Enhancement on Model Interpretability and Sleep Stage Scoring Performance with A Novel Pipeline Based on Deep Neural Network

Zheng Chen, Student Member, IEEE, Ziwei Yang, Student Member, IEEE, Ming Huang*, Member, IEEE, Toshiyo Tamura, Senior Member, IEEE, Naoaki Ono, MD Altaf-Ul-Amin, Shigehiko Kanaya

Abstract—Considering the natural frequency characteristics in sleep medicine, this paper first proposes a time-frequency framework for the representation learning of the electroencephalogram (EEG) following the definition of the American Academy of Sleep Medicine. To meet the temporal-random and transient nature of the defining characteristics of sleep stages, we further design a context-sensitive flexible pipeline that automatically adapts to the attributes of data itself. That is, the input EEG spectrogram is partitioned into a sequence of patches in the time and frequency axes, and then input to a delicate deep learning network for further representation learning to extract the stage-dependent features, which are used in the classification step finally. The proposed pipeline is validated against a large database, i.e., the Sleep Heart Health Study (SHHS), and the results demonstrate that the competitive performance for the wake, N2, and N3 stages outperforms the state-of-art works, with the F1 scores being 0.93, 0.88, and 0.87, respectively, and the proposed method has a high inter-rater reliability of 0.80 kappa. Importantly, we visualize the stage scoring process of the model decision with a Layer-wise Relevance Propagation (LRP) method, which shows that the proposed pipeline is more sensitive and perceivable in the decision-making process than the baseline pipelines. Therefore, the pipeline together with the LRP method can provide better model interpretability, which is important for clinical support.

Index Terms—Sleep stage scoring, EEG, feature representation, deep neural network, model interpretability

I. INTRODUCTION

Sleep is a human function, the characteristics of which are manifested by a sequence of physiological alterations. Sleep screening is not only a tool in the assessment of pathophysiology [1]–[3] but an essential ingredient in the exploration of neuroscience [4]–[6]. Experimental validations show that some electroencephalogram (EEG) features shown in different sleep stages have extraordinary physiological significance [7]. For instance, slow waves contribute to memory consolidation [8], and the sleep spindle is highly correlated to intellectual ability [9]. Therefore, determining sleep stages and consequent macrostructures is indispensable in sleep care and sleep science.

The American Academy of Sleep Medicine (AASM) defines sleep into five different stages. There is a constant cyclic pattern of sleep stages from wake to non-rapid eye movement (NREM) to REM that repeats several times a night, where NREM consists of three stages, i.e., N1, N2, and N3 [10]. Sleep staging clinically relies on overnight electrophysiological recordings using the polysomnography (PSG).

This research and development work was supported by the Grant-in-Aid for Early-Career Scientists #20K19923.
Zheng Chen, Ziwei Yang, Ming Huang, Naoaki Ono, MD Altuf-Ul-Amin and Shigehiko Kanaya are with Graduate School of Science and Technology, Nara Institute of Science and Technology, Takayamacho 8916-5, Ikoma, 6300192 Japan. (e-mail: {chen.zheng.bn1, yang.ziwei.ya3, alex-mhuang, nono, amino, skanaya}@is.naist.jp)

Naoaki Ono and Shigehiko Kanaya are with Data Science Center, Nara Institute of Science and Technology, Takayamacho 8916-5, Ikoma, 6300192 Japan.

Toshiyo Tamura is with Institute for Healthcare Robotics, Waseda University, Japan. (e-mail: tmurata1949@gmail.com)

Corresponding author: Ming Huang (alex-mhuang@is.naist.jp)

[11]. Clinically, experts are required to inspect the PSG recording and then assign a sleep stage to each 30-second epoch on the basis of distinct features. This laborious handling inevitably limits the batch scoring of stages. Moreover, given that the recent advances in portable monitoring with fewer sensors could provide technical support for daily sleep screening [12], [13]. Therefore, a reliable automatic sleep stage scoring is crucial for the sleep community.

Although the performance of the automatic scoring algorithm has been greatly improved by deep learning models as well as free access to large-scale sleep databases [14]–[22], the optimization of the learning pipeline is still at the midway point. Specifically, the representation strategy for EEG signals and the learning structure of the network need precise optimization to adapt to the intrinsic traits of autonomic sleep scoring based on EEG signals.

With the optimized combination on these two aspects, we hope to improve the overall performance of automatic sleep stage scoring on one hand, and to shed light on clinical/physiological significance by retaining the interpretability of the model as much as possible. In the following two sub-sections, the rationale for why the aforementioned special care should be given will be explained, whereafter the motivation of this paper will be elicited.

A. Notes on EEG representation

Quantitative electrophysiology, which allows for the sophisticated processing of EEG signals, has revealed how widespread human neuronal systems generate the characteristic electrical rhythms of different sleep stages [6]. The EEG-based technical specifications of AASM define the staging rule with a 0.5 to 30-35 Hz range in frequency oscillation [10]. By inferring stage-dependent occurrences (or increases) in EEG features, the clinical setting assigns a stage to each 30-second epoch.

Specifically, wakefulness can be identified by a dominant (>50%) alpha rhythm (8-13 Hz), and beta waves (16-32 Hz) are the states associated with the wake stage [23]. Theta waves (4-8 Hz) that consist of low-amplitude and mixed frequency waves are the criteria for stage N1. When one or more K-complex and sleep spindles under low beta waves (12-16 Hz) appear, the corresponding epoch will be scored as stage N2, while slow delta waves (1-4 Hz) in EEG are the dominant feature of stage N3. Although the significant characteristic of REM is movement of the eyes and muscles, saw-tooth waves are an alternative criteria. A summary is shown in TABLE I.

Besides these clinical features, effort has been put into the searching for more salient features in sleep scoring. The proposed features include spectrogram [14]–[19], power spectral density [24]–[26],
functions and downsampling operators, proposed scoring frameworks in frequency or time-frequency domain [20]–[22]. By interleaving a classification task, CNNs have been used in feature extraction from the recent years has made feature generation more flexible and problem-specific [17], [20]. The subject-wise generalization gap, especially in models with complicated representation learning, can be problematic [39]. Therefore, where to draw a line between predefined and adjustable representation generation should be discussed. We can think of predefined-feature-only input and raw EEG input to a learning model as two extremes and assume that a better solution lies between them.

B. Notes on learning structure

The need to delicately select learning structures is rooted in the origin of the EEG signal and the intrinsic characteristics of neuron activities in the cortex, that is, the spatial heterogeneity of relevant features in the cortex [10], [40], which means that the corresponding features may not always be seen from one lead, while stage-irrelevant features/disturbance may be seen [41]. This situation is even made tougher by the transient and temporally random natures of some dominant characteristics [42], [43]. For instance, sleep spindle is the defining characteristic of N2 stage, however, it is a spontaneously burst and only last for a short duration of 0.5–1.5 seconds. Therefore, in N2 stage, sometimes the stage-irrelevant features dominate the current epoch [23]. As a consequence, the stage-dependent features are not always unequivocal among the same staged EEG epochs. Therefore, in addition to the feature representation, stage-specific feature refinement in the global content is needed.

The sleep community has witnessed significant progress with the merging of deep neural networks in automatic sleep scoring [16]–[21]. It is expected that the separated feature engineering study and classification study to be superseded by the end-to-end framework [44], [45].

