IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 19 February 2025, accepted 5 March 2025, date of publication 10 March 2025, date of current version 20 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3549436

RESEARCH ARTICLE

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

AOZORA ITO[1] AND TOSHIHISA TANAKA[1,2], (Senior Member, IEEE)
[1]Department of Electrical Engineering and Computer Science, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan
[2]RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan

Corresponding author: Toshihisa Tanaka (tanakat@cc.tuat.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST) Program under Grant JPMJCR1784, and in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 23K26110.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

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 lightweight model that achieves comparable performance to state-of-the-art models with only one-third of their parameters. Based on the rules for sleep scoring, self-attention is applied to each of the time- and frequency-domain inputs, a raw EEG signal and its amplitude spectrum. The simple method of continuously connecting the intermediate outputs of the epoch-wise model has resulted in a highly lightweight architecture. On the Sleep-EDF-78 dataset, our model achieves an accuracy of 84.8% and a kappa coefficient of 0.787 while requiring significantly fewer parameters (0.47×10^6) compared to existing models (1.3–4.54×10^6). The visualization of feature importance obtained from self-attention confirms that the proposed model learns representative waveform features, including K-complexes and sleep spindles.

INDEX TERMS Deep learning, electroencephalography, interpretability, self-attention mechanism, sleep stage classification, transfer learning.

I. INTRODUCTION

Sleep is critical to physical and mental well-being. Long-term sleep disruption is associated with an increased risk of a variety of diseases, including cardiovascular disease, obesity, and diabetes mellitus [1], [2]. Even acute sleep deprivation adversely affects daily life, including impaired judgment and cognitive abilities [2].

Polysomnography (PSG) is needed for diagnosing sleep disorders. PSG involves simultaneous recordings of various biological signals throughout the night, including electroencephalograpy (EEG), electrooculography, and electromyography signals. Based on recorded signals, experts can evaluate the sleep stages in 30 s epochs according to the American Academy of Sleep Medicine (AASM) sleep scoring manual [3]. AASM defines five sleep stages: Wake (W), rapid eye movement (REM or R), and three non-REM stages (N1, N2, N3) [3]. A complete sleep cycle takes roughly 90 to 110 minutes, with each cycle typically comprising five stages: W ($\sim$ 5% of sleep), N1 ($\sim$ 5%), N2 ($\sim$ 45%), N3 ($\sim$ 25%), and REM ($\sim$ 25%) [4]. Wake is characterized by high-frequency beta waves, while N1, a transitional stage,

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

is marked by low-voltage theta waves. N2, the most prevalent stage, features sleep spindles and K-complexes, which are crucial for sensory processing. N3, deep sleep, is dominated by high-amplitude delta waves, essential for physiological restoration. REM sleep exhibits low-amplitude beta waves, similar to wakefulness, but includes rapid eye movements and muscle atonia [4]. The decision rules to identify the sleep stages are based on features specific to each stage in the acquired biological signals. In particular, temporal and frequency features of EEG signals are listed in the scoring manual as features that define each stage in all five sleep stages. The recommended EEG electrode positions for PSG are frontal, central, and occipital. 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 [5]. Even a skilled expert can take up to 2 hours to score approximately 8 hours of sleep data [6]. Manual sleep scoring impedes suitably handling the millions of patients with sleep disorders [7], rendering automated scoring required. Automation of sleep scoring is expected to reduce the burden on specialists by several thousand hours per year [8].

Sleep scoring based on EEG follows predefined rules, making it suitable for automation using machine learning [8]. Machine learning models have been proposed to determine the sleep stage from physiological signals such as EEG, electrooculography, and electromyography [9], [10], [11]. Additionally, several models have been proposed to determine sleep stages based on only a single-channel EEG. Automatic sleep staging based on single-channel EEG can simplify PSG, thereby reducing the burden on physicians and patients. For these models, various types of inputs have been utilized, including raw EEG signals [12], [13] and spectrograms [14], [15], [16]. Deep learning models for sleep staging based on single-channel EEG have achieved equivalent or better performance than expert judgment [17].

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. Classical machine learning methods for sleep stage classification are relatively easy to interpret in terms of feature and sample contributions, but their classification performance is generally inferior to deep learning methods [18]. 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 [19]. The interpretation of incorrect inferences can also contribute to improve the model performance.

The self-attention mechanism [20] 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. SleepTransformer [15] applies self-attention to spectrograms, while cross-modal transformers [21] apply attention to EEG and EOG signals separately, enabling the interpretation of inference based on their weights. In this paper, we extend this idea by applying self-attention to the time- and frequency-domains of EEG. This allows us to quantify the contribution of each time point and frequency to the inference.

