Sleep and Breathing (2019) 23:1159–1167
https://doi.org/10.1007/s11325-019-01789-4

SLEEP BREATHING PHYSIOLOGY AND DISORDERS • ORIGINAL ARTICLE

Sleep staging from single-channel EEG with multi-scale feature and contextual information

Kun Chen[1] · Cheng Zhang[2] · Jing Ma[2] · Guangfa Wang[2] · Jue Zhang[1,3]

Received: 31 October 2018 / Revised: 16 January 2019 / Accepted: 26 January 2019 / Published online: 12 March 2019
© Springer Nature Switzerland AG 2019

Abstract

Purpose Portable sleep monitoring devices with less-attached sensors and high-accuracy sleep staging methods can expedite sleep disorder diagnosis. The aim of this study was to propose a single-channel EEG sleep staging model, SleepStageNet, which extracts sleep EEG features by multi-scale convolutional neural networks (CNN) and then infers the type of sleep stages by capturing the contextual information between adjacent epochs using recurrent neural networks (RNN) and conditional random field (CRF).

Methods To verify the feasibility of our model, two datasets, one composed by two different single-channel EEGs (Fpz-Cz and Pz-Oz) on 20 healthy people and one composed by a single-channel EEG (F4-M1) on 104 obstructive sleep apnea (OSA) patients with different severities, were examined. The corresponding sleep stages were scored as four states (wake, REM, light sleep, and deep sleep). The accuracy measures were obtained from epoch-by-epoch comparison between the model and PSG scorer, and the agreement between them was quantified with Cohen's kappa ( ).

Results Our model achieved superior performance with average accuracy (Fpz-Cz, 0.88; Pz-Oz, 0.85) and (Fpz-Cz, 0.82; Pz-Oz, 0.77) on the healthy people. Furthermore, we validated this model on the OSA patients with average accuracy (F4-M1, 0.80) and (F4-M1, 0.67). Our model significantly improved the accuracy and compared to previous methods.

Conclusions The proposed *SleepStageNet* has proved feasible for assessment of sleep architecture among OSA patients using single-channel EEG. We suggest that this technological advancement could augment the current use of home sleep apnea testing.

Keywords Sleep staging · Single-channel EEG · Multi-scale feature · Recurrent neural network · Conditional random field

Introduction

Sleep plays an essential role in human health, both physical and mental. Outside of the wake state, sleep commonly occurs in four repeating stages: rapid eye movement (REM), non-REM (NREM) stages 1, 2, and 3 [1]. Sleep disorders like obstructive sleep apnea (OSA) are a global health problem [2]. The overall prevalence of OSA estimated from a systematic review is up to 9–38% in all regions of the world [3]. Another study reported that the prevalence of OSA in Asia in middle-aged men and women is 4.1–7.5% and 2.1–3.2%, respectively [4]. Remarkably, with the increase of risk factors for OSA (obesity, aging) and the improvement of diagnostic techniques, these figures may underestimate the true prevalence of the disease [5]. Owing to frequent apnea, OSA patients may suffer from sleep fragmentation, resulting in a lack of deep sleep (N3) and REM sleep. Moreover, apnea-related alpha is an EEG characteristic of OSA patients due to the fact that apnea is often accompanied by arousal, i.e., the invasion and increase of alpha wave activity [6]. Thus, high-accuracy sleep staging methods are essential for identifying and managing sleep-related diseases such as OSA.

Polysomnography (PSG) is a multi-parametric test to diagnose OSA. However, the messy cables and sensors of PSG limit the availability of home sleep apnea testing. Many studies have tried to develop portable sleep monitoring devices based on the actigraphy [7], cardiorespiratory signals [8], or radio frequency [9]. These measurements increased the comfort and operation simplicity more than PSG but immediately faced the drawback of less accuracy (within the range of 0.65

✉ Jue Zhang
zhangjue@pku.edu.cn

[1] Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[2] Department of Respiratory and Critical Care Medicine, Peking University First Hospital, Beijing 100034, China

[3] College of Engineering, Peking University, Beijing 100871, China

and 0.8). Clinically, manual scoring sleep stages is a labor-intensive work by human visual inspection. Most of previous automatic sleep stage scoring methods require a mass of prior knowledge of sleep analysis. These methods often first extract experience-based features and then train a classifier to identify sleep stages [10, 11]. However, the hand-engineered features based on the characteristics of the available dataset may not well generalize to a larger population on account of the heterogeneity among subjects and recording devices. Recently, Biswal S. et al. [12] achieved human-expert level sleep staging performance on PSG by utilizing convolutional neural network (CNN) to extract features from single independent epoch and then fed the extracted features to the recurrent network (RNN). The similar model, DeepSleepNet, adopted by Supratak A et al. [13] has shown state-of-the-art results of sleep staging on single-channel EEG. However, their network's ability to extract EEG features is limited by single or two in scale. However, although RNN encodes the temporal information in the extracted features, the dependency of adjacent tag information is ignored, such as explicit sleep stage transition rules [14]. Importantly, most of the previous methods are limited on healthy individuals, and the external heterogeneous population (e.g., OSA) validity remains uncertain.