Due to CNNs becoming the de facto standard for visual classification tasks, CNNs have been used in feature extraction from the frequency or time-frequency domain [20]–[22]. By interleaving a collection of multi-size convolutional filters with non-linear activation functions and downsampling operators, proposed scoring frameworks expect to capture sophisticated local features and attain an inductive representation of the EEG epoch [46]. Since CNN does not consider the local features against the global context, it will result in a local inductive bias [47]. Consequently, it may draw more attention from the classifier onto a pattern of the neighboring area while distracting the classifier from really important but transient information and its relevance. Meanwhile, feature extraction in each EEG epoch is handled by the sharing weights of filters. This mechanism expects key rhythms to occur in a relatively inherent position. Therefore, there is a mismatching between the translation-invariant constraint of CNNs and the temporally random and transient nature of the relevant defining characteristics.

Recurrent neural networks (RNNs), such as long short-term memory (LSTM) and gated recurrent neural networks, have been established as the state-of-the-art (SOTA) in automatic sleep scoring [17], [20], [25], [45]. Different from CNNs, the recurrent-based models pay attention to the information of the global context. By generating a set of hidden states, the decision-making considers the influences of a sequence of previous time steps or future steps (by the Bidirectional LSTM) [48]. Given the issue of transiently bursting rhythms, such as non-periodic sharp-wave ripples and spindle activities [21], [49], RNNs cannot be regarded as the best choice for sleep scoring. Further, the inherently sequential processing flow of RNNs precludes the possibility of parallelization within feature capture [50].

Although some advanced deep frameworks have been proposed, for instance, the sequence-to-sequence framework that can handles a set of consecutive sleep epochs by simulating the transition rule of inter-stage [17], epoch-wise automatic scoring frameworks remain the limitations. It implies that a context-sensitive flexible pipeline that automatically adapts attributes of data itself is a requirement.

C. Goal of this study and contributions

Given the considerations introduced above, this work proposes a tailored pipeline for autonomic sleep scoring by proposing a novel way of generating features that alleviate the influence of temporally random and transient nature of the EEG features while retaining the resolution in the frequency domain. This method can find its clinically informative explanation, which will be explained in Section III-A and III-B.

It can also be seen that epoch-wise contextual relating of temporally local features is necessary. The time-frequency feature representation is first organized as a time series of features of the frequency domain and then input into an elaborate attention-based deep learning model. The attention mechanism is designed to extract the global contextual relevance between units of a time series signal or stream of text. Therefore, it is regarded as a suitable learning scheme in this study.
The attention-based model, including the Transformer [51], [52], has been tried in sleep scoring and relevant problems [20], [53], [54]. Among these preceding studies, Qu et al. applied residual blocks to the EEG signal after Hilbert-Transform-like preprocessing and the Transformer on the epoch level for accurate sleep scoring [20]. The results are plausible since the macro sleep structure is relatively constant with cyclic patterns and can be modeled by using sequence-to-sequence strategy. According to the AASM guideline, sleep stages can be generally determined on the basis of intra-epoch signals. Moreover, as has been validated both computationally and experimentally, the structure of shorter signals embeds information related to sleep stages [55], [56].

Specifically, this paper transforms two channels of EEG signals into spectrograms, which are then divided into five parts following the five frequency bands as indicated in TABLE I and the power spectral density is calculated for every 1 Hz (frequency patching). Whereafter, each part is further partitioned temporally into 1-second bins (time patching). Finally, the resultant patches are input into an epoch-wise scoring model. With the attention matrix output by the trained attention-based model, this pipeline is expected to show physiologically interpretable patterns that are important in stage classification. The results also show that our methods reach a new state-of-the-art performance in scoring accuracy of intra-epoch classification under the EEG-signal-only restriction.

Thus, the contributions of this paper are as follows:

- This paper proposes a new framework to extract the relevant features of sleep stage that considers the intrinsic characteristics of stage defining characteristics expressed in EEG signal. The framework is able to elevate the scoring accuracy on one hand, and provides better interpretability on the other hand;
- By constructing a delicate network structure, this paper reaches state-of-the-art performance in scoring accuracy of intra-epoch classification under the EEG-signal-only restriction;
- By using the layer-wise relevance propagation method in the proposed model, the resultant attention-derived matrix can be explained clinically and therefore provides interpretability for the staging decision.

II. SHHS DATABASE

The dataset used in this work is from the Sleep Heart Health Study (SHHS). The purpose of SHHS is to test whether sleep-related breathing is associated with an increased risk of coronary heart disease, stroke, all-cause mortality, and hypertension. Access to the SHHS was permitted via the National Sleep Research Resource [55]. The database consists of two rounds of at-home PSG recordings (SHHS-1 and SHHS-2), and only SHHS-1 is used in our work. Nine institutions cooperatively created SHHS-1, for which the full PSG data of 5793 individuals are collected between 1995 and 1998. The participants were restricted to those who met the recruited criteria, including, age (older than 40 years), no sleep-related diseases, etc.

Sleep stages were scored by consensus between two sleep technologists blind to the condition of the participants for six classes (wake, REM, S1, S2, S3, S4) according to the R&K guidelines [57]. Noteworthy, this work merges S3 and S4 into stage N3 in reference to the AASM criteria. Due to unscored epochs, invalid labels, and misaligned records, 5736 subjects were selected to construct the experimental dataset in this study. Each recording provided two channels (C4-A1 and C3-A2) of EEG, sampled at 125 Hz.

Considering the issue of imbalance in the dataset, a sample balanced subset called “healthy-set” was extracted. Here, this subset served in the pre-training phase in our experiments. Since SHHS-1 provides a personal health description of all subjects, the subjects were selected on the basis of inspection of six clinical criteria. A detailed description of healthy-set can be seen in Appendix I. The healthy-set consists of 684 subject-wise recordings with 26080 epochs for each class. A summary of the experimental dataset used in this work is shown in TABLE II.

III. SLEEP SCORING FRAMEWORK

The proposed framework, the overall illustration of which is shown in Fig. 2, is a sequence-to-target workflow. A spectrogram is transformed into a sequence of patches of 5 frequency bands, augmented with class tokens, and finally input into our attention-based model for sleep classification. This section is divided into...
four sub-sections, the first two of which introduce the generation of time-frequency representations (III-A) and a novel framework for feature organization (III-B). The second half focuses on the network architecture, which includes the sequence augmentation and embedding of the input (III-C) and the architecture of the model (III-D).

### A. Time-Frequency Representation

EEG recordings are typically contaminated with various types of artifacts. An 8th order Butterworth bandpass filter with cutoff frequencies between 0.2 and 32 Hz was applied to all recordings. Whereafter, each epoch is transformed into a spectrogram to manifest its time-frequency characteristics. Given that the advantage of short EEG segments (1-second) in expressing the transient sleep rhythms has been shown [58], a log-power spectrogram is generated for EEG segments (1-second) in expressing the transient sleep rhythms frequencies between 0.2 and 32 Hz was applied to all recordings.

An 8th order Butterworth bandpass filter with cutoff frequencies between 0.2 and 32 Hz was applied to all recordings. Whereafter, each epoch is transformed into a spectrogram to manifest its time-frequency characteristics. Given that the advantage of short EEG segments (1-second) in expressing the transient sleep rhythms has been shown [58], a log-power spectrogram is generated for EEG segments (1-second) in expressing the transient sleep rhythms frequencies between 0.2 and 32 Hz was applied to all recordings.

A 8th order Butterworth bandpass filter with cutoff frequencies between 0.2 and 32 Hz was applied to all recordings. Whereafter, each epoch is transformed into a spectrogram to manifest its time-frequency characteristics. Given that the advantage of short EEG segments (1-second) in expressing the transient sleep rhythms has been shown [58], a log-power spectrogram is generated for EEG segments (1-second) in expressing the transient sleep rhythms frequencies between 0.2 and 32 Hz was applied to all recordings.

