROC, LOC EMG ECG					0.86	0.82
	Sleep-EDF	19	EEG: FpzCz, PzOz EOG					0.86	0.81

Note: Unless specifically indicated, the above EEG channels were referred to the left or the right mastoids (M1 or M2) according to the 10–20 international electrode placement system.

Journal Pre-proof

a significant improvement (+4% on accuracy) between the proposed model and Khalighi et al.'s methods[46]. For studies on the SHHS dataset, our model obtained comparable performance with Sors et al.'s study. However, the deep-learning architecture proposed in Sors et al.'s study[47] employed 12 convolution layers and two fully-connected layers with about 10^6 parameters, while the proposed model exhibited about 10^4 parameters. Note that this was at least two order of magnitude lower than the model proposed by Sors and his colleagues. The compact structure helped saving training time and computational cost, thus facilitating clinical practice.

Moreover, few studies [29], [45]–[48] had tested their model on diverse datasets with different sample attributes, input channels and disease populations. The proposed model shows stable performance on three datasets with completely different attributes, indicating good model generalization in different datasets and sample populations.

Table 8. Model generalizability

Model		Direct predict		Fine tuning with 20 subjects	
		Acc.	K	Acc.	K
Training Testing	SHHS ISRUC	0.73	0.64	0.84	0.79
Training Testing	ISRUC SHHS	0.66	0.55	0.84	0.77

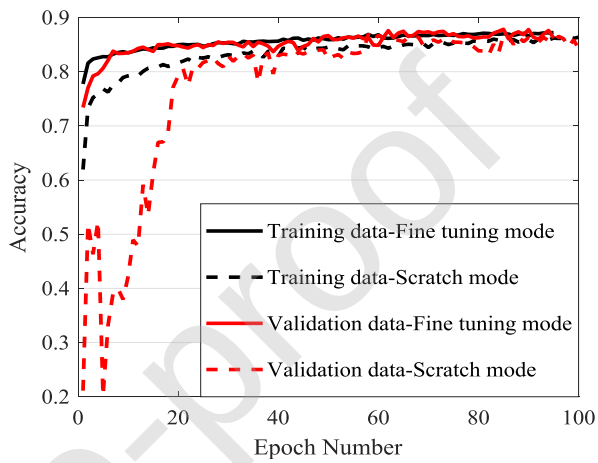

Figure 4. Train curve comparison between fine-tuning mode and scratch mode on the Sleep-EDF dataset

3.4 Evaluation of model transferability

In order to test classification performance of the trained model against data that the model had never seen before, we tested model transferability among three datasets. In terms of channel-matched cases, six matched channels were extracted from the datasets of SHHS and ISRUC in this experiment. After the model was trained on one dataset, the trained model was directly used to predict sleep stages for recordings from the other dataset. It was worth noting that the trained model did not suffer any modification for test data. Table 8 showed the classification results. As can be seen from Table 8, the direct prediction achieved moderate classification accuracy, which may be attributed to the lack of huge training dataset. Nevertheless, fine-tuning the trained model with a small amount of test data, the accuracy can be significantly improved. In addition, the SHHS dataset contained near-healthy participants, while the ISRUC dataset involved patients with complex sleep disturbance. The results indicated good model transferability between different disease populations.

In the case with channel mismatch, the direct prediction was impossible. Here, we tried two classification strategies on the Sleep-EDF dataset: fine-tuning a trained model or training a new model from scratch. For a fair comparison, we used leave-one-out cross-validation and the same set of model parameters for these two classification strategies. The adopted model parameters were the same as those described in Section 2 and those used in previous experiments. The model for fine-tuning was trained on the SHHS dataset. Figure 4 displayed the learning curve of these two classification strategies. As can be seen from Figure 4, the fine-tuning strategy resulted in a faster and smoother convergence curve compared to that of the model trained from scratch. Classification performance improved by 1.6% on accuracy and 2.7% on kappa using the fine-tuning strategy. Table 5 showed the detailed confusion matrix under the fine-tuning strategy.