Following the recent development of neural networks, multi-scale CNN, with the advantage of extracting features from different scales at the same time, have achieved impressive performance in computer vision [15]. In addition, conditional random field (CRF) is favorable to consider the adjacent tag information and jointly decode the best chain of tags for a given input data [16]. The previous studies have demonstrated that CRF can achieve excellent performance in sequence tagging tasks, such as part-of-speech tagging [17].

In this study, we propose a single-channel EEG sleep staging model, termed *SleepStageNet*, which extracts sleep EEG features by multi-scale CNN and then infers the type of sleep stages by capturing the contextual information between adjacent epochs using RNN and CRF. To verify the feasibility of our model, two datasets, one composed by two different single-channel EEGs (Fpz-Cz and Pz-Oz) on 20 healthy people and one composed by a single-channel EEG (F4-M1) on 104 OSA patients with different severities, were examined. In addition, we investigate how the contextual information between epochs affect the performance of sleep staging performance.

Material and methods

Dataset and preprocessing

Sleep-EDF dataset This is a public benchmark database for sleep staging [18]. As Supratak A. et al. [13], the sleep monitoring data on 20 healthy subjects in adults (age, 25–34), were used in this study, without any sleep-related medication. Briefly, each sleep recording contained two scalp-EEG signals from Fpz-Cz and Pz-Oz channels. All EEG channels had the same sampling rate of 100 Hz. Each no overlapping 30-s epoch of these recordings was manually classified into one of the six classes (W, N1, N2, N3, N4, and REM) according to the R&K standard [19]. Sleeping time was recommended to use the annotated lights off and lights on times as start and end times, respectively. In addition, 30 min of such periods before and after the sleep periods were extended. Other time was removed from the analysis, as their patterns were not being related to the events of interest. We merged the N3 and N4 stages into a single stage, deep sleep, according to the AASM standard [1]. Furthermore, owing to transitory nature of N1 stage and difficulty in distinguishing between N1 and N2 stage reported in [10, 11], we followed the same approach used in [8, 9], merging N1 and N2 stage into a single stage, light sleep.

OSA dataset The retrospective analysis of our dataset was approved by the ethics committee of Peking University First Hospital. From May 2016 to October 2017, 104 consecutive patients with suspected OSA were recruited from the sleep laboratory. The subjects were selected to satisfy the following criteria: (1) age older than 18 years. (2) at least 4 h of EEG recording. The overall severity of sleep apnea was described by the apnea-hypopnea index (AHI). Among them, 24 patients were confirmed to have mild OSA ($5 \leq \mathrm{AHI} < 15$), 19 patients had moderate ($15 \leq \mathrm{AHI} < 30$), 39 patients had severe ($\mathrm{AHI} \geq 30$), and the remaining 22 patients were non-OSA. For each subject, the overnight PSG (Siesta 2, Compumedics Ltd., Australia) was performed. Each no overlapping 30-s epoch of the PSG recordings was manually divided into one of the five sleep stages (W, N1, N2, N3, and REM) by a sleep expert in accordance with the AASM standard [1]. We also merged N1 and N2 stages into a single-stage termed light sleep, like the Sleep-EDF dataset. We evaluated our model using the F4-M1 channel, which was resampled to the same sampling rate of 100 Hz.

Table 1 summarizes the basic characteristics of the study population and the number of 30-s epochs for each sleep stage from these two datasets. All EEG recordings from the two datasets were preprocessed with bandpass filters of 0.3–35 Hz for eliminating power frequency (50 or 60 Hz) and baseline wander.

Model structure

As shown in Fig. 1, the architecture of our model, *SleepStageNet*, consists of two parts. The first part is a multi-scale neural network termed SleepFeatureNet, which extracts specific EEG features of different sleep stages from

Table 1 Basic characteristics of the study population and the number of 30-s epochs for each sleep stage from the two datasets

Variable	Sleep-EDF	OSA dataset ($N = 104$)			
	Normal $N = 20$	Normal $N = 22$	Mild $N = 24$	Moderate $N = 19$	Severe $N = 39$
Age	29 ± 3[a]	44 ± 19	49 ± 19	52 ± 12	48 ± 9
Male (%)	50	63.6	60.7	78.9	90.0
AHI (n/h)	< 5	< 5	5–15	15–30	> 30
TST (min)	542 ± 110	367 ± 58	359 ± 68	345 ± 71	364 ± 66
Light (*n* %)	20,603 (49)	8472 (43)	9685 (44)	8173 (45)	20,606 (56)
Deep (*n* %)	5703 (14)	4643 (23)	4210 (19)	2846 (16)	3582 (10)
REM (*n* %)	7717 (18)	3051 (16)	2626 (12)	2121 (12)	3883 (11)
Wake (*n* %)	7927 (19)	3664 (18)	5471 (25)	4916 (27)	8631 (23)

[a] Values are presented as mean±SD

TST = total sleep time

AHI = apnea-hypopnea index

REM = rapid eye movement

each independent 30-s epoch. The second part is a contextual information-learning network, which encodes the temporal information in the extracted features between adjacent epochs by RNN, and then explicitly captures stage transition rules in the sleep states and then infers the type of sleep stages, by CRF.