#### B. Frequency-time patching

In this part, we propose an important processing framework termed frequency-time patching for organizing the input spectrogram. That is, the spectrogram is divided into eight frequency bands every 4 Hz in accordance with the five frequency bands of TABLE I, where the beta band is further divided into four segments (frequency patching). Afterward, time patches are acquired by extracting and rearranging each column (1-sec) of the spectrogram (time patching). This framework embodies two scientific findings, that is, the concurrence of sleep stages in one epoch [40] and the fact that most of the clinically crucial features can be represented by the time-frequency features at a 1-sec resolution [42]. With the special structure of the network architecture, which will be introduced in (III-D), the frequency-time patching structure will be kept consistent throughout the network to facilitate the discussion of model interpretability. Along with Fig. 3, frequency-time patching is introduced herein.

- **Frequency patching**: Spectrogram $S$ is split into five parts $S = (S_\delta, S_\theta, S_\alpha, S_\beta, S_\gamma)$ in accordance with the five predominant frequency bands (TABLE I) of sleep rhythms. The first four spectrogram parts, $S_\delta, S_\theta, S_\alpha, S_\beta \in \mathbb{R}^{(F/8) \times T \times C}$, have the same bandwidth of 4 Hz. A block of the beta band is quartered into four sub-blocks, $S_\beta = (S_{\beta1}, S_{\beta2}, S_{\beta3}, S_{\beta4})$, where the subscript $\beta1 \sim \beta4$ correspond to the four quarters of the beta band.

- **Time patching**: Each frequency block as shown in the middle of Fig. 3 is divided by column to extract frequency-time patches. Therefore, one time patch is $S_t^i = (S_{\delta1}^i, S_{\theta1}^i, S_{\alpha1}^i, S_{\beta1}^i, S_{\gamma1})$, where the superscript $i$ denotes the i-th second.

- **Sub-block averaging**: For the frequency-time patches in the Beta band, mean values are calculated and used in the following steps, so that $S_{\beta} = \bar{S}_{\beta1}, \bar{S}_{\beta2}, \bar{S}_{\beta3}, \bar{S}_{\beta4}$.

- **Spectrogram transformation**: For each time patch $S_t^i$, the new structure is $S_t^i = (S_{\delta1}^i, S_{\theta1}^i, S_{\alpha1}^i, S_{\beta1}^i, S_{\gamma1}) \in \mathbb{R}^{20}$, resulting in the transformation of the original spectrogram $S_t^i$ into $S_t^i \in \mathbb{R}^{20 \times T \times C}$.

- **Patches rearrangement**: A sequence of frequency-time (FT) patches is generated by column-wise traversal. Therefore, $S_{seq} \in \mathbb{R}^{150 \times D_p}$ (150 = 5 frequency bands $\times$ 30 sec) is the patch sequence for one epoch, where $D_p$ denotes the dimension of each patch $\in \mathbb{R}^4$.

To validate the effect of the frequency patching, we compare its performance with time patches (1-sec) input to the same network architecture. The results of the baseline model can be found in TABLE V, the (Time patching + proposed model) pipeline.

### C. Patches Sequence Embedding

As indicated in above, physiologically, the individual FT patches contain uneven information about sleep stages and the concurrence of patches along with its sequential distance (time difference) is informative. Therefore, the self-attention, is a suitable mechanism for the process of feature extraction, and the attention-based classification model have been used computer vision [51] [59].

To construct our model, an extra class token $S_{Clss}$ is inserted to each sample at the beginning of the FT patches sequence. With the processing and propagation of the input through the model layers, this token are retained as the stage indicator, and will be used in the final stage scoring/classification step. These two steps can be formulated as below:

$$S_{seq}^t = \text{Concat}(S_{Clss}, E \cdot S_{seq}),$$  \hspace{1cm} (1)

where $E$ is a patch-wise linear projection that project the two FT patches of two EEG channels of the same time into higher dimension. Consequently, $S_{seq} \in \mathbb{R}^{(150d+1) \times D_p}$ is the output sequence, where $D$ is the dimension of output of the linear projection.

Because The attention mechanism does not differentiate the position of the keys when calculating the attention values of a query, the position embedding is a routine supplementary. In adapting the mechanism for sleep staging, where the global position could be a
nusiance, we make use of the positional encoding technique [51], by which a suitable positioning scheme can be generated by training. After, the position embedding is merged into the $S'_{seq}$ to form final input

$$X = S'_{seq} + \text{Parameterize}(E_{pos}),$$

(2)

where $E_{pos} \in \mathbb{R}^{(150+1) \times D}$ with the same shape of $S'_{seq}$ represents the parameterized embedding sequence.

D. Scoring network

The frequency-time patches go through the network architecture as shown in Fig. 4. Augmented by Multi-heading (see the red box in Fig. 4) the attention layer is applied to delineate the contextual relevance of FT patches in the Multi-layer perceptron (MLP) module, which includes the patch-wise connection (PaC) layer and attention layer (Fig. 6). The basic encoder blocks of the \{Layer Norm(1), Multi-head Attention, Layer Norm(2), PaC layers\} with two residual connections are stacked to construct the network architecture (Fig. 4). In the PaC layers, information propagation is restricted to the same patch. The purpose is to render the network to calculate the relevance of patches by self-attention only. We give details on the three important manipulations/layers herein.

Attention mechanism: The attention mechanism associates the individual patches and maps the relevance to the ground truth stage $y$ with three components: the query ($Q$), key ($K$), and value ($V$) matrices, which are the matrix of linear projections produced by the input $X$.

$$Q, K, V = \text{Linear}(X), X \in \mathbb{R}^{(L+1) \times D}$$

(3)

Here, the matrix $Q$ (in the case of using mini-batch) represents the query that comprises of a query sequence with basic units (FT patches here). In the case of self-attention, the $K$ is the same as $Q$, and the attention is utilized to calculated relevance among the patches. The resultant relevance values are further used to calculate $V$. Ultimately, the weighted value matrix that encompasses the importance of FT patches extracted via a global reference will be fed into a further encoder or final staging module. The mathematical processing of $Q, K, V$ can be summarized as below:

$$W_A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

(4)

$$\text{AttentionScore} = W_A \cdot V$$

(5)

where layer $\sqrt{d}$ denotes a normalization-like scale that is applied to each $Q \cdot K$ computation, and the softmax layer is used to functionally obtain the weight vector of attention $W_A$ for $V$.

Multi-head attention: Similar to the way that a CNN increases the number of filters to enrich the expressiveness of the feature space, the attention mechanism can be extended to multi-head attention to prevent losing the manifold expression of the features. At the beginning of each building block, $h$ (the number of the heads) set of $Q$ and $K$ is generated and mapped by the linear projection. Then, the self-attention implements $h$ times in parallel to calculate relevance representations, where each operation is called a “head.” Eventually, a linear layer projects their concatenated outputs and summarizes the attention result. The multi-head attention is defined as follows:

$$Z_{\text{Multihead}}(Q, K, V) = \text{Concat}(head_1, head_2, \ldots, head_h)W^\alpha,$$

(6)

where $W^\alpha \in \mathbb{R}^{h \cdot D \times (150+1)}$ is a weight matrix. It is used for head-wise attention, while a linear projection is applied after the output of the multi-head attention for each round.

MLP&Staging module: As introduced above, the output of the multi-head layer is fed to the PaC layers with GELU being their activation function. Simultaneously, a residual connection skip-connects to the output of the building block to avoid the gradient vanishing. As the information passes through all stacked blocks, the class patch $S_{\text{cls}}$ has absorbed information about the relevance of FT patches extracted from the global context and is used solely in the scoring step. As shown in Fig. 6, a linear projection finally compresses the flattened class token to neurons that have the same
number of sleep stages.

\[ y' = \text{Linear}(\text{LayerNorm}(S'_{\text{C1a}})), \]  

(7)

where \( y' \in \{W, N1, N2, N3, REM\} \), and \( S'_{\text{C1a}} \) is also normalized before the final classification layer.

IV. ATTENTION VISUALIZATION

The multi-head attention relies heavily on the multiplication operation in the attention calculation, and the relevance scores of the resultant attention matrices might play different roles in the network. Unlike the conventional gradient-based visualization [60], we use an attention-oriented visualization similar to the works of Chefer et al. [61] to highlight the FT patches that the model is attending to by inferring both the gradient and the relevance from the final classification decision for each attention layer. Hence, the output of the visualization is ideally reconfigured as a spectrogram-like attention graph (\( \hat{V} \)). That is, the size of the attention graph is maintained with the 1-channel processed spectrogram \( S' \) and is defined as:

\[ \hat{V} = \hat{A}^{(1)}, \hat{A}^{(2)}, \ldots, \hat{A}^{(B)}, \]  

(8)

where \( \hat{V} \in \mathbb{R}^{N \times T} \) consists of a set of sub-graphs of \( B \) encoder modules. Since each row of \( \hat{A}^{(b)} \) in Eq. (4) is normalized to the attention coefficients of each embedded patch with respect to the others, \( \hat{A}^{(b)} \) can be treated as an attention map. Each sub-graph \( \hat{A}^{(b)} \) in encoder \( b \) has a gradient of the attention map \( \nabla W_{\hat{A}}^{(b)} \) and its relevance diffusion \( R(n_b) \), which can be formulated in:

\[ \hat{A}^{(b)} = I + \text{Mean}_b(\nabla W_{\hat{A}}^{(b)} \odot R(n_b)), \]  

(9)

where \( \odot \) is the Hadamard product. With the multi-head mechanism, a mean operation is applied across the \( h \) dimension. In addition, the identity matrix \( I \) is used to avoid the self-inhibition of each patch [61].

The process of relevance propagation starts from the class \( y \) of the staging module and iteratively diffuses to the each layer \( L(n) \), where \( n \in (1, \cdots N) \) is the layer index for the whole Transformer, here, the staging module, i.e., a linear projection in Eq. 11 is defined as \( L^{(1)} \). Suppose \( L(n)(X^{(n)}, Y^{(n)}) \) describes the layer \( n \) function to the corresponding input \( X^{(n)} \) and the weights \( W^{(n)} \), the relevance propagation is similar to the chain rule and follows the generic Deep Taylor Decomposition [62]:

\[ R_j^{(n)} = \sum_i X_j^{(n)} \frac{\partial L_j^{(n)}(X^{(n)}, W^{(n)})}{\partial X_j} \frac{R_i^{(n-1)}}{L_i^{(n)}(X^{(n)}, W^{(n))}}, \]  

(10)

where the subscript \( j \) denotes the elements in \( R^{(n)} \). Because \( L^{(1)} \) corresponds to the first layer of the network, the index \( i \) represents the elements in \( R^{(n-1)} \). Moreover, this relevance propagation will stop at the first layer of the block for each round \( b \).

To reveal the decision process of the model for different channels, each sample generates two attention graphs. Unit resolution of the attention graph corresponds to 1 FT patch and manifests the intensity it is attended to throughout the pipeline.

Meanwhile, an entropy-based statistical analysis is utilized for each two attention graphs to quantify the causality between the attention visualization and the model decision. Considering the transient attribute of stage-dependent features, the attention intensities in one frequency band distributed homogeneously might contain more stage-dependent information. Otherwise, the band within continuous lower intensity or highlighting should lead to a lower sample entropy value.

V. EXPERIMENT

A. Training Strategy

Since a healthy dataset (mentioned in Section II) was extracted, we conducted a pre-training-to-fine-tuning strategy in the training process. Specifically, during the pre-training phase, the healthy-set, which contained balanced samples for each of the stages, was used to optimize the model parameters. Here, we utilized the Adam optimizer with a biggish learning rate of \( 10^{-3} \) to spur the model to converge fast and adjust the parameters along the broadly right direction. In the fine-tuning phase, the remaining training data was used to further optimize the pre-trained parameters. Moreover, the AdamW optimizer [63] is used. The learning rate was set to \( 10^{-4} \) to meticulously optimize the cross-entropy loss function.

To alleviate the overevaluation of the performance, we implemented a subject-wise 7-fold cross-validation by splitting the data into seven subject-wise subsets. In one trial, six subsets were used in the training step, while the remaining subset (roughly 800 subjects) was used for testing.

B. Parameter settings

The details of the parameter settings are shown in TABLE III. To make the utmost of the model, a grid search of hyperparameters was implemented in this work to seek the best combination. Note that the optimal settings (the bold values in TABLE III) were used both for pre-training and the subsequent training. Additionally, a dropout layer was added after each linear projection and attention layer to further avoid the overfitting problem. The model was implemented with the Pytorch v1.4 framework, and all experiments were conducted on a server with an NVIDIA GeForce RTX 2080Ti GPU.

C. Baseline Networks

As introduced above, there are special treatments in the feature generation and network structure in this research. The validate their efficacy, seven baseline models with changes in the following three aspects are also constructed for comparison.

Data processing: (a) The argument for using the spectrogram was supported by repeating the process of feature generation that appeared in the preceding works of [22], [45], where the pipeline tried to learn how to generate features with raw EEG signals (see the [Inception + proposed model] pipeline in TABLE IV). By using the multi-scale Inception architecture [64], the pipeline is expected to find more latent features in parallel [64]. Specifically, one convolutional layer containing five parallel filters corresponding to the five frequency bands of sleep EEG was constructed. Therefore, the output of this layer can be regarded as the multi-scale components of the original EEG signal. Details on this framework can be seen in Appendix. (b) To evaluate the frequency-time patching of the spectrogram, we prepared input that consisted of 1-sec time patches without

| Parameter                        | Value                     |
|----------------------------------|---------------------------|
| #Stacked encoder                | [6, 8, 12]                |
| #Heads (h)                      | [2, 4, 8, 12]             |
| Dimension of linear projection of \( D \) | [16, 32, 64] |
| Normalization-like scale (\( \sqrt{d} \)) | [2, 4]                 |
| Dimension of MLP output         | {64, 128, 256}           |
| Dropout rate                     | {0.2, 0.5, 0.8}          |
| #Training epoch                  | 200                       |
| Batch size                       | 32                        |
| #Parameters                     | \( 1.3 \times 10^5 \)     |

| TABLE III GRID search of the model parameter setting in experiments and the optimal combination is in bold. |
frequency patching [58]. After, a sequence of 30 time patches of 32 dimensions was embedded to a higher dimension of 64 and fed into the network with the same settings in TABLE III (see the Time patching+proposed model) pipeline.

**Sequential model**: The argument over the advantage of using attention is validated by replacing the attention mechanism with RNN networks. Four RNN models using LSTM and Bi-LSTM were used as the replacement (see the four pipelines using LSTM or Bi-LSTM in TABLE IV that cover both the raw EEG and time-patch input).

