

Article

Multimodal Hybrid CNN-Transformer with Attention Mechanism for Sleep Stages and Disorders Classification Using Bio-Signal Images

Innocent Tujiyina^{1,*} , Bessam Abdulrazak¹  and Rachid Hedjam^{1,2} ¹ Ambient Intelligence Lab, Department of Computer Science, Faculty of Science, University of Sherbrooke, Sherbrooke, QC J1K 2R1, Canada; bessam.abdulrazak@usherbrooke.ca (B.A.); rhedjam@ubishops.ca (R.H.)² Department of Computer Science, Bishop's University, Sherbrooke, QC J1M 1Z7, Canada

* Correspondence: innocent.tujiyina@usherbrooke.ca

Abstract

Background and Objective: The accurate detection of sleep stages and disorders in older adults is essential for the effective diagnosis and treatment of sleep disorders affecting millions worldwide. Although Polysomnography (PSG) remains the primary method for monitoring sleep in medical settings, it is costly and time-consuming. Recent automated models have not fully explored and effectively fused the sleep features that are essential to identify sleep stages and disorders. This study proposes a novel automated model for detecting sleep stages and disorders in older adults by analyzing PSG recordings. PSG data include multiple channels, and the use of our proposed advanced methods reveals the potential correlations and complementary features across EEG, EOG, and EMG signals. **Methods:** In this study, we employed three novel advanced architectures, (1) CNNs, (2) CNNs with Bi-LSTM, and (3) CNNs with a transformer encoder, for the automatic classification of sleep stages and disorders using multichannel PSG data. The CNN extracts local features from RGB spectrogram images of EEG, EOG, and EMG signals individually, followed by an appropriate column-wise feature fusion block. The Bi-LSTM and transformer encoder are then used to learn and capture intra-epoch feature transition rules and dependencies. A residual connection is also applied to preserve the characteristics of the original joint feature maps and prevent gradient vanishing. **Results:** The experimental results in the CAP sleep database demonstrated that our proposed CNN with transformer encoder method outperformed standalone CNN, CNN with Bi-LSTM, and other advanced state-of-the-art methods in sleep stages and disorders classification. It achieves an accuracy of 95.2%, Cohen's kappa of 93.6%, MF1 of 91.3%, and MGm of 95% for sleep staging, and an accuracy of 99.3%, Cohen's kappa of 99.1%, MF1 of 99.2%, and MGm of 99.6% for disorder detection. Our model also achieves superior performance to other state-of-the-art approaches in the classification of N1, a stage known for its classification difficulty. **Conclusions:** To the best of our knowledge, we are the first group going beyond the standard to investigate and innovate a model architecture which is accurate and robust for classifying sleep stages and disorders in the elderly for both patient and non-patient subjects. Given its high performance, our method has the potential to be integrated and deployed into clinical routine care settings.



Academic Editor: Hugo Fernando Posada-Quintero

Received: 28 October 2025

Revised: 9 December 2025

Accepted: 16 December 2025

Published: 8 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: PSG data; RGB spectrogram image; feature extraction and fusion; deep learning; sleep stage classification; sleep disorder classification

1. Introduction

Sleep is a fundamental aspect of our life, and ensuring high-quality sleep is important. The significance of sleep becomes evident when considering that individuals spend approximately one-third of their lives undergoing this vital physiological process [1]. The sleep architecture is composed of distinct sleep stages, each marked by specific physiological changes. Originally, Rechtschaffen and Kales (1968) divided sleep into seven stages using the R&K method. According to their guidelines, these stages are wakefulness (W), four stages of non-REM (NREM) sleep (S1 to S4), and rapid eye movement (REM), in addition to body movement (MT) [2]. Subsequently, in 2007, the American Academy of Sleep Medicine (AASM) revised their sleep scoring manual. As per AASM guidelines, the S3 and S4 were combined into a single stage, labeled as S3, due to their similar characteristics [3]. Inadequate sleep quality in older adults is scientifically linked to numerous health problems, including obesity, diabetes, heart diseases, mood disorders, impaired immune system, elevated risk of mortality, and more [4]. Research reveals that one-third (33%) of Canadian men experience sleep deprivation [5]. Similarly, a study found that over one-third of American adults consistently face challenges in obtaining sufficient sleep [6]. The early identification of sleep pattern changes can help prevent sleep disorders from worsening. Although polysomnography (PSG) is the gold standard for diagnosing sleep conditions, it is expensive, time-consuming [7–10], and generates extensive amounts of data, making the manual scoring of CAP cycles impractical and error-prone [11]. To address the challenges and problems mentioned, researchers have been developing automated methods to assist sleep experts in detecting sleep stages and disorders. This is particularly relevant in aging populations, who exhibit a higher prevalence of sleep disorders and comorbidities, highlighting the need for more robust multimodal approaches.

Previous studies mainly used single-modality EEG signals to classify sleep stages [12–19] and disorders [20], highlighting its simplicity, reduced model complexity, and suitability for extended home-based sleep monitoring. However, accurately identifying distinct sleep stages using only EEG signals is challenging as certain stages, such as REM, N1, and N2, exhibit somewhat similar EEG patterns [21]. Thus, in addition to EEG brain activities, trained sleep experts also examine eye movements (EOG) and muscle activity levels (EMG) when annotating a 30 s PSG epoch [3]. These additional signals play a crucial role in identifying certain sleep stages and diagnosing sleep disorders [22,23]. Furthermore, EOG and EMG have also been proven to be useful additional sources, complementing EEG in other multimodal automatic systems for sleep staging [24–30] and disorder detection [31–34]. Therefore, this study uses multimodal signals (EEG, EOG, and EMG) to better align with the manual approach used by sleep experts, aiming to improve accuracy in classifying sleep stages and diagnosing sleep disorders for broader clinical adoption.

To achieve better results, some recent deep learning studies have utilized time–frequency images of physiological signals to classify sleep stages [16,17,24,26,35,36] and disorders [32,34]. To transform time-domain signals into time–frequency images, some studies used short-time Fourier transform (STFT) [16,17,24,26,32,34,35], while others employed continuous wavelet transform (CWT) [36]. These studies demonstrate the effectiveness of deep learning models in image-based sleep stage classification and disorder detection. In addition, time–frequency images are essential for sleep staging and disorder classification as they capture both time- and frequency-domain features, and they serve as a high-level representation of raw signals [35]. While CWT is a valuable substitute for STFT, in this study, we opt for STFT due to its real-time efficiency and widespread industry adoption [34].

Recent research studies on sleep scoring have widely investigated traditional machine learning algorithms, such as SVM [12,37], RF [13], and KNN [38]. Building on this foundation, Rui et al. (2019) proposed a multi-modality approach for automatic sleep staging

using EEG, EOG, EMG, and ECG signals. Features selected using the ReliefF algorithm were classified with a random forest model, achieving an accuracy of 86.24% [30]. Sharma et al. (2021) focused on detecting sleep disorders by using optimal triplet half-band filter bank (THFB) wavelet-based features from two EEG channels. The extracted Hjorth parameters were then fed into an ensemble boosted trees classifier, achieving an accuracy of 91.3% [20]. In a subsequent study, Sharma et al. (2022) improved accuracy to 94.3% by applying a biorthogonal wavelet filter bank to EOG and EMG signals and using ensemble bagged trees [33]. However, these approaches rely heavily on hand-crafted features and domain expertise, and often suffer from the curse of dimensionality, leading to potential information loss during data reduction [9]. To address these challenges, recent studies are increasingly turning to deep learning, which automatically extracts features and performs well on large sleep datasets for sleep stage and disorder detection.

Deep learning methods, including CNNs, RNNs, transformers, and their combinations, have gained popularity for sleep stage classification and the diagnosis of sleep disorders. Several studies have employed CNNs [24,32,34,39,40] for these tasks. Cheng et al. (2023) proposed a distributed multimodal and multilabel decision-making system (MML-DMS) based on VGG16 CNN architectures for the automatic identification of sleep stages and sleep disorders [34]. They used EEG, ECG, and EMG recordings, computing spectrograms for each channel. These images were then fed into separate VGG16 CNN models, and the concatenated probability vectors from all classifiers were passed through a shallow perceptron neural network for final classification. The proposed approach achieved an average classification accuracy of 94.34% for sleep staging and 99.09% for sleep disorder detection [34]. Overall, standalone CNN-based methods effectively extract features from sleep data without the need for manual feature selection. However, they fall short in modeling the transitional relationships of intra-epoch features, which can carry subtle yet clinically relevant temporal patterns, limiting their ability to fully exploit the temporal complexity present, even within a single epoch.

Given that sleep follows a sequential pattern, several studies have employed recurrent neural networks (RNNs) and long short-term memory (LSTM) networks to capture temporal dependencies in sleep data [14,26]. These studies suggest that accounting for the relationships between sleep epochs improves sleep staging performance. However, RNNs increase model complexity, are difficult to train in parallel, and are susceptible to vanishing or exploding gradients during backpropagation. Recent studies have been inspired by [41] and adopted attention mechanism-based transformer encoder models [17,35]. In line with this, standalone transformer-based models are advantageous as they can be trained in parallel and are less complex than RNNs. They also excel at capturing global features but struggle to extract fine-grained local features, as emphasized in [15]. Therefore, combining CNNs with transformers enables the model to first extract local features effectively, and the resulting feature maps are then fused column-wise, allowing the transformer to capture global dependencies across those features within a single epoch, boosting performance in sleep staging and disorder diagnosis.

Hybrid models combining CNNs with RNNs or LSTM architectures have been used in the current literature [16,25,42–44]. For instance, Li et al. (2022) introduced a sleep stage classification method using EEG spectrograms [16]. Their model, EEGSNet, used CNNs to extract time and frequency features from the spectrogram and two Bi-LSTMs to capture transition patterns between features from adjacent epochs, facilitating accurate sleep stage classification. Their method achieved the highest accuracy of 94.17% and an F1-score of 70.16% for the N1 stage. Similarly, Almutairi et al. (2021) proposed three architectures for detecting obstructive sleep apnea (OSA) from ECG signals [43]: CNN, CNN + LSTM, and CNN + GRU. Using consecutive R-R intervals and QRS complex amplitudes as inputs, their

results showed that the CNN with LSTM outperformed the other models, achieving an average classification accuracy of 89.11% for OSA detection. These studies, which combine the features of CNNs and LSTMs, have enhanced machine-based sleep scoring and sleep disorder diagnosis, bringing their performance somewhat closer to that of human scoring. However, due to their reliance on sequential processing and serialization, they still face challenges with training efficiency and parallelization.

Hybrid models that combine CNNs with transformers have gained attention in recent studies for sleep stage classification [1,15,18,19]. For instance, Yao et al. (2023) proposed a sleep stage classification approach that leverages single-channel EEG, employing four convolutional filter layers for feature extraction and transformers to model temporal variations, achieving a testing accuracy of 80% [19]. These studies have achieved promising results using single-channel EEG signals. However, integrating signals from electromyography (EMG) and electrooculography (EOG) alongside EEG provide additional informative features. For example, the eye movements recorded by EOG are typically more frequent in stage W and REM but are less common in NREM stages [22,23]. This makes EOG features effective at distinguishing NREM stages from W and REM. Also, stage W is characterized by the highest muscle activity, whereas REM shows the lowest or absent EMG activity [22,23]. This makes EMG features useful for distinguishing between stage W and stage REM. Moreover, typically, muscles are temporarily paralyzed during REM sleep; however, in REM sleep behavior disorder (RBD), this paralysis is lacking [22]. This makes EMG features useful for distinguishing between RBD and healthy individuals or those with other disorders such as narcolepsy and PLMD [45]. This highlights the potential for developing CNN with transformer architectures that integrate EEG, EMG, and EOG signals by leveraging their distinct characteristics to enhance classification performance not only in sleep staging but also in sleep disorder diagnosis.

In this study, we propose a multimodal time-domain signal to time–frequency image conversion, where 30 s epochs of EEG, EOG, and EMG raw signals are individually transformed into spectrograms using short-time Fourier transform (STFT) and then converted into RGB spectrograms to prepare the input data for deep learning models. We also proposed three different architectures for classifying sleep stages and sleep disorders: (1) CNNs, (2) CNNs with a Bi-LSTM layer, and (3) CNNs with a transformer encoder. For each method, independent CNN layers with identical parameters (CNN architecture modules) extract unique features from each signal individually. The extracted features from all channels are then fused in a column-wise manner using a feature fusion block, which plays a key role in sleep stage and disorder classification. In Method-1, the fused features are directly classified. In Method-2, they are processed through a Bi-LSTM layer, while in Method-3, they are processed through a transformer encoder with a multi-head attention mechanism. Experiments were conducted using K-fold cross-validation to evaluate and compare the performances of the three model architectures against other advanced state-of-the-art methods. The CNN with a transformer encoder method achieved the best performance, with the highest average classification accuracy in detecting sleep stages and disorders using the CAP sleep dataset.

Overall, the main contributions, features, and advantages of our proposed model can be summarized as follows:

1. We adopt three novel architectures: (1) CNNs, (2) CNNs with Bi-LSTM, and (3) CNNs with a transformer encoder utilizing a multi-head attention mechanism. Each method extracts local features from RGB spectrogram images of EEG, EOG, and EMG signals separately, followed by column-wise feature fusion block to capture intra-epoch information. A residual connection is also applied to preserve the characteristics of the original joint feature maps and prevent gradient vanishing.

2. To the best of our knowledge, we investigate and innovate the CNNs with a transformer-based model architecture which is accurate and robust for classifying five sleep stages, particularly stage N1, and six sleep disorders by using RGB spectrogram images of EEG, EOG, and EMG signals of both patient and non-patient subjects.
3. We use modified L2 regularization to add a penalty term to the CCE loss to prevent overfitting. Moreover, our robust method classifies sleep stages and disorders in the elderly and could serve as a stepping stone for future research and a potential alternative to questionnaire-based diagnostic tools.

The rest of this article is organized as follows: Section 2 describes the dataset, the proposed methods, and their processing steps. Section 3 presents the ablation study and experimental results. Section 4 provides a comparison, discussion of the results, and future directions. Finally, Section 5 concludes the paper with a summary.

2. Materials and Methods

2.1. Data Description

This study used the publicly available PhysioNet database called the Cyclic Alternative Pattern (CAP) sleep database [46] to evaluate the proposed model architectures. The PSG recordings were collected by the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy, available through the PhysioNet CAP sleep database [47,48]. This dataset has 108 PSG recordings. This database was originally recorded from a few minutes before the start of sleep to a few minutes after awaking. The length of each recording data is at least 8 hours. This database is a multimodal database containing at least 3 EEG channels, 2 EOG channels, EMG of the submentalis muscle, bilateral anterior tibial EMG, ECG, and other respiration signals. The trained expert neurologists at the sleep center provided the annotations for all participants, and the recordings were labeled with six sleep stages, wake (W), sleep stages (S1 to S4), rapid eye movement (REM), and MT = body movement, for each 30 s epoch according to the Rechtschaffen and Kales rules [2]. The data includes additional details besides sleep stages, such as body position (left, right, prone, or supine), time of day, events, and durations. The dataset contains recordings of sixteen healthy subjects and ninety-two pathological recordings: forty from patients diagnosed with nocturnal frontal lobe epilepsy (NFLE), twenty-two with REM behavior disorder (RBD), ten with periodic leg movements (PLMs), nine with insomnia, five with narcolepsy, four with sleep-disordered breathing (SDB), and two with bruxism. We used 83 recordings containing EEG, EOG, and EMG signals. Other sleep recordings were excluded from our analysis due to the absence of either signal modality. Out of the 83 subjects, 59% are male (49) and 41% are female (34). The average ages of subjects in the healthy, insomnia, narcolepsy, NFLE, PLM, and RBD categories are 32, 60, 32, 31, 55, and 71 years, respectively. The age of all subjects ranges from 16 to 82 years, with an average of 46 years—males (aged 49 ± 21 years) and females (aged 42 ± 17 years). The CAP sleep database primarily reports data on elderly subjects. Therefore, this work can play a significant role in detecting sleep stages and disorders in older subjects [46].

2.2. Data Preparation

In this study, the S3 and S4 stages were combined into the stage called N3, as per the AASM sleep scoring rules [3]. For each subject, we employed the C4-A1 [20,28,30,32], ROC-LOC [28,30,32,33], and EMG1-EMG2 [28,30,32–34] channels, which have also been used in the recent studies mentioned and have consistently demonstrated good performance. These three channels are commonly included in most CAP PSG recordings. All three data modalities (EEG, EOG, and EMG) were uniformly sampled at 512 Hz. Furthermore, their bandwidths were standardized to 0.2–128 Hz. The signals were normalized by mapping

their mean to 0 and their standard deviation to 1 [49]. According to the guidelines outlined by AASM [3] and clinical protocols, we segmented the EEG, EOG, and EMG signals from each participant into 30 s epochs, where each epoch represents a distinct sleep stage, as seen in Table 1, which shows the epochs corresponding to different sleep stages and disorders. Each epoch consists of a duration equivalent to 30 times the sampling frequency, resulting in 15,360 samples per epoch in every channel signal.

Table 1. Total epoch distribution for the data used in this study.

Sleep Stage	Healthy	Sleep Disorder					Total Epochs	
		NFLE	RBD	PLM	Insomnia	Narcolepsy	Number	%
W	451	3784	5116	1407	3383	1321	15,462	18.34
N1	280	1558	966	284	306	302	3696	4.38
N2	2172	13,807	6886	2845	2434	1707	29,851	35.4
N3	1757	9238	5218	1943	1171	1044	20,371	24.16
R	1409	6431	3415	1341	1081	1257	14,934	17.71
Total	6069	34,818	21,601	7820	8375	5631	84,314	
%	7.2	41.3	25.62	9.27	9.93	6.68		

2.3. RGB Spectrogram Image Representation

The three methods process RGB spectrogram images as input. To achieve this, each 30 s signal epoch (e.g., EEG, EOG, or EMG) was transformed into a spectrogram using the short-time Fourier Ttransform (STFT) [50,51] with a window size of 0.6 s and a 50% overlap. A Hanning window and a 512-point fast Fourier transform (FFT) were applied. The resulting spectrogram $S \in \mathbb{R}^{F \times T}$, with $F = 257$ frequency bins and $T = 101$ time frames, was log-transformed. The log-spectrograms of each signal were min-max normalized and converted into RGB images using the Jet colormap. Finally, the images were resized to $92 \times 61 \times 3$ in preparation for subsequent processing. A sensitivity analysis over a small grid of window, overlap, and FFT configurations confirmed that the chosen STFT parameters preserve key physiologically relevant sleep phenomena, and the model performance remains stable even when higher-resolution inputs are used. Figure 1 presents examples of the original waveforms for EEG, EOG, and EMG, along with their corresponding RGB spectrogram images for insomnia (ins2.edf) patients.

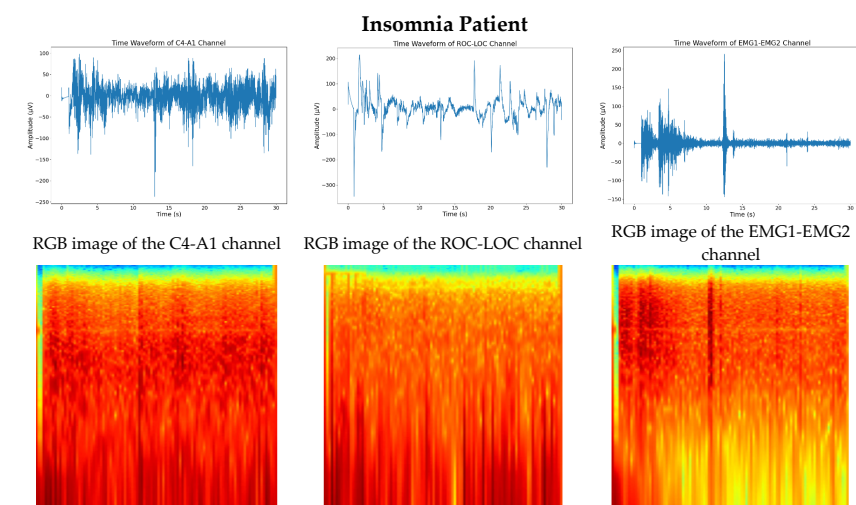


Figure 1. Examples of the original waveforms for EEG, EOG, and EMG, along with their corresponding RGB spectrogram images for insomnia (ins2.edf) patients.

2.4. Proposed Methodology

We propose three different methods for classifying sleep stages and sleep disorders: (1) CNNs, (2) CNNs with Bi-LSTM, and (3) CNNs with a transformer. The three methods are presented as part of an ablation study, which shows the contribution of each component (e.g., Bi-LSTM or transformer) to the performance improvements. Comparing all three systematically demonstrated the added value of temporal modeling (Bi-LSTM) and attention mechanisms (transformer), thereby justifying the performance of the final model. In all methods, the models are trained from scratch, enabling a thorough evaluation of their performance and adaptability to the classification tasks.

(1) Method-1: CNN-Based Model

Method-1 involves training a CNN-based model from scratch to classify sleep stages and disorders, as shown in Figure 2. Method-1 consists of two blocks with identical parameters: one for sleep stage classification and the other for sleep disorder classification. Each block is responsible for extracting single-channel features using CNNs, applying a multichannel feature fusion mechanism, and performing classification. It consists of three identical independent modules: Module 1 extracts features from EEG images, Module 2 from EOG images, and Module 3 from EMG images. Each module receives its respective RGB spectrogram image as input and independently extracts features. Each feature extraction module consists of 5 convolutional layers, and their corresponding hyperparameters are detailed in Table 2. Among them, the convolution operation Conv2D₁ filters the input image using 32 kernels, each with a size of 3×3 . The rectified linear unit (ReLU) activation function is applied to each output of the convolution operation to introduce non-linearity and enhance the model's ability to capture complex patterns within the signals. Batch normalization normalizes mini-batches to stabilize activations, prevent vanishing or exploding gradient issues, enhance feature learning, and reduce overfitting. MaxPool2D₁ (2×2) downsamples the input image by selecting the maximum value in each 2×2 window to preserve the essential features. Dropout (0.3) randomly removes 30% of input units to prevent the model from becoming overly reliant on specific features, improve model generalization and reduce overfitting. Recent studies relied on the simple concatenation of multimodal PSG data, which did not mine the full available information within the single epoch. In contrast, the single-channel features (EEG, EOG, EMG) extracted by the three independent modules are fused in a column-wise manner and then passed through two fully connected layers, each with 512 units and ReLU activation function, followed by a Softmax layer. Block 1 classifies five sleep stages, while Block 2 classifies six sleep disorders.

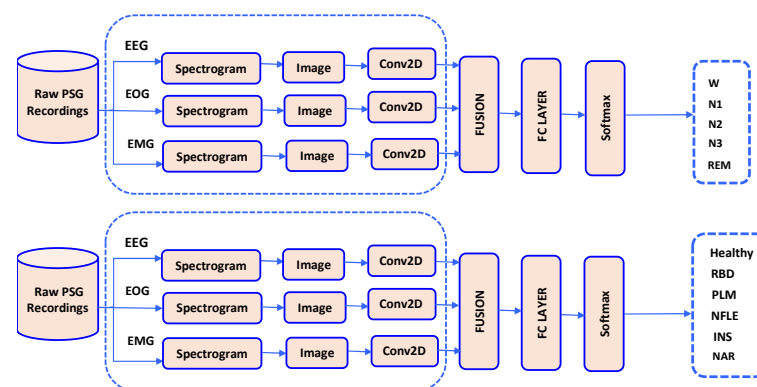


Figure 2. Method-1: CNNs for sleep stages and disorders classification using bio-signal Images.

Table 2. Hyperparameters of each CNN.

Layer	Filters	Kernel Size	Output Dimension
Input			$92 \times 61 \times 3$
Conv2D1 + ReLu1 + BatchNorm1 + MaxPool2D1 (2×2) + Dropout1 (0.3)	32	3×3	$46 \times 31 \times 32$
Conv2D2 + ReLu2 + BatchNorm2 + MaxPool2D2 (2×2) + Dropout2 (0.4)	64	3×3	$23 \times 16 \times 64$
Conv2D3 + ReLu3 + BatchNorm3 + MaxPool2D3 (2×2) + Dropout3 (0.4)	128	3×3	$12 \times 8 \times 128$
Conv2D4 + ReLu4 + BatchNorm4 + MaxPool2D4 (2×2) + Dropout4 (0.3)	256	3×3	$6 \times 4 \times 256$
Conv2D5 + ReLu5 + MaxPool2D5	256	3×3	$3 \times 2 \times 256$

(2) *Method-2: CNN with LSTM-Based Model*

Method-2 involves training a CNN with Bi-LSTM from scratch to classify sleep stages and disorders. In this approach, a Bi-LSTM layer is added to the CNN-based method described in Method-1, as shown in Figure 3. A Bi-LSTM layer is then applied to the fused features to capture temporal dependencies and transition patterns within the intra-epoch joint features before the final classification. RNNs learn sequence transitions between sequence data but suffer from gradient exploding and vanishing in long sequences. LSTMs address this issue by preserving relevant information and filtering out irrelevant data [16]. The LSTM unit consists of cell states, hidden states, an input feature, and three gates. The forget gate, using a sigmoid function, selectively retains or discards information from the previous cell state. The input gate, utilizing both sigmoid and tangent functions, determines what new information should be added to the cell state. The output gate, also employing sigmoid and tangent functions, regulates which information from the cell state, previous hidden state, and input feature contributes to generating the new hidden state, which serves as the final output. LSTMs effectively handle long-term tasks but capture only forward information. In this study, we adopted Bi-LSTM [16], which enhances this by processing sequences in both forward and backward directions. Each layer has a sequence length of 10, with each LSTM cell having a hidden size of 128. The outputs from both directions were then concatenated and passed through two fully connected layers, each with 512 units and a ReLu activation function, followed by a Softmax layer. Block 1 classifies five sleep stages while Block 2 classifies six sleep disorders.

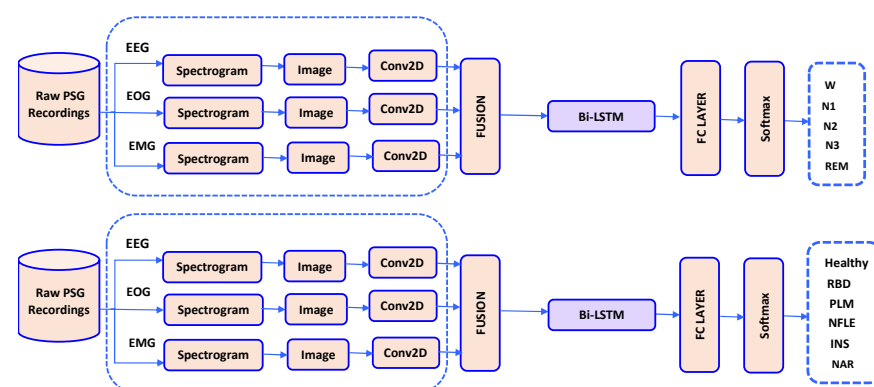


Figure 3. Method-2: CNN + Bi-LSTM for sleep stages and disorders classification using bio-signal images.

(3) *Method-3: CNN with Transformer-Based Model*

In Method-3, we replaced the Bi-LSTM with the transformer encoder (see Figure 4). In this approach, a transformer encoder is added to the CNN-based method described in

Method-1, as shown in Figure 5. The transformer encoder is applied to process joint features, enhance feature fusion, and capture intra-epoch complex contextual dependencies among feature tokens in all positions. After the transformer encoder, a residual connection [1] was used, where the joint features before being fed into transformer encoder were added to the output features from the transformer encoder, which preserves the input joint features and completely prevents the risk of gradient vanishing.

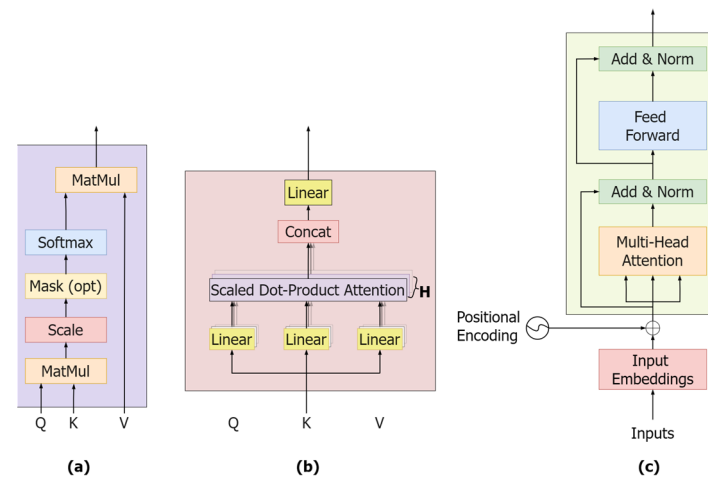


Figure 4. (a) Scaled dot-product attention, (b) multi-head attention, and (c) transformer encoder.

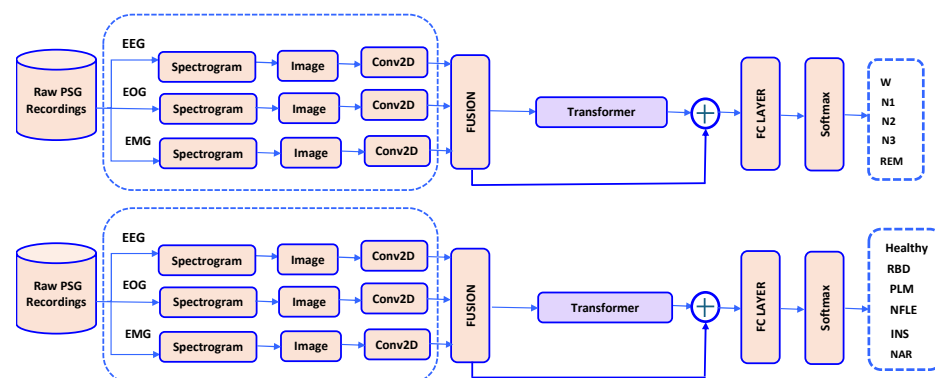


Figure 5. Method-3: CNN + transformer for sleep stages and disorders classification using bio-signal images.

The transformer [41] consists of an encoder and decoder, with the decoder focusing on generation tasks. In this study, only the encoder part is used, which includes two key components: multi-head attention and feedforward networks (see Figure 4). The multi-head attention mechanism (MHA) module employs the scaled dot-product attention mechanism, enabling the model to capture the dependencies among token vectors at different positions of the input joint feature epoch when making predictions. MHA operates based on queries (Q), keys (K), and values (V), which are processed by H scaled dot-product attention modules [41]. The outputs from these H attention heads are then concatenated and transformed through a linear projection to generate the final attention output [41].

The joint features of each epoch $O \in \mathbb{R}^{T \times CF}$ are treated as a sequence of T tokens, where each token is a feature vector of dimension CF. The O is segmented into T tokens of features, meaning that each token vector $\in \mathbb{R}^{1 \times CF}$ consists of features from the same time step. The features of each token vector are treated as token embedding values, which serve as the inputs for our model to learn and process. As the transformer encoder lacks

inherent order information, relative positional encoding is applied to track the order of the embedded tokens and introduce temporal dependencies.

$$\tilde{O} = O + PE \quad (1)$$

where $PE \in \mathbb{R}^{T \times CF}$ denotes the positional encoding matrix and the sine function is applied to even positions, while the cosine function is used for odd positions [41].

$$PE(pos, 2i) = \sin(pos/10000^{\frac{2i}{CF}}) \quad (2)$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{\frac{2i}{CF}}) \quad (3)$$

In this context, ‘pos’ refers to the token position, where $0 \leq pos < T$, and T is the number of tokens. The indices $2i$ or $2i + 1$ represent the dimension of the input, $0 \leq 2i < 2i + 1 < CF$, and CF is the number of features per token. The $(pos, 2i)$ and $(pos, 2i + 1)$ entries refer to specific positions in the positional encoding matrix, where values are computed using sine and cosine functions. Each of these entries follows a sinusoidal pattern across time positions.

We applied the principles of the multi-head self-attention mechanism [41] to capture intra-epoch patterns, dependencies, and similarities among the feature tokens, including their internal contextual relationships. To achieve this, position-encoded inputs are linearly transformed using learnable weight matrices and biases to generate the corresponding queries, keys, and values for all tokens. The scaled-dot product between queries and keys is then computed to obtain similarity scores, which are passed through a softmax function to produce attention weights. These weights determine the contribution of each token to the representation of every other token. The resulting attention probabilities are then used to weight the corresponding value vectors. Finally, the weighted values are summed to produce the self-attention outputs, which integrate information from all input tokens based on their contextual similarities. Scaled dot-product attention is performed in parallel across the mapped queries, keys, and values. Each self-attention cell module, known as an attention head, performs these computations on queries, keys, and values, and captures features from various learned representations in this study. The outputs from the H attention heads are then concatenated and passed through a linear projection to generate the final attentive output. All of these steps can be formulated as follows:

The features of each token vector are linearly mapped to obtain queries (Q), keys (K), and values (V), as follows

$$Q_i = \tilde{O}W_i^Q, \quad 1 \leq i \leq H \quad (4)$$

$$K_i = \tilde{O}W_i^K, \quad 1 \leq i \leq H \quad (5)$$

$$V_i = \tilde{O}W_i^V, \quad 1 \leq i \leq H \quad (6)$$

Then, they are fed into H scaled dot-product attention modules for further processing.

$$X_i = \mathbf{Attention}(Q_i, K_i, V_i) = \mathbf{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{CF}}\right) V_i \quad (7)$$

The outputs from H attention heads are concatenated and projected as follows:

$$M_s = \mathbf{Concat}(X_1, X_2, X_3, \dots, X_H)W^s \quad (8)$$

where H is the total number of attention heads. The $\tilde{O}$ contains the position-encoded values of the tokens. W_i^Q , W_i^K , and $W_i^V \in \mathbb{R}^{(CF \times (\frac{CF}{H}))}$ are the learnable weight matrices. Q_i , K_i ,

and $V_i \in \mathbb{R}^{(TX(\frac{CF}{H}))}$ are the mapped queries, keys, and values. $X_i \in \mathbb{R}^{(TX(\frac{CF}{H}))}$, X_i is the attention output of the i th attention head. $W^s \in \mathbb{R}^{CF \times CF}$ are the learnable weight matrices. $M_s \in \mathbb{R}^{TXCF}$, M_s is the concatenated output from all attention heads.

The multi-head attention (MHA) modules capture relationships among the joint features within the sleep stage and sleep disorder epoch. The encoder also incorporates residual connections and normalization layers, which are essential for maintaining training stability and enhancing convergence speed. The second component of the transformer encoder is the feedforward neural network, which performs nonlinear transformations and extracts meaningful features from the (MHA) output. It consists of one fully connected (FC) layer with a ReLU activation function. The overall process can be formulated as follows:

$$M_s = \text{MHA}(\tilde{O}) \quad (9)$$

$$L_N = \text{LayerNorm}(\tilde{O} + M_s) \quad (10)$$

$$Z_{FFNN} = \text{ReLU}(L_N W + b) \quad (11)$$

$$L_O = \text{LayerNorm}(L_N + Z_{FFNN}) \quad (12)$$

where Z_{FFNN} represents the output of the feedforward neural network, and W and b are weights and bias in the FC layer. L_N and L_O are the outputs of the first and second normalization layers, respectively. We used one transformer encoder. In the transformer encoder, we used $H = 8$ attention heads, and the feedforward layer contained 512 hidden units. The fully connected (FC) layer also had 512 hidden units and a dropout rate of 0.1 was applied throughout the transformer, including in the self-attention and feedforward layers, while a dropout rate of 0.5 was used after the FC layer. Finally, to accurately classify sleep stages and disorder scores, a dense layer with a softmax activation function was used to determine the most probable categories. Block 1 classifies five sleep stages, and Block 2 classifies six sleep disorders.

3. Experiments

3.1. Parameter Setting

In our models, we applied five-fold cross-validation. To prevent model performance degradation caused by poor division of healthy and disordered patient data, we divide the dataset into three sets (training, validation, and testing). In each iteration, the validation set, composed of subjects from each group (healthy, NFLE, RBD, PLM, insomnia, and narcolepsy), was left out from the training set to form the validation set. All folds were constructed in a subject-independent manner, ensuring that all epochs from each subject were assigned to only one fold (train/validation/test), and each fold contained subjects from every diagnostic group to maintain stratification by diagnosis. The objective function used is the categorical cross-entropy (CCE) loss function, and modified L2 regularization was used to add a penalty term to the CCE loss to prevent overfitting, where the regularization factor (weight decay) was set to $\lambda = 10^{-3}$. The models were optimized using the ADAM with $\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The training process was conducted with a batch size of 64 samples. Early stopping was applied to halt training if validation loss did not improve for 20 consecutive epochs. The learning rate was halved (factor = 0.5) if no improvement in validation loss occurred for 15 consecutive epochs (patience = 15), with a minimum value of 10^{-6} . This study used 100 training epochs. All these parameters were set after multiple rounds of parameter tuning during the experiments.

3.2. Performance Evaluation Metrics

To evaluate the performance of our models, we employed the key evaluation metrics commonly used in the recent studies reported, allowing for a comprehensive analysis of our models' strengths and limitations, as well as a fair comparison with existing work. We calculated true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs), which were then used to derive the following metrics: per-class precision (Pre), per-class sensitivity (Sen), per-class specificity (Spe), per-class F1 score (F1), accuracy (Acc), Cohen's kappa (κ) [30,52,53], macro-averaged F1-score (MF1), and the macro-averaged G-mean (MGm) [15]. The Pre, Sen, Spe, F1, Acc, κ , MF1, and MGm are computed as follows:

$$\text{Pre}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (13)$$

$$\text{Sen}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (14)$$

$$\text{Spe}_i = \frac{\text{TN}_i}{\text{TN}_i + \text{FP}_i} \quad (15)$$

$$\text{F1}_i = 2 \times \frac{\text{Pre}_i \times \text{Sen}_i}{\text{Pre}_i + \text{Sen}_i} \quad (16)$$

$$\text{Gm}_i = \sqrt{\text{Spe}_i \times \text{Sen}_i} \quad (17)$$

$$\text{MF1} = \frac{1}{M} \sum_{i=1}^M \text{F1}_i = \frac{1}{M} \sum_{i=1}^M 2 \times \frac{\text{Pre}_i \times \text{Sen}_i}{\text{Pre}_i + \text{Sen}_i} \quad (18)$$

$$\text{MGm} = \frac{1}{M} \sum_{i=1}^M \text{Gm}_i = \frac{1}{M} \sum_{i=1}^M \sqrt{\text{Spe}_i \times \text{Sen}_i} \quad (19)$$

$$\text{Acc} = \frac{\sum_{i=1}^M \text{TP}_i}{N} \quad (20)$$

$$P_o = \text{Acc} = \frac{\sum_{i=1}^M \text{TP}_i}{N} \quad (21)$$

$$P_e = \frac{\sum_{i=1}^M (\text{TP}_i + \text{FP}_i) \times (\text{TP}_i + \text{FN}_i)}{N^2} \quad (22)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (23)$$

where i denotes a sleep stage or sleep disorder, P_o represents the proportion of observed agreements between predictions and true labels, (P_e) represents the expected agreement by chance based on class distributions, N represents the total number of samples across sleep stages or sleep disorders, and M refers to the number of classes for sleep stages or sleep disorders.

3.3. Experimental Results and Ablation Study

The sleep staging and disorders performance: To emphasize the performance of the final CNN-transformer model, we conducted three experiments with the CNN, CNN with Bi-LSTM, and CNN with transformer to evaluate both sleep staging and disorder classification. Tables 3–10 illustrate the normalized confusion matrices obtained from test subjects from the CAP sleep database, offering insight into the distribution of correctly and incorrectly classified samples, thereby interpreting the performance of the models. In Tables 3–10, the rows represent the true labels, and the columns represent the predicted labels. Each cell in the confusion matrix indicates the number of samples for which a true

label was predicted as a particular stage/disorder. The diagonal numbers represent the correctly classified samples, while off-diagonal numbers represent misclassifications. On the right side of Tables 3–10, we provide the per-class metrics, and below these tables, we highlight the overall metrics.

Table 3. Performance comparison of two-channel inputs using the baseline model: Method-1 for sleep stages.

Input Channels	Per-Class F1-Score					Overall Metrics			
	W	N1	N2	N3	REM	Acc.	κ	MF1	MGm
C4-A1 + ROC-LOC	92.1	51.5	81.8	85.8	81.1	83.7	78.4	78.5	86.4
C4-A1 + EMG1-EMG2	90.7	47.9	80.8	84.8	80.9	82.6	76.9	77.0	85.6
ROC-LOC + EMG1-EMG2	83.3	36.3	68.0	74.2	75.7	73.1	64.3	67.5	78.6

Table 4. Performance comparison of two-channel inputs using the baseline model: Method-1 for sleep disorders.

Input Channels	Per-Class F1-Score					Overall Metrics				
	Healthy	INS	NARCO	NFLE	PLM	RBD	Acc.	κ	MF1	MGm
C4-A1 + ROC-LOC	94.7	95.8	94.5	93.8	94.0	96.1	94.9	93.9	94.8	96.9
C4-A1 + EMG1-EMG2	92.9	96.2	91.5	92.1	93.7	96.9	94.0	92.8	93.9	96.3
ROC-LOC + EMG1-EMG2	92.3	92.4	93.5	93.6	89.0	92.3	92.3	90.8	92.2	95.3

Table 5. Confusion matrix of the baseline model (Method-1) for sleep stages.

	Predicted					Per-Class Metrics			
	W	N1	N2	N3	REM	Pre	Sen	F1	Gm
W	93.9	14.6	2.6	0.3	2.4	93.9	91.3	92.6	94.8
N1	1.9	53.7	2.1	0.1	2.3	53.7	58.9	56.2	76.0
N2	2.7	21.4	84.9	13.8	8.0	84.9	82.9	83.9	87.4
N3	0.2	0.0	6.1	85.0	0.4	85.0	89.8	87.3	92.7
REM	1.3	10.3	4.3	0.8	86.9	86.9	86.3	86.6	91.7
Acc. = 86.0		κ = 81.3			MF1 = 81.3			MGm = 88.5	

Table 6. Confusion matrix of the baseline model (Method-1) for sleep disorders.

	Predicted						Per-Class Metrics			
	Healthy	INS	NARCO	NFLE	PLM	RBD	Pre	Sen	F1	Gm
Healthy	96.6	0.4	0.7	0.7	0.7	0.4	96.6	96.9	96.8	98.2
INS	0.6	97.6	1.2	0.6	1.1	0.6	97.6	96.8	97.2	98.1
NARCO	1.0	0.7	95.9	0.8	1.4	0.6	95.9	95.9	95.9	97.5
NFLE	0.7	0.4	0.7	96.7	0.7	0.6	96.7	97.0	96.8	98.2
PLM	0.5	0.5	0.8	0.7	95.2	0.7	95.2	95.7	95.5	97.5
RBD	0.6	0.4	0.7	0.5	0.9	97.1	97.1	97.3	97.2	98.3
Acc. = 96.6		κ = 96.0				MF1 = 96.6		MGm = 98.0		

Table 7. Confusion matrix of Method-2 for sleep stages.

	Predicted					Per-Class Metrics			
	W	N1	N2	N3	REM	Pre	Sen	F1	Gm
W	94.5	15.5	2.7	0.4	2.4	94.5	91.1	92.8	94.7
N1	1.9	55.5	2.1	0.1	2.2	55.5	58.9	57.2	76.1
N2	2.4	20.2	88.1	12.5	7.3	88.1	84.3	86.1	88.9
N3	0.2	0.0	3.9	86.6	0.4	86.6	93.4	89.8	94.7
REM	1.0	8.8	3.2	0.4	87.7	87.7	89.9	88.8	93.6
Acc. = 87.8		$\kappa = 83.8$			MF1 = 82.9			MGm = 89.6	

Table 8. Confusion matrix of Method-2 for sleep disorders.

	Predicted						Per-Class Metrics			
	Healthy	INS	NARCO	NFLE	PLM	RBD	Pre	Sen	F1	Gm
Healthy	98.2	0.2	0.4	0.6	0.4	0.3	98.2	97.9	98.1	98.8
INS	0.5	98.8	0.5	0.4	0.9	0.3	98.8	98.0	98.4	98.8
NARCO	0.3	0.2	97.7	0.2	0.6	0.1	97.7	98.8	98.2	99.1
NFLE	0.4	0.2	0.5	98.1	0.5	0.5	98.1	97.9	98.0	98.8
PLM	0.2	0.4	0.6	0.5	97.0	0.5	97.0	96.9	96.9	98.2
RBD	0.4	0.2	0.3	0.2	0.6	98.3	98.3	98.5	98.4	99.1
Acc. = 98.1		$\kappa = 97.7$				MF1 = 98.0		MGm = 98.8		

Table 9. Confusion matrix of the Method-3 for sleep stages.

	Predicted					Per-Class Metrics			
	W	N1	N2	N3	REM	Pre	Sen	F1	Gm
W	97.8	11.3	1.6	0.1	0.5	97.8	95.2	96.5	97.3
N1	0.7	69.1	1.6	0.1	1.0	69.1	73.3	71.1	85.1
N2	1.3	17.0	94.6	4.2	0.3	94.6	94.7	94.6	95.9
N3	0.1	0.0	1.8	95.4	0.1	95.4	97.0	96.2	97.8
REM	0.1	2.6	0.4	0.2	98.1	98.1	98.1	98.1	98.9
Acc. = 95.2		$\kappa = 93.6$			MF1 = 91.3			MGm = 95.0	

Table 10. Confusion matrix of the model (Method-3) for sleep disorders.

	Predicted						Per-Class Metrics			
	Healthy	INS	NARCO	NFLE	PLM	RBD	Pre	Sen	F1	Gm
Healthy	99.5	0.0	0.1	0.2	0.0	0.0	99.5	99.5	99.5	99.7
INS	0.1	99.7	0.4	0.1	0.5	0.2	99.7	99.1	99.4	99.5
NARCO	0.1	0.2	99.0	0.2	0.4	0.1	99.0	99.1	99.0	99.4
NFLE	0.2	0.0	0.2	99.2	0.1	0.1	99.2	99.4	99.3	99.6
PLM	0.0	0.1	0.1	0.2	98.6	0.2	98.6	99.1	98.8	99.4
RBD	0.1	0.0	0.2	0.1	0.4	99.4	99.4	99.4	99.4	99.6
Acc. = 99.3		κ = 99.1				MF1 = 99.2		MGm = 99.6		

(1) Experiment 1

To investigate and quantify the contribution of fusing features from three channels, in experiment 1 we first evaluated feature fusion for all possible channel pairs using

Method-1, shown in Figure 2, for both sleep staging and disorder classification to analyze their performance.

Among the two-channel inputs presented in Tables 3 and 4, the C4-A1 + ROC-LOC demonstrated the best performance, achieving an accuracy of 83.7%, Cohen's kappa of 78.4%, MF1 of 78.5%, and MGm of 86.4% for sleep staging, and an accuracy of 94.9%, Cohen's kappa of 93.9%, MF1 of 94.8%, and MGm of 96.9% for disorder detection. Using two-channel inputs, Method-1 performs better in sleep disorder detection than in sleep stage detection, with improvements of more than 11.2% in accuracy, 15.5% in Cohen's kappa, 16.3% in MF1, and 10.5% in MGm. This performance difference is primarily attributed to the greater class imbalance present in sleep stage samples compared to sleep disorder samples. Also, spectral patterns exhibit more distinguishable frequency characteristics in sleep disorders than in sleep stages, contributing to improved sleep disorder detection performance.

For channel fusion, which also aligns with the clinical practice of sleep scoring and disorder diagnosis and requires the integration of multiple channels, we evaluated three approaches: data-level fusion, feature-level fusion, and decision-level fusion. Among these, feature-level fusion outperformed the others, an accuracy of 86%, Cohen's kappa of 81.3%, MF1 of 81.3%, and MGm of 88.5% for sleep staging, and an accuracy of 96.6%, Cohen's kappa of 96%, MF1 of 96.6%, and MGm of 98% for disorder detection. A comparison of the experimental results using the baseline model, Method-1, shows that three-channel inputs (Tables 5 and 6) result in better performance than two-channel inputs (Tables 3 and 4). For sleep staging, accuracy improved by 2.3%, Cohen's kappa by 2.9%, MF1 by 2.8%, and MGm by 2.1%. For sleep disorder detection, accuracy increased by 1.7%, Cohen's kappa by 2.1%, MF1 by 1.8%, and MGm by 1.1%.

(2) Experiment 2

In Experiment 2, we enhanced the performance of Method-1 by introducing Method-2, which integrates a CNN followed by the Bi-LSTM layer, as shown in Figure 3. The addition of the Bi-LSTM layer enables the model to capture intra-epoch temporal dependencies and transition patterns within the extracted joint features, thereby improving the discriminative capacity of the model prior to classification.

From the results of Method-2, presented in Tables 7 and 8, the CNNs with Bi-LSTM achieved an accuracy of 87.8%, Cohen's kappa of 83.8%, MF1 of 82.9%, and MGm of 89.6% for sleep staging, and an accuracy of 98.1%, Cohen's kappa of 97.7%, MF1 of 98.0%, and MGm of 98.8% for disorder detection. It can be observed that the Method-2 CNNs with Bi-LSTM outperformed the standalone CNNs in Method-1 with improvements of 1.8% in accuracy, 2.5% in Cohen's kappa, 1.6% in MF1, and 1.1% in MGm in sleep staging. For sleep disorder detection, accuracy increased by 1.5%, Cohen's kappa by 1.7%, MF1 by 1.4%, and MGm by 0.8%. The diagonal elements of the confusion matrices in Tables 7 and 8 also indicate notable improvements, particularly for stage N2, as well as for narcolepsy and PLM disorders, with respective increases of 3.2%, 1.8%, and 1.8%. These results highlight the Bi-LSTM's ability to strongly capture intra-epoch patterns such as sleep spindles, K-complexes, minimal eye movements, and low muscle tone, which are distinctive features of stage N2.

(3) Experiment 3

In experiment 3, we improved Method-2 by replacing the Bi-LSTM with a transformer encoder, resulting in Method-3, which integrates a CNN with a transformer encoder, as shown in Figure 5. The transformer effectively captured intra-epoch temporal dynamics and contextual dependencies among learned feature maps extracted by CNN. A residual connection was added to preserve the original joint features and the risk of gradient

vanishing. We performed a sensitivity analysis on the number of attention heads (H) in the transformer encoder and the number of transformer encoder layers in Method-3 for both sleep staging and disorder detection. Adding a single transformer encoder layer improved performance in both tasks. However, increasing the number of layers to two, three, or four did not yield further improvements and instead led to a decline in performance and increased model complexity. Similarly, performance improved slightly as the number of heads increased from 2 and 4 to 8, but began to decline when using 16 or 32 heads. In our preliminary results for Method-3, the performance in sleep disorder detection was good compared to that in sleep stage classification. Motivated by the approach in [34], we briefly evaluated Method-3 using weights internally trained, specifically for the sleep staging task. This internal initialization boosted the performance; however, it was used for exploratory analysis. All the reported results reflect models trained entirely from scratch.

From the results of Method-3, presented in Tables 9 and 10, the CNNs with transformer achieved an accuracy of 95.2%, Cohen's kappa of 93.6%, MF1 of 91.3%, and MGm of 95.0% for sleep staging, and an accuracy of 99.3%, Cohen's kappa of 99.1%, MF1 of 99.2%, and MGm of 99.6% for disorder detection. It can be observed that the Method-3 CNNs with transformers outperformed the CNNs with Bi-LSTM in Method-2 with improvements of 7.4% in accuracy, 9.8% in Cohen's kappa, 8.4% in MF1, and 5.4% in MGm in sleep staging. For sleep disorder detection, accuracy increased by 1.2%, Cohen's kappa by 1.4%, MF1 by 1.2%, and MGm by 0.8%. The diagonal elements of the confusion matrices in Tables 9 and 10 also indicate notable improvements. A detailed analysis of Table 9 shows that four sleep stages (W, N2, N3, and REM) were well classified. However, the classification of stage N1 posed a challenge due to its limited number of samples and strong similarity to the neighboring stages, Wake and N2, which have larger sample sizes. As N1 lies between them, there is a high probability of it being misclassified as either W or N2 [15,17]. However, the results of the CNN with transformer model highlight its advantages and robustness in capturing intra-epoch patterns and handling imbalanced datasets in both sleep staging and disorder detection.

4. Comparison and Discussion

The results of the ablation experiments are summarized in the bar graphs shown in Figures 6 and 7, where each bar corresponds to a different classification approach tested in our experiments for both sleep stages and disorders. The multichannel feature fusion CNN with multi-head attention outperforms all the other methods—including C4-A1 + ROC-LOC CNN, C4-A1 + EMG1-EMG2 CNN, ROC-LOC + EMG1-EMG2 CNN, C4-A1 + ROC-LOC + EMG1-EMG2 CNN Method-1 (Concatenation), and C4-A1 + ROC-LOC + EMG1-EMG2 Method-2 (CNN+Bi-LSTM)—in terms of both per-class metrics and overall metrics.

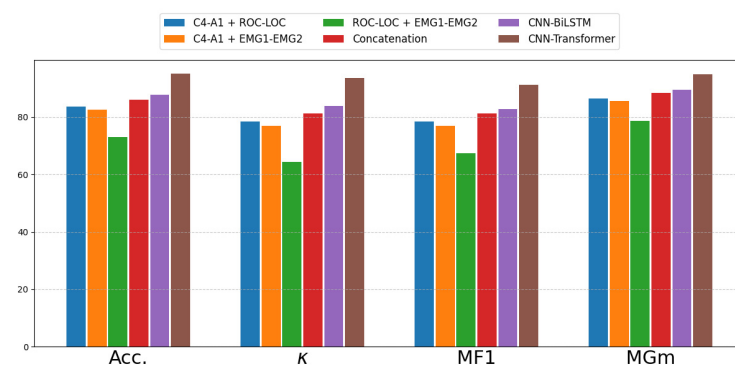


Figure 6. Performance comparison of the methods tested for sleep stages classification.

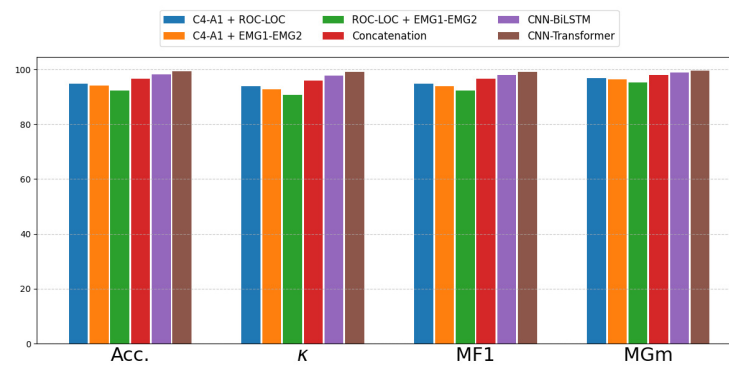


Figure 7. Performance comparison of the methods tested for sleep disorders classification.

The addition of ROC-LOC and EMG1-EMG2 to C4-A1 led to clear sleep stage improvements in the overall performance metrics. According to the per-class F1-scores, in both Method-1 and Method-3, the contributions of ROC-LOC and EMG1-EMG2 features were particularly significant for enhancing the classification performance of stage N1 and REM; specifically, in Method-1, improvements of 4.7% and 5.5% were observed for N1 and REM, respectively, while Method-3 showed increases of 14.9% for N1 and 11.5% for REM. These results are expected, as stage N1 is not only associated with alpha and high-amplitude theta brain waves but also characterized by slow eye movements and moderate muscle tone. Similarly, REM sleep is distinguished by rapid eye movements in various directions and relaxed muscle activity, making the ROC-LOC and EMG1-EMG2 features particularly informative for its detection [25]. The number of misclassifications between N3 and N1 or W, as well as between N3 and REM, is nearly zero. This is a good indication of model performance, because N3 represents deep sleep, whereas REM, N1, and W correspond to REM, light sleep, and Wakefulness, respectively. Additionally, the other stages also showed slight but consistent improvements in their classification from Method-1 to Method-3, confirming the potential correlation and complementarity of features across the EEG, EOG, and EMG channels, in a manner similar to how sleep experts manually assess sleep data [35].

As shown above, the features across EEG, EOG, and EMG help in the detection of the REM stage, which is essential for diagnosing sleep disorders, particularly narcolepsy and RBD [25]. Adding ROC-LOC and EMG1-EMG2 to C4-A1 improved the classification of narcolepsy and RBD. According to the per-class F1-scores, in Method-1, improvements of 1.4% and 1.1% were observed for narcolepsy and RBD, respectively, while Method-3 showed increases of 3.1% for narcolepsy and 2.2% for RBD. These results are expected due to the physiological traits of the disorders, as narcolepsy is associated with abnormal transitions into REM sleep. RBD, on the other hand, is characterized by abnormal muscle activity during REM sleep. The inclusion of EOG and EMG modalities enhances the model's ability to capture the characteristic features of these disorders. Other disorders also showed consistent classification improvements from method-1 to Method-3, highlighting the complementary value of multimodal features, akin to expert sleep scoring.

Lastly, we aim to compare the consistency of our results with previous state-of-the-art methods. Tables 11 and 12 present a comparative analysis of our CNN with transformer method against existing state-of-the-art approaches for sleep staging and sleep disorder classification, respectively. It can be observed from Tables 11 and 12 that the CNN with transformer method outperformed previous state-of-the-art methods. This outperformance is due to the strong capability of CNN, which excels in single-channel feature extraction, the effective fusion of multichannel features, and the use of multi-head attention mechanisms, which capture intra-epoch temporal dynamics and contextual dependencies among the joint features. This helped our method to effectively mine useful features within the joint single-epoch features.

Table 11. Performance comparison with previous methods on sleep stages based on the CAP dataset.

Study	Modality	Features	Classifier	Acc.	κ	MF1	MGM
This study	EEG, EOG, EMG	RGB Spectrogram	CNN + Transformer	95.20%	93.60%	91.30%	95.00%
[34]	EEG, ECG, EMG	Log Spectrogram	MML-DMS (VGG16)	94.34%	-	91.00%	-
[30]	EEG, EOG, EMG, ECG	Multi-Domain	Random Forest	86.24%	81.80%	79.62%	89.14%
[54]	ECG-HRV	DFA—Detrended Fluctuation Analysis Alpha 1	k-Fold Cross-Validation (k = 13)	73.60%	-	-	-

Table 12. Performance comparison with previous methods on sleep disorders based on the CAP dataset.

Study	Modality	Features	Classifier	Acc.	κ	MF1	MGM
This study	EEG, EOG, EMG	RGB Spectrogram	CNN + Transformer	99.30%	99.10%	99.20%	99.60%
[34]	EEG, ECG, EMG	Log Spectrogram	MML-DMS (VGG16)	99.09%	-	99.00%	-
[32]	EEG, EMG, ECG, EOG	Spectrogram	DL-AR	95.00%	-	-	97.16%
[33]	EOG, EMG	Hjorth Parameters	Ensemble Bagged Trees	94.30%	91.81%	93.83%	-
[20]	EEG	Hjorth Parameters	Ensemble Bagged and Boosted Trees	91.30%	-	82.43%	-
[55]	ECG	Spectral Features and Sleep Quality Parameters	Ensemble of Bagged Trees	86.27%	70.00%	-	88.75%

Our CNN with transformer method outperforms the best-performing study [34] by 0.86% and 0.21% for sleep stages and sleep disorders, respectively. In addition, our method has a smaller model footprint and lower computational costs than [34], which used the VGG16 CNN structures. Our model achieves superior performance compared to other state-of-the-art approaches, especially in the classification of N1, a stage known for its classification difficulty. The CNN with transformer method has several advantages over the existing studies; notably, it demonstrates superior performance in terms of MF1 and MGM for both sleep stages and sleep disorders, underscoring its better processing strategy in handling imbalances across all data categories. The highest Cohen’s kappa values, 93.6% for sleep staging and 99.1% for disorder classification, achieved by our model demonstrate an almost perfect agreement with human sleep experts, reflecting the highest level of concordance, as defined by Landis and Koch. Our model shows strong potential, delivering a performance comparable to or exceeding all previously reported state-of-the-art results, as shown in Tables 11 and 12, and can be integrated into clinical care settings.

Regarding future work, some datasets contain more than 136 channels [56], which are common across all subjects. For such datasets, we plan to apply channel selection techniques to identify a minimal set of informative channels necessary for automated sleep scoring and sleep disorder diagnosis, thereby enhancing the feasibility of clinical application. We will also explore other modalities, such as ECG and respiratory signals, for the effective detection of other disorders, such as SDB, in datasets with a sufficient number of SDB subjects. Furthermore, we plan to investigate inter-epoch contextual dependencies and explore other advanced transformer architectures to further enhance model

performance. Finally, to enhance model generalization while reducing computational costs and improving processing speed, we will apply our approach to additional datasets, including the SHHS, MASS, MIT-BIH, Apnea-ECG Database, and St. Vincent's University Hospital/University College Dublin Sleep Apnea Database (UCDDb).

5. Conclusions

In this paper, we proposed a novel RGB spectrogram-based model for sleep stage and disorder detection using deep learning algorithms. We implemented three novel deep learning methods: CNN, CNN with LSTM, and CNN with transformer. We used three PSG channels: EEG, EOG, and EMG. We validated and tested the proposed methods using the PhysioNet CAP sleep database. After extensive ablation experiments, the experimental results demonstrated that our proposed CNN with transformer method outperformed standalone CNN, CNN with Bi-LSTM, and other advanced state-of-the-art methods in sleep staging and disorder detection. In this model, CNN was used to extract features from the RGB spectrogram images of each channel individually. The extracted multichannel features were fused in a column-wise manner using a feature fusion block, and a multi-head attention mechanism was used to capture intra-epoch temporal dynamics and contextual dependencies among the joint features. In addition, we added the residual connections to preserve the original joint features. Notably, our model achieved superior performance compared to other advanced models for the N1 stage, which is known to be particularly difficult to classify. Furthermore, we demonstrated that multimodal feature-level fusion significantly contributed to improving classification performance. This advanced model enables accurate sleep staging, facilitates comprehensive sleep quality assessment, and provides valuable support for the clinical diagnosis and treatment of sleep-related disorders.

Author Contributions: Conceptualization, I.T., B.A. and R.H.; methodology, I.T., B.A. and R.H.; software, I.T., B.A. and R.H.; validation, I.T., B.A. and R.H.; formal analysis, I.T., B.A. and R.H.; investigation, I.T., B.A. and R.H.; resources, I.T., B.A. and R.H.; data curation, I.T., B.A. and R.H.; writing—original draft preparation, I.T., B.A. and R.H.; writing—review and editing, I.T., B.A. and R.H.; visualization, I.T., B.A. and R.H.; supervision, B.A. and R.H.; project administration, B.A. and R.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used fully anonymized and publicly available data from the PhysioNet CAP sleep database, provided by the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy, downloaded via [physionet.org](https://physionet.org/content/capslpdb/1.0.0/) at [46] <https://physionet.org/content/capslpdb/1.0.0/> (accessed on 30 May 2025). As the dataset contains only de-identified information and involves no interaction with human participants, Institutional Review Board approval and informed consent were not required.

Acknowledgments: The authors acknowledge the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy, for providing the CAP sleep database available through PhysioNet (physionet.org).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qu, W.; Wang, Z.; Hong, H.; Chi, Z.; Feng, D.D.; Grunstein, R.; Gordon, C. A residual based attention model for EEG based sleep staging. *IEEE J. Biomed. Health Inf.* **2020**, *24*, 2833–2843. [\[CrossRef\]](#)
2. Hori, T.; Sugita, Y.; Koga, E.; Shirakawa, S.; Inoue, K.; Uchida, S.; Fukuda, N. Proposed supplements and amendments to ‘A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects’, the Rechtschaffen & Kales (1968) standard. *Psychiatry Clin. Neurosci.* **2001**, *55*, 305–310. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Iber, C.; Ancoli-Israel, S.; Chesson, A.L.; Quan, S.F. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specification*; American Academy of Sleep Medicine: Darien, IL, USA, 2007.
4. Kwon, S.; Kim, H.; Yeo, W.-H. Recent advances in wearable sensors and portable electronics for sleep monitoring. *iScience* **2021**, *24*, 102461. [\[CrossRef\]](#)
5. Canadian Men’s Health Foundation (CMHF) and Intensions Consulting. Third of Canadian Men Are Sleep Deprived. 2016. Available online: <https://menshealthfoundation.ca/news/study-finds-third-canadian-men-sleep-deprived/> (accessed on 25 June 2025).
6. Chattu, V.K.; Sakhamuri, S.M.; Kumar, R.; Spence, D.W.; BaHammam, A.S.; Pandi-Perumal, S.R. Insufficient Sleep Syndrome: Is it time to classify it as a major noncommunicable disease? *Sleep Sci.* **2018**, *11*, 56. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Surantha, N.; Kusuma, G.P.; Isa, S.M. Internet of things for sleep quality monitoring system: A survey. In Proceedings of the 11th 2016 International Conference on Knowledge, Information and Creativity Support Systems, KICSS 2016, Yogyakarta, Indonesia, 10–12 November 2016. [\[CrossRef\]](#)
8. Siyanbade, J.; Abdulrazak, B.; Sadek, I. Unobtrusive Monitoring of Sleep Cycles: A Technical Review. *BioMedInformatics* **2022**, *2*, 204–216. [\[CrossRef\]](#)
9. Loh, H.W.; Ooi, C.P.; Vicnesh, J.; Oh, S.L.; Faust, O.; Gertych, A.; Acharya, U.R. Automated detection of sleep stages using deep learning techniques: A systematic review of the last decade (2010–2020). *Appl. Sci.* **2020**, *10*, 8963. [\[CrossRef\]](#)
10. Mohamed, M.; Mohamed, N.; Kim, J.G. Advancements in Wearable EEG Technology for Improved Home-Based Sleep Monitoring and Assessment: A Review. *Biosensors* **2023**, *13*, 1019. [\[CrossRef\]](#)
11. Mendonça, F.; Fred, A.; Mostafa, S.S.; Morgado-Dias, F.; Ravelo-García, A.G. Automatic detection of cyclic alternating pattern. *Neural Comput. Appl.* **2018**, *34*, 11097–11107. [\[CrossRef\]](#)
12. Alickovic, E.; Subasi, A. Ensemble SVM method for automatic sleep stage classification. *IEEE Trans Instrum. Meas.* **2018**, *67*, 1258–1265. [\[CrossRef\]](#)
13. Zhao, S.; Long, F.; Wei, X.; Ni, X.; Wang, H.; Wei, B. Evaluation of a single-channel EEG-based sleep staging algorithm. *Int. J. Env. Res. Public Health* **2022**, *19*, 2845. [\[CrossRef\]](#)
14. Michielli, N.; Acharya, U.R.; Molinari, F. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Comput. Biol. Med.* **2019**, *106*, 71–81. [\[CrossRef\]](#)
15. Zhang, W.; Zhang, S.; Wang, Y.; Li, C.; Peng, H.; Chen, X. A CNN-Transformer-ConvLSTM-CRF Hybrid Network for Sleep Stage Classification. *IEEE Sens. J.* **2024**, *24*, 29018–29029. [\[CrossRef\]](#)
16. Li, C.; Qi, Y.; Ding, X.; Zhao, J.; Sang, T.; Lee, M. A deep learning method approach for sleep stage classification with EEG spectrogram. *Int. J. Env. Res. Public Health* **2022**, *19*, 6322.
17. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 2456–2467. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwok, C.-K.; Li, X.; Guan, C. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 809–818. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Yao, Z.; Liu, X. A cnn-transformer deep learning model for real-time sleep stage classification in an energy-constrained wireless device. In Proceedings of the 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER), Baltimore, MD, USA, 25–28 April 2023; pp. 1–4.
20. Sharma, M.; Tiwari, J.; Patel, V.; Acharya, U.R. Automated identification of sleep disorder types using triplet half-band filter and ensemble machine learning techniques with eeg signals. *Electronics* **2021**, *10*, 1531. [\[CrossRef\]](#)
21. Kim, H.; Lee, S.M.; Choi, S. Automatic sleep stages classification using multi-level fusion. *Biomed. Eng. Lett.* **2022**, *12*, 413–420. [\[CrossRef\]](#)
22. Altevogt, B.M. *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*; National Academies Press: Cambridge, MA, USA, 2006.
23. U.S. Department of Health and Human Services and National Heart Lung and Blood Institute. Your Guide to Healthy Sleep. NIH Publication No. 11-5271. 2011. Available online: <https://www.nhlbi.nih.gov/sites/default/files/publications/11-5271.pdf> (accessed on 29 June 2025).
24. Phan, H.; Andreotti, F.; Cooray, N.; Chen, O.Y.; De Vos, M. Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 1285–1296. [\[CrossRef\]](#)

25. Almutairi, H.; Hassan, G.M.; Datta, A. Classification of sleep stages from EEG, EOG and EMG signals by SSNet. *arXiv* **2023**, arXiv:2307.05373. [[CrossRef](#)]
26. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 400–410. [[CrossRef](#)]
27. Fernández-Varela, I.; Hernández-Pereira, E.; Moret-Bonillo, V. A convolutional network for the classification of sleep stages. *Proceedings* **2018**, *2*, 1174. [[CrossRef](#)]
28. Andreotti, F.; Phan, H.; Cooray, N.; Lo, C.; Hu, M.T.M.; De Vos, M. Multichannel sleep stage classification and transfer learning using convolutional neural networks. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 171–174.
29. Chambon, S.; Galtier, M.N.; Arnal, P.J.; Wainrib, G.; Gramfort, A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 758–769. [[CrossRef](#)]
30. Yan, R.; Zhang, C.; Spruyt, K.; Wei, L.; Wang, Z.; Tian, L.; Li, X.; Ristaniemi, T.; Zhang, J.; Cong, F. Multi-modality of polysomnography signals’ fusion for automatic sleep scoring. *Biomed. Signal Process. Control.* **2019**, *49*, 14–23. [[CrossRef](#)]
31. Stephansen, J.B.; Olesen, A.N.; Olsen, M.; Ambati, A.; Leary, E.B.; Moore, H.E.; Carrillo, O.; Lin, L.; Han, F.; Yan, H.; et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **2018**, *9*, 5229. [[CrossRef](#)]
32. Zhuang, D.; Rao, I.; Ibrahim, A.K. A Machine Learning Approach to Automatic Classification of Eight Sleep Disorders. *arXiv* **2022**, arXiv:2204.06997. [[CrossRef](#)]
33. Sharma, M.; Darji, J.; Thakrar, M.; Acharya, U.R. Automated identification of sleep disorders using wavelet-based features extracted from electrooculogram and electromyogram signals. *Comput. Biol. Med.* **2022**, *143*, 105224. [[CrossRef](#)] [[PubMed](#)]
34. Cheng, Y.H.; Lech, M.; Wilkinson, R.H. Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning. *Sensors* **2023**, *23*, 3468. [[CrossRef](#)]
35. Dai, Y.; Li, X.; Liang, S.; Wang, L.; Duan, Q.; Yang, H.; Zhang, C.; Chen, X.; Li, L.; Li, X.; et al. Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with psg. *IEEE J. Biomed. Health Inf.* **2023**, *27*, 4204–4215.
36. Wei, Y.; Zhu, Y.; Zhou, Y.; Yu, X.; Luo, Y. Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel EEG. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 801–811. [[CrossRef](#)]
37. Sharma, M.; Goyal, D.; Achuth, P.V.; Acharya, U.R. An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank. *Comput. Biol. Med.* **2018**, *98*, 58–75. [[CrossRef](#)]
38. Phan, H.; Do, Q.; Do, T.-L.; Vu, D.-L. Metric learning for automatic sleep stage classification. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 5025–5028.
39. Sors, A.; Bonnet, S.; Mirek, S.; Vercueil, L.; Payen, J.-F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed. Signal Process. Control.* **2018**, *42*, 107–114. [[CrossRef](#)]
40. Yan, R.; Li, F.; Zhou, D.; Ristaniemi, T.; Cong, F. A deep learning model for automatic sleep scoring using multimodality time series. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–22 January 2021; pp. 1090–1094.
41. Vaswani, A. Attention is all you need. *Adv. Neural. Inf. Process. