SleepSatelightFTC: A Lightweight and Interpretable Deep Learning Model for Single-Channel EEG-Based Sleep Stage Classification

Aozora Ito[*]and Toshihisa Tanaka[†]

August 2, 2024

Abstract

Sleep scoring by experts is necessary for diagnosing sleep disorders. To this end, electroencephalography (EEG) is an essential physiological examination. As manual sleep scoring based on EEG signals is time-consuming and labor-intensive, an automated method is highly desired. One promising automation technology is deep learning, which has performed well or better than experts in sleep scoring. However, deep learning lacks adequate interpretability, which is crucial for ensuring safety and accountability, especially for complex inference processes. We propose SleepSatelightFTC, a model that employs self-attention to visualize feature importance for inference and transfer learning on continuous epoch data to reflect the inference context. This model achieves a higher accuracy (84.8%) and kappa coefficient (0.787) with fewer parameters than state-of-the-art models for sleep stage classification on the 2018 version of the Sleep-EDF Database Expanded. The visualization of feature importance obtained from self-attention confirms that the proposed model learns representative waveform features, including K-complexes and sleep spindles.

1 Introduction

Sleep is critical to physical and mental well-being. Like during the awake state, the brain shows characteristic activity during sleep [1]. Sleep progresses through distinct stages, each characterized by specific patterns of brain activity, eye movement, and muscle tone. Understanding the sleep stages is essential for diagnosing and treating various sleep disorders as well as investigating the role of sleep on overall health and disease.

[*] Aozora Ito is with the Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan (e-mail: ito22@sip.tuat.ac.jp).

[†] Toshihisa Tanaka is with the Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan and with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: tanakat@cc.tuat.ac.jp).

This work was supported by JST CREST under Grant JPMJCR1784 and JSPS KAKENHI 23K26110. (Corresponding author: Toshihisa Tanaka.)

Electroencephalography (EEG) is commonly used to investigate brain activity and is necessary for diagnosing sleep-related disorders. In fact, polysomnography involves simultaneous recordings of various biological signals throughout the night, including EEG, electrooculography, and electromyography signals, aiming to diagnose sleep disorders. Additionally, sleep recordings are essential in routine EEG examinations because they may unveil valuable information for diagnosing epilepsy and other neurological disorders. Based on recorded signals, experts can evaluate the sleep stages in 30 s epochs according to the sleep scoring manual of the American Academy of Sleep Medicine [2]. The manual defines five sleep stages: Wake (W), rapid eye movement (REM or R), non-REM 1 (N1), non-REM 2 (N2), and non-REM 3 (N3). The decision rules to identify the sleep stages are based on temporal and frequency features specific to each stage in the acquired biological signals. In addition, some sleep stages can be determined by considering context from adjacent stages.

Manual sleep scoring is time-consuming and labor-intensive for knowledgeable and experienced experts [3]. Even a skilled expert can take up to 2 hours to score approximately 8 hours of sleep data [4]. Manual sleep scoring impedes suitably handling the millions of patients with sleep disorders [5], rendering automated scoring required. Automation of sleep scoring is expected to reduce the burden on specialists by several thousand hours per year [6].

Sleep scoring based on EEG follows predefined rules, making it suitable for automation using machine learning [6]. Several models have been proposed to divide and classify EEG signals into sleep stages using machine learning, with models based on deep learning achieving equivalent or better performance than expert judgment [7]. However, such deep learning models lack interpretability due to the complexity of inference compared with classical machine learning models, hindering experts to judge the validity of inferences. This lack of interpretability may increase the difficulty to identify reasons underlying misclassifications, the inability to explain the model decisions to patients and healthcare providers, and potential bias in model predictions. The interpretability of models is also essential to ensure safety, ethics, and accountability [8]. The interpretation of incorrect inferences can also contribute to improve the model performance.

The self-attention mechanism [9] allows to interpret machine learning models through the visualization of feature importance. Self-attention automatically adjusts the feature importance in learning, indicating strong attention to specific data features.

We propose an EEG-based sleep stage classification model called SleepSatelightFTC that includes self-attention for interpretability. Based on the rules for sleep scoring, self-attention is applied to time- and frequency-domain features considered as inputs. To reflect the sleep context, we apply transfer learning to continuous epoch data.

2 Methods

2.1 Dataset and Preprocessing

We used a public dataset, Sleep-EDF Database Expanded [10, 11] in this study. Sleep-EDF Database Expanded has two versions (2013 and 2018) and two sub-

Table 1: Number of epochs per label in Sleep-EDF-20 and Sleep-EDF-78

Dataset	W	N1	N2	N3	R
Sleep-EDF-20	8285	2804	17,799	5703	7717
Sleep-EDF-78	69,824	21,522	69,132	13,039	25,835

sets (Sleep Telemetry and Sleep Cassette). Most research on sleep stage classification has used either the 2013 or 2018 version of Sleep Cassette. We used both versions separately. The 2013 Sleep Cassette (Sleep-EDF-20) consists of polysomnography data over 39 nights acquired from 20 healthy participants (10 males and 10 females) aged 25–34 years. The 2018 Sleep Cassette (Sleep-EDF-78) consists of polysomnography data over 153 nights acquired from 78 healthy participants (37 males and 41 females) aged 25–101 years. The polysomnography data were annotated by an expert according to the R & K manual [12], which defines six sleep stages. Only the Fpz–Cz EEG channel was used for evaluation. The sampling frequency of the acquired EEG signals was 100 Hz.