**Learning strategy**: Although the model was pre-trained within a balanced dataset, the imbalance issue existed in fine-tuning training. Since previous work on deep learning has tried to overcome the imbalance problem with weighted loss function, we implemented a class-wise weighted cross-entropy loss function [65] in another baseline model while all the other aspects are the same (see the FT patching+proposed model + weighted learning) pipeline in TABLE V.

For each fold, we counted the proportion of each class (cls) from the total sample \( T_{cls} \). To provide more weight to the class that had fewer samples, the weight vector \( P_w \) was derived from the inverse of the class proportion. This step is defined as below.

\[
P_w = \frac{\sum_{cls=1}^{N} T_{cls}}{T_{cls}}
\]

To normalize the sample distributions of different folds to the same distribution, \( P_w \) was scaled by its maximum element. Afterwards, the resulting vector \( W_{wce} \) as the parameter was transferred to the following loss function.

\[
W_{wce} = P_w / \text{Max}(P_w)
\]

Since the maximum element divided by itself was one, \( W_{wce} \) in this implementation was registered to a range of \([0, 1]\). The final expression of the loss is described as:

\[
\mathcal{L} = -W_{wce} \sum_{cls=1}^{N} y \log(y'),
\]

where \( y \) was the ground truth of the input samples, while \( y' \) was the predicted label that resulted in the network referring to Eq. 7.

**D. Evaluation metrics**

To evaluate the staging performance of each class, three metrics were used in the experiments, i.e., the stage-specific precision (Pre), recall (Rec), and F1-score (F1).

Precision is the proportion of positive prediction that was actually correct, while recall is the proportion of actual positives that were successfully predicted. F1-score reflects the overall metrics based on these two. Moreover, we adopted overall accuracy to evaluate the training effusiveness and Cohen’s Kappa coefficient (\( k \)) to measure the inter-rater reliability (IRR) [66]. The formula for these metrics can be found in Appendix.

**E. Result**

**Observation 1**: The proposed model had a small generalization gap. Fig. 7(a) depicts the average training loss and the 7-fold validation loss in the fine-tuning step. Since the cost gap between the training and the validation can be viewed as a loose measure of the generalization gap, the small sample-wise gap (0.06 on average) and the narrow variation among folds suggests the model could be extended to new dataset without a plunge in performance. As the training loss converged gradually, the validation loss showed a similar trend without obvious fluctuation in the latter training epochs, which implies that the over-fitting of the trained model was not severe in the proposed model.

Fig. 7(b) shows the trend in the accuracy of the 7-fold validations. The accuracy reached a plateau in a fewer epochs than the decrease of loss and was steady (0.84 – 0.86) in the following training epochs.

**Observation 2**: The proposed model is better in generating the stage-dependent features. Looking closer into the performance of the baseline models in TABLE V, two conclusions can be drawn. The first one is that the models based on the attention mechanism, including the Time patching+Attention and the Inception+Attention pipelines, outperformed the RNN based ones. Given that the best of the RNN models comes from FT Patching+Bi-LSTM, which is different from the proposed pipeline in terms of the model architecture only, it is reasonable to conclude that the proposed model is a better architecture in generating stage-dependent features.

**Observation 3**: The proposed frequency-time patching is an ideal representation of sleep stages. The conclusion is drawn by comparing the performance of the model with different inputs, namely the time patching and FT patching of the spectrogram and the raw EEG signal input to the Inception module. Among the three pipelines (weighted learning excluded here), the retention of the frequency-band information, that is the Inception+proposed model and the FT patching+proposed pipelines, showed its effectiveness in a comparison of the overall performances with the time patching+proposed pipelines. Given that the design of the Inception module serves the same purpose of retaining the resolution in the frequency domain, the combination of spectrogram and FT patching is more appropriate than the data-driven approach for feature generation. This situation may be caused by the nature of the high randomness in EEG signals and can be mitigated by transformation to the frequency domain and the following integral process each 1 Hz.

**Observation 4**: The proposed method reached the new SOTA for most of the stages. We compared the performances of the proposed method with related works that experimented with the same or different database (SHHS) in TABLE VI. We can observe that the classification of the wake stage had the best performance when compared with the other works. Meanwhile, the proposed pipeline
TABLE V
COMPARISON OF PERFORMANCE AMONG BASELINE PIPELINES AND OUR PIPELINES. WE MAKE THE BEST STAGE-WISE PERFORMANCE OF EACH EVALUATION METRIC.

| Comparative Domain | Pipeline | Evaluation Metrics | Wake | N1 | N2 | N3 | REM | Overall |
|--------------------|----------|--------------------|------|----|----|----|------|---------|
| Data Processing    | Time patching + proposed model | Pre | 0.91 | 0.42 | 0.85 | 0.84 | 0.73 | 0.75 |
|                    |          | Re | 0.93 | 0.20 | 0.89 | 0.83 | 0.76 | 0.72 |
|                    |          | F1 | 0.92 | 0.31 | 0.87 | 0.84 | 0.75 | 0.74 |
|                    | Inception + proposed model | Pre | 0.89 | 0.38 | 0.85 | 0.85 | 0.75 | 0.77 |
|                    |          | Re | 0.89 | 0.32 | 0.80 | 0.84 | 0.77 | 0.72 |
|                    |          | F1 | 0.90 | 0.34 | 0.82 | 0.84 | 0.76 | 0.74 |
| Model              | Time patching + LSTM | Pre | 0.85 | 0.31 | 0.80 | 0.83 | 0.65 | 0.69 |
|                    |          | Re | 0.89 | 0.18 | 0.79 | 0.79 | 0.64 | 0.66 |
|                    |          | F1 | 0.87 | 0.22 | 0.80 | 0.82 | 0.64 | 0.68 |
|                    | FT patching + LSTM | Pre | 0.86 | 0.32 | 0.82 | 0.83 | 0.66 | 0.70 |
|                    |          | Re | 0.88 | 0.22 | 0.80 | 0.80 | 0.65 | 0.67 |
|                    |          | F1 | 0.87 | 0.27 | 0.81 | 0.82 | 0.65 | 0.69 |
|                    | Time patching + Bi-LSTM | Pre | 0.84 | 0.31 | 0.79 | 0.83 | 0.67 | 0.68 |
|                    |          | Re | 0.89 | 0.23 | 0.86 | 0.80 | 0.63 | 0.68 |
|                    |          | F1 | 0.87 | 0.26 | 0.83 | 0.80 | 0.65 | 0.68 |
|                    | FT patching + Bi-LSTM | Pre | 0.85 | 0.33 | 0.82 | 0.84 | 0.68 | 0.70 |
|                    |          | Re | 0.89 | 0.24 | 0.87 | 0.81 | 0.63 | 0.69 |
|                    |          | F1 | 0.87 | 0.29 | 0.85 | 0.83 | 0.66 | 0.70 |
| Learning Strategy  | FT patching + proposed model + weighted learning | Pre | 0.90 | 0.40 | 0.81 | 0.82 | 0.76 | 0.74 |
|                    |          | Re | 0.91 | 0.30 | 0.85 | 0.80 | 0.75 | 0.72 |
|                    |          | F1 | 0.90 | 0.35 | 0.83 | 0.81 | 0.75 | 0.73 |
| Proposed Pipeline  | FT patching + proposed model | Pre | 0.93 | 0.42 | 0.87 | 0.89 | 0.80 | 0.78 |
|                    |          | Re | 0.93 | 0.33 | 0.90 | 0.84 | 0.79 | 0.76 |
|                    |          | F1 | 0.93 | 0.38 | 0.89 | 0.87 | 0.80 | 0.77 |