We propose a single-channel 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 each of the time- and frequency-domain inputs, a raw EEG signal and its amplitude spectrum. The amplitude spectrum has lower dimensionality compared to spectrograms and can concisely represent the overall frequency characteristics of the signal. This is expected to improve interpretability while reducing the dimensionality of the input data to the model, enabling the development of lightweight models through a reduction in the number of parameters and computational cost. To reflect the sleep context, we apply transfer learning to continuous epoch data.

II. METHODS

A. DATASET AND PREPROCESSING

We used a public dataset, Sleep-EDF Database Expanded [22], [23] in this study. Sleep-EDF Database Expanded has two versions (2013 and 2018) and two subsets (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 PSG 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 PSG data over 153 nights acquired from 78 healthy participants (37 males and 41 females) aged 25–101 years. The PSG data were annotated by an expert according to the R & K manual [24], 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.

The preprocessing pipeline followed standard procedures for EEG-based sleep stage classification. First, an anti-aliasing brick-wall filter was applied with a Nyquist frequency of 25 Hz before downsampling to 50 Hz. No additional bandpass filtering or denoising was performed. To ensure consistency in data labeling, we excluded epochs marked as “MOVEMENT” or “?”, which indicate excessive body motion artifacts or noise, as per the R & K manual used in the Sleep-EDF dataset. These labels denote epochs where sleep staging was deemed unreliable by expert annotators. This exclusion approach aligns with previous studies, including TinySleepNet [17]. Unlike some previous works, no additional normalization or rescaling procedures (e.g., z-score normalization) were applied to the EEG signals. To adhere to the manual of the American Academy of Sleep

TABLE 1. Number of epochs per label in Sleep-EDF-20 and Sleep-EDF-78. Numbers in parentheses indicate the percentage of each label in the total data set.

Dataset	W	N1	N2	N3	R	Total
Sleep-EDF-20	8285 (20%)	2804 (7%)	17,799 (42%)	5703 (13%)	7717 (18%)	42,308
Sleep-EDF-78	69,824 (35%)	21,522 (11%)	69,132 (35%)	13,039 (7%)	25,835 (13%)	199,352

Medicine, which defines five sleep stages, all the N4 labels were merged into the N3 label. 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.

B. 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 [25]. Furthermore, gamma waves are likely to be affected by artifacts and noise. Thus, we downsample the EEG signals to 50 Hz to establish a time-domain input. By reducing the input size of the model, we expect to reduce the number of parameters in the overall model. In spectral analysis for identifying differences in sleep EEG signals, the multi-taper method outperforms the single-taper method [26]. Thus, we calculate the amplitude spectrum in 0–25 Hz by applying the multi-taper method [27] 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, such as RNNs [14], [28] and Transformers [15], [29]. 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 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.

C. 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 [30] 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.

D. EVALUATION METRICS

We used the following evaluation metrics of model performance: accuracy (ACC), macro-F1 score (MF1) [31], kappa coefficient (Cohen's *kappa* κ) [32], 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