Journal Pre-proof

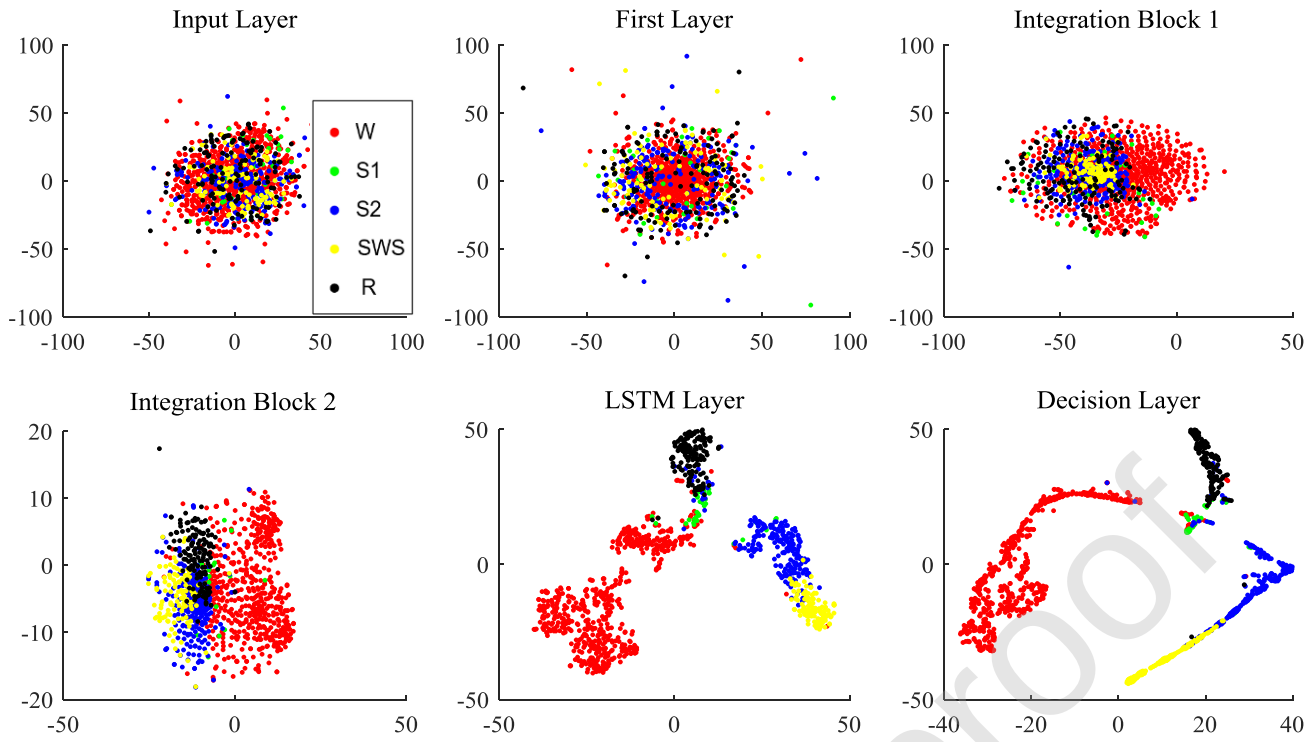

Figure 5. Model visualization using t-SNE method.

3.5 Model visualization

In order to illustrate how well each layer distinguished sleep stages, we visualized layer outputs using t-Distributed Stochastic Neighbor Embedding (t-SNE)[49]. The t-SNE can transform high-dimensional data into two-dimensional data to facilitate data visualization. Figure 5 displayed compressed layer outputs when the trained model predicted subject "shhs1-204846" from the SHHS dataset.

It can be seen from the first map in Figure 5 that the distribution of input data was random. The first layer with linear activation functions regularized the inputs. As moving forward from the integration block 1 to the decision layer, five sleep stages were more clearly separated. In particular, the LSMT layer led to a significant increase in separability, and the decision layer resulted in further clear separation.

4. Discussion

In this work, we develop a deep learning network for the automatic classification of sleep stages. Most of the automatic methods reported so far are based on human-engineered features or designed for a specific dataset. Thus, these models are hard to generalize correctly and easily to other datasets, especially when the channels do not match. To address these problems, we propose a compact and versatile end-to-end architecture to automate sleep scoring. We think two characteristics propel our model better than state-of-the-art methods. The first is good generalization and transferability. The above experiments have demonstrated that our model achieves strong classification performance on three disparate datasets, no matter whether it is from the healthy or the patients with sleep disturbance. This indicates good model transferability and generalization among different datasets and disease populations. The characteristic avoids cumbersome task-specific adjustments to model architecture and hyper-parameters, thereby facilitating clinical applications. Moreover, the proposed structure is conducive to the fine-tuning strategy, especially in the cases with limited training data and mismatched channels, which can significantly improve classification accuracy. Secondly, the proposed model exhibits a relatively low number of parameters, which drastically reduces training time, thereby saving computational resources.