TABLE VI
PERFORMANCE OBTAINED BY PROPOSED PIPELINE AND EXISTING WORKS USING SAME SHHS DATABASE.

| Method | # Experimental SHHS subjects | Wake | N1 | N2 | N3 | REM | k |
|--------|------------------------------|------|----|----|----|-----|---|
| Proposed pipeline | EEG + proposed model | 5736 | 0.93 | 0.42 | 0.87 | 0.89 | 0.80 |
|          |                              | Pre  | 0.93 | 0.33 | 0.90 | 0.84 | 0.79 | 0.80 |
|          |                              | Re   | 0.93 | 0.38 | 0.88 | 0.87 | 0.80 |
| Enrique Fernandez et al., 2021 [67] | EEG&EMG + Separable CNN | 5793 | 0.89 | 0.57 | 0.85 | 0.88 | 0.83 |
|          |                              | Pre  | 0.93 | 0.23 | 0.89 | 0.77 | 0.85 | 0.80 |
|          |                              | Re   | 0.91 | 0.40 | 0.87 | 0.83 | 0.84 |
| Huy Phan et al., 2021 [68] | EMG, EEG, EOG + GRU, LSTM | 5791 | 0.90 | 0.30 | 0.87 | 0.80 | 0.81 |
|          |                              | Pre  | 0.93 | 0.36 | 0.86 | 0.87 | 0.83 | 0.81 |
|          |                              | Re   | 0.86 | 0.33 | 0.87 | 0.82 | 0.81 |
| Emadeldeen Eldele et al., 2021 [69] | EEG + CNN | 329 | 0.90 | 0.30 | 0.87 | 0.80 | 0.81 |
|          |                              | Pre  | 0.92 | 0.50 | 0.84 | 0.67 | 0.89 | 0.79 |
|          |                              | Re   | 0.92 | 0.40 | 0.84 | 0.76 | 0.89 |
| Shreyasi Pathak et al., 2021 [70] | EEG, EOG, EMG + CNN, bi-LSTM | 5793 | 0.92 | 0.42 | 0.85 | 0.85 | 0.87 |
|          |                              | Pre  | 0.88 | 0.47 | 0.90 | 0.86 | 0.86 | 0.80 |
|          |                              | Re   | 0.89 | 0.45 | 0.87 | 0.85 | 0.86 |
| Hogeon Seo et al., 2020 [71] | EEG + RCNN | 5791 | 0.86 | 0.74 | 0.68 | 0.76 | 0.66 |
|          |                              | Pre  | 0.80 | 0.40 | 0.68 | 0.54 | 0.73 |
|          |                              | Re   | 0.82 | 0.78 | 0.57 | 0.78 |
| Niranjan Srídhar et al., 2020 [72] | ECG + CNN | 561 | 0.80 | 0.34 | 0.75 | 0.80 | 0.79 |
|          |                              | Pre  | 0.82 | 0.50 | 0.84 | 0.66 | 0.89 |
|          |                              | Re   | 0.81 | 0.67 | 0.78 | 0.76 | 0.74 |
| Qiao Li et al., 2018 [73] | ECG + CNN | 5793 | 0.80 | 0.74 | 0.54 | 0.73 |
|          |                              | Pre  | 0.82 | 0.67 | 0.56 | 0.65 |
|          |                              | Re   | 0.85 | 0.68 | 0.81 | 0.78 |
| Siddharth Biswal et al., 2018 [18] | EEG, Resp, EMG + RCNN | 15804 | 0.89 | 0.55 | 0.75 | 0.84 | 0.86 |
|          |                              | Pre  | 0.81 | 0.58 | 0.68 | 0.54 | 0.78 |
|          |                              | Re   | 0.85 | 0.56 | 0.71 | 0.66 |
| Foroozan Karimazadeh et al., 2018 [27] | EEG + KNN | 140 | 0.89 | 0.55 | 0.75 | 0.84 | 0.86 |
|          |                              | Pre  | 0.85 | 0.56 | 0.71 | 0.66 |
|          |                              | Re   | 0.85 | 0.56 | 0.71 | 0.66 |

also outperformed the other works in terms of stages N2 and N3, with the F1 score being 0.88 and 0.87, respectively. Pathak’s paper [70], reach the highest performance for the REM stage by fusing the EOG and EMG signals, which are clinically important for REM stage. As we will mention in the Discussion, although the fusing of EMG gets the new best performance of REM using our pipeline, this paper focus on the EEG-only situation. In a similar way, Biswal et al. [18] gets the best performance for N1 by fusing the EEG, waveforms of the chest belt (respiration) and EMG and capturing the intra- and inter-epoch context using the Bidirectional LSTM. We will further discuss the results in in TABLE VI in Discussion.

Observation 5: Attention map of the proposed FT patching is...
to discuss the interpretability of the pipeline proposed in Section IV in Fig. 8, we attempted to visualize the attention scores of two different input sequences, which were FT patches and time patches. For simplicity and clarity, we normalized the intensity of each reconstructed spectrogram-like attention map. Gradient-based visualization (GbV) using Grad-CAM in column (e) in Fig. 8 was also generated for a direct comparison.

**Fig. 8.** Visualization of different pipelines.

**Spectrogram & FT patches:** While most of the bright spots (higher power of FT patches) in the spectrograms can find their correspondences in the attention maps of FT patches, the latter contain more clues. For instance, regarding the wake stage, the bright patch around 2 sec of the spectrogram (C4-A1 channel) of the Alpha band finds its correspondence in the FT patch map (see the stride in the ellipse), and the bright patches of the Alpha band can also be found around 22
sec in C3-A2. The re-organization of the FT patches can also be seen in stage N1, where the last patches of the Theta band were closely attended to (ellipse of C4-A1). The fusion of the two spectrograms can also be seen around 2–5 sec, where C4-A1 is given more attention than the counterpart in C3-A2 (see the two red boxes of N1). When the stage turned to deep sleep (N2, and N3), it can be seen that the patches in the Delta band gradually became patches in the spotlight. For the N2 stage, most of the K-complex shown in both the EEG signal and the spectrogram was closely attended to. Besides, a spindle-alike patch (see the red box in C3-A2 channel) was given high attention as well. Regarding the N3 stage, while the spectrograms were very similar to those of N2, the attention maps of the FT patch features themselves were generally bright patches of the Delta band.

**Attention map of FT patches & GbV:** The GbV suggests different FT patches [bright spots in Fig. 8 (e)] that contributed to the identification of the sleep stages. However, some of the highlighted FT patches were not authentic. The solely bright stride in the Delta band of C4-A1 for the wake stage and the bright strides in the Delta band of the two channels of the N1 stage were not concretely supported by clinical findings. For direct visualization, it is reasonable to conclude that an attention map of FT patches results in a distinct view of the input that is more informative than a network without the attention mechanism.

**Attention map of FT patches & time patches:** The homogeneity of the time patching in the frequency domain drove the attention to

---

**Fig. 9.** Visualization of the attention derived values of each frequency band of the proposed pipeline.

**Fig. 10.** Examples of hypnogram manually scored by human expert (a) and hypnogram automatically scored by our method (b) for one subject from SHHS dataset. Misclassification is marked in red. The sticks in the bottom figure (c) mark the wrong labels. Blue sticks represent the regular sleep stage transitions that can not be detected; while the red sticks represent the falsely detected irregular transitions.
invariably focus on the time domain solely. Referring to the corresponding spectrogram, the bright time patches shown in Fig. 5(d) find themselves unobtrusive against a generally bright background, e.g., the 2-sec time patch of C4-A1 of the wake, or without concrete clinical support, e.g., the bright time patches of the two channels of the N3 stage. Some other samples can be found in Appendix.