Briefly, the SleepFeatureNet is a modified CNN by replacing a single scale convolutional layer with a multi-scale convolution layer, which was inspired by the popular GoogLeNet [15]. The contextual information learning network consists of a bi-directional gated recurrent unit (Bi-GRU) [20] and a chain-structured conditional random field (CRF) [16]. Part II in Fig. 1 illustrates the architecture in a simple way. Here, the Bi-GRU layer adapted to aggregating the features, which enables the networks to take into account the features EEG extracted in adjacent frames. The length of the input epochs is five, and both the number of input neurons and output neurons of Bi-GRU is four. Importantly, a chain CRF is finally used to capture explicit stage transition rules and infers the type of sleep stages. The architecture of SleepFeatureNet, technical details of CRF, model training and evaluation, and key hyperparameters were formed in the Appendix.

Evaluation metrics

We evaluated the performance metrics of different models using accuracy, macro-averaging F1-score (MF1) [21], and Cohen's kappa [22]. All metrics are commonly used in automated sleep staging.

(a) Accuracy: accuracy is defined as the percent of correct labels predicted by the model out of a total number of annotations.

(b) Macro averaging F1-score (MF1): F1-score is the weighted average of precision and recall, which takes both false positives and false negatives into account. The macro-averaging F1 score is the average of the F1 scores obtained for each category. The MF1 score reaches its best value at 1 and worst score at 0.

(c) Kappa: Cohen's kappa coefficient ( ) is a statistic which measures the degree of inter-rater agreement between model prediction and annotations by a sleep technologist. The values are typically categorized as follows: 0–0.4 are considered low, 0.4–0.6 are moderate, 0.6–0.8 are high, and above 0.8 are near perfect agreement.

Performance analysis

The results of different sleep staging methods in this study are expressed as mean±SD. Two-tailed paired *t* test were used for comparison of the metrics between *SleepStageNet* and DeepSleepNet. Values of $p < 0.05$ were considered to be significant. The calculations were performed with SPSS Software System version 19.0 (SPSS Inc., USA). Moreover, to precisely evaluate the feasibility of our proposed model, we summarized the performance of other two EEG-based automatic scoring methods using hand-engineered features [10, 11], and three representative non-EEG-based scoring methods [7–9].

Results

Sleep staging performance in healthy individuals

Table 2 shows a comparison between our proposed method and other available sleep stage scoring methods. It is readily observed that the EEG-based methods show better

