Toward Interpretable Sleep Stage Classification Using Cross-Modal Transformers

Jathurshan Pradeepkumar©, Graduate Student Member, IEEE, Mithunjha Anandakumar©, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L. Kappel©, Member, IEEE, Anjula C. De Silva©, and Chamira U. S. Edussooriya©, Member, IEEE

Abstract—Accurate sleep stage classification is significant for sleep health assessment. In recent years, several machine-learning based sleep staging algorithms have been developed, and in particular, deep-learning based algorithms have achieved performance on par with human annotation. Despite improved performance, a limitation of most deep-learning based algorithms is their black-box behavior, which has limited their use in clinical settings. Here, we propose a cross-modal transformer, which is a trained method for sleep stage classification. The proposed cross-modal transformer consists of a cross-modal transformer encoder architecture along with a multi-scale one-dimensional convolutional neural network for automatic representation learning. The performance of our method is on-par with the state-of-the-art methods and eliminates the black-box behavior of deep-learning models by utilizing the interpretability aspect of the attention modules. Furthermore, our method provides considerable reductions in the number of parameters and training time compared to the state-of-the-art methods. Our code is available at https://github.com/Jathurshan00330/Cross-Modal-Transformer. A demo of our work can be found at https://bit.ly/Cross_modaTransformer_demo.

Index Terms—Automatic sleep stage classification, interpretable deep learning, transformers, deep neural networks.

I. INTRODUCTION

A CCURATE sleep stage classification plays a crucial role in sleep medicine and human health. In general, sleep experts use polysomnography (PSG) recordings to diagnose sleep related disorders. PSG mainly comprises of electroencephalogram (EEG) and electrooculogram (EOG). The PSG recordings are typically segmented into 30 s epochs and manually annotated by sleep experts based on guidelines such as Rechtschaffen and Kales (R&K) [1] or American Academy of Sleep Medicine (AASM) [2]. The manual annotation process is tedious, prone to human errors, labour intensive and time consuming. To overcome these drawbacks, multiple studies have proposed automatic sleep stage classification algorithms as the alternative.

Recent works that employ deep learning based algorithms achieved impressive results in sleep stage classification [3] than the conventional machine learning algorithms such as support vector machines (SVMs) [4], k-nearest neighbors (KNNs), decision trees [5] and random forest classifiers [4]. These include convolutional neural networks (CNNs) [6], [7], [8], [9], recurrent neural networks (RNNs) [10], [11], [12], deep belief networks (DBNs) [13], autoencoders [14] and hybrid architectures, such as CNN with RNN [15], [16] and deep neural networks with RNN [17]. Furthermore, different learning techniques such as transfer learning [18], multi-view learning [19], meta learning [20], knowledge distillation [21] and model personalization have improved the performance of automatic sleep staging. With these advances, automatic sleep staging have been able to achieve a performance on par with manual annotation.

Despite the improved performance, a significant limitation lies in their black-box behavior. A general concern when it comes to application of artificial intelligence in healthcare and medicine is the underlying mechanism of deep-learning algorithms [22]. This is a major drawback of existing deep learning based sleep staging algorithms that keeps them from being adopted towards clinical settings. To address this issue, we focus on developing interpretable sleep staging algorithms in our work an interpretable model has the ability to provide explanations for the decisions made concerning specific inputs, thereby offering a key solution to this problem.

Transformers [23] has become de facto in natural language processing (NLP) tasks. Following the seminal work [23],
there have been exceptional works on transformers such as BERT [24], vision transformers (ViTs) [25], which have improved the state-of-the-art in both NLP and computer vision domains. Transformers have introduced interpretability aspects in these domains without compromising high performance. To leverage the capability of attention mechanism in transformers for interpretability, we employ transformer based architectures on physiological signals in our method, specifically for automatic classification of sleep stages to achieve high prediction performance together with interpretability. To the best of our knowledge, SleepTransformer [26] and SleepViTransformer [27] are the only works which has explored transformers for sleep stage classification. Transformer encoders used in SleepTransformer [26] are inspired from the seminal transformer architecture proposed in [23], and on top of the transformer encoder they utilize an additional attention layer to achieve a compact representation for each epoch of the PSG signals. We hypothesized that a better compact representation can be learned with less overhead by employing an additional learnable vector named \textit{[CLS]} vector similar to BERT [24] and ViTs [25]. The vector representation learned by the transformer encoder corresponding to the \textit{[CLS]} vector can be utilized as the compact vector representation for each epoch. This learned vector representation aggregates all the sequence information in a PSG epoch. Another drawback of SleepTransformer [26] is that they only utilize EEG signals for sleep staging, whereas in our method we explore the capability of transformers in learning from multiple signals. The SleepTransformer [26] learns only intra and inter epoch relationships in an EEG signal whereas in our method, we explore cross-modal relationships between EEG and EOG signals along with intra-modal and inter epoch relationships, by modifying our transformer encoder architecture.

To overcome the limitations of past works and to improve interpretability of deep learning based sleep staging algorithms, we propose cross-modal transformers. Cross-modal transformers consist of a cross-modal transformer encoder architecture along with a multi-scale one-dimensional convolutional neural network (1D CNN) for automatic representation learning. The major contributions of our work presented in this paper are summarized below:

- We propose two cross-modal transformers: 1) Epoch cross-modal transformer and 2) Sequence cross-modal transformer to solve the problem of sleep stage classification under two classification schemes which are one-to-one and many-to-many classification.
- Our method employs a two-stage process: initially, it learns representations directly from the raw signals of each modality, which is then followed by our cross-modal transformer architecture. We utilize existing multi-scale 1D-CNN architecture and adapted towards our specific task in the initial stage to learn an optimal feature representation by considering both local and global features.
- A cross-modal transformer encoder architecture to learn both intra-modal temporal attention, i.e., attention between time steps within a feature representation of a modality, and cross-modal attention, i.e., attention between each modalities. The sequence cross-modal transformer consists of an additional block to learn inter-epoch attention, i.e., attention between adjacent epochs.
- A simple yet an effective method based on attention mechanisms to enable interpretation of the model’s predictions. The proposed method is capable of learning and interpreting: 1) intra-modal relationships, 2) cross-modal relationships and 3) inter-epoch relationships.
- The performance of our sequence cross-modal transformers is on-par with the state-of-the-art methods with a comparable reduction in the number of parameters as can be seen in Fig. 1.
- Our method requires less time to train and has a smaller model footprint in terms of parameters compared to the state-of-the-art methods.