Another view of the attention maps of the FT patches can be seen in radar graphs (Fig. 7), which sums up the sample entropy of the attention intensities for the FT patches in the five frequency bands. In stage wake sample, the entropy of Alpha band reaches relatively higher values, that is, 6.21 and 5.51 in C4-A1 and C3-A2. From wake to N1, the dominant Alpha band attenuated in N1 accompanying the increase in the Theta band. As the sleep went deeper, the attention given to the Alpha band turned stable gradually at a relatively low state, while the Delta band came into the foreground. Although the Theta dominates the REM case in appearance, the Beta frequency band indicates certain quantities of information. That result meets the sleeping truth that the brain becomes active again and starts to dream in REM.

VI. DISCUSSION

As introduced above, by proposing a new feature processing framework for EEG signal called FT patching and associating the FT patches through the multiplications of the Query, Key, and Value matrices, the attention scores were generated as alternatives of the conventional features. The proposed model attained the best performance with a lighter network architecture compared with the baseline models (see #parameters in TABLE III and IV).

One of the main purpose of this paper is to push the epoch-wise automatic stage scoring algorithm with EEG signal to a new level. From the comparison with the preceding researches, it is reasonable to conclude that this purpose is fulfilled. Of note, the N1 stage stand in the midway of a dynamic process from conscious to a real sleep stage. According to the definition of N1, comparison with the preceding epoch, i.e., decrease of Alpha band component, is required. However, such kind of information is inevitably lacking in epoch-wise classification. For this reason, in the research that also takes in the inter-epoch relation [71], significant improvement on the Recall can be seen.

Besides, it is well acknowledged that EOG signal is indispensable in identifying the REM stage, and the results of the Pathak et al [70] can be regarded as experimental evidence. By taking the EOG signal as part of the input, the identification of REM reach the best performance. We have tried to include the EOG in our pipeline, and the results get the highest record for the REM stage with 0.91 precision, 0.90 Recall and 0.91 F1 score. The corresponding confusion matrix has shown in Fig. 11(b). In the main body of this paper the results are discussed under the restriction of EEG signal in pursuing a potential extension to home use, where the simplicity on the sensor attachment is advantageous.

In spite of the difficulties mentioned above, the pipeline still get the best performance for the Wake and deep sleep stages with the least information in terms of information sources (EEG only versus sensors fusion) and abundance (epoch-wise versus inter-epoch).

The augmented class patches that were trained in the model absorbs the pair-wise relevance of patches and conclude the patches each class needs to attend to. As mentioned in the Observation 3 and 5 direct connection between the clinical standard and the attention map can be seen in our pipeline. We consider this exposure an important enhancement of the interpretability of an autonomic scoring algorithm and will facilitate clinical/physiological discussion.

By visually comparing the manual annotations of the sleep stages with the autonomic scoring of our pipeline (Fig. 10), it can be seen that the misclassification tends to occur in between {N2, N3} pair and {Wake, REM} pair (The counting results have shown in TABLE VII). Moreover, misclassification often occurs when the sleep stages transition frequently in a relatively short interval (see the red dots in the middle hypnogram of Fig. 11). This situation may be caused by the incompleteness of sleep relevant information of the EEG signal compared with polysomnography used in manual annotation. In contrast, our model can recognize the transitions of stage with relatively low frequency accurately. Furthermore, irregular misclassification pairs that the inter-epoch transitions violate the regular sleep cyclic pattern can be seen, occupying about 18% of the total misclassification. For instance, our pipeline may output assignments of {N1, REM} (N1 \(\rightarrow\) REM or REM \(\rightarrow\) N1), a sharp change of stage that skips N2 and N3 stages. Noteworthy, for the irregular pairs of {N2, REM}, N2 sometimes changes to REM without the deep sleep phase may happen occasionally (around 200 in Fig. 10). However once the body becomes stable at REM stage, this kind of change seldom happen. Given the issue mentioned above, introducing of the constraint on inter-epoch transition is considerable in future works.

VII. CONCLUSION

Developing automatic sleep staging alternatives is an important task to reduce the risk of sleep-related disorders. Although the recent work achieves promising results using deep learning-based methods, the pipeline has not been optimized to adapt to the nature of this task. In our study, we introduced a novel feature processing framework that reflect the nature of the defining characteristics of EEG and then explored the feasibility of attention-based sleep staging model with a large-scale database.

The model evaluation showed that the proposed pipeline gets the best performance for of the wake, N2, and N3. Moreover, our work shows the encouraging potential of the attention-based model in providing evidence that the transient characteristics of sleep stages can be captured in parallel from different 1-sec time positions [74].
ACKNOWLEDGMENT
This research and development work was supported by the Grant-in-Aid for Early-Career Scientists #20K19923.