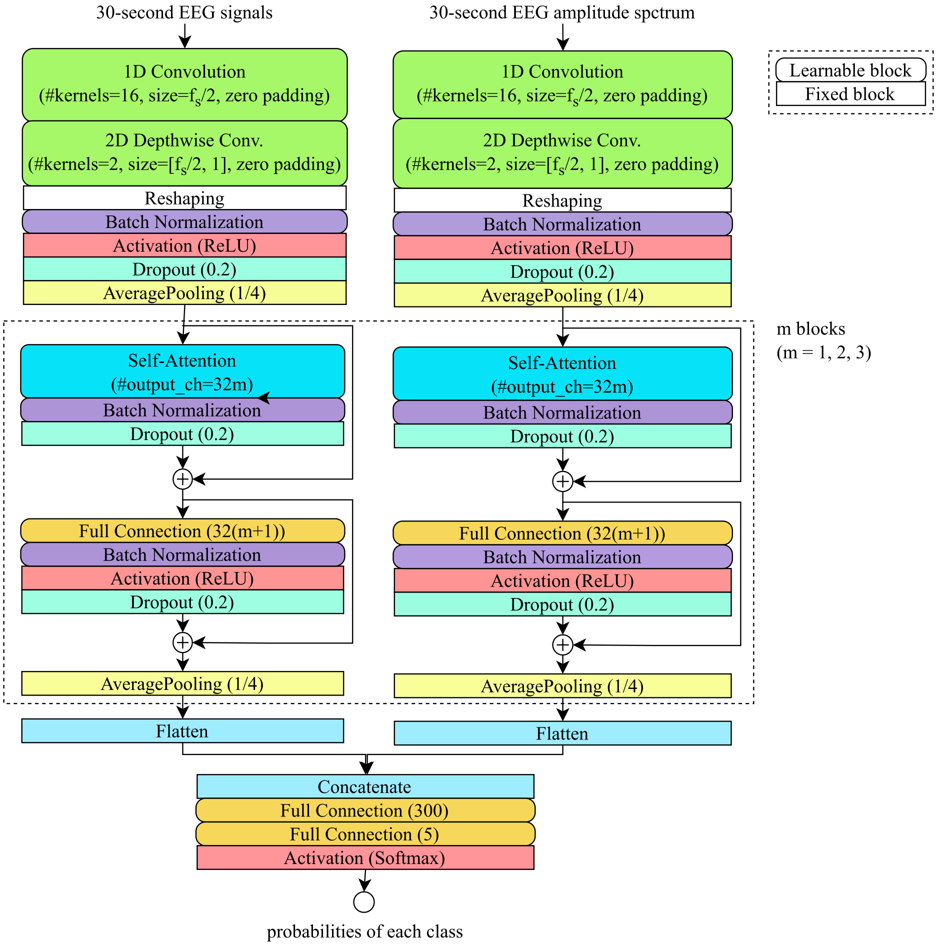


FIGURE 1. Epoch-wise classification model. The time-domain input is a 30 s raw EEG signal with 1500 samples (sampling rate f_S of 50 Hz), and the frequency-domain input is the 0–25 Hz amplitude spectrum with 750 samples. The model consists of convolutional layers with 16 kernels of size $f_S/2$ for both 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.

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 [33]. We compared SleepSatelightFTC with existing models for sleep stage classification in terms of the abovementioned metrics and number of model parameters.

III. RESULTS

A. 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.

B. 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. The number of AttnSleep parameters is retrieved from [34]. SleepEEGNet, IITNet, DeepSleepNet-lite, CTCNet, WASR + LCNN, SleepTransformer, EEGSNet, and FFTCN were trained under the same conditions as SleepSatelightFTC. AttnSleep, TSA-Net, SeriesSleepNet, and TinySleepNet used weighted cross-entropy, and L-seqsleepnet used cross-entropy averaged over the sequence length. TinySleepNet used data augmentation during training.

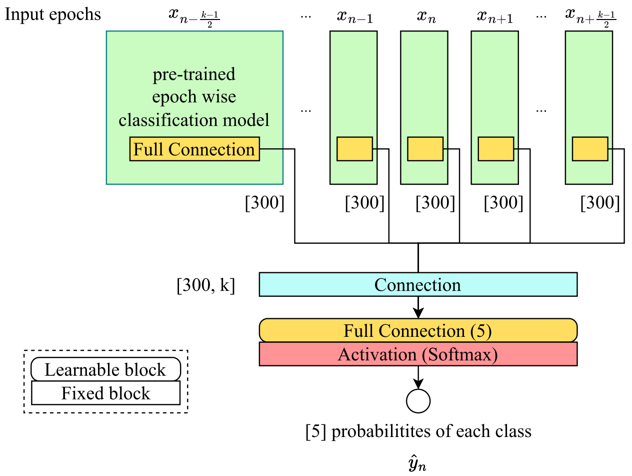


FIGURE 2. **Transfer learning in sequential epoch data. The pretrained epoch-wise classification model is applied to *k* consecutive epochs, where *k* is an odd number ranging from 3 to 29. The outputs of the fully connected layers from the epoch-wise classification model are combined for the *k* epochs 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.**

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

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 nearly as high as the other models but slightly lower, with MF1 being particularly low.

On the larger Sleep-EDF-78, with EEG signals downsampled to a sampling frequency of 50 Hz, 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. Additionally, when using the original EEG signals with a sampling frequency of 100 Hz, the model achieves an overall ACC of 82.2%, MF1 of 74.0%, and κ of 0.751.

C. 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.

IV. DISCUSSION

A. 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 0.6–4.54 $\times 10^6$, being approximately 1.3–9.66 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 Fig. 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 [8]. By suppressing overlearning, lightweight models like SleepSatelightFTC are less likely to reflect subjective biases, possibly increasing the consistency and reliability of sleep stage classifications.

Given the lightweight architecture of SleepSatelightFTC, it is well-suited for real-time applications, particularly in clinical environments where quick decisions are necessary. The reduced number of parameters ensures faster inference times compared to larger models. In terms of computational efficiency, SleepSatelightFTC achieves a processing speed of

TABLE 3. Sleep stage classification performance of evaluated models. F_s is the sampling frequency of input signals and * indicates methods that use different loss functions from that of SleepSatelightFTC. The number of AttnSleep parameters is retrieved from [34].

