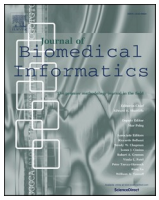
Journal of Biomedical Informatics 157 (2024) 104689

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

A novel deep learning model based on transformer and cross modality attention for classification of sleep stages

Sahar Hassanzadeh Mostafaei [a], Jafar Tanha [a,*], Amir Sharafkhaneh [b]

[a] Faculty of Electrical and Computer Engineering, University of Tabriz, P.O. Box 51666-16471, Tabriz, Iran
[b] Professor of Medicine, Section of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, Baylor College of Medicine, Houston, TX, USA

ARTICLE INFO

Keywords:
Sleep stage classification
Multimodal learning
Transformer encoder decoder
Cross-modality attention
Physiological channels

ABSTRACT

The classification of sleep stages is crucial for gaining insights into an individual's sleep patterns and identifying potential health issues. Employing several important physiological channels in different views, each providing a distinct perspective on sleep patterns, can have a great impact on the efficiency of the classification models. In the context of neural networks and deep learning models, transformers are very effective, especially when dealing with time series data, and have shown remarkable compatibility with sequential data analysis as physiological channels. On the other hand, cross-modality attention by integrating information from multiple views of the data enables to capture relationships among different modalities, allowing models to selectively focus on relevant information from each modality. In this paper, we introduce a novel deep-learning model based on transformer encoder-decoder and cross-modal attention for sleep stage classification. The proposed model processes information from various physiological channels with different modalities using the Sleep Heart Health Study Dataset (SHHS) data and leverages transformer encoders for feature extraction and cross-modal attention for effective integration to feed into the transformer decoder. The combination of these elements increased the accuracy of the model up to 91.33% in classifying five classes of sleep stages. Empirical evaluations demonstrated the model's superior performance compared to standalone approaches and other state-of-the-art techniques, showcasing the potential of combining transformer and cross-modal attention for improved sleep stage classification.

1. Introduction

Sleep medicine is a specialized field in health care dedicated to diagnosing and treating sleep-related disorders. It encompasses various medical disciplines and addresses disorders such as sleep apnea, insomnia, and circadian rhythm disorders. The importance of sleep medicine in promoting general health and well-being, diagnosing and treating sleep disorders, improving sleep quality, and preventing health risks related to sleep disorders. This field is critical to optimizing health outcomes and improving the quality of life for individuals with sleep-related challenges. A sleep study, or a polysomnography, is a medical examination that involves monitoring and recording various physiological parameters during sleep to evaluate sleep patterns and diagnose sleep disorders [1]. It typically takes place in a sleep center or laboratory, where individuals spend a night or more while being monitored by specialized equipment. The important collected data during a sleep study includes brain activity, eye movements, muscle activity, heart rate, respiratory effort, and oxygen saturation. Sleep studies help healthcare professionals assess the quality and structure of sleep, identify abnormalities, and treat sleep disorders.

A sleep disorder refers to a condition characterized by disruptions in the normal progression of sleep states during sleep. The sleep cycle typically consists of distinct stages, including wakefulness, non-rapid eye movement (NREM) and rapid eye movement (REM) sleep, each serving specific functions in the overall sleep process. Sleep cycle disorders involve abnormalities in the timing, duration, or transitions between these stages. Sleep staging, also known as sleep classification, refers to the process of categorizing different stages of sleep based on the analysis of physiological channels recorded during sleep. The sleep staging system divides sleep into several distinct stages, each characterized by specific patterns of brain activity, eye movements, muscle tone, and other physiological parameters. According to the American Academy of Sleep Medicine (AASM), the five main categories of sleep stages are Wake, N1, N2, N3, and REM [2].

* Corresponding author at: Faculty of Electrical and Computer Engineering, University of Tabriz, P.O. Box 51666-16471, Tabriz, Iran.
E-mail addresses: S.h.mostafaei@tabrizu.ac.ir (S.H. Mostafaei), Tanha@tabrizu.ac.ir (J. Tanha), Amirs@bcm.edu (A. Sharafkhaneh).

https://doi.org/10.1016/j.jbi.2024.104689
Received 29 February 2024; Received in revised form 13 June 2024; Accepted 10 July 2024
Available online 18 July 2024

Recently, deep learning methods have shown significant success by providing accurate and efficient models for sleep stage classification [3–6]. Advanced deep learning models in sleep staging include the use of neural network architectures, particularly recurrent neural networks (RNN), convolutional neural networks (CNN), and more recently, attention-based and transformer-based models [7–11].

The attention mechanism in the framework of deep learning models helps to focus on different parts of the input sequence and assign different degrees of importance to each element [12]. This allows the model to pay attention to features or parts of physiological data collected during sleep studies, increasing its ability to capture patterns and related dependencies in the sleep data. There are different types of attention mechanisms, each of which can handle specific requirements [13]. By considering the nature of physiological signals in classifying sleep stages, these techniques highlight the importance of different channels and related features in sleep data by capturing temporal or long-range dependencies [6,8,14,15].

The transformer architecture is based on the self-attention mechanisms present in the encoder and decoder, which are used in the context of sequence-to-sequence tasks such as natural language processing [16]. In the context of language translation, the goal is to produce a sentence in the target language corresponding to a given sentence in the source language. However, in different tasks, different types of input sequences are used [17]. A study introduced an audiovisual transformer model designed to classify audio events using video and audio [18]. In this model, video and audio sequences act as different input modalities for the encoder and decoder, respectively. This approach shows versatility and can be extended to handle other types of sequential inputs, including physiological signal processing. The Efficiency of transformers in capturing dependencies within sequences has made it an effective architecture in deep learning, which is also used to classify sleep stages [19,20]. Since several essential physiological channels in the task of sleep staging provide different insights into sleep data, it can be effective to use them simultaneously for multi-modal learning [21–23].

In the field of multi-modal learning, cross-modality attention enables the model to selectively focus on information from different types of data during training [24–27]. Unlike simple combinations such as feature concatenation of different modalities, these methods help to identify and use the strengths of each modality according to the relevant features and lead to a stronger and more informative presentation. Cross-attention mechanisms in sleep stage classification can help the model to understand the complex relationships between different types of physiological channels and provide an efficient way to process and combine features on sleep channels, reducing the risk of information loss or redundancy [28,29].

According to the mentioned information, to benefit from the advantages of multimodal learning, we propose a novel deep learning model based on transformer encoder-decoder architecture and cross-modal attention to classify sleep stages. The proposed deep learning model uses various important physiological channels in the two different modalities including raw signal data and handcraft features. In the proposed model we have several main steps:

1. Preprocessing, standardizing, and augmenting sleep data samples across various channels to generate two distinct input modalities: one comprising raw signal data and the other comprising handcrafted features.
2. Developing CNN-Attention blocks tailored to the characteristics of individual physiological channels within each input modality.
3. Employing transformer encoders specifically designed for each input modality to encode the extracted features.
4. Integrating high-level features through cross-modal attention mechanisms.
5. Utilizing the transformer decoder to generate output sequences based on the basic and advanced features for predicting five classes of sleep stages.

Following the model development, a comprehensive evaluation is conducted using various metrics, including Accuracy, Macro F1 Score (MF1), Sensitivity, Specificity, Cohen's kappa (κ), Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). Comparative analyses and statistical tests are performed between our proposed model and other state-of-the-art techniques. The structure of the remainder of this paper is outlined as follows: Section II describes the related works, and Section III provides detailed insights into the materials and methods utilized in this study, encompassing discussions on the dataset, channels, methodologies, and the novel introduced model. In Section IV, the outcomes from the proposed model are presented, accompanied by extensive comparisons with established state-of-the-art techniques. To encapsulate our study, Sections V and VI provide discussion and conclusions.

2. Related works

This section provides an overview of related works in the domain of sleep staging tasks leveraging deep learning techniques. The methodologies discussed encompass state-of-the-art approaches, including transformer architectures, attention mechanisms, and multi-modal learning models. These methods are tailored to tackle the classification of sleep stages into five distinct classes, reflecting advancements in the field.

The problems with automated sleep staging were the main emphasis of this work [30], especially the difficulties of correctly identifying stage N1 because of its low resource status in sleep staging. The objective of the suggested approach is to attain expert-level performance in the categorization of N1 stages and the overall accuracy of sleep staging. The process entails creating a neural network model that combines a classifier with two branches with an attention-based convolutional neural network. To achieve a compromise between contextual reference and universal feature learning, a transitive training technique is used. After a thorough evaluation on a large-scale dataset, the model is assessed on seven datasets spanning five cohorts.

FullSleepNet is a unique multi-task learning technique for employing fully convolutional neural networks (CNNs) to identify arousals and sleep phases. It was first published in this study [31]. After processing single-channel EEG data for the whole night, FullSleepNet creates segmentation masks for labels indicating the arousal and sleep stages. It consists of four modules: an attention mechanism to concentrate on pertinent portions of the input, a recurrent module to capture long-range relationships, a convolutional module for extracting local features, and a segmentation module for generating final predictions.

CTCNet is a new neural network architecture intended for sleep stage categorization that is first presented in this research [32]. A multi-scale convolutional neural network (MSCNN) for extracting low- and high-frequency features, adaptive channel feature recalibration (ACFR) to improve sensitivity to significant channel features, and a multi-scale dilated convolutional block (MSDCB) to capture various characteristics among feature channels are some of the essential components that CTCNet integrates to overcome prior limitations. Furthermore, global temporal context information is extracted using a transformer encoder, and spatial location correlations between EEG variables are captured and refined via a capsule network module.

This study [33] presented SleepEGAN is an ensemble deep learning model with generative adversarial network (GAN) enhancements designed for the categorization of unbalanced sleep stages. A unique GAN architecture called EGAN was designed specifically for supplementing EEG signal properties in order to reduce class imbalance. The training procedure incorporates generated samples for minority classes. Furthermore, a free ensemble learning approach is developed to improve prediction accuracy and resilience by mitigating the variation in model estimate brought about by heterogeneity across test and validation sets.

A sequence-by-sequence sleep staging model called

SleepTransformer is presented in this work [34] with the goal of improving interpretability and measuring uncertainty. SleepTransformer removes convolutional and repetitive components in favor of transformer encoder design, which is different from typical approaches. Interpretability is offered by SleepTransformer through the use of self-attention scores at the sequence and period levels. Heat maps, which emphasize significant aspects in the EEG data, are used to show attention ratings at the course level. Similar to human expert scoring, attention ratings at the sequence level represent the impact of nearby epochs on the identification of a target epoch.