REFERENCES
[1] A. Krystal, J. Edinger, W. Wohlgemuth, and G. Marsh, “NREM sleep eeg frequency spectrum correlates of sleep complaints in primary insomnia subtypes,” SLEEP, vol. 25, pp. 626–636, 09 2002.
[2] M. Dresler, V. Spooemaker, P. Beiting, M. Czisch, M. Kimura, A. Steiger, and F. Holsober, “Neuroscience-driven discovery and development of sleep therapeutics,” Pharmacology & Therapeutics, vol. 141, no. 3, pp. 300–334, 2014.
[3] A. Bhattacharjee, S. Saha, S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, “Sleep apnea detection based on rician modeling of feature variation in multiband eeg signal,” IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 3, pp. 1066–1074, 2019.
[4] R. Stuckgold, J. Hobson, R. Fosse, and M. Fosse, “Sleep, learning, and dreams: Off-line memory reprocessing,” Science (New York, N.Y.), vol. 294, pp. 1052–7, 12 2001.
[5] H.-V. Ngo, T. Martinetz, J. Born, and M. M¨olle, “Auditory closed-loop stimulation of the sleep slow oscillation enhances memory,” Neuron, vol. 78, 04 2013.
[6] J. Hobson and E. Pace-Schott, “The cognitive neuroscience of sleep: neuronal systems, consciousness and learning,” Nature reviews. Neuroscience, vol. 3, pp. 679–93, 10 2002.
[7] D. Aeschbach, T. Postolache, L. Sher, J. Matthews, M. Jackson, and J. Wehr, “Evidence from the waking electroencephalogram that short sleepers live under higher homeostatic sleep pressure than long sleepers,” Neuroscience, vol. 102, no. 3, pp. 493–502, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304652200005182
[8] L. Marshall, H. Helgadottir, M. Mølle, and J. Born, “Boosting slow oscillations during sleep potentiates memory,” Nature, vol. 444, pp. 610–3, 12 2006.
[9] S. Fogel and C. Smith, “The function of the sleep spindle: A physiological index of intelligence and a mechanism for sleep-dependent memory consolidation,” Neuroscience and biobehavioral reviews, vol. 35, pp. 1154–65, 12 2010.
[10] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, “The AASM Manual for the Scoring of Sleep and Associated Events,” American Academy of Sleep Medicine, vol. 53, no. 9, pp. 1689–1699, 2013.
[11] E. Pace-Schott and J. Hobson, “Pace-schott ef, hobsone ja. the neurobiol-... of polynomography,” Sleep and Biological Rhythms, vol. 15, pp. 39–48, 09 2016.
[12] J. Shambrook, S. Fährbregas, and J. Johnstone, “Validation of an automated wireless system to monitor sleep in healthy adults,” Journal of sleep research, vol. 21, pp. 221–30, 08 2011.
[13] Z. Xu, X. Yang, J. Sun, P. Liu, and W. Qin, “Sleep stage classification using time-frequency spectra from consecutive multi-time points,” Frontiers in neuroscience, vol. 14, 01 2020.
[14] S. Narisi and G. D. Clifford, “Attentive adversarial network for large-scale sleep staging,” in Proceedings of the 5th Machine Learning for Healthcare Conference, ser. Proceedings of Machine Learning Research, vol. 126. Virtual: PMLR, 07–08 Aug 2020, pp. 457–478.
[15] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “Joint classification and prediction cnn framework for automatic sleep stage classification,” IEEE Transactions on Biomedical Engineering, vol. 66, no. 5, pp. 1285–1296, 2019.
[16] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “Segsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, pp. 400 – 410, 01 2019.
[17] N. Banluesombatkul, P. Ouppaphan, P. Leelaaporn, P. Lakhan, B. Chaitusaney, N. Jainchiratayatam, E. Chungsunwanich, W. Chen, H. Phan, N. Phokhandakul, and T. Wilaiprasitpong, “Deepsleepnet: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning(provide datasets),” IEEE Journal of Biomedical and Health Informatics, vol. PP, 11 2020.
[18] P. Jadhav, G. Rajgur, D. Datta, and S. Mukhopahadhyay, “Automatic sleep stage classification using time–frequency images of cwt and transfer learning using convolution neural network,” Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 494–504, 2020.
[19] M. Peker, “A new approach for automatic sleep scoring: Combining taguchi based complex-valued neural network and complex wavelet transform,” Computer Methods and Programs in Biomedicine, vol. 25, no. 2, pp. 362–370, 2018.
[20] M. Peker, “A new approach for automatic sleep scoring: Combining taguchi based complex-valued neural network and complex wavelet transform,” Computer Methods and Programs in Biomedicine, vol. 25, no. 2, pp. 362–370, 2018.
[21] H. Koch, J. Christensen, R. Frandsen, M. Zoetmulder, L. Arvastson, S. Christensen, P. Jadhav, and H. Sorensen, “Automatic sleep classification using a data-driven topic model reveals latent sleep states.” Journal of neuroscience methods, vol. 235, 07 2014.
[22] T. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, “Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification,” in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 1324–1330, main track. [Online]. Available: https://www.ijcai.org/Proceedings/2020/184
[23] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, “Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification,” in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 1324–1330, main track. [Online]. Available: https://www.ijcai.org/Proceedings/2020/184
[24] M. Dikykh, Y. Li, and S. Abdulla, “Eeg sleep stages identification based on weighted undirected complex networks,” Computer Methods and Programs in Biomedicine, vol. 184, p. 105116, 10 2019.
[25] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, and G.-Q. Zhang, “Hyclass: A hybrid classifier for automatic sleep stage scoring,” IEEE Journal of Biomedical and Health Informatics, vol. PP, pp. 1–1, 02 2017.
[26] D. Jiang, Y.-n. Lu, Y. Ma, and Y. Wang, “Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement,” Expert Systems with Applications, vol. 121, 12 2018.
[27] Z. Chen, N. Ono, M. Alatt-Ui-Amin, S. Kanaya, and M. Huang, “Ieva: An improved deep learning structure for eeg signal characterization and...
reconstruction,” in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Seoul, Korea (South): IEEE Computer Society, dec 2020, pp. 1909–1913.

[39] B. Rodríguez Gálvez, R. Thobaben, and M. Skoglund, “The convex information bottleneck lagrangian,” *Entropy*, vol. 22, no. 1, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/1/98

[40] A. Ecker, P. Berens, G. Keliris, M. Bethge, N. Logothetis, and A. Tolias, “Decorrelated neuronal firing in cortical microcircuits,” *Science (New York, N.Y.),* vol. 327, pp. 584–7, 01 2010.

[41] H. Phan, K. B. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. D. Vos, “Sleeptraformer: Automatic sleep staging with interpretability and uncertainty quantification,” *CoRR*, vol. abs/2105.11043, 2021. [Online]. Available: http://arxiv.org/abs/2105.11043

[42] V. Crunelli, M. Lorincz, W. Connelly, F. David, S. Hughes, R. Lambert, N. Leresche, and A. Errington, “Dual function of thalamic low-vigilance state oscillations: Rhythm-regulation and plasticity,” *Nature Reviews Neuroscience*, 01 2018.

[43] A. Kaplan, A. F. Finkelkurs, S. Borisov, and B. Darkhovsky, “Nonstationary nature of the brain activity as revealed by eeg/meg: Methodological, practical and conceptual challenges,” *Signal Processing*, vol. 85, pp. 2190–2112, 11 2005.

[44] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, “Mixed neural network approach for temporal sleep stage classification,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2018.

[45] I. N. S. M., X. Zhu, Y. Chen, and W. Chen, “Sleep stage classification based on eeg, eog, and cnn-gru deep learning model,” in *2019 10th International Conference on Awareness Science and Technology (iCAST)*, 2019, pp. 1–7.

[46] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1925–1941, 2020.

[47] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, “Are convolutional neural networks or transformers more like human vision?” *CoRR*, vol. abs/2105.07197, 2021. [Online]. Available: https://arxiv.org/abs/2105.07197

[48] S. M. Lake, M. Cettolo, and M. Federico, “A comparison of transformer and recurrent neural networks on multilingual neural machine translation,” in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 641–652. [Online]. Available: https://www.aclweb.org/anthology/C18-1054

[49] M. F. Can, M. Karlsson, and L. Frank, “Transient slow gamma synchrony underlies hippocampal memory replay,” *Neuron*, vol. 75, pp. 700–713, 2012.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[51] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[52] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[53] E. Fernandez-Blanco, C. Fernandez-Lozano, A. Pazos, and D. Rivero, “Ensemble of convolution neural networks on heterogeneous signals for sleep stage scoring,” *CoRR*, vol. abs/2107.11045, 2021.

[54] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, “Xsleepnet: Multi-view sequential model for automatic sleep staging,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[55] 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*, vol. 29, pp. 809–818, 2021.

[56] S. Pathak, C. Lu, S. B. Nagaraj, M. van Putten, and C. Seifert, “Stqs: Interpretable multi-modal spatial-temporal-sequential model for automatic sleep scoring,” *Artificial Intelligence in Medicine*, vol. 114, p. 102038, 2021.

[57] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, “Intra- and inter-epoch temporal context network (iinet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg,” *Biomedical Signal Processing and Control*, vol. 61, p. 102037, 2020.

[58] N. Siddhar, P. S. Ali Shoeb, D. B. S. Alaa Kharbour, and J. Burkart, “Deep learning for automated sleep staging using instantaneous heart rate,” *npj Digital Medicine*, no. 1, 2020.

[59] L. C. Li Q, N. S. Shashikumar SP, and C. GD, “Deep learning in medicine,” in *Artificial Intelligence in Medicine*, vol. 13, p. 598, 2020.

[60] M. Langkvist and A. Lotti, “A deep learning approach with an attention mechanism for automatic sleep stage classification,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[61] M. Sundararajan, A. Talay, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML’17. JMLR.org, 2017, p. 3319–3328.

[62] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” *arXiv preprint arXiv:2012.09838*, 2020.

[63] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decompositions,” *Pattern Recognition*, vol. 65, p. 211–222, 05 2017.

[64] I. Loschilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