The proposed architecture takes raw PSG signals as input without any human-engineered features, thereby preserving the coherence among multi-modality signals. There is no elaborate processing on raw PSG signals, except for a simple filtering process to improve the signal-to-noise ratio. Besides, the whole PSG recording is fully included in the analysis without discarding any recorded segments, even severely contaminated segments. The crude pre-processing enhances model robustness, and therefore the proposed model is more easily adaptable to noisy clinical applications. We have noticed that some studies claimed the network using raw PSG signals as inputs showed inferior performance[22] and was more prone to overfitting [32], compared with that using spectrograms as inputs. Therefore, we adopt several strategies to control overfitting, such as the L2-regularization

Journal Pre-proof

in the first CNN layer, dropout layer and batch-normalization. To improve model performance, CNN modules with different kernel sizes and LSTM modules are employed to capture information across spatial and temporal scales. Experimental results prove the feasibility of these strategies.

The proposed model is capable of coping with multiple PSG signals. Experiments have demonstrated that the input of multi-modality signals is conducive to the improvement of model performance. This conclusion is consistent with the findings of our previous research[15] and the manual scoring standards[4], [5]. Sleep experts inspect multiple PSG channels, including EEG (records of brain activity), EOG (records of eye movement) and EMG (records of muscle activity). The additional EOG and EMG channels usually provide important information to distinguish sleep stages, especially when EEG activity is ambiguous, such as wakefulness and REM stages. The results in Figure 3 show that the addition of EMG and EOG produces a better and more stable model performance.

Although ECG is not recommended for manual sleep scoring in the scoring standards of AASM or R&K, it is undeniable that ECG is one of the most commonly used tools in clinical to monitor vital signs. In sleep scoring, the application of ECG channels facilitates to distinguish signal artifacts. Besides, according to our previous research[15], ECG signals perform well in distinguishing sleep and wakefulness. Given that, we train the model to recognize ECG signals so that it can contribute to the discrimination of sleep stages in different classification problems, for example, binary classification of sleep segments. In addition, by changing the number of units in the decision layer, the proposed model can be easily applied to different classification problems of sleep stages, such as distinguishing sleep state and awake state, the recognition of light sleep and deep sleep.

Few studies tested their model on PSG recordings collected from a variety of recording environments and hardware platforms. Zhang et al. [22] did so, where they trained a model on 461 recordings from the SOF dataset and then tested the trained model on the SHHS dataset, achieving a kappa value of 0.53. In the present article, direct testing of a trained model on different datasets yielded moderate accuracy, which was less satisfactory. A possible reason is the lack of sufficient training data since we cannot train the model on a huge dataset due to limited computation resource. In addition, model performance on independent datasets depends on the similarity between the training set and the test set, while the employed three datasets have disparate attributes, as shown in Table 1. Nevertheless, our model is promising and worthwhile to train it on a huge and high-quality dataset in our future research, which helps to improve model generalization. Moreover, it would be interesting to explore model performance on large populations with diverse sleep problems, given the complex and diverse clinical symptoms of suspected patients.

5. Conclusion

The present paper proposed a deep learning model for automatic sleep scoring, which took raw PSG signals as input without any human-engineered features. The model employed two parallel convolution layers with different filter sizes and one LSTM layer to exploit information across spatial and temporal scales, thereby enhancing model performance. Moreover, the unique structure allowed the model to cope with various input channels and several signal modalities from different datasets without task-specific modifications to model architecture and hyper-parameters. Model generalization and model transferability were tested on participants with distinct attributes, even subjects with complex sleep disturbances. Results evaluated on three public datasets showed that the model achieved a comparable or better performance compared to the state-of-the-art methods, and the highest classification accuracy was achieved by the fusion of multiple PSG signals. Future work will require huge and high-quality datasets to improve the robustness and generalization of the proposed model.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No.91748105), National Foundation in China (No.JCKY2019110B009 & 2020-JCJQ-JJ-252), the Fundamental Research Funds for the Central Universities [DUT2019, DUT20LAB303] in Dalian University of Technology in China, and the China Scholarship Council (Nos. 201606060227). This study was to memorize Prof. Tapani Ristaniemi for his great help to the authors, Fengyu Cong, Rui Yan, Fan Li and DongDong Zhou.