L-SeqSleepNet is a new deep learning model that was introduced in this paper [35]. Its purpose is to enhance sleep staging performance by effectively capturing whole-cycle sleep information. In contrast to traditional sequential modeling techniques, L-SeqSleepNet uses a long sequence modeling method that is effective. In order to recover the original sequence size, it splits lengthy sequences into many subsequences, performs intra- and inter-subsequence sequential modeling, and then unfolds the subsequences. This method makes use of the structural information found in sleep cycles to improve staging speed while successfully addressing the shortcomings of current models in managing lengthy sequences.

Because single-channel EEG follows sleep staging criteria, it is frequently used in sleep staging research. Nevertheless, current deep learning techniques frequently ignore the possible advantages of integrating long-term context, such as sleep stage transition rules, which might improve staging performance, in favor of only considering short-term temporal context information. By creating SleepContextNet, a temporal context network intended to capture long-term context between EEG sleep phases, this study [36] attempted to overcome this problem. SleepContextNet integrates long- and short-term context information in chronological order by using a mix of Recurrent Neural Network (RNN) layers and Convolutional Neural Network (CNN) layers to learn representative features from each sleep stage. Furthermore, a unique approach for data augmentation is shown to maintain long-term context information in EEG data without altering the sample count.

AttnSleep is a deep learning architecture based on attention that is proposed in this research [37] and is intended to categorize sleep phases based on single-channel EEG inputs. The two primary components of AttnSleep are the temporal context encoder (TCE), which uses a multi-head attention mechanism with causal convolutions, and the feature extraction module, which is based on multi-resolution convolutional neural networks (MRCNN) and adaptive feature recalibration (AFR). While AFR improves feature quality by modeling interdependencies between features, MRCNN collects both low-frequency and high-frequency features from various frequency bands. Using multi-head self-attention combined with causal convolutions, the TCE effectively captures temporal relationships in the derived features. Moreover, a class-aware loss function is intended to handle problems with data imbalance without introducing additional computing complexity.

While graph neural networks have demonstrated potential in classifying sleep stages, there are still issues with efficiently using epoch information from adjacent EEG channels, obtaining representative features from transitional data, and improving classification accuracy. A multi-layer graph attention network (MGANet) is designed in order to overcome these restrictions [38]. With the use of graph attention convolution and GRU algorithms, MGANet integrates node-level attention to concentrate on channel interactions in the time–frequency and spatial domains. A multi-head spatial–temporal mechanism dynamically modifies characteristics and balances channel weights, while various layers of a graph attention network precisely record spatial sleep data. By applying stage-level attention to confusing sleep phases, graph convolutional networks' drawbacks are addressed.

3. Materials and methods

3.1. Dataset

In this paper, we utilize a sleep dataset sourced from the Sleep Heart Health Study (SHHS), encompassing multiple physiological channels recorded in the European Data Format (EDF) from 6441 participants [39,40]. Conducted by the National Heart, Lung, and Blood Institute, the SHHS is a multicenter cohort dataset designed to investigate cardiovascular, respiratory, and sleep disorders. Each file in the dataset is segmented into 30-second intervals called an epoch and labeled according to Rechtschaffen and Kales (R&K) rules, categorizing the data into six stages: Wakefulness, S1, S2, S3, S4, and REM [41].

For our analysis, we randomly select 120 participants from the initial part of the dataset (SHHS1) and specifically focus on key channels, including electroencephalogram (EEG, EEG sec), electrooculogram (EOG/R, EOG/L), electrocardiogram (ECG), electromyogram (EMG), abdominal and thoracic efforts, and airflow. The data from these channels are recorded at different sampling rates. Specifically, EEGs, EMG, and ECG channels are sampled at 125 Hz, while EOG channels are sampled at 50 Hz. Additionally, the respiratory channels have a sampling rate of 10 Hz. We then transform the sleep stage labels into five classes according to AASM rules by combining S3 and S4 into N3, resulting in the classification of five distinct sleep stages: Wake, N1, N2, N3, and REM [2].

3.2. Preprocessing

In this section to prepare input data for deep learning models, we follow a series of steps. Initially, we allocated 80 % of the selected participants to the training dataset, reserving the remaining 20 % for the test dataset. Subsequently, to augment the training dataset and enhance its diversity, we employ various augmentation methods. These techniques encompass jittering, window slicing, and magnitude warping which help generate diverse samples of the training dataset. we also assign 10 % of the augmented training data to validation dataset [42]. Table 1 presents the total epochs and the number of epochs in each sleep class of the original and augmented data.

Within the context of this paper, our input data consists of two modalities: signal data input and their handcrafted features input. The next step involves deriving handcrafted features from the signal data inputs. This process requires extracting various types of features from each selected channel, which includes EEG, EEG sec, EOG/L, EOG/R, ECG, EMG, abdominal and thoracic efforts, and airflow. Feature extraction methods are designed to include common features and advanced features based on the characteristics of channels, which help to comprehensively represent the input data. Table 2 shows the details of feature extraction methods related to each channel.

Before feature extraction, it's crucial to conduct signal analysis to unveil meaningful insights. Signal processing offers a plethora of methods to delve into signal characteristics and extract valuable information. Among these techniques, discrete wavelet transform (DWT) is used, which can decompose the input signal into multiple sub-bands [43]. This decomposition process yields both time and frequency-domain information, facilitating a comprehensive analysis of the signal's characteristics. DWT boasts advantages such as multiresolution analysis, allowing for detailed examination of signal details at different

Table 1
Total number and number of epochs in each sleep stage.

Participants	Sleep class					Total
	Wake	N1	N2	N3	REM	
120	13,332	4704	46,416	15,258	17,490	97,200
Augmented	6666	2352	23,258	7629	8745	48,650
						145,850

Table 2
Feature extraction methods.

Feature extraction						Channel
Basic features						EEG, EEG sec, EOG/L, EOG/R,
Statistical	Entropy				Time	ECG, EMG, Abdomen, Thoracic, and Airflow
Frequency features						EEG and EEG sec
Alpha	Beta	Theta	Delta	Gamma		
R peak features						ECG
R-Peaks Envelope (RPE)			R-R Interval (RRI)			

scales, making it a versatile tool for various signal-processing tasks.

Following this step, basic features are derived from all chosen physiological channels, encompassing statistical, entropy, and time features [44]. Statistical features shed light on the distributional properties of the signal, while entropy measures its complexity, and time domain features describe temporal characteristics. The extracted basic features include mean, standard deviation, skewness, kurtosis, variance, and median. Entropy features involve Shannon entropy, spectral entropy, and approximate entropy, quantifying different aspects of the signal's characteristics. Time features encompass zero crossing rate and root mean square.

Specifically, from EEG channels, we extract frequency features related to five bands: alpha, beta, theta, delta, and gamma [45]. These frequency bands capture crucial information about the neural activity present in EEG signals. Analyzing the power or amplitude distribution within each frequency band provides insights into various aspects of brain activity associated with different sleep stages. In the ECG channel, we capture R peak features, involving the identification and extraction of peaks corresponding to the R wave. These features include R-R Interval (RRI) and R Peak Envelope (RPE) [46,47]. The R-R interval represents the time duration between successive R peaks in the ECG signal, while the R peak envelope reflects the amplitude or magnitude of the R peaks over time, computed through smoothing or envelope detection techniques. The integration of signal data and their associated handcrafted features forms a comprehensive framework for input, facilitating the classification of various sleep data stages.

Additionally, we introduced another input modality derived from raw signal data using the Short-Time Fourier Transform (STFT) method [48]. STFT is a signal processing technique utilized to analyze non-stationary signals across time, offering a time–frequency representation. This enables the detection of transient events, characterization of periodic components, and identification of changes in frequency components over time. Indeed, in addition to handcrafted features, we extracted STFT features to serve as another input modality alongside the raw signal data.

The primary objective behind collecting these diverse features is twofold: firstly, to unveil the valuable information encapsulated within them, and secondly, to analyze the efficacy of the proposed model with various types of input modalities.

3.3. Cnn-attention

Convolutional neural networks (CNNs) are effective deep learning tools applicable to time series data, including physiological channels [49]. These networks use sliding filters that act as kernels to scan the input data and identify distinct features and complex patterns. On the other hand, attention mechanisms strategically focus on diverse parts of the input sequence and assign different degrees of importance to each part [12]. This focused attention facilitates the extraction of critical information from the data. Integrating Convolutional Neural Networks (CNN) with attention mechanisms is a powerful approach to capturing hierarchical features and focusing on important regions of the input sequence. In this paper, we develop different CNN-Attention blocks to extract effective patterns of each physiological channel according to the input modality. The selected channels encompass EEG, EEG sec, EOG/L, EOG/R, ECG, EMG, abdominal and thoracic efforts, and airflow, with various sampling rates of 125, 50, and 10 Hz, in two modalities: signal data and handcrafted features. CNN layers are designed based on input modality and channel characteristics with residual connections. Residual connections, inspired by ResNet architectures, help address the vanishing gradient problem and facilitate the training of deep learning networks [50]. Attention mechanisms, on the other hand, allow the model to selectively attend to relevant parts of the input, enhancing its discriminative capabilities. We employed an attention mechanism known as additive attention [51]. This mechanism computes the context vector of attention scores within the sleep sequence and is proper for dealing with high-dimensional raw signal data according to the paper "Attention is all you need" [16]. Fig. 1 shows the structure of the additive attention mechanism used in this paper.

A combination of these components can increase the model's robustness in capturing local and global features, along with selective attention, and help the model generalize well across different inputs. Fig. 2 demonstrates the developed CNN-Attention blocks for each input modality in more detail. As shown in this figure, (a), (b), and (c) correspond to the first input modality, while (d) corresponds to the second input modality. Since different physiological channels have different sampling rates, we designed CNN-Attention blocks based on their natural inherent. For EEG, ECG, and EMG channels from the first input modality with a sampling rate of 125 Hz, these blocks have eight CNN layers with tuned kernel sizes, followed by max pooling and dropout layers with multiple residual connections. These residual connections combine low-level and high-level features that maintain local and global patterns. For subsequent channels with sampling rates of 50 and 10 Hz, these blocks have six layers. In the second input modality, for each channel, we designed five layers of CNN-Attention blocks in the same way. We incorporate Gaussian noise, global averaging, and additive attention in our approach. Additive attention takes features as input and produces their attention scores as output sequences. Finally, we merge the features extracted from the last CNN layer with the corresponding attention scores to generate a sequence of informative features. This process continues for all channels of each input modality, especially the handcrafted features with differences in the size of kernel and stride. This integration helps to achieve complex features in various channels from two different modalities.