We applied the preprocessing steps to the EEG signals described in [7] for a fair comparison. First, to adhere to the manual of the American Academy of Sleep Medicine, which defines five sleep stages, all the N4 labels were merged into the N3 label. Next, we excluded the epochs with label "MOVEMENT" or "?" from the analysis. Finally, we extracted the section from 30 min before the start of sleep to 30 min after the end of sleep to exclude periods unrelated to sleep. The number of epochs per label in Sleep-EDF-20 and Sleep-EDF-78 are listed in Table 1.

2.2 Proposed SleepSatelightFTC Model

Most rules for sleep scoring are based on temporal or frequency features extracted from EEG signals. Accordingly, we propose a sleep stage classification model that processes EEG epochs as shown in Fig. 1. High-frequency activity (gamma waves with frequency > 30 Hz) in an EEG signal is likely related to the sleep–wake cycle but represents less than 1% of the total power spectrum [13]. Thus, we downsample the EEG signals to 50 Hz to establish a time-domain input. In spectral analysis for identifying differences in sleep EEG signals, the multi-taper method outperforms the single-taper method [14]. Thus, we calculate the amplitude spectrum in 0–25 Hz by applying the multi-taper method [15] to the EEG signals. The amplitude spectrum is expressed in decibel–microvolts (dB µV) by taking the logarithm to establish the input in the frequency domain. The proposed model applies self-attention to each input. Self-attention highlights the input that contributes to inference and allows to visualize the attention strength. Thus, self-attention is expected to enable the visualization of the model features in the time and frequency domains during training.

Sleep shows long-term context, and most existing models for sleep stage classification employ architectures that consider the context before and after every evaluated epoch [6]. The proposed sleep stage classification model, SleepSatelightFTC, has the architecture shown in Fig. 2. This model applies transfer learning to the base epoch-wise classification model shown in Fig. 1. The outputs of fully connected layers in the epoch-wise classification model are combined to