Declaration of interest

None.

References

[1] M. Dattilo *et al.*, “Sleep and muscle recovery: endocrinological and molecular basis for a new and promising hypothesis,” *Med. Hypotheses*, vol. 77, no. 2, pp. 220–222, 2011.

[2] R. Stickgold and M. P. Walker, “Sleep-dependent memory consolidation and reconsolidation,” *Sleep Med.*, vol. 8, no. 4, pp. 331–343, 2007.

[3] L. Xie *et al.*, “Sleep drives metabolite clearance from the adult brain,” *Science (80-.).*, vol. 342, no. 6156, pp. 373–377, 2013.

[4] 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.

[5] 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.

Journal Pre-proof

[6] 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.

[7] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009.

[8] 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.

[9] 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.

[10] A. Procházka, J. Kuchyňka, O. Vyšata, P. Cejnar, M. Vališ, and V. Mařík, "Multi-Class Sleep Stage Analysis and Adaptive
 Pattern Recognition," *Appl. Sci.*, vol. 8, no. 5, p. 697, 2018.

[11] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, "Sleep stage classification using single-channel EOG," *Comput. Biol. Med.*, vol. 102, no. August, pp. 211–220, 2018.

[12] S. I. Dimitriadis, C. Salis, and D. Linden, "A novel, fast and efficient single-sensor automatic sleep-stage classification based on complementary cross-frequency coupling estimates," *Clin. Neurophysiol.*, vol. 129, no. 4, pp. 815–828, 2018.

[13] S. Sheykhivand, T. Y. Rezaii, A. Farzamnia, and M. Vazifehkhahi, "Sleep Stage Scoring of Single-Channel EEG Signal based on RUSBoost Classifier," in *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2018, pp. 1–6.

[14] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, 2018.

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

[16] S. Güneş, K. Polat, and Ş. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7922–7928, 2010.

[17] 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.

[18] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.

[19] 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, p. 108312, 2019.

[20] 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, p. zsy041, 2018.

[21] 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.

[22] 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, p. zsz159, 2019.

[23] J. Zhang, R. Yao, W. Ge, and J. Gao, "Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 183, p. 105089, 2020.

[24] X. Zhang, W. Kou, E. I.-C. Chang, 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.

[25] S. Wang, J. Cao, and P. S. Yu, "Deep Learning for Spatio-Temporal Data Mining : A Survey," pp. 1–21.

[26] Y. Liu, R. Fan, and Y. Liu, "Deep Identity Confusion for Automatic Sleep Staging Based on Single-Channel EEG," *Proc. - 14th Int. Conf. Mob. Ad-Hoc Sens. Networks, MSN 2018*, pp. 134–139, 2018.

[27] X. Chen, J. He, X. Wu, W. Yan, and W. Wei, "Sleep staging by bidirectional long short-term memory convolution neural network," *Futur. Gener. Comput. Syst.*, vol. 109, pp. 188–196, 2020.

[28] P. Fonseca *et al.*, "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population," *Sleep*, no. April, pp. 1–10, 2020.

[29] 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.

[30] 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.

[31] A. Malafeev *et al.*, "Automatic Human Sleep Stage Scoring Using Deep Neural Networks," *Front. Neurosci.*, vol. 12, p. 781, 2018.

[32] H. Phan *et al.*, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," *arXiv Prepr. arXiv1907.13177*, 2019.

[33] 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.

[34] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

Journal Pre-proof

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 7132–7141, 2018.
[41] 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.
[42] F. Chollet, "Keras: Deep learning library for theano and tensorflow," *URL https//keras. io/k*, vol. 7, no. 8, p. T1, 2015.
[43] 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.
[44] 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.
[45] 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.
[46] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 7046–7059, 2013.
[47] 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.
[48] 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, p. 1174, 2018.
[49] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE Laurens," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Journal Pre-proof