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)	SleepEEGNet [35]	84.3	79.7	0.790	89.2	52.2	86.8	85.1	85.0	2.60
	IITNet [36]	83.9	77.6	0.780	87.7	43.4	87.7	86.7	82.5	
	DeepSleepNet-lite [37]	84.0	78.0	0.780	87.1	44.4	87.9	88.2	82.4	0.60
	AttnSleep [38]*	85.6	80.9	0.800	90.3	47.9	89.8	89.0	85.0	4.54 [34]
	L-SeqSleepNet [28]*	86.3	79.3	0.813	91.6	45.3	88.5	86.2	85.2	0.63
	SeriesSleepNet [39]*	84.8	79.8	0.792	89.7	49.4	87.9	88.2	83.6	
	TSA-Net [13]*	86.6	80.4	0.816	90.5	46.9	89.2	**90.1**	85.4	
	CTCNet [29]	86.2	**82.5**	0.82	**92.2**	**57.6**	89.8	89.6	82.8	
	WASR + LCNN [40]	**87.6**	82.1	**0.83**	90.2	53.9	**92.0**	88.0	**86.3**	1.54
	SleepSatelightFTC (proposed), F_s: 50 Hz	85.7	77.7	0.800	89.4	42.8	87.6	84.4	84.2	0.47
Sleep-EDF-78 (Fpz–Cz EEG)	SleepEEGNet [35]	80.0	73.6	0.730	91.7	44.1	82.5	73.5	76.1	2.60
	TinySleepNet [17]*	83.1	78.1	0.770	92.8	51.0	85.3	81.1	80.3	1.3
	DeepSleepNet-lite [37]	80.3	75.2	0.730	91.5	46.0	82.9	79.2	76.4	0.60
	AttnSleep [38]*	82.9	78.1	0.770	92.6	47.4	85.5	83.7	81.5	4.54 [34]
	SleepTransformer [15]	81.4	74.3	0.743	91.7	40.4	84.3	77.9	77.2	3.7
	EEGSNet [14]	83.0	77.3	0.77	93.2	50.0	84.2	74.4	**83.5**	0.6
	TSA-Net [13]*	81.7	74.2	0.740	91.4	35.7	84.3	79.0	80.6	
	CTCNet [29]	82.5	**79.1**	0.78	92.5	53.8	86.8	**87.3**	74.8	
	FFTCN [16]	82.6	77.1	0.76	92.2	47.3	84.8	80.0	81.0	
	WASR + LCNN [40]	84.3	78.6	**0.79**	92.1	**58.4**	**90.2**	74.6	77.7	1.54
	SleepSatelightFTC (proposed), F_s: 50 Hz	**84.8**	77.8	0.787	**93.8**	47.4	86.5	79.5	82.1	0.46
	SleepSatelightFTC (proposed), F_s: 100 Hz	82.2	74.0	0.751	92.7	39.6	84.8	77.2	75.6	0.78

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

167 s per 1000 training steps, significantly outperforming other Transformer-based models such as SleepTransformer (308 s/1000 steps) [15] and L-SeqSleepNet (450 s/1000 steps) [28]. This substantial reduction in training time suggests that the model can also achieve faster inference speeds, making it a viable candidate for real-time or near-real-time deployment in sleep monitoring systems. While a more detailed speed analysis across different hardware platforms is necessary, these results indicate that SleepSatelightFTC provides a promising balance between computational efficiency and classification accuracy, making it highly suitable for practical applications.