II. RELATED WORK

A. Transformers

Since the introduction by Vaswani et al. [23], Transformers and its variants have been successful in NLP and computer vision tasks in terms of performance and powerful representation learning. Currently, Transformers have achieved the state-of-the-art performance in several NLP tasks. Pre-training have been leveraged in Transformers to achieve better performance in NLP tasks, where BERT [24] utilized self-supervised pretraining strategies and fine-tuning towards a supervised downstream task, whereas GPT [28] focused on language modeling in pre-training. Although CNNs have dominated the field of computer vision, there have been several efforts to explore the benefits of Transformers in computer vision domain. Initially, self-attention was introduced along with CNNs for computer vision tasks [29]. Then the focus was on completely replacing convolutions in the architecture. ViTs [25] were able to achieve this feat, by directly applying them on images with very few modifications. Furthermore, self-supervised pre-training [30] was leveraged to improve the performance, which emerged several properties of the ViTs. Likewise, Transformers have been explored in several medical applications [31] such as medical image segmentation, detection and registration, whereas SleepTransformer [26] and SleepViTransformer [27] are the only works in the domain of sleep staging. SleepTransformer was able to achieve a competitive accuracy of 81.4% on sleep-EDF-expanded 2018 without any pre-training.

B. Deep Learning Based Sleep Stage Classification

Recent work that employ deep-learning based algorithms achieved impressive results in sleep stage classification [3] than conventional machine learning algorithms. The conventional machine learning algorithms generally use the handcrafted features as the input [4], [5], [32]. Furthermore, this approach may not generalize to a large population due to the heterogeneity of subjects and recording devices [16].

The existing automatic sleep stage algorithms can be classified into two categories based on the input to the network: raw signals and time-frequency maps. A time-frequency map requires prior knowledge of the dataset as well as signal processing, as it heavily relies on the preprocessing steps. In contrast, deep-learning based algorithms are capable of
performing automatic feature extraction, thus the drawbacks of handcrafted features are eliminated.

Several previous works have focused on utilizing the single channel EEG recordings for sleep stage classification. However, EOG is capable of detecting eye movements, which is a fundamental indicator for differentiating rapid eye movement (REM) and non-rapid eye movement (NREM) stages [33]. Therefore, the information in EOG channels, i.e., multi-modal scheme can be exploited to improve the performance of sleep stage classification [4], [34].

Phan et al. [19], have used both raw signals and time-frequency representations of the signals as the inputs to a sequence-to-sequence network. The overall accuracy achieved was 84% with only single channel EEG, as well as with both single channel EEG and EOG. XSleepNet is the current state-of-the-art in automatic sleep staging. The inclusion of EOG signal is insignificant in terms of the performance of the XSleepNet model. Supratak and Guo [16] have proposed CNN + LSTM to capture time invariant features and transition characteristics among sleep stages from raw single channel EEG. Mousavi et al. [35] have proposed SleepEEGNet, composed of CNN to extract features and sequence-to-sequence model to capture long short-term context dependencies between epochs of raw single channel EEG. The sequence-to-sequence model composed of bidirectional RNN and attention mechanisms. Even though RNN and LSTM based approaches yields a competitive performance with state-of-the-art methods, the computational cost of LSTMs and RNNs should be taken into consideration. RNNs have limitations due to their recurrent nature and high complexity, thus difficult to train them in parallel. Eldele et al. [36] have proposed AttnSleep which consists of a multi-resolution CNN with adaptive feature re-calibration and temporal context encoder utilizing multi-head self attention mechanism to capture the temporal dependencies from single channel EEG. AttnSleep achieved accuracy of 82.9%, which is higher than most of the existing algorithms. SeqSleepNet, proposed by Phan et al. [10], is a sequence-to-sequence hierarchical RNN classification model trained on multi-channel time-frequency maps as input. SeqSleepNet uses attention based bidirectional RNN for short-term sequence modeling and bidirectional RNN for long-term sequence modeling. SeqSleepNet achieved accuracy of 82.6% with single channel EEG and 83.8% with both single channel EEG and EOG channels demonstrating the significance of including EOG to improve the accuracy. TinySleepNet proposed by Supratak et al. [37] is an efficient model for sleep stage classification based on raw single channel EEG, which achieved competitive performance with 83.1% accuracy. TinySleepNet is an improved version of DeepSleepNet model with improved efficiency, i.e., less number of parameters, thus reducing the computational resource requirement.

### III. METHODOLOGY

In this section, our proposed cross-modal transformers are presented. First, we formulate the problem definition for one-to-one and many-to-many sleep stage classification tasks. Then we introduce our cross-modal transformers: the epoch cross-modal transformer for one-to-one classification and the sequence cross-modal transformer for many-to-many classification. We further explain the multi-scale 1D CNN based representation learning and cross-modal transformer encoder under the epoch cross-modal transformer.

#### A. Problem Definition

We address the problem of classifying the sleep stage of a 30 s epoch of PSG signals which were acquired in the experiment described in section IV-A. Our training set with the size of $N$, consists of labeled 30 s PSG epochs $\{x_i, y_i\}_{i=1}^{N}$, where $\{x_i, y_i\} \in X \times Y$. Here $X \in \mathbb{R}^{T \times C}$ denotes the input space of recorded PSG signals, where $T$ represents the time steps in an epoch and $C \in \{\text{EEG (Fpz-Cz), EOG}\}$ represents the modalities in the recorded PSG signals. $Y \in \{\text{WAKE, N1, N2, N3, REM}\}$ represents the output space of sleep stages. Our goal is to learn a function $f_{\theta} : X \rightarrow Y$ by minimizing the error $\mathbb{E}_{(x,y)}(L(f_{\theta}(x), y))$ on the given training dataset. Here, $L$ denotes the loss function.
We solve the aforementioned problem using cross-modal transformers under the one-to-one and many-to-many classification schemes [10] illustrated in Fig. 2. In the one-to-one classification, we consider a single PSG epoch to predict the corresponding sleep stage [10]. In many-to-many classification, we consider a sequence of PSG epochs and predict their corresponding sleep stages at once [10].