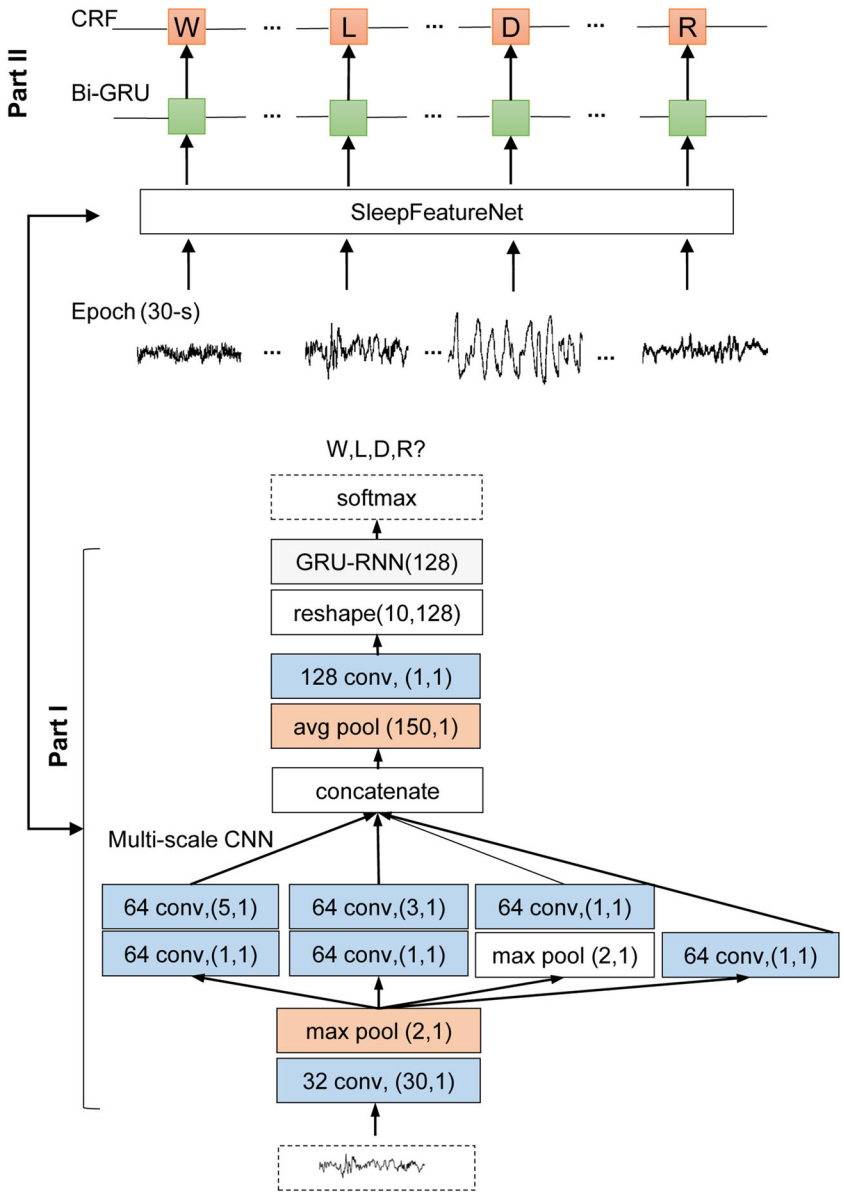

Fig. 1 An overview architecture of *SleepStageNet* consisting of two main parts: multi-scale feature learning (part I) and contextual information learning (part II). The feature vector of each epoch is computed by SleepFeatureNet in part I, then feed it into the part II that combines bi-directional GRU layer and CRF layer to capture temporal contextual information and infers the sleep stage

performance metrics than those of non-EEG-based methods, and especially, for each EEG-based approach, the staging performance based on Fpz-Cz channel is better than that based on Pz-Oz channel.

On Fpz-Cz channel, the DeepSleepNet performs an average ACC of 0.84, MF1 of 0.82, and κ of 0.74. Our SleepFeatureNet performs an average ACC of 0.86, MF1 of 0.85, and κ of 0.80. Our *SleepStageNet* improves the performance considerably by taking the average ACC of 0.88, MF1 of 0.87, and κ of 0.82. The similar trend was on the Pz-Oz channel with best performs (ACC of 0.85, MF1 of 0.83, and κ of 0.77). In particular, our *SleepStageNet* showed a significantly higher ACC ($p < 0.05$) and κ ($p < 0.05$) than those of DeepSleepNet, both on Fpz-Cz and Pz-Oz channels.

Sleep staging performance in OSA individuals

In Table 3, we report the extended validation in the OSA patients by the F4-M1 channel. The results suggest that our *SleepStageNet* exhibits the best performance by taking the average ACC of 0.80, MF1 of 0.72, and κ of 0.67. As expected, we realize that by adding the RNN-CRF layer to the baseline SleepFeatureNet, the performance is improved substantially with 7% increase in ACC, 7% increase in MF1, and 10% increase in κ on OSA dataset. Of note, both the accuracy (ACC) and consistency (k) using our *SleepStageNet* have been significantly ($p < 0.01$) improved than the reference model DeepSleepNet. Our performance was comparable to the results of Sun et al. [23] based on large-scale six-channel EEG data.

Table 2 Comparison between our proposed method and other sleep staging methods across accuracy (ACC), macro-f1 score (MF1), and Cohen's kappa ( ) in healthy individuals

Approach	Signal source	Overall metrics		
		ACC	MF1	
Ref. [7]	Actigraphy	0.65[ab]	–	–
Ref. [8]	Cardiorespiratory	0.57 ± 0.13	–	0.71 ± 0.86
Ref. [9]	Radio frequency	0.80[a]	–	0.70[a]
Ref. [10]	EEG/Cz-Pz	0.85[ac]	–	0.75[a]
Ref. [10]	EEG/Pz-Oz	0.75[ac]	–	0.63[a]
Ref. [11]	EEG/Pz-Oz	0.82[ac]	–	–
DeepSleepNet	EEG/Fpz-Cz	0.84 ± 0.08	0.82 ± 0.09	0.74 ± 0.11
SleepFeatureNet	EEG/Fpz-Cz	0.86 ± 0.04	0.85 ± 0.05	0.80 ± 0.06
SleepStageNet	EEG/Fpz-Cz	0.88 ± 0.04	0.87 ± 0.04	0.82 ± 0.06
DeepSleepNet	EEG/Pz-Oz	0.81 ± 0.07	0.78 ± 0.09	0.70 ± 0.12
SleepFeatureNet	EEG/Pz-Oz	0.84 ± 0.03	0.82 ± 0.04	0.76 ± 0.05
SleepStageNet	EEG/Pz-Oz	0.85 ± 0.04	0.83 ± 0.05	0.77 ± 0.06

[a] Overall metric of the test set

[b] Three-class classification based on 5-min segment

[c] Data corrected to two decimal places

Compared with the DeepSleepNet, our architecture can reduce remarkably parameters and computation from 24.7 M to 0.18 M. Meanwhile, the variances of both the ACC and κ results using our model are much smaller, which show that our method provides more stable performance in sleep staging task. Moreover, we analyzed the overall performance of our *SleepStageNet* for each sleep stage on the test dataset. It is easy to observe that many misclassifications occur between the pairs of Deep-Light, Light-REM, and REM-Wake in Fig. 2b, as is visualized in Fig. 2a using *t* SNE algorithm [18]. In addition, we observed that the *SleepStageNet* can distinguish the light, deep, and wake stage with high accuracy, about 0.80, and the accuracy of the REM stage was reasonably low at 0.71.