3.4. Transformer architecture

The transformer, introduced in the paper "Attention is All You Need," is a neural network architecture designed for sequence-to-sequence tasks [16]. Its architecture relies on the self-attention mechanism, enabling it to assign varying importance to different segments of the input sequence during prediction. This model consists of an encoder-decoder structure. The encoder processes the input sequence, while the decoder produces the output sequence. Each component of this model is composed of various sub-parts, illustrated in Fig. 3.

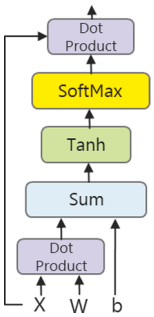


Fig. 1. Additive attention architecture.

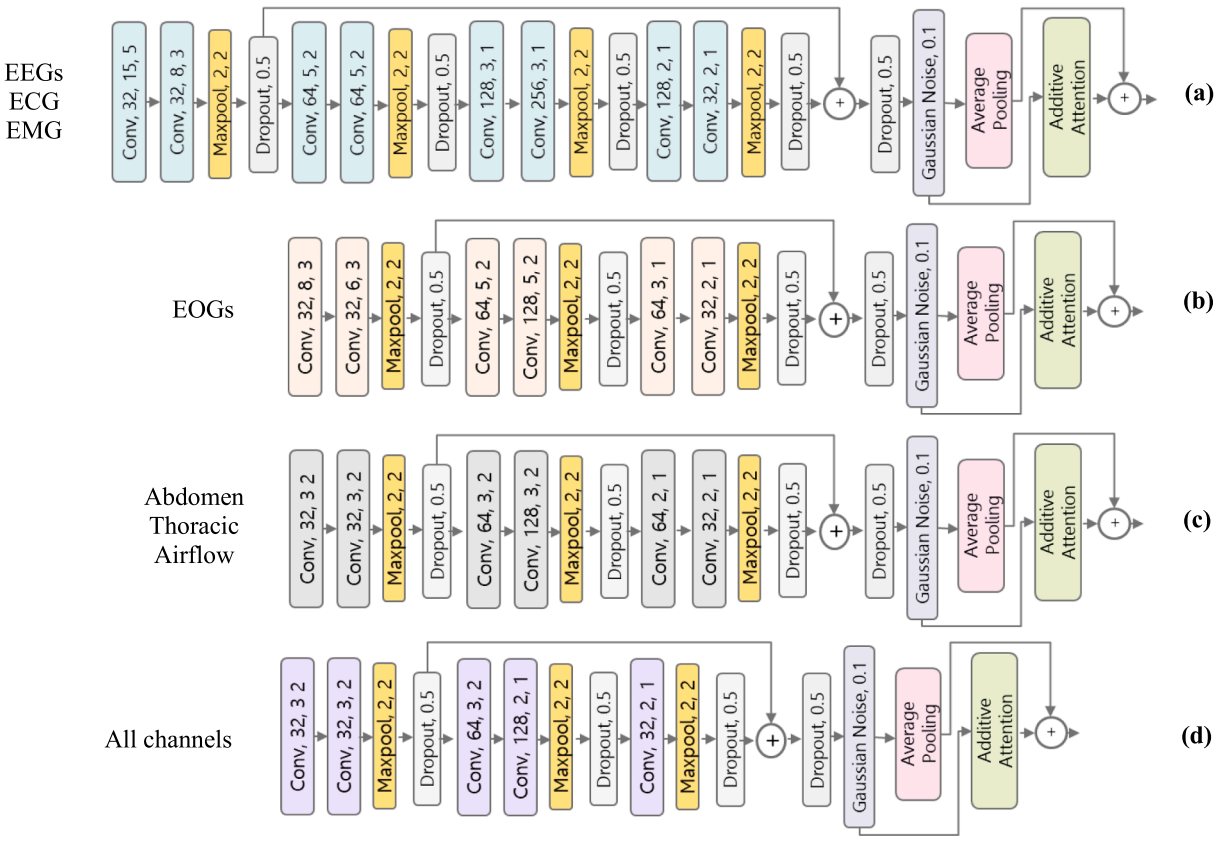


Fig. 2. Developed CNN-Attention blocks for each input modality: (a), (b), and (c) for raw signal data and (e) for handcrafted features data.

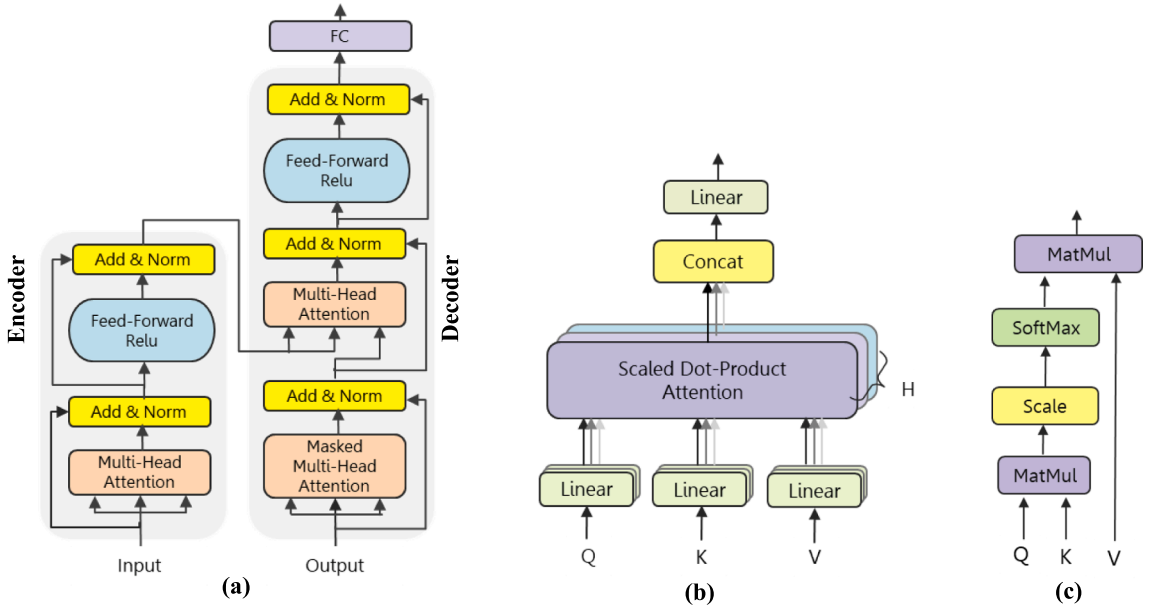


Fig. 3. Architecture of the Transformer: (a) Encoder-Decoder, (b) Multi-head attention, (c) Scaled dot-product attention.

3.4.1. Encoder

According to Fig. 3. (a), the encoder of this model uses the multi-head attention and feed-forward network with residual connections followed by normalization layers to process the input sequence and capture diverse patterns and relationships [16]. Multi-head attention that is based on scaled dot product attention illustrated in Fig. 3. (b) and (c), allow the model to jointly attend to information from different positions in the input sequence f. Each feature is mapped as input f to three data types, the query, the key, and value of size d, which d is the dimension of the key embeddings. The attention weights for each position in the input are calculated as follows:

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})\,V \tag{1}$$

In multi-head attention, this process is repeated for H attention

heads, where H is the number of heads and the outputs of all heads are concatenated to form the final output. Each attention head has its own set of learned linear projections for queries W_h^Q, keys W_h^K, values W_h^V, and W^O which represents the output projection. The output is a concatenation of the weighted sum of values from each head. The steps of multi-head attention can be formulated as follows:

$MultiHead\ (f) = Concat\ (h_1, \ldots, h_H)W^O$ (2).

$h_i = Attention(Q_i, K_i, V_i),\ 1 \le i \le H$ (3).

Where $Q_i = fW_i^Q$, $K_i = fW_i^K$, *and* $V_i = fW_i^V$. This multi-head attention mechanism is a powerful tool for capturing complex dependencies in the input sequence and provides the model with the ability to analyze information from multiple perspectives.

3.4.2. Decoder

The decoder in the transformer model follows a similar structure to the encoder, consisting of repeated decoder layers. Each decoder layer has a specific structure, as illustrated in Fig. 3. In contrast to the encoder layers, the decoder layers include an additional sublayer known as masked self-attention during the decoding process that prevents positions from attending to subsequent positions. This is crucial to avoid "cheating" by looking at future tokens during generation. This sublayer sits between the self-attention sublayer and the feedforward sublayer. It accepts K keys and V values from the encoder output, while Q queries are derived from the previous self-attention sublayer. This design allows the decoder to attend to relevant information from the encoder's output while also considering self-attention within the decoder itself. It is also capable of efficient parallelization during training and inference, leading to faster convergence and inference times compared to sequential models. This parallelization is particularly beneficial when dealing with large datasets and complex models.

In this study, we employ two transformer encoders corresponding to distinct input modalities, alongside a transformer decoder tasked with generating output feature sequences for sleep stage classification. However, we will not address specific aspects such as masking and shifting of output tokens and specific implementation details such as positional encoding techniques that are not used in this work. For a comprehensive understanding of the machine translation transformer, we recommend referring to the original articles of this model [16].

Given the dual input modalities of raw signal and handcrafted data, it becomes essential to fuse the encoded features and attention scores of both input modalities. This integration is achieved through the use of the cross-modal attention mechanism, which is described in the next subsection.

3.5. Cross modality attention

Cross-modality attention serves as a mechanism within neural network architectures, specifically designed for multimodal learning scenarios [26,27]. It becomes particularly valuable when there's a necessity to integrate information from diverse modalities (distinct types of data). In this paper, we use this mechanism to efficiently integrate the information obtained from two transformer encoders of each input modality, including signal data and corresponding handcrafted features. With this approach, attention weights are calculated independently for each modality using dot product attention to generate the context vector and then combined through multiplication as a fusion strategy. Fig. 4 describes the integration steps of the cross-modal attention mechanism. This mechanism enables the model to focus on specific aspects of one modality while concurrently processing information from another modality.

3.6. The proposed deep learning model

In this section, we introduce a novel deep-learning model that uses the transformer encoder-decoder and cross-modal attention to classify sleep stages. Our proposed Cross-modal SleepTransformer network, utilizing various components and multiple physiological channels in two distinct modalities, is designed to handle the complexities inherent in sleep data. According to Fig. 5 which shows the architecture of Cross-modal SleepTransformer, we explain our proposed model in more detail.