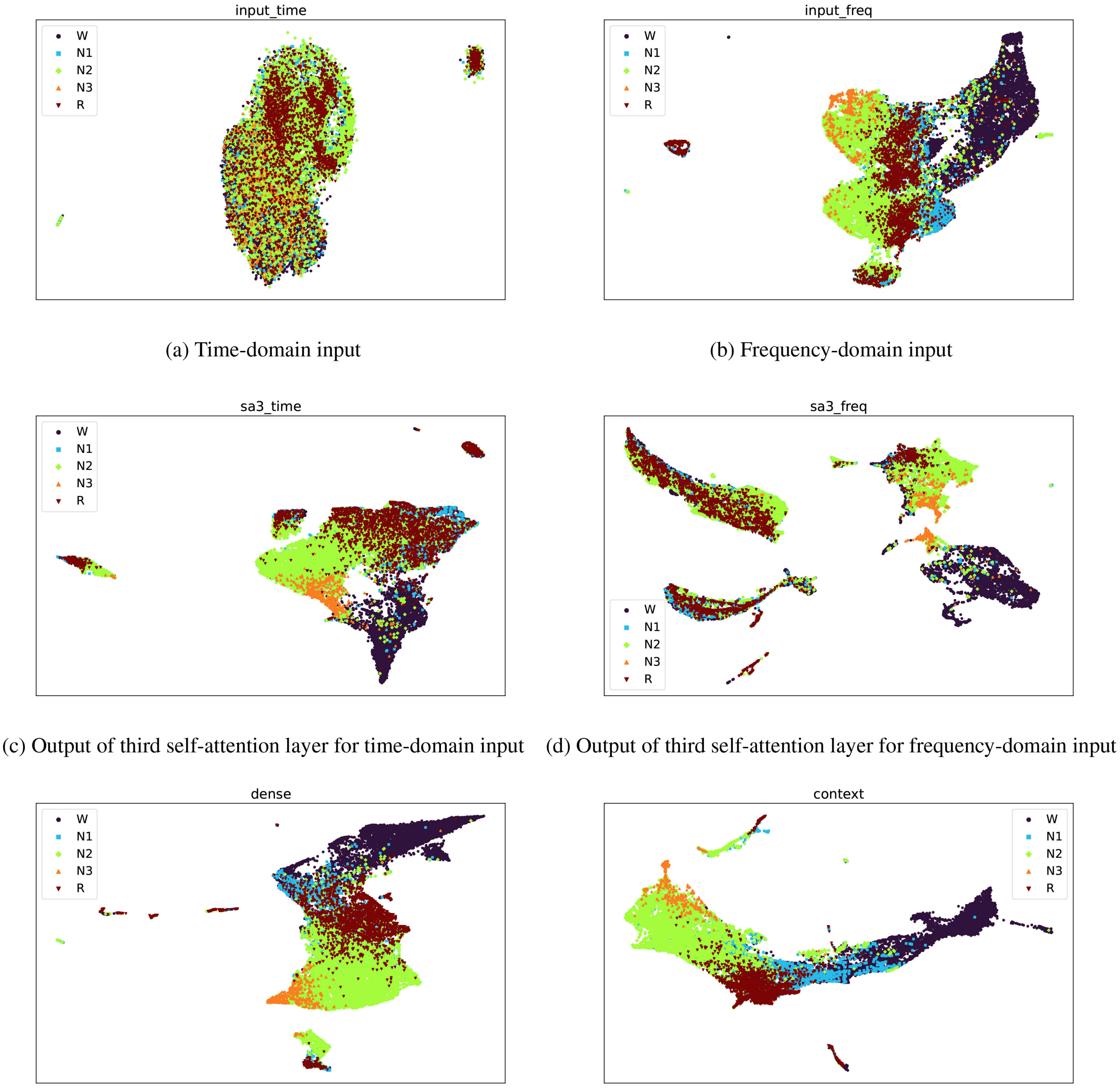
SleepSatelight FTC classification performance on Sleep-EDF-20 was lower than in previous studies. The F1 scores per class for N1 and N3 were the lowest among the models we compared. This may be due to the limited number of samples for these stages in the smaller Sleep-EDF-20 dataset, which prevented the model from effectively learning their features. As a result, this contributed to the overall lower performance observed on this dataset.

SleepSatelightFTC achieves higher F1 scores for sleep stages N2 and W but lower F1 scores for sleep stage N1 than existing models on Sleep-EDF-20 and Sleep-EDF-78. 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.

The number of SleepSatelightFTC parameters increased by 1.8 when using EEG with a sampling frequency of 100 Hz, compared to a sampling frequency of 50 Hz. The overall performance decreased by 2.6% in the accuracy, 3.8% in the macro F1 score, and 0.036 in the Kappa coefficient. This is likely due to the fact that adding gamma waves as input, which are less relevant for determining sleep stage, prevented the model from learning the basis for inference. This result suggests that not including gamma waves as input may be effective in inferring sleep stages.

B. 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



(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.

studies, context is considered from various epoch lengths, such as adjacent epochs [14], 15 epochs [41], and 100 epochs [42]. 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 approximately 45% of the total sleep time and is longer in late sleep stages [4]. During intervals of identical sleep stages, transfer learning is expected to compensate for out-of-context inferences.