Effects of contextual information on sleep staging performance

To further explore the effect of context information, which was introduced for improving the performance for sleep staging in this study, we compared the hypnograms predicted by our models with that of a sleep expert in Fig. 3. Although the predicted hypnograms look very similar in general, the results are a bit different in specifics. For instance, the hypnogram predicted using SleepFeatureNet leads to more oscillations than that of *SleepStageNet* at the 2nd to 3rd hours of the subject's sleep. Specifically, the true REM and light stages are prone to be mistakenly scored as wake stage. It can be observed that the *SleepStageNet* further improves the accuracy by taking into account contextual information. Subsequently, we investigated whether using contextual information can improve sleep staging performance in each subject on the testing set. As shown in Fig. 4, the stacked bar graph is the values of kappa for each testing subject on the test dataset. It illustrates that the contextual information increased (0.01–0.26) the kappa in the most of subjects (23/26), but a slight decrease (< 0.03) in a tiny number of subjects (3/26).

Discussion

In this paper, we proposed an automated sleep staging model, *SleepStageNet*, that benefited from both multi-scale feature

Table 3 Comparison between our proposed method and other sleep stage scoring methods on the comprehensive dataset that include OSA patients with different severity

Approach	Severity AHI/h	Test set			
		ACC	MF1		Parameters
Ref. [23]	1.2–16.2	–	–	0.68[a]	–
DeepSleepNet	< 5	0.75 ± 0.07	0.67 ± 0.12	0.59 ± 0.15	24.7 M
SleepFeatureNet	0.2–109.1	0.73 ± 0.07	0.65 ± 0.09	0.57 ± 0.10	0.18 M
SleepStageNet	0.2–109.1	0.80 ± 0.05	0.72 ± 0.10	0.67 ± 0.08	0.18 M

[a] Overall Cohen's kappa of the test set

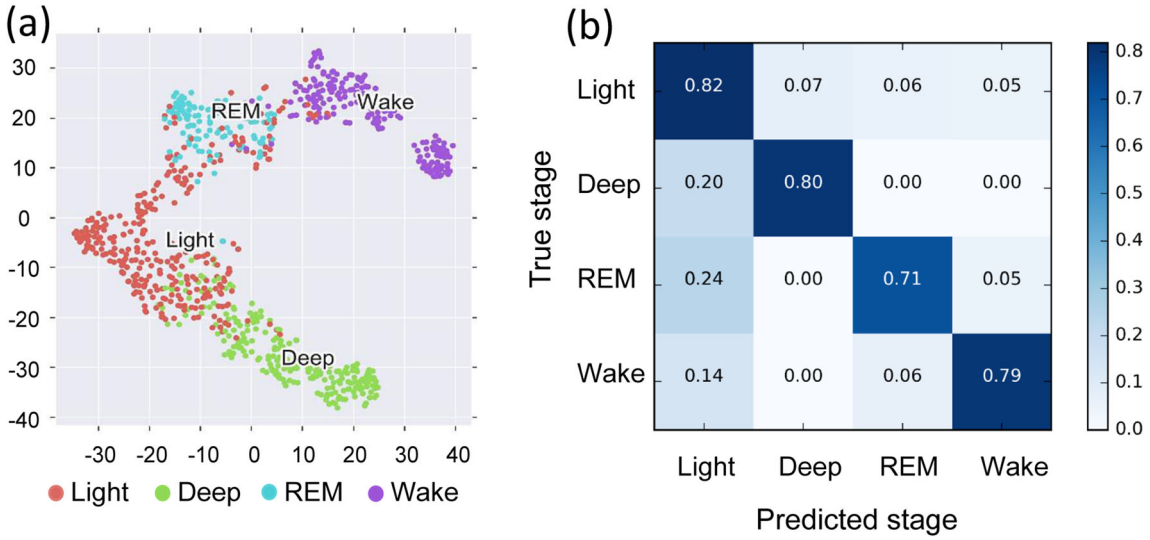

Fig. 2 The performance of *SleepStageNet* for each sleep stage. (a) The distribution of the features extracted by SleepFeatureNet form a typical subject. Axes are in arbitrary units. (b) Confusion matrix of prediction from *SleepStageNet* on the test dataset

and contextual information. Our model was validated on different single-channel EEGs (Fpz-Cz, Pz-Oz, and F4-M1) and the OSA patients with different severity. The results suggested that the *SleepStageNet* significantly improved the overall performance more than the previous method on average ACC and kappa ( ). Furthermore, our proposed model requires a much less computational cost.