Beginning with an array of nine physiological channels including EEG, EEG sec, EOG/L, EOG/R, ECG, EMG, abdominal and thoracic efforts, and airflow, along with their corresponding handcrafted features as multi-modal inputs, our approach involves a multi-step process. Initially, we employ the CNN-Attention blocks, as depicted in Fig. 1, to extract relevant features individually for each channel of each input modality. These CNN-Attention blocks are designed based on the characteristics of the channels and related sampling rates. In Fig. 1, (a), (b), and (c) correspond to raw signal inputs with sampling rates of 125, 50, and 10 Hz and (d) is related to handcrafted data inputs. For the attention mechanism, we used the additive attention that is applied during the sequence, and the resulting context vector is calculated based on the weighted sum of the input sequence. These extracted features from all channels of each input modality are then combined to form composite inputs, to feed into dedicated transformer encoders. In fact, we use two transformer encoders for each input modality to extract relationships and dependencies from combinations of physiological channel features.

In this network, $C_i = \{C_1, C_2, \cdots C_n\}$ corresponds to the physiological channels of each input modality, $f_i = \{f_1, f_2, \cdots f_n\}$ is related to extracted features from CNN-attention blocks and n is the number of channels. The composite inputs ($Modalityfeatures_s$, $Modalityfeatures_h$) of the features extracted from each input modality are then fed into transformer encoders. Each transformer encoder consists of multiple layers, encompassing multi-head attention, feedforward neural network, normalization layers, and residual connections. These components adeptly process the input sequence, capturing intricate dependencies by computing attention scores for each element across the entire sleep sequence. In the subsequent phase, the encoded features (O_{enc_s}, O_{enc_h}) from the two distinct input modalities undergo harmonious integration in a cross-modality attention mechanism. In this mechanism, attention weights for each input modality are calculated based on the K of the next modality and then the context vector is achieved through dot product

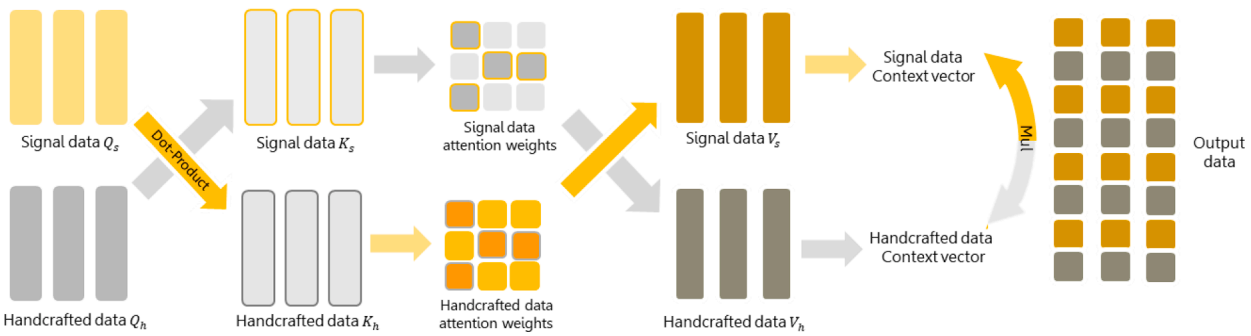


Fig. 4. Cross-modal attention mechanism.

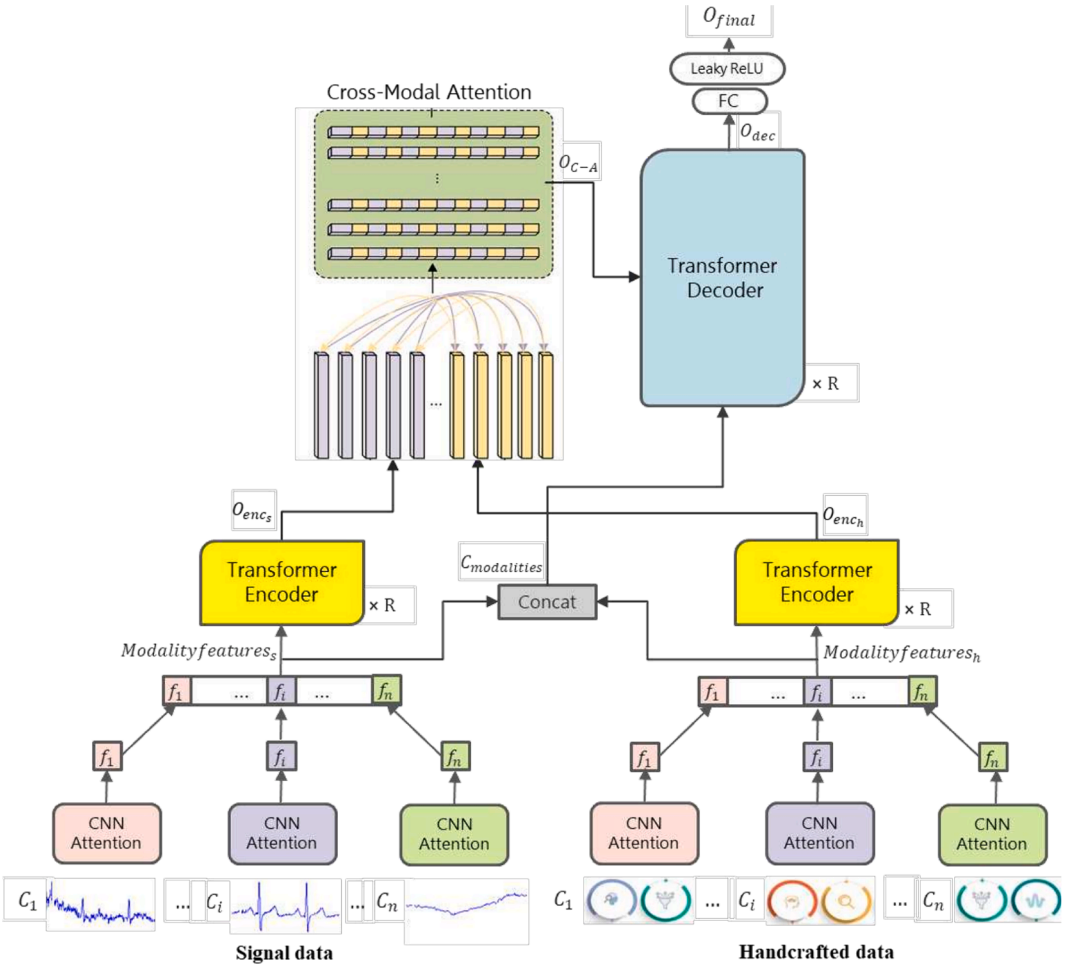


Fig. 5. Architecture of the proposed deep learning model.

with V. Finally the output of the cross attention (O_{C-A}) is determined by multiplying context vectors of each input modality. This pivotal step facilitates the effective integration of attention scores, encapsulating crucial information about each channel and its associated handcrafted features. The cross-modality attention mechanism imparts a comprehensive understanding of sleep stages by selectively focusing on salient aspects across modalities.

Moving to the transformer decoder with two inputs, we have two different combinations of features. The first input to the decoder $(C_{modalities})$ is a simple fusion of raw signal data and handcrafted features, while the next input (O_{C-A}) involves a complex fusion using cross-modal attention mechanism. This complex fusion is fed to the decoder as the output of the transformer encoder. This design aims to compel the decoder to generate an output sequence guided by the integrated attention scores, promoting a nuanced understanding of the relationships between different modalities and enhancing the model's capacity to produce semantically informed predictions.

Finally, the output of the transformer decoder (O_{dec}), passes through a fully connected layer and the overall result benefits from the robust activation capabilities of the Leaky Rectified Linear Unit (ReLU). The formula of these steps can be described as follows:

$$f_i = CNN - Attention(C_1, \cdots, C_n), 1 \le i \le n \tag{4}$$

$$Modality features_{s,h} = \text{Concat}(f_1, \cdots, f_n), where\ s: \ Signal\ data\ inputs,\ h : \ Handcrafted\ data\ inputs \tag{5}$$

$$C_{modalities} = \text{Concat}(Modality features_s,\ Modality features_h \tag{6}$$

$$O_{enc_s} = \text{Encoder}(Modality features_s \tag{7}$$

$$O_{enc_h} = \text{Encoder}(Modality features_h \tag{8}$$

$$O_{C-A} = \text{Cross} - \text{Attention}(O_{enc_s}, O_{enc_h} \tag{9}$$

$$O_{dec} = \text{Decoder}(C_{modalities},\ O_{C-A}) \tag{10}$$

$$O_{final} = Leaky\_ReLU\ (ffn\ (O_{dec})) \tag{11}$$

$$where \begin{cases} Attention\_weights_S = \text{Softmax}\ (\frac{Q_S K_h^T}{\sqrt{d_k}}) & (12) \\ context_S = Attention\_weights_S . V_S & (13) \\ Attention\_weights_h = \text{Softmax}\ (\frac{Q_h K_s^T}{\sqrt{d_k}}) & (14) \\ context_h = Attention\_weights_S . V_h & (15) \\ O_{C-A} = context_s * context_h & (16) \end{cases}$$

The method effectively handles diverse physiological information from nine channels, with their corresponding handcrafted features. Using the multi-modal approach, allows the model to obtain complementary information from both raw signal data and handcrafted features. Individual feature extraction using CNN-Attention blocks for each channel enhances the model's ability to capture channel-specific patterns and differences. On the other hand, transformer encoders enable the extraction of relationships and dependencies from combinations of

physiological channels, capturing intricate dependencies in the input sequence and effective integration of attention scores. It provides a comprehensive understanding of sleep stages by selectively focusing on salient aspects across different modalities. Finally, the complex fusion using cross-attention guides the decoder to generate output sequences based on integrated attention scores, improving the model's ability to generate predictions that are semantically meaningful. The selection of leaky ReLU in the final layer is significant in addressing the challenge of the class imbalance problem. In situations where there is an unequal distribution of samples across classes, softmax may struggle to adequately address the imbalance, often leading to biased predictions favoring the majority class. By using leaky ReLU, the model can better manage class imbalance and ensure improved gradient flow during backpropagation, helping to mitigate issues such as the vanishing gradient problem. Finally, the proposed method combines the strengths of multimodal input processing, channel-specific feature extraction, attention mechanisms, and transformer architecture to achieve accurate and fine-grained classification of sleep stages. This approach is designed to use the diverse information available in physiological channels to improve model performance.