C. VISUALIZATION OF INFERENCE PROCESS

We visualized the inference process of SleepSatelightFTC using a dimensionality reduction method called uniform manifold approximation and projection [43]. The data distributions per class after time-domain input, frequency-domain input, self-attention layer output, epoch-wise inference, and

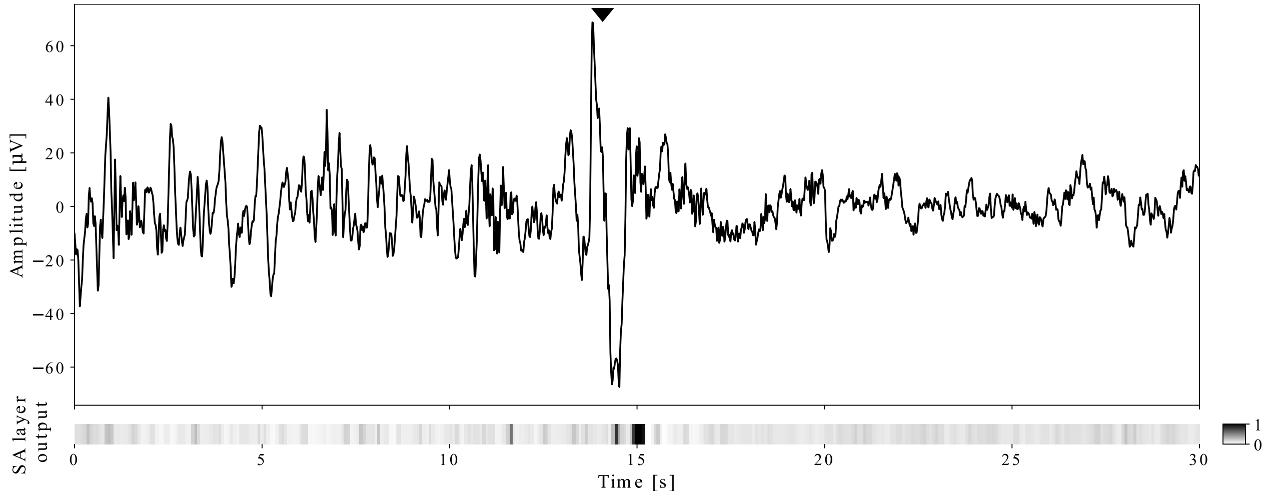


FIGURE 4. EEG signal for sleep stage N2 and its self-attention layer output heatmap.The heatmap shows the importance for classification of each timepoint in the EEG signal. The self-attention layer assigns higher importance to the waveform resembling a K-complex (▼), which is a characteristic feature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

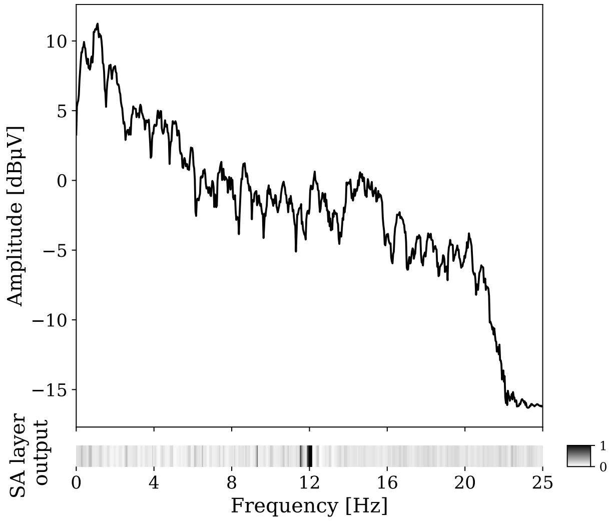


FIGURE 5. Amplitude spectrum of sleep stage N2 and its self-attention layer output heatmap. The heatmap shows the importance for classification of each frequency component in the amplitude spectrum. The self-attention layer assigns higher importance to the 12 Hz frequency component, which is associated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

transfer learning are shown in Fig. 3. Fig. 3a and 3b show that the model inputs are more coherently distributed by class in the frequency domain than in the time domain. On the other hand, Fig. 3b and 3c 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 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.

D. 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 [44] 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 Fig. 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 [3]. 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 [3]. 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.

While recent Transformer-based models, such as SleepTransformer [15] and CTCNet [29], have demonstrated strong performance in sleep stage classification, SleepSatelightFTC offers several advantages in terms of model efficiency and interpretability. Transformer models generally rely on self-attention mechanisms applied across long temporal sequences, requiring substantial computational resources and a large number of parameters. In contrast, SleepSatelightFTC employs a lightweight self-attention architecture that focuses on epoch-wise time- and frequency-domain representations, maintaining interpretability while reducing computational costs. Specifically, CTCNet integrates a Transformer-based backbone to model sequential dependencies in sleep stages, but this comes at the cost of increased model complexity and computational requirements. Our results show that SleepSatelightFTC achieves comparable accuracy with significantly fewer parameters, making it a more practical choice for real-time applications. Additionally, the ability to visualize self-attention mechanisms in both time and frequency domains enhances its transparency, a critical aspect for clinical use.