Intrinsically, scoring sleep stage is a sequence-labeling task, as sleep technologists have always need to explore not only temporally local features, but also neighboring epochs. For instance, scoring stage N2 takes into account whether K complex or sleep spindles occurs early or in the last half of the previous epoch [1]. Therefore, we added the RNN-CRF layers to capture the contextual information of adjacent epochs,

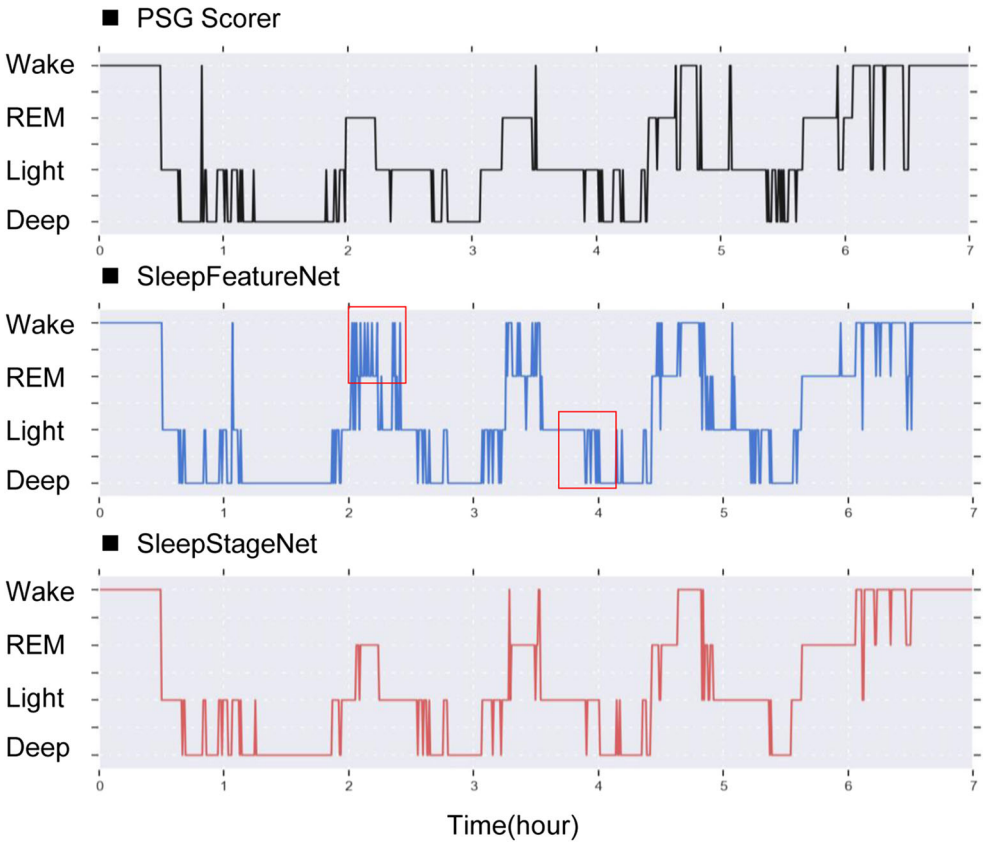

Fig. 3 Examples of the hypnogram manually scored by a sleep expert based on PSG data (top), and the hypnograms predicted using the SleepFeatureNet (middle) and the *SleepStageNet* (bottom), respectively. The positions of the red box indicate that the SleepFeatureNet are prone to make mistakes, but are improved in the *SleepStageNet*

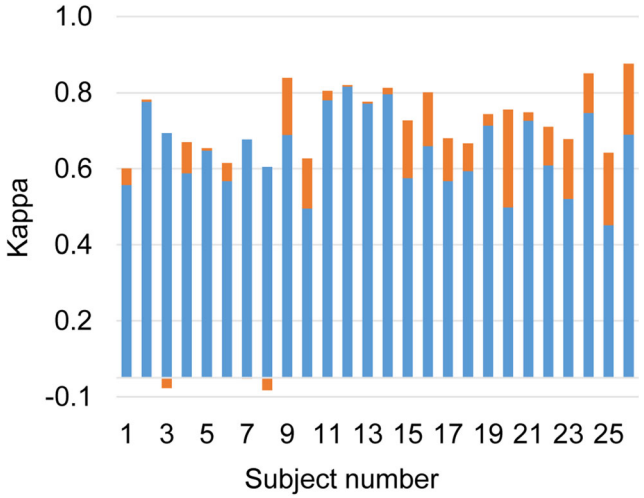

Fig. 4 The stacked bar graph is the values of kappa for each subject on the test dataset. The blue represents the results of SleepFeatureNet, and the orange represents the effect of considering context information on kappa values. It illustrates that the contextual information increased the kappa in the most of subjects (23/26), but a slight decrease in a tiny number of subjects (3/26)

instead of decoding each label independently. Our model considerably improved the overall performance more than the model based on independent epochs, as shown in Table 2 and Table 3. Moreover, we found that the majority of misclassifications between the pairs of deep-light and REM-wake were substantially improved. This improvement might well be due to our network learning the difference of transition probability matrices of sleep stages, as outlined in [14, 24]. For instance, if the current epoch is the REM stage, the next frame is most likely REM stage, but it is unlikely to be deep sleep. In Fig. 2, for our *SleepStageNet*, the true REM instances were prone to be mistakenly scored as light stages. Also, there is no way to distinguish N1 and N2 using our method, due to combining N1 and N2 sleep into a single stage, light sleep. In particular, the limitations could be improved by supplementing information about EMG and EOG [25]. We also found that, as mentioned by Sun et al. [23], a few subjects on the entire OSA dataset considering context information lead to a slight drop (< 0.03) of kappa. The different phenotype is likely due to individual differences such as age, BMI. There is a reason to believe that with the richness of patient data, the result will be improved.