4. Experiments and results

In this section, our objective is to assess the performance of our proposed model by employing diverse evaluation metrics. Subsequently, we compare its results with other state-of-the-art techniques based on deep learning models for the sleep staging task. Furthermore, we conduct statistical tests to discern the comparative performance of our model against others. First, we describe the evaluation metrics utilized, parameter configurations, and the experimental setup adopted in this paper. Additionally, we have made the codes of our model available in the GitHub repository for reference. https://github.com/saharhzm/CrossModalSleepTransformer.

4.1. Evaluation metrics

There are various evaluation metrics for classification models [52], some of which we selected to investigate the performance of our proposed model. The evaluation metrics used include Accuracy (Acc), Specificity (Spec), Sensitivity (Sens), Cohen's kappa (κ), Micro F1 Score (MF1), Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). Given the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the computation of all metrics is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{19}$$

$$Specificity = \frac{TN}{TP + FP} \tag{20}$$

$$MF1 = \frac{1}{M}\sum_{i=1}^{M}\frac{2^{*}\ \text{Precision}\ ^{*}\text{Recall}}{\text{Precision}\ +\ \text{Recall}} \tag{21}$$

$$Cohen's kappa(\kappa) = \frac{p_0 - p_c}{1 - p_c} \tag{22}$$

$$AUROC = \int_0^1 TPR(FPR^{-1}(t))dt \tag{23}$$

$$AUPRC = \int_0^1 Precision(Recall(t))dt \tag{24}$$

Where Accuracy measures the overall correctness of the model. Sensitivity, also known as True Positive Rate or Recall, quantifies the proportion of correctly predicted positive observations relative to all actual positives. Specificity, on the other hand, measures the ratio of correctly predicted negative observations to all actual negatives. The F1 Score is a metric that combines precision and recall into a single value, representing their harmonic mean [53]. It is particularly useful when there is an imbalance between the number of positive and negative instances. The Macro Average F1 Score provides an overall evaluation of the model's performance across multiple classes. In this context, M represents the number of sleep classes. It is calculated by averaging the F1 scores for each class, giving equal weight to all classes. Cohen's Kappa assesses the agreement between predicted and actual classifications while considering the possibility of agreement occurring by chance [54]. In this metric p_0 is the observed agreement, and p_c is the expected agreement. AUROC assesses the model's capability to differentiate between positive and negative classes across a range of thresholds. True Positive Rate (TPR) represents the proportion of correctly identified positive samples, while False Positive Rate (FPR) indicates the proportion of incorrectly identified negative samples. On the other hand, AUPRC evaluates the balance between precision and recall across various threshold values. These metrics are computed for each model through fivefold cross-validation where in each fold we conducted an 80–20 train-test split and the validation set is created by randomly sampling 10 % of the augmented train set. The final results are obtained by averaging them in all experiments.

4.2. Parameter settings and experimental setup

The proposed model is configured with specific parameter settings as follows: We employed the Adam optimizer with a learning rate of 5e-4 for the initial ten epochs, then reduced the learning rate to 1e-4 for the remainder of the training. The choice of these specific learning rates is determined by preliminary experiments and observing the improvement in model performance and stability. Mean Squared Error (MSE) is chosen as the loss function and the rationale behind employing it in our model is closely tied to the use of the leaky ReLU activation function in the final layer. Given that leaky ReLU outputs values ranging from negative infinity to positive infinity without directly computing class probabilities, we opted for MSE as it aligns better with this activation function compared to categorical cross-entropy. It's worth noting that the choice of leaky ReLU in the final layer can adversely impact the performance of categorical cross-entropy loss. For the CNN layers, we initialized the weights using a scaled Gaussian distribution, incorporated the ReLU activation, and added a Gaussian noise of 0.1 after each CNN block. We use the random normal to initialize the trainable weights in additive attention. For cross-modal attention, we employed the model dimensionality of 512. In the transformer encoder applied to the first input, the head size and the number of heads are configured as 128 and 8, respectively. Conversely, for the second input, these parameters are set to 64 and 4. The Transformer decoder is also equipped with 4 heads and 64 of head size. The feed-forward output dimension of the transformer is set to 512 and the Leaky ReLU activation with an alpha of 0.1 is applied in the last layer. A batch size of 512 is also applied during training. We implemented the model using Keras 2.9.0 and used Kaggle TPU VM v3-8 to train it.

4.3. Experimental results

We initiated the training of our proposed model by conducting 400 epochs on the augmented training data obtained from the SHHS1 dataset. Subsequently, we assessed the model's performance on the test

data and summarized the results in the following tables and figures. At first, we conducted ablation studies to evaluate the effectiveness of different channels and components in our proposed model. Table 3 represents the performance of our model across different combinations of channels, evaluated based on accuracy, AUC and macro F1 score metrics. Given that the core of sleep staging relies on three key channels including EEG, EMG, and EOG, we initially evaluated the model using these channels. Next, we included the ECG channel in the input data, which significantly enhanced performance by providing valuable insights into cardiovascular function during sleep. Following this, we incorporated the remaining channels involving abdominal, thoracic, and airflow due to their similar nature, which further enhanced the model's performance.

In the subsequent experiment, we investigated the impact of different components on our proposed model. Table 4 elucidates the effect of key parts on the model's performance. The top part of the table examines the performance when each component is removed. Notably, removing the transformer decoder resulted in the most significant reduction in the model's ability to identify samples of all sleep classes. In this experiment, we ignored the simple combination of channels $C_{modalities}$ and continued with O_{C-A}. The reason for this choice is the presence of more effective features in O_{C-A} than $C_{modalities}$. It is noteworthy that the model failed to converge even after 500 epochs, underscoring the effectiveness of transformer architectures with parallelization processes. Eliminating the cross-attention mechanism, which facilitates the integration of high-level features from different modalities, also led to decreased performance in classifying sleep stages, particularly in minority classes. Additionally, performance suffered when the attention mechanism used in CNN-attention blocks was removed. The lower part of the table compares the efficacy of dot-product attention and additive attention mechanisms. As outlined in the paper "Attention is All You Need" [16], both dot-product attention and additive attention have similar theoretical complexity. However, in practical implementation, dot-product attention tends to be faster and more space-efficient due to its utilization of highly optimized matrix multiplication algorithms. Although both mechanisms perform comparably for small values of d_k, where d_k represents the dimension of the key embeddings, additive attention demonstrates superior performance over dot-product attention without scaling for larger values of d_k [55]. The results of our experiments also indicated that while dot-product attention is faster, additive attention is more efficient for high-dimensional raw signal data.

In the following experiment, we conduct an additional analysis utilizing AUROC and AUPRC metrics to evaluate our model's performance across all classes, considering each class individually. AUROC assesses the model's capability to differentiate between positive and negative classes across different thresholds. The ROC curve illustrates this by plotting the true positive rate against the false positive rate.

Conversely, AUPRC quantifies the balance between precision and recall across various threshold values. Fig. 6 presents the plotted curves of these metrics based on our model's performance on the test dataset. As evident from the results, the model demonstrates satisfactory performance across all classes according to both AUROC and AUPRC metrics, except the N1 class, which exhibits more misclassifications compared to others.

To demonstrate the efficacy of our model, we conducted supplementary analysis incorporating an additional input modality as spectral features alongside raw signal data. This spectral information is derived from the raw signal data through the Short-Time Fourier Transformation (STFT) method, providing insights into the magnitude and phase of frequency components over time [42]. Fig. 7 depicts the confusion matrix of our proposed model utilizing different input modalities. In Fig. 7(a), where raw signal data and handcrafted features are utilized, we achieved an accuracy rate of 0.9133. The recall rates for sleep classes, N1, N2, N3, N4, and REM are 0.92, 0.54, 0.94, 0.91, and 0.93, respectively. Fig. 7(b) illustrates the utilization of raw signal data and STFT features, resulting in an accuracy rate of 0.9027, with recall rates of 0.91, 0.47, 0.94, 0.91, and 0.92 for the respective classes. These results imply that handcrafted features contain more informative details regarding sleep patterns compared to spectral data. Additionally, it showcases the versatility of our model in predicting sleep stages using diverse modalities.

Table 5. provides the classification results of our proposed model using different modalities and its performance comparison with other state-of-the-art techniques. The methods are listed in Table 5 represent the current state-of-the-art approaches for sleep staging tasks. These methods leverage advanced techniques such as attention mechanisms, transformer architectures, multi-modal learning, and hybrid models, all within the realm of deep learning. They managed sleep stage classification utilizing various channels by the SHHS1 dataset. It's worth noting that most of these methods focus on utilizing some channels for classification. In contrast, our proposed method employs a unique combination of multiple channels with their handcrafted and spectral features, which none of the referenced methods utilize. Furthermore, the utilization of the transformer decoder sets our model apart from the others, as none of them incorporate this component. These distinctions highlight the uniqueness of our model compared to existing approaches. However, we recalculated the results for some of the methods marked with an asterisk using our settings and used the results reported in the original papers for the other methods. The provided table contains the used channels, overall metrics, and class-wise MF1 for each model. In the context of overall metrics, including ACC, κ, MF1, Sensitivity, and Specificity, our proposed model achieved the highest values of 0.913, 0.873, 0.860, 0.849, and 0.973, respectively, when utilizing input modalities involving raw signal data and handcrafted features (Cross-modal SleepTransformer1). In addition, our proposed model obtained the second-best results with values of 0.902, 0.858, 0.846, 0.829, and 0.970 for ACC, κ, MF1, Sensitivity, and Specificity using raw signal data and spectral features (Cross-modal SleepTransformer2). Indeed, the distinction between Cross-modal Sleep Transformer1 and Cross-modal Sleep Transformer2 lies in their input data modalities. The right side of the table displays the class-wise Macro F1 (MF1) scores for each class, where our proposed model excelled by achieving F1 values surpassing 90 % in the most of classes. Based on the results our proposed model outperforms most of the other state-of-the-art methods in classifying sleep stages across the five distinct classes. According to the table, the models that use multimodal learning methods with different channels get relatively higher classification performance than other models. On the other hand, using the transformer architecture for these channels with sequential dependencies can increase the ability of the model to identify the complexities of sleep patterns in the classification of different stages.

Additionally, we conducted statistical tests, including the Friedman and Post-hoc methods, to better assess the performance of our proposed

Table 3
The results of our proposed model using different combination of channels.

Different channels	Accuracy	Avg AUC	MF1 | AUC				
			Wake	N1	N2	N3	REM
EEG, EMG, EOG	0.894	0.974	0.873 | 0.977	0.423 | 0.943	0.931 | 0.982	0.923 | 0.991	0.904 | 0.981
EEG, EMG, EOG, ECG	0.907	0.979	0.904 | 0.984	0.501 | 0.951	0.933 | 0.983	0.951 | 0.997	0.917 | 0.983
EEG, EMG, ECG, EOG, Abdomen, Thorax and Airflow	**0.913**	**0.984**	0.932 | 0.995	0.590 | 0.966	0.933 | 0.983	0.925 | 0.993	0.938 | 0.986

Table 4
The effect of different parts on the performance of the proposed model.

Different parts	Accuracy	Avg AUC	MF1 | AUC				
			Wake	N1	N2	N3	REM
Without transformer decoder	0.887	0.961	0.901 | 0.983	0.156 | 0.881	0.913 | 0.968	0.886 | 0.981	0.920 | 0.984
Without cross-attention	0.892	0.970	0.894 | 0.981	0.232 | 0.911	0.933 | 0.983	0.923 | 0.993	0.903 | 0.981
Without attention	0.896	0.975	0.927 | 0.991	0.407 | 0.941	0.928 | 0.975	0.891 | 0.986	0.932 | 0.985
With dot-product attention	0.909	0.980	0.872 | 0.978	0.529 | 0.952	0.942 | 0.988	0.950 | 0.997	0.914 | 0.983
With additive attention (The proposed model)	**0.913**	**0.984**	0.932 | 0.995	0.590 | 0.966	0.933 | 0.983	0.925 | 0.993	0.938 | 0.986

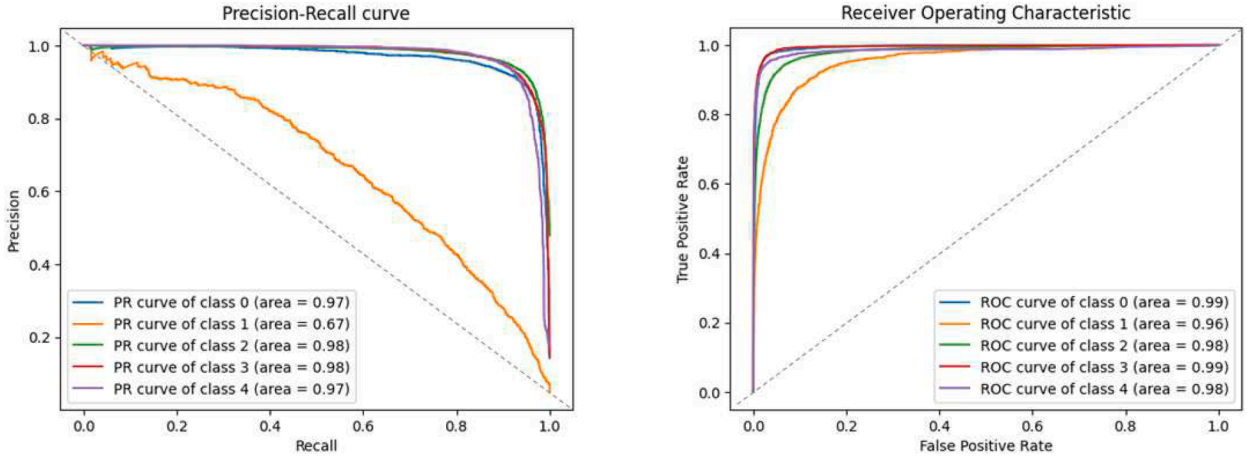


Fig. 6. AUROC and AUPRC curves of the proposed model.

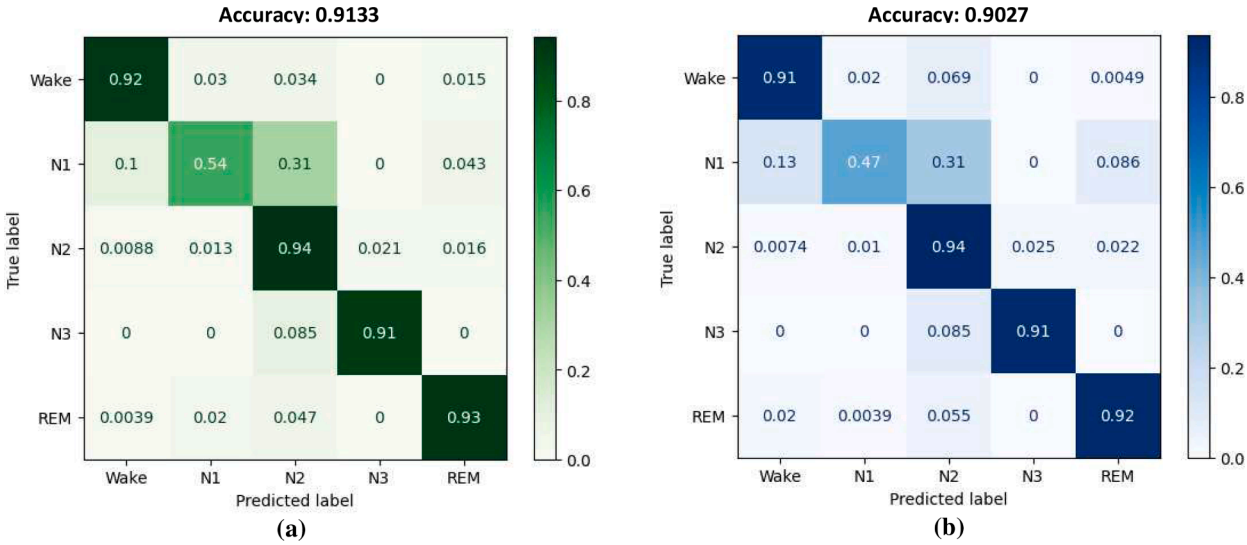


Fig. 7. Confusion matrix of the proposed deep learning model using different input modalities. (a): input modalities as raw signal data and handcrafted features, (b) input modalities as raw signal data and spectral features.

model compared to other techniques. The Friedman test revealed significant differences in performance across multiple models, while the post-hoc comparisons offered detailed insights into specific model pairs with significant performance disparities compared to the control method. In these tests, we selected models listed in Table 5 that have the most reported metrics. Tables 6 and 7 summarize the outcomes of these tests, conducted at a significance level (α) of 0.05, with Cross-modal SleepTransformer1 as the control method.

Table 6 illustrates that our models achieve the highest ranking among all methods evaluated. Furthermore, Table 7 indicates that for most models, null hypothesis H0 is rejected based on the associated p-values. These findings from the statistical tests suggest that our model outperforms the majority of models with statistically significant differences in performance.

5. Discussion

Accurately classifying sleep stages is crucial for diagnosing and managing sleep disorders, providing valuable insights into individuals' sleep patterns and overall well-being. However, the manual process of polysomnogram sleep staging is labor-intensive and requires specialized expertise, which can limit access to care amidst growing global demand.

Table 5
Classification performance of the proposed model and comparison of other methods.

Deep learning model	Used channels	Overall metrics					Class-wise MF1				
		Acc	κ	MF1	Sens	Spec	W	N1	N2	N3	REM
Cross-modal SleepTransformer1	EEG, EMG, ECG, EOG, Abdomen, Thorax and Airflow	**0.913**	**0.873**	**0.860**	**0.849**	**0.973**	0.93	0.59	**0.93**	**0.92**	**0.93**
Cross-modal SleepTransformer2		0.902	0.858	0.846	0.829	0.970	0.91	0.57	0.92	0.92	0.92
Zhang et al. [30] (2023)	EEG, EOG, EMG	0.881	0.836	0.818	–	–	0.91	0.61	0.86	0.86	0.93
*FullSleepNet [31] (2023)	EEG	0.876	0.827	0.805	0.796	0.965	0.92	0.48	0.89	0.83	0.89
Olesen et al. [56] (2021)	EEG, EMG, EOG	0.871	0.816	0.788	0.777	0.963	**0.94**	0.47	0.87	0.74	0.89
CTCNet [32] (2024)	EEG	0.857	0.81	0.772	–	–	0.87	0.33	0.89	0.90	0.83
SeriesSleepNet [57] (2023)	EEG	0.842	0.78	0.778	–	–	0.84	0.48	0.86	0.83	0.86
*XSleepNet [58] (2021)	EEG, EMG, EOG	0.893	0.842	0.814	0.816	0.967	0.88	0.44	0.92	0.86	0.89
SleepEGAN [33] (2023)	EEG	0.880	0.830	0.821	–	–	0.89	0.54	0.89	0.87	0.90
Xiao et al. [59] (2022)	EEG	0.883	0.823	–	–	–	–	–	–	–	–
SleepTransformer [34] (2022)	EEG	0.877	0.828	0.801	0.787	0.965	0.92	0.46	0.88	0.85	0.88
L-SeqSleepNet [35] (2023)	EEG	0.884	0.838	0.814	0.804	0.867	0.93	0.51	0.89	0.84	0.89
Sors et al. [60] (2018)	EEG	0.870	0.810	0.780	–	–	0.91	0.42	0.87	0.85	0.85
FCNN+RNN [58] (2018)	EEG, EMG, EOG	0.881	0.832	0.809	0.797	0.966	0.91	0.48	0.88	0.82	0.87
MGANet [38] (2022)	EEG	0.873	0.827	0.801	–	–	–	–	–	–	–
SeqSleepNet [61] (2019)	EEG	0.884	0.838	0.801	0.785	0.967	0.91	0.43	0.87	0.82	0.87
IITNet [62] (2020)	EEG	0.867	0.801	0.798	–	–	–	–	–	–	–
SleepContexNet [36] (2022)	EEG	0.864	0.810	0.805	–	–	0.89	0.52	0.87	0.84	0.89
CNN [63] (2020)	EEG	0.868	0.810	0.785	–	0.950	–	–	–	–	–
LightSleepNet [64](2021)	EEG	0.867	0.813	–	–	–	–	–	–	–	–
Yan et al. [65] (2021)	EEG, EMG, EOG, ECG	0.850	0.790	0.760	–	–	0.93	0.37	0.88	0.81	0.87
*AttnSleep [37] (2021)	EEG	0.841	0.778	0.751	0.748	0.942	0.85	0.32	0.90	0.87	0.80
Liu et al. [66] (2022)	EEG	0.868	–	0.835	–	–	0.91	**0.66**	0.87	0.83	0.89
EEGSNet [67] (2022)	EEG	0.851	0.79	0.785	–	–	0.88	0.47	0.85	0.82	0.88
Fernandez et al. [68] (2020)	EEG	0.852	0.790	0.760	–	–	–	–	–	–	–
Kong et al. [69] (2023)	EEG	0.819	0.740	0.735	–	–	0.84	0.36	0.84	0.80	0.80
Pei et al. [70] (2022)	EEG, EMG, EOG	0.831	0.760	–	–	–	0.89	0.31	0.85	0.78	0.80
SleepFCN [71] (2022)	EEG	0.817	0.740	0.720	–	–	0.81	0.29	0.85	0.84	0.79

* The results of these models are calculated based on our specific settings. For the other models, we have reported the results as presented in their original papers.

Table 6
The results of Friedman test (significance level of 0.05).

Statistic	P-value	Result
12.87790	0.00000	H0 is rejected

Ranking			
Rank	Algorithm	Rank	Algorithm
1.50000	Cross-modal SleepTransformer1	10.62500	SeqSleepNet [61]
3.18750	Cross-modal SleepTransformer2	10.75000	Olesen et al. [56]
5.68750	L-SeqSleepNet [35]	12.62500	Sors et al. [60]
6.00000	XSleepNet [58]	12.68750	CTCNet [32]
6.25000	SleepEGAN [33]	13.31250	Yan et al. [65]
6.56250	Zhang et al. [30]	14.25000	EEGSNet [67]
9.18750	SleepTransformer [34]	15.25000	SeriesSleepNet [57]
9.25000	FullSleepNet [31]	15.50000	AttnSleep [37]
9.50000	FCNN+RNN [58]	18.62500	SleepFCN [71]
10.37500	SleepContexNet [36]	18.87500	Kong et al. [69]

Our Cross-modal SleepTransformer model fills this gap by offering precise and dependable sleep stage classification using comprehensive physiological data. This has significant clinical implications, as the model provides objective assessments of sleep stages, facilitating diagnosis and monitoring. We believe that utilizing more channels, particularly the ECG channel, provides critical information that enhances our understanding of sleep stages and improves the model's performance. ECG data can significantly aid in the classification of sleep stages by providing valuable information about the autonomic nervous system and cardiovascular function during sleep. Different sleep stages are characterized by distinct patterns in Heart Rate Variability (HRV) [72]. For example, HRV tends to be higher during NREM sleep and lower during REM sleep. By analyzing the time between successive R-wave peaks (RRI and RPE features), we can help differentiate between sleep stages. Short-term increases in heart rate and changes in HRV can signal arousals or transitions between sleep stages. These subtle changes can be detected through careful analysis of ECG data. Furthermore, ECG can improve the quality of other modalities by helping to remove artifacts [73]. Additionally, measurements of respiratory channels such as abdominal and thoracic effort provide insights into respiratory patterns, which can vary significantly during different sleep stages, including N1 [74]. Airflow data also captures breathing irregularities that might indicate transitions between wakefulness and N1 sleep.

By analyzing multiple biological channels along with handcrafted features, our model offers detailed insights into sleep architecture, assisting clinicians in identifying abnormalities. These features capture essential characteristics of the physiological signals, providing additional information beyond raw data, which leads to improved performance.

Accurate sleep stage classification is vital for diagnosing narcolepsy, insomnia, and obstructive sleep apnea (OSA). Identifying REM and NREM stages in OSA patients guides treatment decisions, while accurate staging helps insomnia patients tailor treatment. Challenging stages like N1 and N2 require precise identification to understand sleep architecture and detect abnormalities. High classification accuracy in these stages aids in diagnosing insomnia and assessing sleep quality. Finally, our model optimizes treatment strategies and streamlines workflows, making it suitable for large-scale clinical applications and improving patient outcomes.

6. Conclusion

In this study, we introduced a novel deep-learning model based on transformer architecture and cross-modality attention mechanism to classify sleep stages using key physiological channels in distinct input modalities. The model utilized CNN-attention blocks for feature extraction, followed by the application of two transformer encoders to process modality-specific features. Subsequently, we employed cross-modality attention to integrate the information of modalities and generate the final output sequence by the transformer decoder to predict

Table 7
The results of Finner Post-hoc test hoc using Cross-modal SleepTransformer1 as control method.

Comparison	Statistic	p-value	Result
Cross-modal SleepTransformer1 vs Kong et al. [69]	5.87382	0.00000	H0 is rejected
Cross-modal SleepTransformer1 vs SleepFCN [71]	5.78931	0.00000	H0 is rejected
Cross-modal SleepTransformer1 vs AttnSleep [37]	4.73286	0.00001	H0 is rejected
Cross-modal SleepTransformer1 vs SeriesSleepNet [57]	4.64835	0.00002	H0 is rejected
Cross-modal SleepTransformer1 vs EEGSNet [65]	4.31029	0.00006	H0 is rejected
Cross-modal SleepTransformer1 vs Yan et al. [65]	3.99335	0.00021	H0 is rejected
Cross-modal SleepTransformer1 vs CTCNet [32]	3.78207	0.00042	H0 is rejected
Cross-modal SleepTransformer1 vs Sors et al. [60]	3.76094	0.00042	H0 is rejected
Cross-modal SleepTransformer1 vs Olesen et al. [56]	3.12707	0.00372	H0 is rejected
Cross-modal SleepTransformer1 vs SeqSleepNet [61]	3.08481	0.00387	H0 is rejected
Cross-modal SleepTransformer1 vs SleepContexNet [36]	3.00030	0.00465	H0 is rejected
Cross-modal SleepTransformer1 vs FCNN+RNN [58]	2.70449	0.01081	H0 is rejected
Cross-modal SleepTransformer1 vs FullSleepNet [31]	2.61998	0.01283	H0 is rejected
Cross-modal SleepTransformer1 vs SleepTransformer [34]	2.59885	0.01283	H0 is rejected
Cross-modal SleepTransformer1 vs Zhang et al. [30]	1.71144	0.10889	H0 is accepted
Cross-modal SleepTransformer1 vs SleepEGAN [33]	1.60579	0.12728	H0 is accepted
Cross-modal SleepTransformer1 vs XSleepNet [58]	1.52128	0.14215	H0 is accepted
Cross-modal SleepTransformer1 vs L-SeqSleepNet [35]	1.41563	0.16484	H0 is accepted
Cross-modal SleepTransformer1 vs Cross-modal SleepTransformer2	0.57048	0.56835	H0 is accepted

sleep stages.

We assessed the performance of our proposed model using various metrics and conducted comparative analyses with state-of-the-art methods. We also performed statistical tests to show the effectiveness of our model against other methods. The findings revealed that our approach, using a combination of transformer encoder-decoder architecture and attention mechanism with multi-modal inputs, outperforms most of other techniques in the challenging task of sleep staging.

The flexibility of the proposed model makes it suitable for managing different input modalities and adapting to diverse physiological channels. This versatility increases its use in a wide range of sleep-related tasks. It provides a robust and versatile solution for sleep staging and uses advanced deep-learning techniques to effectively process multi-modal inputs and capture complex patterns in physiological data.

CRediT authorship contribution statement

Sahar Hassanzadeh Mostafaei: Writing – original draft, Methodology. **Jafar Tanha:** Writing – review & editing. **Amir Sharafkhaneh:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.V. Rundo, R. Downey III, *Polysomnography*, Handb. Clin. Neurol. 160 (2019) 381–392.

[2] "American Academy of Sleep Medicine, (2021), AAST PSG Guideline, AAST," [Online]. Available: https://www.aastweb.org/Portals/0/Docs/Resources/Guidelines/AAST%20PSG%20Guideline%20Final.pdf.

[3] Y.J. Ma, J. Zschocke, M. Glos, M. Kluge, T. Penzel, J.W. Kantelhardt, R.P. Bartsch, Automatic sleep-stage classification of heart rate and actigraphy data using deep and transfer learning approaches, Comput. Biol. Med. 163 (2023) 107193.

[4] N. Michielli, U.R. Acharya, F. Molinari, Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals, Comput. Biol. Med. 106 (2019) 71–81.

[5] X. Ji, Y. Li, P. Wen, P. Barua, U.R. Acharya, MixSleepNet: A multi-type convolution combined sleep stage classification model, Comput. Methods Programs Biomed. (2023) 107992.

[6] C. Liu, Y. Yin, Y. Sun, O.K. Ersoy, Multi-scale ResNet and BiGRU automatic sleep staging based on attention mechanism, PLoS One 17 (6) (2022) e0269500.

[7] J. Lu, C. Yan, J. Li, C. Liu, Sleep staging based on single-channel EEG and EOG with Tiny U-Net, Comput. Biol. Med. (2023) 107127.

[8] X. Xu, C. Chen, K. Meng, L. Lu, X. Cheng, H. Fan, NAMRTNet: Automatic classification of sleep stages based on improved ResNet-TCN network and attention mechanism, Appl. Sci. 13 (11) (2023) 6788.

[9] H. Wang, C. Lu, Q. Zhang, Z. Hu, X. Yuan, P. Zhang, W. Liu, A novel sleep staging network based on multi-scale dual attention, Biomed. Signal Process. Control 74 (2022) 103486.

[10] Y. Liu, Z. Jia, "Bstt: A bayesian spatial-temporal transformer for sleep staging", *The Eleventh International Conference on Learning Representations* (2022).

[11] Y. Wang, Z. Mei, Q. Wu, Z. Huang, Convolutional transformer with domain adversarial learning for multi-channel sleep stage classification, in: *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2023, pp. 1–6.

[12] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.

[13] J.-W. Liu, J.-W. Liu, X.-L. Luo, Research progress in attention mechanism in deep learning, Chinese Journal of Engineering 43 (11) (2021) 1499–1511.

[14] Y. Hu, W. Shi, C.-H. Yeh, Spatiotemporal convolution sleep network based on graph attention mechanism with automatic feature extraction, Comput. Methods Programs Biomed. 244 (2024) 107930.

[15] Z. Jin, K. Jia, A temporal multi-scale hybrid attention network for sleep stage classification, Med. Biol. Eng. Compu. (2023) 1–13.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).

[17] R. Hu, J. Chen, L. Zhou, A transformer-based deep neural network for arrhythmia detection using continuous ECG signals, Comput. Biol. Med. 144 (2022) 105325.

[18] W. Boes, H.V. Hamme, "Audiovisual transformer architectures for large-scale classification and synchronization of weakly labeled audio events", *Proceedings of the 27th ACM International Conference on Multimedia* (2019) 1961–1969.

[19] G. Shi, Z. Chen, R. Zhang, A transformer-based spatial-temporal sleep staging model through raw EEG, in: *2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, IEEE, 2021, pp. 110–115.

[20] Z. Yang, D. Wang, Z. Chen, M. Huang, N. Ono, Md. Altaf-Ul-Amin, S. Kanaya, Exploring feasibility of truth-involved automatic sleep staging combined with transformer, in: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2021, pp. 2920–2923.

[21] M. Kim, K. Jung, W. Chung, Automatic sleep stage classification method based on transformer-in-transformer, in: *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, IEEE, 2023, pp. 1–5.

[22] X. Huang, F. Schmelter, M.T. Irshad, A. Piet, M.A. Nisar, C. Sina, M. Grzegorzek, Optimizing sleep staging on multimodal time series: Leveraging borderline synthetic minority oversampling technique and and supervised convolutional contrastive learning, Comput. Biol. Med. 166 (2023) 107501.

[23] Y. Yu, S. Chen, J. Pan, MCASleepNet: Multimodal channel attention-based deep neural network for automatic sleep staging, in: International Conference on Artificial Neural Networks, Springer Nature Switzerland, Cham, 2023, pp. 308–319.

[24] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, "Multi-modality cross attention network for image and sentence matching", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 10941–10950.

[25] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, Adv. Neural Inf. Proces. Syst. 32 (2019).

[26] W.-Y. Lee, L. Jovanov, W. Philips, Cross-modality attention and multimodal fusion transformer for pedestrian detection, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 608–623.

[27] L. Zou, Z. Huang, F. Wang, Z. Yang, G. Wang, CMA: Cross-modal attention for 6D object pose estimation, Comput. Graph. 97 (2021) 139–147.

[28] C.K. Chieh, M. Hasegawa-Johnson, N.L. McElwain, B. Islam, Classification of infant sleep/wake states: cross-attention among large scale pretrained transformer networks using audio, ECG, and IMU Data, in: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 2370–2377.

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

[30] D. Zhang, J. Sun, Y. She, Y. Cui, X. Zeng, L. Lu, C. Tang, N. Xu, B. Chen, W. Qin, A two-branch trade-off neural network for balanced scoring sleep stages on multiple cohorts, Front. Neurosci. 17 (2023).
[31] H. Zan, A. Yildiz, Multi-task learning for arousal and sleep stage detection using fully convolutional networks, J. Neural Eng. 20 (5) (2023) 056034.
[32] W. Zhang, C. Li, H. Peng, H. Qiao, X. Chen, CTCNet: A CNN Transformer capsule network for sleep stage classification, Measurement 114157 (2024).
[33] X. Cheng, K. Huang, Y. Zou, S. Ma, SleepEGAN: A GAN-enhanced ensemble deep learning model for imbalanced classification of sleep stages, Biomed. Signal Process. Control 92 (2024) 106020.
[34] P. Huy, K.B. Mikkelsen, O. Chen, P. Koch, A. Mertins, M. De Vos, SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification, IEEE Trans. Biomed. Eng. (2022).
[35] H. Phan, K.P. Lorenzen, E. Heremans, O.Y. Chén, M.C. Tran, P. Koch, A. Mertins, M. Baumert, K.B. Mikkelsen, M. De Vos, "L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging", IEEE Journal of Biomedical and Health. (2023).
[36] C. Zhao, J. Li, Y. Guo, SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG, Comput. Methods Programs Biomed. 220 (2022) 106806.
[37] E. Eldele, et al., An attention-based deep learning approach for sleep stage classification with single-channel eeg, IEEE Trans. on Neural Systems and Rehabilitation Engineering 29 (2021) 809–818.
[38] Q. Wang, Y. Guo, Y. Shen, S. Tong, H. Guo, Multi-layer graph attention network for sleep stage classification based on EEG, Sensors 22 (23) (2022) 9272.
[39] Q. Stuart, F.B.V. Howard, C. Iber, J.P. Kiley, F.J. Nieto, G.T. O'Connor, D. M. Rapoport, The sleep heart health study: design, rationale, and methods, Sleep 20 (12) (1997) 1077–1085.
[40] G.Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S.D. Mariani, S. Redline, The national sleep research resource: towards a sleep data commons, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1351–1358.
[41] A. Rechtschaffen, A. Kales, A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects, Public Health Service, US Government Printing Office, Washington DC, 1968.
[42] B. Iwana, S. Uchida, Time series data augmentation for neural networks by time warping with a discriminative teacher, in: 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 3558–3565.
[43] G. Chen, K. Li, Y. Liu, Applicability of continuous, stationary, and discrete wavelet transforms in engineering signal processing, J. Perform. Constr. Facil 35 (5) (2021) 04021060.
[44] M. Wamidh, K.A. Shaker, K.Z.M. Aydam, B.H. Taher, "Feature extraction methods: a review", *J. Phys. Conf. Ser.* 1591 (1) (2020) 012028. IOP Publishing.
[45] I. Stancin, M. Cifrek, A. Jovic, A review of EEG signal features and their application in driver drowsiness detection systems, Sensors 21 (11) (2021) 3786.
[46] P. Hamilton, "Open source ECG analysis", *Comput. Cardiol.* (2002) 101–104. IEEE.
[47] S.H. Mostafaei, J. Tanha, A. Sharafkhaneh, Z.H. Mostafaei, M.H.A. Al-Jaf, A. F. Babaei, An ensemble model for sleep stages classification, in: 31st International Conference on Electrical Engineering (ICEE), IEEE, 2023, pp. 327–332.
[48] L. Li, H. Cai, H. Han, Q. Jiang, H. Ji, Adaptive short-time Fourier transform and synchrosqueezing transform for non-stationary signal separation, Signal Process. 166 (2020) 107231.
[49] B. Zhao, H. Lu, S. Chen, J. Liu, D. Wu, Convolutional neural networks for time series classification, J. Syst. Eng. Electron. 28 (1) (2017) 162–169.
[50] F. He, T. Liu, D. Tao, Why resnet works? residuals generalize, IEEE Trans. Neural Networks Learn. Syst. 31 (12) (2020) 5349–5362.
[51] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.," *arXiv preprint arXiv:1409.0473, 2014.*
[52] G. Naidu, T. Zuva, S.E. Mmbongeni, "A review of evaluation metrics in machine learning algorithms", in: *Computer Science on-Line Conference*, Springer International Publishing, Cham, 2023, pp. 15–25.
[53] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 1–13.
[54] M.L. McHugh, Interrater reliability: the kappa statistic, Biochemia Medica 22 (3) (2012) 276–282.
[55] Denny Britz, Anna Goldie, Minh-Thang Luong, V Le Quoc, Massive exploration of neural machine translation architectures, CoRR abs/1703.03906 (2017).
[56] A. Olesen, P. Jørgen Jennum, E. Mignot, H. Sorensen, Automatic sleep stage classification with deep residual networks in a mixed-cohort setting, Sleep 44 (1) (2021) p.zsaa161.
[57] M. Lee, H.-G. Kwak, H.-J. Kim, D.-O. Won, S.-W. Lee, SeriesSleepNet: an EEG time series model with partial data augmentation for automatic sleep stage scoring, Front. Physiol. 14 (2023).
[58] P. Huy, O.Y. Chén, M.C. Tran, P. Koch, A. Mertins, M. De Vos, XSleepNet: Multi-view sequential model for automatic sleep staging, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[59] W. Xiao, R. Linghu, H. Li, F. Hou, Automatic sleep staging based on single-channel EEG signal using null space pursuit decomposition algorithm, Axioms 12 (1) (2022) 30.
[60] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, J. Payen, A convolutional neural network for sleep stage scoring from raw single-channel EEG, Biomed. Signal Process. Control 42 (2018) 107–114.
[61] P. Huy, F. Andreotti, N. Cooray, O.Y. Chén, M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging", *EEE Transactions on Neural Systems and Rehabilitation Engineering* 27 (3) (2019) 400–410.
[62] H. Seo, S. Back, S. Lee, D. Park, T. Kim, K. Lee, Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG, Biomed. Signal Process. Control 61 (2020) 102037.
[63] K. Mikkelsen and M. De Vos, "Personalizing deep learning models for automatic sleep staging," *arXiv preprint arXiv:1801.02645, 2018.*
[64] D. Zhou, Q. Xu, J. Wang, J. Zhang, G. Hu, L. Kettunen, Z. Chang, F. Cong, LightSleepNet: A lightweight deep model for rapid sleep stage classification with spectrograms. *43rd Annual International Conference of the IEEE*, 2021.
[65] R. Yan, F. Li, D.D. Zhou, T. Ristaniemi, F. Cong, Automatic sleep scoring: A deep learning architecture for multi-modality time series, J. Neurosci. Methods 348 (2021) 108971.
[66] Z. Liu, S. Luo, Y. Lu, Y. Zhang, L. Jiang, H. Xiao, Extracting multi-scale and salient features by MSE based U-structure and CBAM for sleep staging, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 31–38.
[67] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, M. Lee, A deep learning method approach for sleep stage classification with eeg spectrogram, Int. J. Environ. Res. Public Health 19 (10) (2022) 6322.
[68] E. Fernandez-Blanco, D. Rivero, A. Pazos, EEG signal processing with separable convolutional neural network for automatic scoring of sleeping stage, Neurocomputing 410 (2020) 220–228.
[69] G. Kong, C. Li, H. Peng, Z. Han, H. Qiao, EEG-based sleep stage classification via neural architecture search, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 1075–1085.
[70] W. Pei, Y. Li, S. Siuly, P. Wen, A hybrid deep learning scheme for multi-channel sleep stage classification, Computers, Materials and Continua 71 (1) (2022) 889–905.
[71] N. Goshtasbi, R. Boostani, S. Sanei, SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 2088–2096.
[72] C. Yan, P. Li, M. Yang, Y. Li, J. Li, H. Zhang, C. Liu, Entropy analysis of heart rate variability in different sleep stages, Entropy 24 (3) (2022) 379.
[73] X. Jiang, G.B. Bian, Z. Tian, Removal of artifacts from EEG signals: a review, Sensors 19 (5) (2019) 987.
[74] F. Ebrahimi, I. Alizadeh, Automatic sleep staging by cardiorespiratory signals: a systematic review, Sleep Breath. (2021) 1–17.