While some Transformer-based models, such as MultiChannelSleepNet [11] or Cross-Modal Transformers [21], leverage multimodal data for improved classification performance, our study focuses on EEG-only models. Therefore, SleepSatelightFTC provides a more direct comparison to models like CTCNet and SleepTransformer, demonstrating the feasibility of an efficient, interpretable, and computationally lightweight sleep staging approach.

V. 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. The model employed self-attention to time- and frequency-domain inputs, raw EEG signals and amplitude spectrum, and transfer learning to sequential epochs to consider temporal context. 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 and judge the validity of inference.

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. EEG patterns in individuals with sleep disorders may differ significantly from those of healthy subjects, affecting the model's generalizability. In future work, we will evaluate the model performance based on data from patients with sleep disorders and explore domain adaptation techniques to address this issue. One potential approach is to fine-tune only the final fully connected layer of the trained model. Since sleep architecture varies between healthy individuals and those with sleep disorders, adapting the final classification layer while retaining the learned feature representations from healthy subjects may help mitigate domain discrepancies and improve model performance in clinical applications. Additionally, we plan to investigate methods to handle inter-rater variability and improve performance on underrepresented sleep stages.

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.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. M. Bertisch, B. D. Pollock, M. A. Mittleman, D. J. Buysse, L. A. Bazzano, D. J. Gottlieb, and S. Redline, "Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: Sleep heart health study," *Sleep*, vol. 41, no. 6, Jun. 2018, Art. no. zsy047.

[2] G. Medic, M. Wille, and M. Hemels, "Short- and long-term health consequences of sleep disruption," *Nature Sci. Sleep*, vol. Volume 9, pp. 151–161, May 2017.

[3] 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," Amer. Acad. Sleep Med., Darien, CT, USA, 2018.

[4] A. K. Patel, V. Reddy, K. R. Shumway, and J. F. Araujo, *Physiology, Sleep Stages*. Treasure Island, FL, USA: StatPearls Publishing, Sep. 2022.

[5] H. R. Colten and B. M. Altevogt, "Institute of medicine (U.S.) committee on sleep medicine and research," in *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. Washington, DC, USA: National Academy Press, 2006.

[6] 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.

[7] 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*, vol. 7, no. 1, p. 1, Dec. 2018.

[8] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: Recent development, challenges, and future directions," *Physiological Meas.*, vol. 43, no. 4, Apr. 2022, Art. no. 04TR01.

[9] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," 2021, *arXiv:2105.13864*.

[10] Y. Li, Z. Xu, Y. Zhang, Z. Cao, and H. Chen, "Automatic sleep stage classification based on a two-channel electrooculogram and one-channel electromyogram," *Physiolog. Meas.*, vol. 43, no. 7, Apr. 2022, Art. no. 07NT02.

[11] Y. Dai, X. Li, S. Liang, L. Wang, Q. Duan, H. Yang, C. Zhang, X. Chen, L. Li, X. Li, and X. Liao, "MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with PSG," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 9, pp. 4204–4215, Sep. 2023.

[12] D. Zhou, J. Wang, G. Hu, J. Zhang, F. Li, R. Yan, L. Kettunen, Z. Chang, Q. Xu, and F. Cong, "SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103592.

[13] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1008–1018, 2023.

[14] 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," *Int. J. Environ. Res. Public Health*, vol. 19, no. 10, p. 6322, May 2022.

[15] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[16] J. Bao, G. Wang, T. Wang, N. Wu, S. Hu, W. Hee Lee, S.-L. Lo, X. Yan, Y. Zheng, and G. Wang, "A feature fusion model based on temporal convolutional network for automatic sleep staging using single-channel EEG," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 11, pp. 6641–6652, Nov. 2024.

[17] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 641–644.

[18] D. Palogiannidis, Z. Mihoub, and H. Solieman, "Comparing ML and DL sleep staging methods with the effect of PCA using EEG signals," in *Proc. Conf. Russian Young Res. Electr. Electron. Eng. (ElConRus)*, Jan. 2022, pp. 1387–1390.

[19] S.-C. Lu, C. L. Swisher, C. Chung, D. A. Jaffray, and C. Sidey-Gibbons, "On the importance of interpretable machine learning predictions to inform clinical decision making in oncology," *Frontiers Oncol.*, vol. 13, Feb. 2023, Art. no. 1129380.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[21] J. Pradeepkumar, M. Anandakumar, V. Kugathasan, D. Suntharalingham, S. L. Kappel, A. C. De Silva, and C. U. S. Edussooriya, "Towards interpretable sleep stage classification using cross-modal transformers," 2022, *arXiv:2208.06991*.

[22] 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, Sep. 2000.