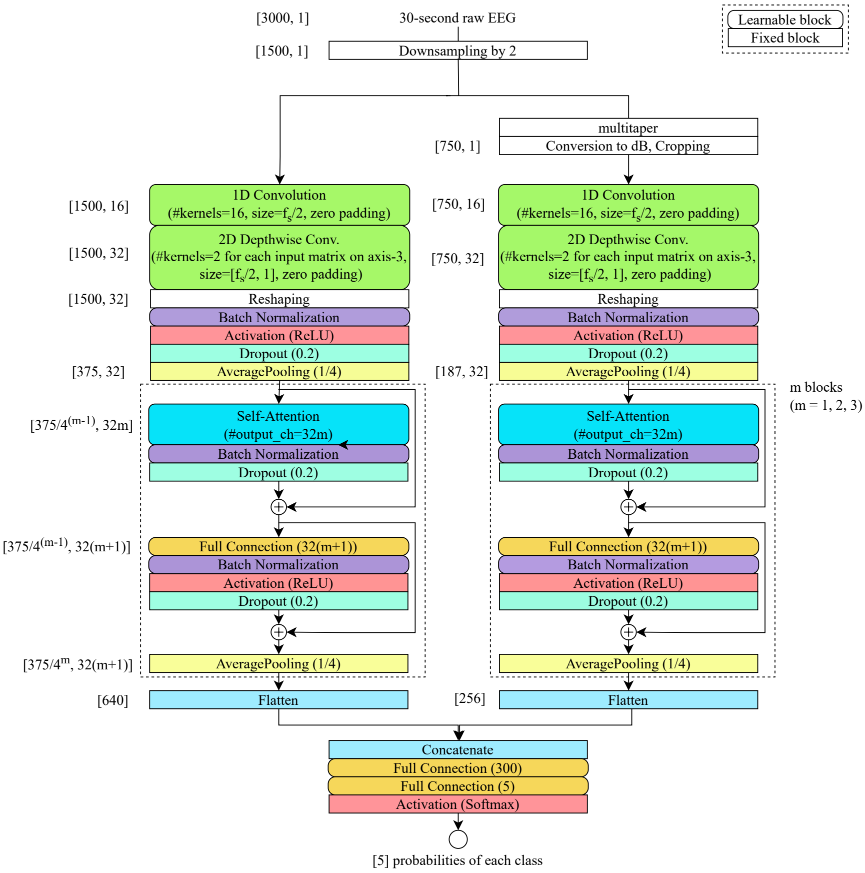


Figure 1: Epoch-wise classification model. The time-domain input is a 30 s raw EEG signal with 1500 samples (sampling rate f_s of 50 Hz), and the frequency-domain input is the 0–25 Hz amplitude spectrum with 750 samples. The model consists of convolutional layers with 16 kernels of size $f_s/2$ for both the time- and frequency-domain inputs, followed by 2D depthwise convolutional layers with 2 kernels per input matrix, batch normalization, rectified linear unit (ReLU) activation, dropout of 0.2, and average pooling of $1/4$. The model then applies m blocks ($m = 1, 2, 3$) of simple self-attention layers, dense layers, batch normalization, activation, dropout of 0.2, and average pooling of $1/4$. The self-attention layer has $32m$ output channels. The outputs of the two input branches are flattened, concatenated, and passed through two fully connected layers with 300 and 5 units, followed by softmax activation.

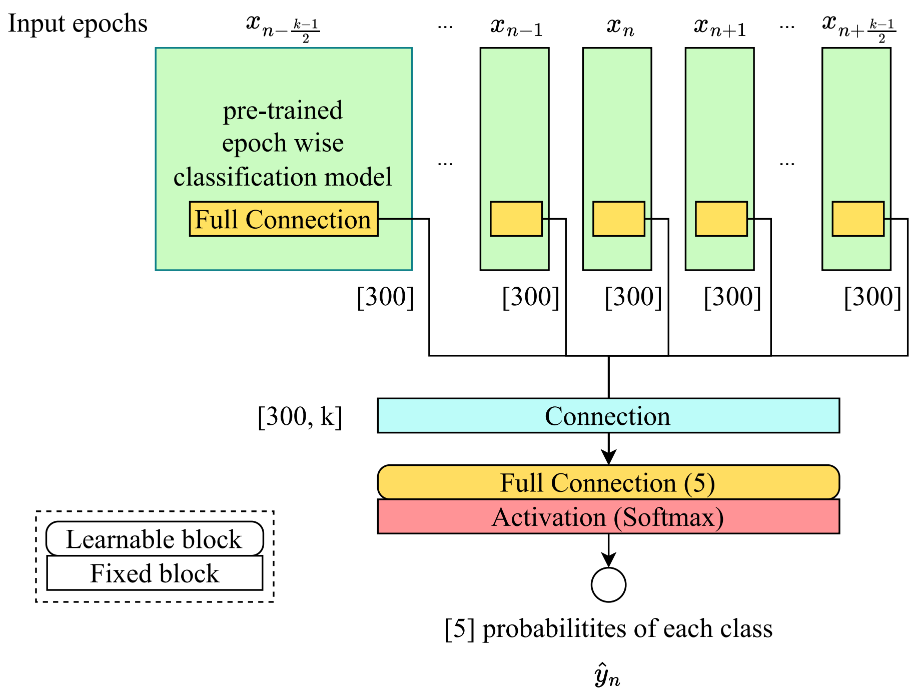


Figure 2: Transfer learning in sequential epoch data. The pretrained epoch-wise classification model is applied to k consecutive epochs, where k is an odd number ranging from 3 to 29. The outputs of the fully connected layers from the epoch-wise classification model are combined for the k epochs and passed through an additional fully connected layer with five units, followed by softmax activation to predict the probability of each sleep stage. The model is trained to predict the sleep stage of the central epoch in the sequence.

obtain sequential epochs and used as input for transfer learning. The model output is a one-epoch sleep stage, and its input is an odd number of sequential epochs centered on the target epoch for inference. The number of epochs is selected as an odd number between 3 and 29.

2.3 Learning Method

Multiclass cross-entropy was used as the loss function. Learning terminated when the loss in the validation data had not improved over five consecutive iterations by using early stopping. Adam [16] was used as the optimizer, and the learning rate was set to 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. The batch size was set to 32. For Sleep-EDF-20, leave-one-subject-out (20-fold) cross-validation was conducted to validate the model's classification performance. For Sleep-EDF-78, all participants were randomly divided into 10 groups, and then subject-wise 10-fold cross-validation was conducted.

2.4 Evaluation Metrics

We used the following evaluation metrics of model performance: accuracy (ACC), macro-F1 score (MF1) [17], kappa coefficient (Cohen's *kappa* κ) [18], and number of model parameters. Details of every metric are provided below.

ACC is the ratio of correctly predicted sleep stages to the total number of predictions. It provides an overall measure of the model performance but does not account for class imbalance. Sleep stage classification is a five-class classification problem, where every class is considered positive, and the other four classes are considered negative. The confusion matrix per class comprises four components: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates.

MF1 is the unweighted mean of the F1 scores per class, with each F1 score being the harmonic mean of precision and recall. Precision is the ratio of true positive predictions to the total positive predictions, and recall is the ratio of true positive predictions to the total actual positives. MF1 treats each class equally regardless of its prevalence in a dataset, being suitable for imbalanced datasets. It emphasizes the performance of each class over the overall percentage of correct responses because it neglects the number of samples per class.