There are two aspects that help us build efficient models. Firstly, we discarded extraneous information such as eliminating power frequency (50 or 60 Hz) and baseline wander, which could reduce task complexity. Secondly, our multi-scale CNN architecture controlled the trade-off between temporal and spatial resolution in the feature extraction process, and this is especially suitable for sleep staging tasks with significant differences in frequency and local characteristics.

For future potential clinical applications, our work could be furtherly improved in the following aspects. Firstly, our model could augment the performance of recently emerging portable sleep monitors with frontal EEG channels. Secondly, this model needs to prompt the predicted sleep stage that has low confidence or was changed by contextual information and allows sleep technologists to determine manually.

Conclusion

Based on single-channel EEG, we proposed a fully automated sleep staging model, *SleepStageNet*, which combines multi-scale feature and contextual information. The sleep staging results suggested that the *SleepStageNet* could achieve superior overall performance on different levels of OSA patients with much less computational cost. We suggest this technological advancement could augment the current use of home sleep apnea testing.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing of interests.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required.

Appendix

1. Architecture of SleepFeatureNet

As demonstrated in Fig. 1, the SleepFeatureNet consists of four convolutional layers and a gated recurrent neural network layer (GRU). Its structure is briefly described from bottom to top, in which each 30-s EEG epoch is the input of the model. Each convolutional layer sequentially executes three operations: convolution with its filters, batch normalization, and using the rectified linear unit activation. Each pooling layer performed a down-sampling operation along the spatial dimensions (width, height). The specifications of the number of filters, the filter sizes and pooling operation and sizes are described in Fig. 1. Each “conv” block shows the number of filters, a filter size. Each “pool” block shows a pooling operation and size. The above stride size is fixed at one. Importantly, the multi-scale convolutions in the second and third convolutional layers provide a more efficient contribution for capturing the temporal-spatial characteristics of single-channel EEG signals in each 30 s. Then, we concatenated the outputs of the previous layer by a “concatenate” layer that takes a list of vectors as the input and simply outputs the concatenation of the vectors. The output activations of the last convolutional layer are further

resized by "reshape" layer, and the reshaped active outputs then are fed into a following gated recurrent unit (GRU) to summarize the local features.

2. Technical details of CRF

The input of the CRF layer is a sequence of hidden states $z = \{z_1, z_2, ..., z_n\}$ (where i is ith epoch, $n = 5$), while the corresponding sleep stage sequence is $y = \{y_1, y_2, ..., y_n\}$. $Y(z)$, which represents the set of possible sleep stage sequences for z.

The conditional probability $p(y|\, z; w, b)$ over all possible sleep stage sequence y given z is described as the following form:

$$\mathrm{p}(y|z; w, b) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, z)}{\sum_{y' \in Y(z)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, z)} \quad (1)$$

Where $\psi_i(y', y, z) = \exp(w^T_{y',y} z_i + b_{y',y})$ are the potential functions, $w^T_{y',y} z_i$ and $b_{y',y}$ are the weight and bias corresponding to sleep stage pair (y', y), respectively.

The maximum conditional likelihood estimation is used for CRF training. For a training set $\{(z_i, y_i)\}$, the log-likelihood is given by:

$$\mathrm{L}(w, b) = \sum_{\mathrm{i}}^{n} \log p(y|z; w, b) \quad (2)$$

Decoding in CRF is to search for the sleep stage sequence $\hat{y}$ with the highest conditional probability:

$$\hat{y} = \underset{\mathrm{y} \in Y(z)}{\operatorname{argmax}}\, p(y|z; w, b) \quad (3)$$

In this work, the CRF training and decoding were solved efficiently by adopting the Viterbi algorithm [26].

3. Model training and evaluation

Our training process consists of two steps. The first step is to train the SleepFeatureNet part independently. Explicitly, the "GRU-RNN" layer is followed by a softmax layer (see Fig. 2) for supervised training. Here, the parameter optimization is performed with stochastic gradient descent (SGD) with an initial learning rate of 10^{-3}, a fixed momentum value of 0.9 and a mini-batch size of 64. Besides, we add a dropout layer [27] with a fixed dropout rate at 0.5 before connecting the softmax layer to help prevent over-fitting problems. The popular cross-entropy loss is applied to train the model to output probabilities for mutually exclusive classes. The second step is to train the *SleepStageNet* model with a sequential training set. Precisely, as a feature extractor, the sub-model SleepFeatureNet is frozen using the best parameters in the first step. In our training procedure, the sequences of 30-s EEG epochs from each subject data were split into sub-sequences of length 5, then we fed 10 sub-sequences per one training, and the optimized parameters of SGD were identical to the first step. For each step, our models were trained for up to 30 iterations and chose the best performing one on validation sets. Our proposed model was implemented using Keras with TensorFlow as an underlying computation engine. All experiments were performed on a computer with one 2.5 GHz Intel Core™ processor and one NVIDIA GPU (GTX 1080).

We evaluated our model using 20-fold cross-validation for each EEG channel on Sleep-EDF dataset. For each fold, we used the recordings from one subject for testing and the remaining recordings of 19 subjects for training model. The predicted sleep stages from all folds were combined to compute the overall performance. For a fair comparison, the reference model DeepSleepNet and dataset described in [13] were unchanged except the corresponding labels were merged into the four sleep stages (wake, light, deep, and REM). Our OSA dataset is a relatively large data set, including 104 OSA patients with different severity, which has 96,580 30-s epochs. We randomly split 66 subjects as the training set (12 normal, 18 mild, 12 moderate, and 24 severe), 12 subjects as verification set (4 normal, 2 mild, 2 moderate, and 4 severe), and 26 subjects as testing set (6 normal, 4 mild, 5 moderate, and 11 severe). Finally, the performances recorded in all subjects on the test set were averaged and considered as the overall performance of different models.

References

1. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus CL, Vaughn BV (2012) The AASM manual for the scoring of sleep and associated events. Rules, terminology and technical specifications. American Academy of Sleep Medicine, Darien
2. Jordan AS, Mcsharry DG, Malhotra A (2014) Adult obstructive sleep apnoea. Lancet 383(9918):736–747
3. Senaratna CV, Perret JL, Lodge C, Lowe A, Campbell BE, Matheson MC, Hamilton GSAP, Dharmage SC (2016) Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev 34:70–81. https://doi.org/10.1016/j.smrv.2016.07.002
4. Lam B, Lam DCL, Ip MSM (2007) Obstructive sleep apnoea in Asia. Int J Tuberc Lung Dis 11(1):2–11
5. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM (2013) Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol 177(9):1006–1014. https://doi.org/10.1093/aje/kws342
6. Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ (1998) Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med 158(158):358–362
7. Gu W, Yang Z, Shangguan L, Sun W, Jin K, Liu Y (2014) Intelligent sleep stage mining service with smartphones. In Proceedings of the 2014 ACM international Joint Conference on pervasive and ubiquitous Computing (pp. 649–660). ACM

8. Tataraidze A, Korostovtseva L, Anishchenko L, Bochkarev M, Sviryaev Y (2016) Sleep architecture measurement based on cardiorespiratory parameters. In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th annual international conference of the (pp. 3478–3481) IEEE
9. Zhao M, Yue S, Katabi D, Jaakkola TS, Bianchi MT (2017). Learning sleep stages from radio signals: a conditional adversarial architecture. In International conference on machine learning (pp. 4100–4109)
10. Berthomier C, Drouot X, Herman-Stoïca M, Berthomier P, Prado J, Bokar-Thire D, Benoit O, Mattout J, D'Ortho M (2007) Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep 30(11):1587–1595. https://doi.org/10.1093/sleep/30.11.1587
11. Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P, Provazník I (2012) Sleep scoring using artificial neural networks. Sleep Med Rev 16(3):251–263. https://doi.org/10.1016/j.smrv.2011.06.003
12. Biswal S, Kulas J, Sun H, Goparaju B, Westover MB, Bianchi MT, Sun J (2017) SLEEPNET: automated sleep staging system via deep learning. arXiv preprint arXiv:1707.08262
13. Supratak A, Dong H, Wu C, Guo Y (2017) DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 25(11):1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116
14. Kim JW, Lee JS, Robinson PA, Jeong DU (2009) Markov analysis of sleep dynamics. Phys Rev Lett 102(17):178104. https://doi.org/10.1103/PhysRevLett.102.178104
15. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A (2015) Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1–9)
16. Lafferty JD, Mccallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of Icml 3(2):282–289
17. Ekbal A, Bandyopadhyay S (2008) Part of speech tagging in bengali using support vector machine. In Information technology, 2008. ICIT'08. International conference on (pp. 106–111). IEEE
18. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605
19. Rechtschaffen A (1968) A manual of standardized terminology, technique and scoring system for sleep stages of human subjects. Public Health Service
20. Zhou GB, Wu J, Zhang CL, Zhou ZH (2016) Minimal gated unit for recurrent neural networks. Int J Autom Comput 13(3):226–234
21. Sasaki Y (2007) The truth of the F-measure. Teach Tutor Mater 1(5):1–5
22. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
23. Sun H, Jia J, Goparaju B, Huang GB, Sourina O, Bianchi MT, Westover MB (2017) Large-scale automated sleep staging. Sleep 40(10). https://doi.org/10.1093/sleep/zsx139
24. Schlemmer A, Parlitz U, Luther S, Wessel N, Penzel T (2015) Changes of sleep-stage transitions due to ageing and sleep disorder. Philos Top 373(2034). https://doi.org/10.1098/rsta.2014.0093
25. Estrada E, Nazeran H, Barragan J, Burk JR, Lucas EA, Behbehani K (2006) EOG and EMG: two important switches in automatic sleep stage classification. Conf Proc IEEE Eng Med Biol Soc 1:2458–2461. https://doi.org/10.1109/IEMBS.2006.260075
26. Forney GDJ (1993) The viterbi algorithm. Proc IEEE 61(5):268–278
27. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.