[23] A. L. Goldberger, L. A. N. 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, Jun. 2000, Art. no. E21520.

[24] A. Rechtschaffen and A. Kales, *Public Health Service*. Washington, DC, USA: U.S. government printing office, 1968.

[25] 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, Nov. 1999.

[26] 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.

[27] *Neuroimaging in Python–Nitime 0.9.Dev Documentation*. Accessed: Apr. 13, 2024. [Online]. Available: https://nipy.org/nitime/api/generated/nitime.algorithms.spectral.html

[28] 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 modeling for automatic sleep staging," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 10, pp. 4748–4757, Oct. 2023.

[29] W. Zhang, C. Li, H. Peng, H. Qiao, and X. Chen, "CTCNet: A CNN transformer capsule network for sleep stage classification," *Meas., J. Int. Meas. Confederation*, vol. 226, Feb. 2024, Art. no. 114157.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, May 2015, pp. 1–11.

[31] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, Aug. 1999, pp. 42–49.

[32] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica, Casopis Hrvatskoga Drustva Medicinskih Biokemicara/HDMB*, vol. 22, no. 3, pp. 276–282, 2012.

[33] 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.

[34] M. Singh, S. Chauhan, A. K. Rajput, I. Verma, and A. K. Tiwari, "EASM: An efficient AttnSleep model for sleep apnea detection from EEG signals," *Multimedia Tools Appl.*, vol. 84, no. 4, pp. 1985–2003, Apr. 2024.

[35] 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, May 2019, Art. no. e0216456.

[36] 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," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102037.

[37] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[38] 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 Trans. Neural Syst. Rehabil. Eng., IEEE Eng. Med. Biol. Soc.*, vol. 29, pp. 809–818, 2021.

[39] M. Lee, H.-G. Kwak, H.-J. Kim, D.-O. Won, and S.-W. Lee, "SeriesSleepNet: An EEG time series model with partial data augmentation for automatic sleep stage scoring," *Frontiers Physiol.*, vol. 14, Aug. 2023, Art. no. 1188678.

[40] K. Fei, J. Wang, L. Pan, X. Wang, and B. Chen, "A sleep staging model on wavelet-based adaptive spectrogram reconstruction and light weight CNN," *Comput. Biol. Med.*, vol. 173, May 2024, Art. no. 108300.

[41] 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 Neurosci.*, vol. 15, Jul. 2021, Art. no. 573194.

[42] 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 Neurol.*, vol. 14, Apr. 2023, Art. no. 1162998.

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

[44] 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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Feb. 2011.

AOZORA ITO received the B.E. degree in electrical engineering and in computer science from Tokyo University of Agriculture and Technology, Tokyo, Japan, in 2023, where he is currently pursuing the M.E. degree with the Graduate School of Engineering. His research interests include machine learning, neuroscience, and biomedical signal processing.

TOSHIHISA TANAKA (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology, in 1997, 2000, and 2002, respectively.

From 2000 to 2002, he was a JSPS Research Fellow. From October 2002 to March 2004, he was a Research Scientist with the RIKEN Brain Science Institute. In April 2004, he joined the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, where he is currently a Professor. In 2005, he was a Royal Society Visiting Fellow with the Communications and Signal Processing Group, Imperial College London, U.K. From June 2011 to October 2011, he was a Visiting Faculty Member of the Department of Electrical Engineering, University of Hawaii at Manoa. He is the Co-Founder and the CTO of Sigron Inc. His research interests include signal processing and machine learning, including brain and biomedical signal processing, brain–machine interfaces, and adaptive systems.

Prof. Tanaka is a member of IEICE, APSIPA, the Society for Neuroscience, and Japan Epilepsy Society. Furthermore, he served as a Member-at-Large on the Board of Governors (BoG) for the Asia–Pacific Signal and Information Processing Association (APSIPA). He was a Distinguished Lecturer of APSIPA. He serves as the Vice-President for APSIPA. He served as an Associate Editor and a Guest Editor for special issues in journals, including IEEE Access, *Neurocomputing*, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *Computational Intelligence and Neuroscience* (Hindawi), IEEE Transactions on Neural Networks and Learning Systems, *Applied Sciences* (MDPI), *Advances in Data Science and Adaptive Analysis* (World Scientific), and *Neural Networks* (Elsevier). He served as the Editor-in-Chief for *Signals* (MDPI). He is a Co-Editor of *Signal Processing Techniques for Knowledge Extraction and Information Fusion* (Mandic, Springer, 2008) and a leading Co-Editor of *Signal Processing and Machine Learning for Brain-Machine Interfaces* (Arvaneh, IET, U.K., 2018).