Cohen's kappa measures the agreement between the predicted and actual sleep stages, considering the possibility of agreement occurring by chance. It is calculated as follows:

$$\kappa = \frac{ACC - p_e}{1 - p_e} \tag{1}$$

where p_e is the degree of coincidence given by

$$p_e = \sum_{i=1}^{5} p_{ei} \tag{2}$$

with the degree of coincidence per class being expressed as

$$p_{ei} = \frac{(TN + FP)(TN + FN) + (TP + FN)(TP + FP)}{(TP + TN + FP + FN)^2} \tag{3}$$

Kappa coefficients are interpreted according to their values, with a value of 0 indicating no agreement, and various ranges indicating slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00) agreement [19]. We compared SleepSatelightFTC with existing models for sleep stage classification in terms of the abovementioned metrics and number of model parameters.

3 Results

3.1 Classification Performance According to Number of Input Epochs

ACC of the proposed SleepSatelightFTC model according to the number of input epochs for transfer learning is listed in Table 2. ACC on Sleep-EDF-20 was the highest at 85.73% for 25 input epochs, and ACC on Sleep-EDF-78 was the highest at 84.83% for 15 input epochs. We used the models with the highest ACC values to compare them with existing models.

Table 2: ACC according to number k of input epochs for transfer learning.

# epochs	Overall accuracy	
k	Sleep-EDF-20	Sleep-EDF-78
3	83.21	81.77
5	84.11	82.56
7	84.61	82.89
9	84.85	83.05
11	85.07	83.17
13	84.86	83.11
15	85.00	**84.83**
17	85.12	84.78
19	85.35	84.69
21	85.62	83.57
23	85.22	83.59
25	**85.73**	83.60
27	85.64	83.49
29	85.61	83.53

3.2 Classification Performance

A comparison of the classification performance between various models for sleep stage classification is presented in Table 3. The overall performances are given by ACC, MF1, and κ, and the class-wise performances are given by F1 scores. The performances for the existing models are those retrieved from the corresponding papers. DeepSleepNet-lite, SleepEEGNet, SleepTransformer, EEGSNet, and IITNet were trained under the same conditions as SleepSatelightFTC. AttnSleep, TSA-Net, and TinySleepNet used weighted cross-entropy, and L-seqsleepnet used cross-entropy averaged over the sequence length. TinySleepNet used data augmentation during training.

On the smaller Sleep-EDF-20, SleepSatelightFTC achieves an overall ACC of 85.7%, MF1 of 77.7%, and κ of 0.800. ACC and κ are higher than those of models trained under the same conditions but lower than those of the other models. On the larger Sleep-EDF-78, SleepSatelightFTC achieves an overall ACC of 84.8%, MF1 of 77.8%, and κ of 0.787. ACC and κ of our model are much higher than those of the state-of-the-art models. In addition, SleepSatelightFTC achieves the highest F1 scores in the classification of sleep stages W and N2.

Table 3: Sleep stage classification performance of evaluated models (* indicates methods that use different loss functions from that of SleepSatelightFTC)

Dataset	Model	Overall performances			F1 scores for each class					# parameters
		ACC	MF1	κ	W	N1	N2	N3	R	$\times 10^6$
Sleep-EDF-20 (Fpz–Cz EEG)	DeepSleepNet-lite [20]	84.0	78.0	0.780	87.1	44.4	87.9	88.2	82.4	0.60
	SleepEEGNet [21]	84.3	79.7	0.790	89.2	**52.2**	86.8	85.1	85.0	2.60
	IITNet [22]	83.9	77.6	0.780	87.7	43.4	87.7	86.7	82.5	
	L-seqsleepnet [23]*	86.3	79.3	0.813	**91.6**	45.3	88.5	86.2	85.2	
	AttnSleep [24]*	85.6	**80.9**	0.800	90.3	47.9	**89.8**	89.0	85.0	
	TSA-Net [25]*	**86.6**	80.4	**0.816**	90.5	46.9	89.2	**90.1**	**85.4**	
	SleepSatelightFTC (proposed)	85.7	77.7	0.800	89.4	42.8	87.6	84.4	84.2	0.47
Sleep-EDF-78 (Fpz–Cz EEG)	DeepSleepNet-lite [20]	80.3	75.2	0.730	91.5	46.0	82.9	79.2	76.4	0.60
	SleepEEGNet [21]	80.0	73.6	0.730	91.7	44.1	82.5	73.5	76.1	2.60
	SleepTransformer [26]	81.4	74.3	0.743	91.7	40.4	84.3	77.9	77.2	3.7
	EEGSNet [27]	83.0	77.3	0.77	93.2	50.0	84.2	74.4	**83.5**	0.6
	TSA-Net [25]*	81.7	74.2	0.740	91.4	35.7	84.3	79.0	80.6	
	AttnSleep [24]*	82.9	**78.1**	0.770	92.6	47.4	85.5	**83.7**	81.5	
	TinySleepNet [7]*	83.1	**78.1**	0.770	92.8	**51.0**	85.3	81.1	80.3	1.3
	SleepSatelightFTC (proposed)	**84.8**	77.8	**0.787**	**93.8**	47.4	**86.5**	79.5	82.1	0.46

3.3 Ablation Study

An ablation study confirmed that each component of the proposed SleepSatelightFTC model contributes to inference on Sleep-EDF-78. SleepSatelightFTC consists of time- and frequency-domain inputs as well as transfer learning. We evaluated the classification performance when one or two of these three components were removed from SleepSatelightFTC, obtaining the results listed in Table 4.

When the frequency-domain input, time-domain input, and transfer learning are removed, ACC drops by 2.3%, 3.8%, and 5.9%, respectively. The F1 scores of sleep stage N3 are lower when the time- or frequency-domain inputs are removed than when transfer learning is removed. Furthermore, when the pair of frequency-domain input and transfer learning and the pair of time-domain input and transfer learning are removed, ACC drops by 6.5% and 9.1%, respectively. MF1 and κ also drop in these cases.

Table 4: Results of ablation study on Sleep-EDF-78

Model variant	Overall performances			F1 score of each class				
	ACC	MF1	κ	W	N1	N2	N3	R
SleepSatelightFTC (time-domain input + frequency-domain input + transfer learning)	**84.8**	**77.8**	**0.787**	**93.8**	**47.4**	**86.5**	**79.5**	**82.1**
time-domain input + transfer learning	82.5	74.4	0.753	92.6	40.8	85.1	77.1	76.2
frequency-domain input + transfer learning	81.0	72.8	0.732	90.7	38.5	84.1	76.6	74.2
time-domain input + frequency-domain input	78.9	69.7	0.704	90.5	30.0	83.5	78.5	66.1
time-domain input	78.3	67.8	0.693	90.5	25.6	82.9	76.5	63.4
frequency-domain input	75.7	64.6	0.654	86.1	20.5	81.7	74.8	60.2

4 Discussion

4.1 Comparison of Proposed and Existing Models

The proposed SleepSatelightFTC model achieves higher ACC than existing models. In addition, the number of parameters in SleepSatelightFTC is 4.7×10^5, while that in the comparison models are 0.6–3.7 $\times10^6$, being approximately 1.3–8 times larger than the number of parameters in SleepSatelightFTC. This can be attributed to the model architecture. Most existing models for sleep stage classification use raw EEG signals or spectrograms as inputs, extract epoch-wise features, and then consider contextual information before and after every evaluated epoch. In contrast, SleepSatelightFTC employs a parallel architecture that extracts features in both the time and frequency domains for subsequent integration. This approach allows the model to use information from multiple perspectives, effectively classifying sleep stages. Furthermore, applying self-attention to each domain enables the model to automatically learn and select essential features. The self-attention output shown in Figs. 4 and 5 confirms that the proposed model focuses on characteristic waveforms and frequency components, such as K-complexes and spindle waves. Additionally, introducing transfer learning using continuous epoch data enables judgments that consider the sleep context, thereby improving ACC in classification.

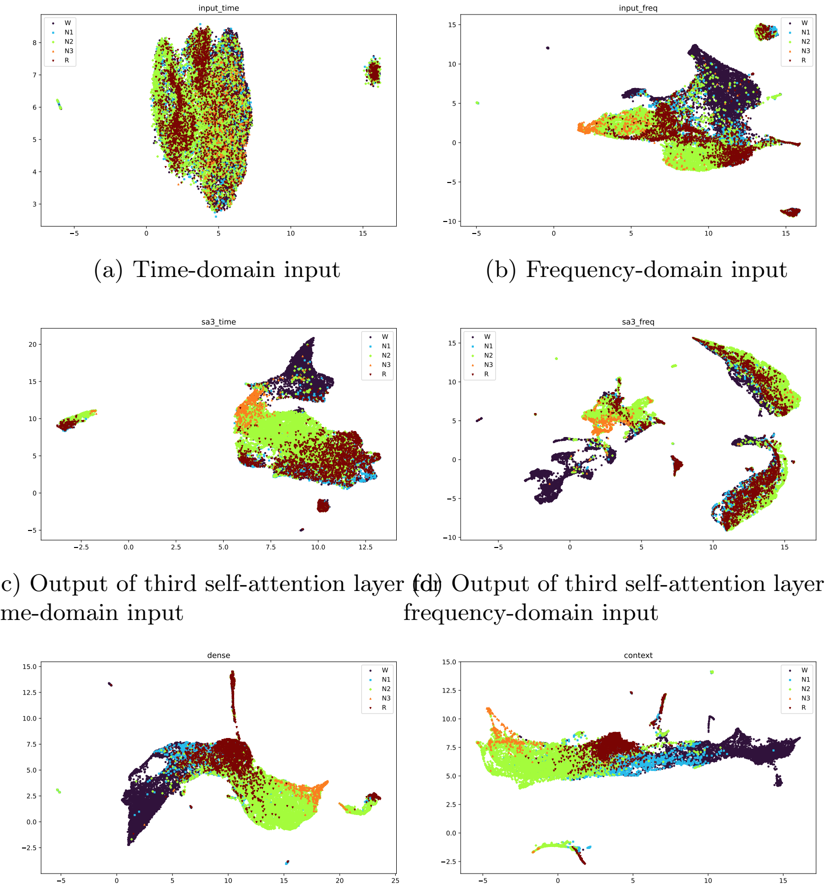
SleepSatelightFTC simplifies the context processing network of existing models by applying transfer learning to continuous epoch data. A single expert usually performs manual EEG-based sleep scoring. Therefore, existing models for sleep stage classification likely imitate the expert's subjective evaluation [6]. By suppressing overlearning, lightweight models like SleepSatelightFTC are less likely to reflect subjective biases, possibly increasing the consistency and reliability of sleep stage classifications.

SleepSatelightFTC achieves higher F1 scores for sleep stages N2 and W but lower F1 scores for sleep stage N1 than existing models. This discrepancy may be due to the smaller number of epochs available for sleep stage N1 compared with those for sleep stages N2 and W on the Sleep-EDF Database Expanded. The use of weighted cross-entropy loss, as in AttnSleep, TSA-Net, and TinySleepNet, may improve the classification performance in sleep stages with scarce training data available.

4.2 Contribution of Model Components to Inference

The proposed SleepSatelightFTC model achieves the highest ACC for 25 input epochs on Sleep-EDF-20 and 15 input epochs on Sleep-EDF-78, as listed in Table 2. In previous studies, context is considered from various epoch lengths, such as adjacent epochs [27], 15 epochs [28], and 100 epochs [29]. The ideal number of epochs to consider for the sleep context needs to be further analyzed, but the optimal number of epochs for the proposed model likely ranges from 15 to 25 epochs.

The ablation study results show that the performance declines the most when transfer learning is removed, followed by the removal of the time- and frequency-domain inputs. Hence, transfer learning as well as time- and frequency-domain inputs contribute to inference in that order. Each sleep stage typically lasts from a few to several tens of minutes, especially sleep stage N2, which accounts for



(a) Time-domain input

(b) Frequency-domain input

(c) Output of third self-attention layer for time-domain input

(d) Output of third self-attention layer for frequency-domain input

(e) After epoch-wise inference

(f) After transfer learning

Figure 3: Visualization of inference process by uniform manifold approximation and projection.

approximately 45% of the total sleep time and is longer in late sleep stages [30]. During intervals of identical sleep stages, transfer learning is expected to compensate for out-of-context inferences.

4.3 Visualization of Inference Process

We visualized the inference process of SleepSatelightFTC using a dimensionality reduction method called uniform manifold approximation and projection [31]. The data distributions per class after time-domain input, frequency-domain input, self-attention layer output, epoch-wise inference, and transfer learning are shown in Fig. 3. Figs. 3 (a) and 3 (b) show that the model inputs are more coherently distributed by class in the frequency domain than in the time domain. On the other hand, Figs. 3 (b) and 3 (c) show that the self-attention layer outputs are more coherently distributed by class in the time domain than in the frequency domain. Overall, Fig. 3 shows the distribution becoming more

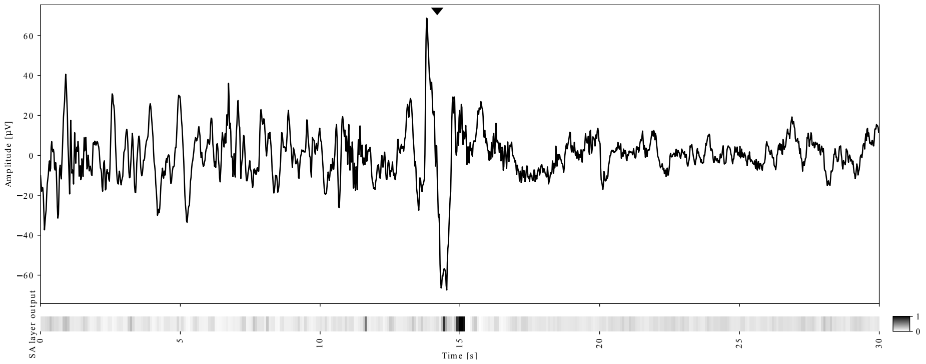


Figure 4: EEG signal for sleep stage N2 and its self-attention layer output heatmap. The heatmap shows the importance for classification of each timepoint in the EEG signal. The self-attention layer assigns higher importance to the waveform resembling a K-complex (▼), which is a characteristic feature of sleep stage N2.

clustered toward the latter half of the model. This suggests that self-attention may not necessarily be more effective for frequency features than for temporal features.

4.4 Self-attention Responses to Characteristics of Each Sleep Stage

We also created a heatmap of the output of the first self-attention layer for every input in SleepSatelightFTC. Because SleepSatelightFTC employs rectified linear unit activation, non-negative values of the output of the self-attention layer were set to 0. The output of the self-attention layer was averaged over time and normalized by using the function minmax_scale [32] from the preprocessing module of the scikit-learn library.

The model inputs and heatmaps of the self-attention layer outputs for an epoch correctly classified as sleep stage N2 are shown in Figs. 4 and 5. In the heatmap, self-attention responds strongly to a waveform that appears to be a K-complex, shown at the 15 s position in the EEG signal and the 12 Hz position in the amplitude spectrum.

The K-complex consists of a distinct negative sharp wave followed immediately by a positive component, observed especially in sleep stage N2 [2]. The K-complex duration is over 0.5 s, and the maximum amplitude is usually recorded in the frontal induction. In addition, spindle waves are features of sleep stages N2 and N3 characterized by a 12 Hz component in the frontal area [2]. The EEG signals of the Fpz–Cz channel considered in this study contain frontal EEG features. This suggests that SleepSatelightFTC learns K-complexes and spindle waves as features of sleep stage N2.

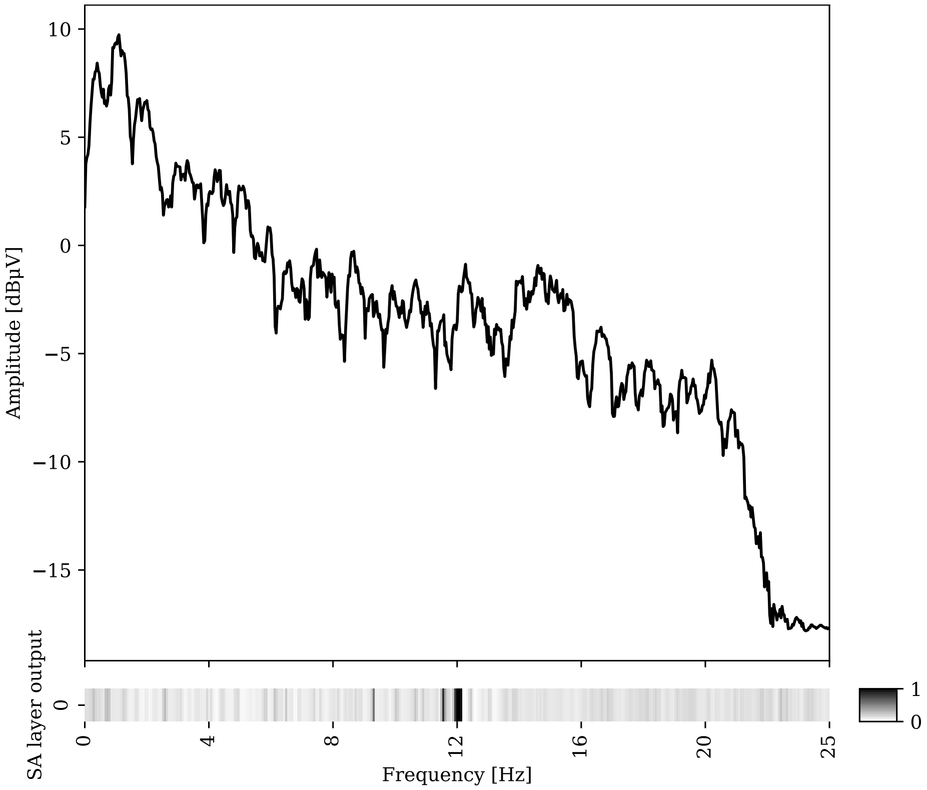


Figure 5: Amplitude spectrum of sleep stage N2 and its self-attention layer output heatmap. The heatmap shows the importance for classification of each frequency component in the amplitude spectrum. The self-attention layer assigns higher importance to the 12 Hz frequency component, which is associated with sleep spindles, another characteristic feature of sleep stage N2.

5 Conclusion

We propose SleepSatelightFTC, a lightweight and interpretable deep learning model for EEG-based sleep stage classification that achieves higher ACC with fewer parameters than state-of-the-art models by applying self-attention to time- and frequency-domain inputs and transfer learning to sequential epochs. The model interpretability through self-attention heatmaps, which highlight essential waveform features consistent with sleep scoring manuals, enhances the model accountability and allows experts to understand the reasons underlying its decisions. Nevertheless, our study has various limitations, such as using polysomnography data from only healthy subjects and not accounting for inter-rater variability in manual scoring. In future work, we will evaluate the model performance based on data from patients with sleep disorders, investigate methods to handle inter-rater variability, and improve the performance on stages with scarce training data available. Despite these limitations, SleepSatelightFTC demonstrates the potential of interpretable deep learning models for automatic sleep stage classification, which may substantially reduce the burden on sleep experts and improve the efficiency of sleep disorder diagnosis and treatment after further development and validation on diverse datasets.

Acknowledgements

We thank Dr. Kosuke Fukumori of RIKEN AIP for his inspiration and advice. We would like to thank Editage (www.editage.jp) for English language editing.

References

[1] J. Allan, "Sleep is of the brain, by the brain and for the brain," http://dx.doi.org/10.1038/nature04283, Oct. 2005, accessed: 2024-4-16.

[2] R. B. Berry, R. Brooks, C. E. Garnaldo, S. M. Harding, R. M. Lloyd, C. L. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications, version 2.5," *Darien, CT: American Association of Sleep Medicine*, 2018.

[3] Institute of Medicine (US) Committee on Sleep Medicine and Research, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, H. R. Colten and B. M. Altevogt, Eds. Washington (DC): National Academies Press (US), 2006.

[4] A. Malhotra, M. Younes, S. T. Kuna, R. Benca, C. A. Kushida, J. Walsh, A. Hanlon, B. Staley, A. I. Pack, and G. W. Pien, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, Apr. 2013.

[5] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, "The global problem of insufficient sleep and its serious public health implications," *Healthcare (Basel, Switzerland)*, vol. 7, no. 1, Dec. 2018.

[6] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: recent development, challenges, and future directions," *Physiological measurement*, vol. 43, no. 4, p. 04TR01, Apr. 2022.

[7] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2020, pp. 641–644, Jul. 2020.

[8] S.-C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Frontiers in oncology*, vol. 13, p. 1129380, Feb. 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, Jun. 2017.

[10] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE transactions on bio-medical engineering*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–20, Jun. 2000.

[12] A. Rechtschaffen and A. Kales, "[引用] washington, DC: US government printing office," *Public Health Service*, 1968.

[13] D. W. Gross and J. Gotman, "Correlation of high-frequency oscillations with the sleep-wake cycle and cognitive activity in humans," *Neuroscience*, vol. 94, no. 4, pp. 1005–1018, 1999.

[14] M. J. Prerau, R. E. Brown, M. T. Bianchi, J. M. Ellenbogen, and P. L. Purdon, "Sleep neurophysiological dynamics through the lens of multitaper spectral analysis," *Physiology*, vol. 32, no. 1, pp. 60–92, Jan. 2017.

[15] "Neuroimaging in python — nitime 0.9.dev documentation," https://nipy.org/nitime/api/generated/nitime.algorithms.spectral.html, accessed: 2024-4-13.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations*, May 2015.

[17] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '99. New York, NY, USA: Association for Computing Machinery, Aug. 1999, pp. 42–49.

[18] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica: casopis Hrvatskoga drustva medicinskih biokemicara / HDMB*, vol. 22, no. 3, pp. 276–282, 2012.

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[20] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-Lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 29, pp. 2076–2085, Oct. 2021.

[21] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, p. e0216456, May 2019.

[22] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, "Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG," *Biomedical signal processing and control*, vol. 61, p. 102037, Aug. 2020.

[23] H. Phan, K. P. Lorenzen, E. Heremans, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, M. Baumert, K. B. Mikkelsen, and M. De Vos, "L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[24] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An Attention-Based deep learning approach for sleep stage classification with Single-Channel EEG," *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 29, pp. 809–818, May 2021.

[25] G. Fu, Y. Zhou, P. Gong, P. Wang, W. Shao, and D. Zhang, "A Temporal-Spectral fused and attention-based deep model for automatic sleep staging," *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. PP, Jan. 2023.

[26] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE transactions on bio-medical engineering*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[27] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, "A deep learning method approach for sleep stage classification with EEG spectrogram," *International journal of environmental research and public health*, vol. 19, no. 10, May 2022.

[28] J. Fan, C. Sun, M. Long, C. Chen, and W. Chen, "EOGNET: A novel deep learning model for sleep stage classification based on Single-Channel EOG signal," *Frontiers in neuroscience*, vol. 15, p. 573194, Jul. 2021.

[29] P. Somaskandhan, T. Leppänen, P. I. Terrill, S. Sigurdardottir, E. S. Arnardottir, K. A. Ólafsdóttir, M. Serwatko, S. Þ. Sigurðardóttir, M. Clausen, J. Töyräs, and H. Korkalainen, "Deep learning-based algorithm accurately classifies sleep stages in preadolescent children with sleep-disordered breathing symptoms and age-matched controls," *Frontiers in neurology*, vol. 14, p. 1162998, Apr. 2023.

[30] A. K. Patel, V. Reddy, K. R. Shumway, and J. F. Araujo, *Physiology, Sleep Stages*. StatPearls Publishing, Sep. 2022.

[31] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, R. J. Weiss, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay,

"Scikit-learn: Machine learning in python," *Journal of machine learning research: JMLR*, vol. abs/1201.0490, Feb. 2011.