B. Epoch Cross-Modal Transformer

In this subsection, we focus on solving sleep stage classification with the one-to-one scheme using a epoch cross-modal transformer. The proposed epoch cross-modal transformer, shown in Fig. 3, consists of two main blocks: 1) multi-scale 1D-CNN for automatic representation learning and 2) a cross-modal transformer encoder architecture to learn both intra-modal temporal attention and cross-modal attention. The epoch transformer initially learns two separate feature representations from the input EEG and EOG signals in an epoch using 1D-CNN. Here, features are learned and extracted using non-overlapping windows from the signals. Then, the cross-modal transformer encoder learns a representation by considering intra-modal temporal attention and cross-modal attention, which is then fed into a linear layer for classification. The two main blocks of the epoch transformer are further elaborated in proceeding subsections. The key advantage of epoch cross-modal transformer is that it is smaller in size and enables faster training and tuning of the hyper-parameters on a resource constraint environment. The same model can easily be scaled into a sequence cross-modal transformers efficiently.

1) Multi-Scale 1D-CNN for Representation Learning:

Inspired by vision transformers using image patches as sequential data, we employ multi-scale 1D-CNN as shown in Fig. 3 to learn the feature representation of non-overlapping windows with the size of 0.5 s from the input 30s epoch of raw PSG signals. Here, the 1D signals $X_c \in \mathbb{R}^{T \times 1}$ will be mapped into a feature space of $X'_c \in \mathbb{R}^{(T/(0.5 \times f_s)) \times E}$, where $c \in C$, $f_s$ is the sampling frequency and $E$ is the embedding size. Let $T/(0.5 \times f_s)$ be $T'$ for convenience. The features are extracted from non-overlapping windows instead of overlapping windows to improve interpretability, such that the extracted feature vectors can be fed into the cross-modal encoder as sequential data to learn the attentions between all windows.

We hypothesize that the global and local features in a window of raw signals will contribute towards the classification of the sleep stages. In order to extract both global and local features in a window, we employed multi-scale 1D-CNN, where the raw input signal goes through three parallel paths as shown in Fig. 3(b): 1) One 1D-CNN with kernel size of 50, 2) Two 1D-CNNs with kernel sizes of 25 and 2 respectively, and 3) Three 1D-CNNs with kernel sizes of 5, 5 and 2. Each 1D convolution layer will be followed by a LeakyReLU
The features extracted at different scales will then be normalized using batch normalization. Finally, the extracted feature representations will be concatenated along the embedding dimension and will undergo an additional 1D-CNN with kernel size of 1, followed by LeakyReLU activation and batch normalization.

2) Cross-Modal Transformer Encoder and Classification:
We propose a cross-modal transformer encoder architecture to learn a powerful feature representation by attending intra-modal temporal information and cross-modal relationships. As illustrated in Fig. 3(a), the cross-modal transformer encoder consists of two main blocks: 1) attention and 2) position wise fully-connected feed-forward network blocks.

Initially, for each modality $C$, a learnable $CLS_c$ vector $\in \mathbb{R}^{(1 \times E)}$, similar to the one proposed in BERT [24], is randomly initiated and concatenated with the output of the multi-scale 1D-CNN block along the time axis. Similar to the seminal work [23], positional encodings are added to the concatenated vector and fed to the intra-modal attention block to learn the relationships between all the time steps in the modality. We extract only the vector representation corresponding to the $CLS_c$ vector from the output from each modality, as it aggregates all the information of the intra-modal temporal information [24]. A new learnable $CLS_{cross}$ vector named $CLS_{cross}$ is randomly initialized and concatenated along with the $CLS_c$ vector representation extracted from each modality. Cross-modal attention block is employed on the aforementioned vector to learn the relationships between the modalities, which get aggregated in the corresponding representation of $CLS_{cross}$. The vector representation corresponding to the $CLS_{cross}$ is extracted and $CLS_c$ representation of each modality is replaced by this vector. Then the concatenated vectors are passed through the feed-forward network. Finally, the vector representations corresponding to $CLS_{cross}$ in each modality are extracted, flattened and passed through a single linear classifier with 5 neurons for classification.

The operations of the intra-modal attention block, cross-modal attention block and feed-forward network are similar to the seminal work by [23]. The attention blocks consist of multi-head attention followed by residual addition and layer normalization. The feed-forward network consists of a feed-forward layer followed by residual addition and layer normalization. An additional linear layer is sufficient for classification of the sleep stages from the extracted representation, which shows the power of the learned representation.

C. Sequence Cross-Modal Transformer
In order to solve sleep stage classification with the many-to-many scheme we employ sequence cross-modal transformer, which is an extension of the epoch cross-modal transformer. The proposed architecture shown in Fig. 4 consists of multiple epoch level block for each 30s epoch of raw PSG signals in a non-overlapping sequence (Fig. 2). The epoch level block is built using multi-scale 1D-CNN block and intra-modal attention block for each modality and a cross-modal attention block. The corresponding vector representation of $CLS_{cross}$ vector of each epoch block is extracted and concatenated. Inter-epoch attention block followed by a feed-forward network is employed on the concatenated vector to learn the relationships between the epochs. Finally, we extract and flatten output representation corresponding to each epoch separately and employ a linear classifier to predict their sleep stages. During inference, we let the model run across the sequence of PSG epochs, where multiple predictions for an epoch will be achieved. Then, we calculate the mean probability to predict the correct sleep stage.
D. Interpretability

In this subsection, we propose a method to interpret the predictions of cross-modal transformer, which have the potential to help sleep experts and enhance their confidence in the system. The main advantage of our architecture is that the attention mechanism can be simply utilized to learn and interpret: 1) intra-modal relationships, 2) cross-modal relationships and 3) inter-epoch relationships.

We extract the outputs of the intra-modal attention block and calculate the scaled dot-product attention between the representation corresponding to the \(CLS\) token and other representations corresponding to the non-overlapping windows of the raw signal \((Z_i)\), as given in

\[
A_i = \text{softmax}\left(\frac{CLS \times Z_i}{\sqrt{d_E}}\right),
\]

to interpret intra-modal relationships. Here, \(A_i\) represents the attention score for the \(i\)th time step in a 30 s PSG epoch, where \(i = 1, 2, \ldots, 60\). Calculating the scaled dot product allows for the identification of time steps in a PSG epoch that are more aligned with the \(CLS\) token and have a greater impact on the final classification. This gives attention weights for each non-overlapping window of the raw signal, which interpret their impact on the prediction.

In a similar procedure, the relationships between modalities is interpreted by the scaled dot-product attention of the representations in the output of the cross-modal attention block corresponding to \(CLS\) of each modality. Finally, we interpret the inter-epoch relationships by calculating the scaled dot-product attention between \(CLS_{\text{cross}}\) of each epoch.

IV. EXPERIMENTS

A. Dataset

1) SleepEDF Expanded Dataset: We used the publicly available sleep-EDF-expanded dataset (Sleep-EDF-78) [41] from Physionet [42] to evaluate the proposed architecture for sleep stage classification. We used the sleep cassette (SC) dataset of sleep-EDF 2018 for our experiments, which consists of 153 whole night PSG recordings from 78 healthy individuals. Each recording comprises of two bipolar EEG channels (Fpz-Cz and Pz-Oz), a horizontal EOG signal and a submental chin EMG signal. For our experiments, we only utilized Fpz-Cz EEG channel and EOG signal. The PSG recordings were annotated with hypnograms annotated by sleep technicians based on Rechtschaffen and Kales (R&K) guidelines [1]. In the hypnogram, each 30 s epoch of the recorded data is assigned to one of the following labels: Stage 0, Stage 1, Stage 2, Stage 3, Stage 4, REM, movement time and '?' (unscored). For this study, we converted these annotations to the AASM standards [2], by combining Stage 3 and Stage 4 to a single stage N3 while Stage 0, Stage 1 and Stage 2 are relabelled as Wake, N1 and N2, respectively. Also, epochs consisting of annotations 'movement time' and '?' were discarded. Altogether 415, 465 epochs of 30s duration of PSG were extracted. We employed a five-fold cross validation method, where the dataset was divided into subject-independent groups to ensure a fair evaluation of the model performance (Subjects in each group are independent).

2) SHHS Dataset: We also utilized the publicly available Sleep Heart Health Study (SHHS) dataset [43], [44] in our research. SHHS is a multicenter, longitudinal study examining the effects of sleep-disordered breathing on cardiovascular diseases. This dataset includes two sets of recordings: SHHS1 (visit 1) and SHHS2 (visit 2), where we used the SHHS1 (visit 1) for our study. Each recording contains two EEG channels (C4-A1 and C3-A2), and we specifically used the C4-A1 EEG channel along with EOG in our experiments similar to past works. SHHS was initially annotated according to R&K guidelines. We converted these annotations to AASM using the same approach applied to sleepEDF. The epochs containing MOVEMENT and UNKNOWN were discarded, and recordings without all sleep stages were also discarded. Following [26], we randomly split the subjects in the dataset into 70% for training and 30% for testing, with 100 subjects from the training set separated as a validation set.

The data processing for sleepEDF and the training setups are detailed in the supplementary materials (Section A and B).

V. RESULTS AND DISCUSSION

A. Sleep Stage Classification

We evaluated the performance of our proposed epoch and sequence cross-modal transformers on the SleepEDF-expanded 2018 and SHHS datasets, comparing it against various state-of-the-art methods. The performance metrics used are accuracy (ACC), Cohen’s kappa (\(\kappa\)), macro-F1 score (MF1), sensitivity (Sens), specificity (Spec), and macro-averaged G-mean (MGM). Further details on the performance metrics are provided in the supplementary materials (Section C). The results of the methods for both classification schemes (one-to-one and many-to-many classification) are presented in Tables I and II. Additionally, we evaluate the class-wise performance, with results provided in Tables III and IV.

In the SleepEDF dataset, our proposed sequence cross-modal transformer with 15 epochs in a sequence achieves comparable performance with the state-of-the-art methods for accuracy and Cohen’s Kappa coefficient (\(\kappa\)). However, our method outperforms past works in macro-averaged F1 (79.4%), sensitivity (80.2%), specificity (95.9%), and macro-averaged G-mean (87.7%). Similarly, the same sequence cross-modal transformer outperforms past works in sensitivity (83.1%) and macro-averaged G-mean (89.6%) in the SHHS dataset. For other metrics, our performance is slightly lower than past works. A key factor to note is that our sequence cross-modal transformer achieves high classification performance with fewer parameters compared to previous methods. While SeqSleepNet [10] has fewer parameters, it relies on RNNs, which require longer training and inference times.

Compared to XSleepNets [19], our sequence cross-modal transformer achieves better performance with fewer PSG epochs in the sequence (5, 10, and 15) in the SleepEDF dataset, though it performs slightly lower in the SHHS dataset. Despite this, our method offers better interpretability and smaller model print compared to XSleepNets. CoRe-Sleep [40] achieves better performance compared to ours, however the method does not focus on interpretability. Furthermore,
TABLE I
PERFORMANCE COMPARISON BETWEEN CROSS-MODAL TRANSFORMER AND PREVIOUS WORKS ON SLEEP-EDF-EXPANDED 2018 DATASET

| Method                  | Channels | Epochs in a Sequence | Number of Parameters | Overall Performance |
|-------------------------|----------|----------------------|----------------------|---------------------|
| SleepEEGNet [35]        | Fpz-Cz   | 10                   | ~ 2.6M               | ACC 80.0, k 0.730, F1 73.6, Sens. 71.2, Spec. 82.5 |
| DeepSleepNet [38]       | Fpz-Cz   | 25                   | ~ 24.7M              | ACC 77.1, k 0.69, F1 71.2, Sens. 72.8, Spec. 85.6 |
| MultitaskCNN [36]       | Fpz-Cz   | -                    | -                    | ACC 79.6, k 0.72, F1 72.8, Sens. 83.7, Spec. 82.5 |
| AttnSleep [36]          | Fpz-Cz   | 3                    | -                    | ACC 82.9, k 0.77, F1 78.1, Sens. 85.6, Spec. 85.6 |
| SleepTransformer [26]   | Fpz-Cz   | 21                   | ~ 3.7M               | ACC 81.4, k 0.743, F1 74.3, Sens. 74.5, Spec. 85.0 |
| SeqSleepNet [38]        | Fpz-Cz   | 20                   | ~ 0.2M               | ACC 82.6, k 0.76, F1 76.4, Sens. 76.3, Spec. 85.32 |
| FCNN+RNN [19]           | Fpz-Cz   | 20                   | ~ 5.6M               | ACC 82.8, k 0.761, F1 76.6, Sens. 73.9, Spec. 85.0 |
| XSleepNet2 [19]         | Fpz-Cz   | 20                   | ~ 5.8M               | ACC 84.0, k 0.778, F1 77.9, Sens. 77.5, Spec. 86.12 |
| XSleepNet1 [19]         | Fpz-Cz   | 20                   | ~ 5.8M               | ACC 83.6, k 0.773, F1 77.8, Sens. 77.7, Spec. 86.23 |
| NaveFusion [19]         | Fpz-Cz   | 20                   | ~ 5.8M               | ACC 82.3, k 0.755, F1 76.2, Sens. 75.7, Spec. 84.94 |
| TinySleepNet [38]       | Fpz-Cz   | 15                   | ~ 1.3M               | ACC 83.1, k 0.77, F1 78.1, Sens. 85.6, Spec. 85.6 |
| SleepPyCo without CL [39]| Fpz-Cz   | 10                   | ~ 3.7M               | ACC 83.5, k 0.772, F1 77.7, Sens. 85.6, Spec. 85.6 |
| SleepPyCo* [39]         | Fpz-Cz   | 10                   | ~ 3.7M               | ACC 84.6, k 0.787, F1 79.0, Sens. 85.6, Spec. 85.6 |
| XSleepNet2 [19]         | Fpz-Cz & EOG | 20                  | -                    | ACC 84.0, k 0.778, F1 78.7, Sens. 77.6, Spec. 86.18 |
| XSleepNet1 [19]         | Fpz-Cz & EOG | 20                  | -                    | ACC 84.0, k 0.777, F1 78.4, Sens. 77.1, Spec. 85.6, Spec. 85.6 |
| SeqSleepNet [19]        | Fpz-Cz & EOG | 20                  | -                    | ACC 83.8, k 0.776, F1 78.2, Sens. 77.4, Spec. 85.6, Spec. 85.6 |
| FCNN+RNN [19]           | Fpz-Cz & EOG | 20                  | -                    | ACC 82.7, k 0.759, F1 76.9, Sens. 75.5, Spec. 84.82 |
| NaveFusion [19]         | Fpz-Cz & EOG | 20                  | -                    | ACC 82.5, k 0.737, F1 76.9, Sens. 75.8, Spec. 84.99 |
| SleepViTransformer* [27]| Fpz-Cz & EOG | 4                    | -                    | ACC 85.0, k 0.792, F1 79.1, Sens. 85.6, Spec. 85.6 |

* refers to the methods that utilize self-supervised pertaining before supervised training for classification. CL indicates contrastive learning.

TABLE II
PERFORMANCE COMPARISON BETWEEN CROSS-MODAL TRANSFORMER AND PREVIOUS WORKS ON SHHS DATASET

| Method                  | Channels | Epochs in a Sequence | Number of Parameters | Overall Performance |
|-------------------------|----------|----------------------|----------------------|---------------------|
| MultitaskCNN [36]       | C4-A1    | -                    | -                    | ACC 81.4, k 0.74, F1 71.2, Sens. 83.6, Spec. 84.0 |
| AttnSleep [36]          | C4-A1    | 3                    | ~ 3.7M               | ACC 87.7, k 0.828, F1 80.1, Sens. 87.15, Spec. 87.15 |
| SleepTransformer [26]   | C4-A1    | 21                   | ~ 3.7M               | ACC 86.5, k 0.811, F1 78.5, Sens. 86.9, Spec. 85.97 |
| SeqSleepNet [19]        | C4-A1    | 20                   | ~ 0.2M               | ACC 86.7, k 0.813, F1 79.5, Sens. 86.2, Spec. 86.86 |
| FCNN+RNN [19]           | C4-A1    | 20                   | ~ 5.6M               | ACC 87.6, k 0.826, F1 80.7, Sens. 90.1, Spec. 87.6 |
| XSleepNet2 [19]         | C4-A1    | 20                   | ~ 5.8M               | ACC 87.5, k 0.826, F1 81.0, Sens. 86.5, Spec. 88.1 |
| XSleepNet1 [19]         | C4-A1    | 20                   | ~ 5.8M               | ACC 87.5, k 0.826, F1 81.0, Sens. 86.5, Spec. 88.1 |
| NaveFusion [19]         | C4-A1    | 20                   | ~ 5.8M               | ACC 87.5, k 0.826, F1 81.0, Sens. 86.5, Spec. 88.1 |
| SleepPyCo* [39]         | C4-A1    | 10                   | ~ 2.37M              | ACC 87.9, k 0.830, F1 80.7, Sens. 88.1, Spec. 88.1 |
| CoRe-Sleep [40]         | C4-A1    | -                    | -                    | ACC 88.2, k 0.834, F1 80.8, Sens. 88.4, Spec. 88.4 |
| XSleepNet2 [19]         | C4-A1 & EOG | 20                  | -                    | ACC 88.8, k 0.843, F1 81.8, Sens. 90.8, Spec. 88.4 |
| XSleepNet1 [19]         | C4-A1 & EOG | 20                  | -                    | ACC 88.8, k 0.843, F1 82.0, Sens. 91.3, Spec. 88.7 |
| SeqSleepNet [19]        | C4-A1 & EOG | 20                  | -                    | ACC 88.4, k 0.837, F1 80.7, Sens. 96.7, Spec. 87.7 |
| FCNN+RNN [19]           | C4-A1 & EOG | 20                  | -                    | ACC 88.0, k 0.831, F1 80.5, Sens. 92.2, Spec. 86.6 |
| NaveFusion [19]         | C4-A1 & EOG | 20                  | -                    | ACC 88.4, k 0.839, F1 81.7, Sens. 96.8, Spec. 88.9 |
| SleepViTransformer* [27]| C4-A1 & EOG | 4                    | -                    | ACC 88.1, k 0.83, F1 83.8, Sens. 86.1, Spec. 86.1 |
| CoRe-Sleep [40]         | C4-A1 & EOG | -                    | -                    | ACC 89.5, k 0.853, F1 82.3, Sens. 90.3, Spec. 89.7 |

* refers to the methods that utilize self-supervised pertaining before supervised training for classification.

we compared the performance with existing transformer-based methods SleepTransformer [26] and SleepViTransformer [27].

Without any pretraining on a larger database, our sequence cross-modal transformer outperforms SleepTransformer on
When considering class-wise performance based on F1-score in Table III and IV, our sequence cross-modal transformer achieves the state-of-the-art performance in the prediction of Wake and N1 in SleepEDF dataset and N1 in SHHS dataset without using external data or pretraining. Our method also outperforms SleePyCo [39] without their contrastive learning strategy and achieves a better macro-F1 score compared to SleePyCo with contrastive learning. We strongly believe that deep learning techniques such as transfer learning, meta learning [20], knowledge distillation [21], large-scale training and self-supervised pretraining could further improve the performance of our method. In our future studies, we plan to explore self-supervised pretraining strategies to further improve the performance of interpretable deep learning models for sleep staging.

When considering class-wise performance based on F1-score in Table III and IV, our sequence cross-modal transformer achieves the state-of-the-art performance in the prediction of Wake and N1 in SleepEDF dataset and N1 in SHHS dataset. This enhancement in the performance can be attributed to the capability of our method to learn cross-modal relationships, where EOG along with EEG makes an impact on their predictions. The performance of our sequence cross-modal transformer in predicting N2, N3 and REM stages are on-par with the previous work.

B. Comparison of Computational Complexity

Table V shows the comparison of model size in number of parameters and training time taken for 1000 steps between the variants of the proposed cross-modal transformers and the previously reported work. Here, our method trains faster compared to current the state-of-the-art XSleepNets because their architecture is based on RNNs which requires to process the data sequentially. Unlike RNNs, transformers are capable of processing data in parallel. The parallelism in the transformers enabled our proposed method to train faster compared to the others.
Fig. 5. The variants of the epoch level model used to conduct the ablation study. (a) shows the architecture with single channel EEG as the input, (b) shows the extended version of (a) with both EEG and EOG as inputs, and (c) shows a version of the epoch cross-modal transformer without cross-modal attention.

C. Ablation Study

Here, we systematically study the importance and advantages of using EOG, cross-modal attention and CLS vectors as aggregated representation in our method. Additionally, we study the impact of different representation learning methods on the performance and the results are provided in the supplementary materials (Section D).

1) Importance of EOG and Cross-Modal Attention: In manual annotation of sleep stages, EOG signals plays a major role in identifying Wake, N1 and REM sleep stages. This shows the importance of developing an algorithm to utilize the features of EOG signals along with EEG for sleep staging. We show the importance of efficiently utilizing EOG and learning the cross-relationship between EEG and EOG modalities by conducting a separate study. In this study, the epoch cross-modal transformer with the model variants shown in Fig. 5 and their corresponding sequence model variants are used. The results of the study are given in Table VI. The proposed model architecture significantly contributes to the performance, as it can be inferred from the performance increase of 1.8% achieved in accuracy by a only EEG variant of sequence cross-modal transformer compared to the SleepTransformer [26]. Also, our single channel EEG model outperforms majority of the existing EEG-only models and it is on-par with the XSleepnets [19].

As anticipated, the model using both EEG and EOG signals had better performance than only EEG, which clearly states that our method is learning from both EEG and EOG signals efficiently. Significant improvement in the prediction of N1 and REM sleep stages can be observed when EOG is added into our method. In order to identify the effective representation learning frameworks two strategies were analysed: (a) learning joint features from concatenated EEG and EOG channels using a single multi-scale 1D-CNN block (EEG + EOG (Jt)) and (b) learn separate features for EEG and EOG using separate multi-scale 1D-CNN blocks (EEG + EOG). As per the results in Table VI, having two separated multi-scale 1D-CNN blocks achieves better performance. Performance was further improved when the cross-modal attention was incorporated into our method. Inclusion of cross-modal attention significantly improved the prediction of Wake, N2, N3 and REM stages.

2) Importance of Using CLS: In our method, utilizing CLS vectors as aggregated representation enabled the reduction of the model size. The size of our initial version of epoch cross-modal transformer without having CLS was ~ 2.1M, which is 6.5 times larger than our proposed epoch cross-modal transformer. Addition of CLS vector enabled scaling of the proposed methods in terms of number of modalities and number of epochs in a sequence. Also, the cross-relationships between any number of modalities can be learned easily.
The Results of the Study on the Importance of EOG and Cross-Modal Relationships for Sleep Staging on Sleep-EDF-Expanded 2018 Dataset

| Method                        | ACC  | \(\kappa\) | Per-class Performance |
|-------------------------------|------|------------|-----------------------|
|                               |      |            | W | N1 | N2 | N3 | RIM |
| Epoch Cross-Modal Transformer |      |            |   |    |    |    |     |
| Only EEG                      | 78.3 | 0.704      | 91.4 | 37.7 | 81.6 | 75.3 | 69.3 |
| EEG + EOG (H)                | 80.0 | 0.727      | 92.1 | 45.5 | 81.9 | 74.6 | 76.7 |
| EEG + BOG                     | 80.4 | 0.731      | 92.4 | 44.1 | 82.4 | 74.5 | 78.8 |
| EEG + EOG + CMA (Proposed)   | 80.8 | 0.736      | 92.3 | 45.3 | 82.9 | 75.5 | 77.4 |

Sequence Cross-Modal Transformer - 15

| Method                        | ACC  | \(\kappa\) | Per-class Performance |
|-------------------------------|------|------------|-----------------------|
|                               |      |            | W | N1 | N2 | N3 | RIM |
| Only EEG                      | 83.2 | 0.770      | 92.9 | 51.6 | 84.8 | 75.4 | 83.5 |
| EEG + EOG (H)                | 84.1 | 0.785      | 93.2 | 55.0 | 85.3 | 77.0 | 86.4 |
| EEG + BOG                     | 84.2 | 0.785      | 93.6 | 53.9 | 85.3 | 76.9 | 86.6 |
| EEG + EOG + CMA (Proposed)   | 84.3 | 0.785      | 93.5 | 54.3 | 85.5 | 77.0 | 86.7 |

*CMA refers to cross-modal attention and H refers to joint features from multi-scale 1D-CNN for combined EEG and EOG as inputs. The model architecture variants for only EEG, EEG + EOG (H) and EEG + BOG are illustrated in Fig. 5(a), Fig. 5(b) and Fig. 5(c), respectively.

because we leverage a simple CLS latent vector for each modality which aggregates all the sequence information. Most importantly, the CLS vectors enabled interpretations of the predictions by employing self-attention, which is simple and effective.

3) Comparison Between Epoch and Sequence Cross-Modal Transformers: It can be clearly observed that the sequence cross-modal transformer significantly outperforms epoch cross-modal transformer, because of the inclusion of inter-epoch attention block and predicting sleep stages of multiple epochs simultaneously. Table VI clearly indicates the effectiveness of sequential modeling in sleep staging, as it consistently improves the performance over epoch models under various settings. Due to smaller model print, epoch cross-modal transformer is suitable for resource constrained environments, where the algorithm can be implemented on edge devices.

4) Ablation on Representation Learning: The importance of analyzing the representation learning component in our method is important for understanding the construction of classification models in this domain. To gain insights into this aspect, we conducted an ablation study comparing the performance of 1D-CNN, short-time Fourier transform (STFT), and multi-scale 1D-CNNs on the SHHS dataset using an epoch cross-modal transformer. The STFT was used directly with our model without an additional linear input layer. To further analyze the effect of cross-modal attention used in our method, we conducted an additional experiment where cross-modal attention was removed from the final setup with the multi-scale 1D-CNN. The results of these experiments are presented in Table VII. Further analysis of the hyperparameters in our method is provided in the supplementary materials.

These results clearly indicate that the representations learned by multi-scale CNNs outperform those learned by STFT and 1D-CNN in terms of performance metrics. Furthermore, there is a noticeable drop in performance when cross-modal attention is removed from the model. The improvement of performance achieved by cross-modal attention in both datasets confirms its effectiveness in sleep stage classification tasks.

Fig. 6. Interpretation results by our sequence cross-modal transformer-15 for a sequence of PSG recordings of subject/recording SC4381 captured between 38820-39270s. Each row in the figure corresponds to a 30s PSG epoch in the sequence. The column A indicates the corresponding sleep stage of the epoch, while the subsequent columns represent B: inter-epoch attentions, C: cross-modal attentions, D: intra-modal attentions in EEG, and E: intra-modal attentions in EOG, respectively. In inter-epoch attention plots, red bar indicates the normalized self-attention of the current epoch. Additionally, we have highlighted some interesting signal patterns identified by our method using light blue dashed circles with numbers.
D. Interpreting the Results

Along with competitive sleep stage classification performance, the major contribution of cross-modal transformers are their interpretability. As mentioned in section III-D, we leverage the attention mechanisms to interpret the results as intra-modal, cross-modal and inter-epoch relationships illustrated in Fig. 6. Here, we interpret the predictions of the sequence cross-modal transformer for a sequence of PSG epochs. Specifically, Fig. 6 depicts the interpretation results for the sequence cross-modal transformer-15, which is designed to predict 15 epochs in a sequence.

The interpretation results in Fig. 6 focus on a sequence of PSG epochs that include transitions from REM, Wake, N1, and N2 stages. The inter-epoch attention graphs in Fig. 6B highlight some significant relationships between these epochs. To illustrate, it is evident that the initial four epochs, corresponding to the REM stage, exhibit a notably stronger interconnection compared to the other epochs. This is depicted by the higher inter-epoch attention scores assigned to these first four bars in the bar-plot for the initial four epochs. Similar observation can be seen for N1 stages in from epoch 6 to 13. Moreover, the prediction of the wake stage is related to its preceding epochs (REM) and following epoch (N1) as shown in the bar-plot of epoch 5. Cross-modal attention graphs in Fig. 6C interprets the impact of each modality on the prediction. Finally, the intra-modal attention maps filters, which time segments in the PSG signal were given more importance by the model for the predictions. In this context, we present a color plot superimposed onto the signal plot, depicting the intra-epoch attention scores associated with various signal regions. The darker regions highlighted within the signals correspond to higher intra-epoch attention scores, illustrating their importance in the final prediction. Some of such patterns are marked in Fig. 6 using light blue dashed line circles with a number. Circle I in Fig. 6E highlights the eye movements which are related to REM sleep. Also, circles II and III in Fig. 6D highlights the K-complex patterns in the N2 sleep stages. By offering interpretability, our method enables clinicians to further validate the predictions, thereby enhancing reliability.

E. Limitations

Despite the potential benefits of transformer-based algorithms for sleep staging, they have several limitations. One major limitation is their high data requirements, often needing large amounts of labeled data for training, which can be challenging with smaller or imbalanced datasets. Strategies like supervised or self-supervised pre-training [24], [45] can help mitigate this issue. Our method combines transformers with CNNs, reducing data requirements. Another limitation is the computational complexity of transformer-based algorithms, which can demand significant resources for training and inference.

VI. Conclusion

In this paper, we present an interpretable, transformer-based deep learning method named cross-modal transformers, for automatic sleep stage classification. Our sequence cross-modal transformer achieved performance on par with the state-of-the-art, with reduction in model print and faster training. In addition to accurate sleep staging, our major contribution is to eliminate the black-box behavior of deep learning by leveraging attention mechanisms to interpret the results. We believe that developing interpretable deep learning methods is the most feasible way forward to use artificial intelligence for clinical applications. In future work, the proposed method can be further improved by employing training strategies in deep learning such as transfer learning and self-supervised pre-training.

ACKNOWLEDGMENT

The authors thank Dr. Ranga Rodrigo and the National Research Council of Sri Lanka for providing computational resources.

REFERENCES

[1] A. Rechtschaffen, A Manual for Standardized Terminology, Techniques and Scoring System for Sleep Stages in Human Subjects. Brain Information Service, Brain Research Institute of Health, Education, and Welfare, Gurgaon, India, 1968.
[2] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, The AASM Manual for Scoring Sleep Associated Events: Rules, Terminology Technical Specification, 1st ed., Westchester, IL, USA: American Academy of Sleep Medicine, 2007.
[3] H. W. Loh et al., “Automated detection of sleep stages using deep learning techniques: A systematic review of the last decade (2010–2020),” Appl. Sci., vol. 10, no. 24, p. 8963, Dec. 2020.
[4] M. M. Rahman, M. I. H. Bhiuian, and A. R. Hassan, “Sleep stage classification using single-channel EEG,” Comput. Biol. Med., vol. 102, pp. 211–220, Nov. 2018.
[5] A. R. Hassan and M. I. H. Bhiuian, “Automatic sleep scoring using statistical features in the EMD domain and ensemble methods,” Biocybern. Biomed. Biotechnol., vol. 36, no. 1, pp. 248–255, 2016.
[6] T. Zhu, W. Luo, and F. Yu, “Convolution- and attention-based neural network for automated sleep stage classification,” Int. J. Environ. Res. Public Health, vol. 17, no. 11, p. 4152, Jun. 2020.
[7] X. Zhang et al., “Automated multi-model deep neural network for sleep stage scoring with unfiltered clinical data,” Sleep Breathing, vol. 24, no. 2, pp. 581–590, Jun. 2020.
[8] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “Joint classification and prediction CNN framework for automatic sleep stage classification,” IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 1285–1296, May 2019.
[9] O. Yildirim, U. B. Bagolu, and U. R. Acharya, “A deep learning model for automated sleep stages classification using PSG signals,” Int. J. Environ. Res. Public Health, vol. 16, no. 4, p. 599, Feb. 2019.
[10] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, “SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, Mar. 2019.
[11] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. D. Vos, “Automatic sleep stage classification using single-channel EEG: Learning sequential features with attention-based recurrent neural networks,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2018, pp. 1452–1455.
[12] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, “A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning,” IEEE J. Biomed. Health Informatics, vol. 24, no. 5, pp. 1351–1366, May 2020.
[13] M. Längkvist, L. Karlsson, and A. Loutfi, “Sleep stage classification using unsupervised feature learning,” Adv. Artif. Neural Syst., vol. 2012, pp. 1–9, Jul. 2012.
[14] O. Tsinalis, P. M. Matthews, and Y. Guo, “Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders,” Ann. Biomed. Eng., vol. 44, no. 5, pp. 1587–1597, Oct. 2015.
J. B. Stephansen et al., “Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy,” *Nature Commun.*, vol. 9, no. 1, pp. 1–15, Dec. 2018.

A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A model for automatic sleep staging based on raw single-channel EEG,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, “Mixed neural network approach for temporal sleep stage classification,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.

H. Phan et al., “Towards more accurate automatic sleep staging via deep transfer learning,” *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 1787–1798, Jun. 2021.

H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, “XSleepNet: Multi-view sequential model for automatic sleep staging,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2022.

N. Banluesombatkul et al., “MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 1949–1963, Jun. 2021.

C. Yoo, H. W. Lee, and J.-W. Kang, “Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1273–1284, Mar. 2022.

O. Lahav, N. Mastronarde, and M. van der Schaar, “What is interpretable? Using machine learning to design interpretable decision-support systems,” 2018, arXiv:1811.10799.

A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

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.

L. Peng, Y. Ren, Z. Luan, X. Chen, X. Yang, and W. Tu, “SleepVi-Transformer: Patch-based sleep spectrogram transformer for automatic sleep staging,” *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105203.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” San Francisco, CA, USA, 2018.

Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, “OCNet: Object context network for scene parsing,” 2018, arXiv:1809.00916.

M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

F. Shamshad et al., “Transformers in medical imaging: A survey,” *Med. Image Anal.*, vol. 88, 2023, Art. no. 102802.

A. R. Hassan and M. I. H. Bhuian, “Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating,” *Biomed. Signal Process. Control*, vol. 24, pp. 1–10, Feb. 2016.

C. Sun, C. Chen, J. Fan, W. Li, Y. Zhang, and W. Chen, “A hierarchical sequential neural network with feature fusion for sleep staging based on EEG and RR signals,” *J. Neural Eng.*, vol. 16, no. 6, Oct. 2019, Art. no. 066020.

K. B. Mikkelsen, H. Phan, M. L. Rank, M. C. Hemmsen, M. de Vos, and P. Kidmose, “Light-weight sleep monitoring: Electrode distance matters more than placement for automatic scoring,” 2021, arXiv:2104.04567.

S. Mousavi, F. Aghaf, and U. R. Acharaya, “SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach,” *PloS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

E. Eldele et al., “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

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.

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.

S. Lee, Y. Yu, S. Back, H. Seo, and K. Lee, “SleePyCo: Automatic sleep scoring with feature pyramid and contrastive learning,” *Expert Syst. Appl.*, vol. 240, Apr. 2024, Art. no. 122551.

K. Kontras, C. Chatzichristos, H. Phan, J. Suykens, and M. De Vos, “CoRe-sleep: A multimodal fusion framework for time series robust to imperfect modalities,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 840–849, 2024.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Obery, “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.

A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, Jul. 2000.

G.-Q. Zhang et al., “The national sleep research resource: Towards a sleep data commons,” *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018.

S. F. Quan et al., “The sleep heart health study: Design, rationale, and methods,” *SLEEP*, vol. 20, no. 12, pp. 1077–1085, 1997.

T. Chen, S. Korntith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” 2020, arXiv:2002.05709.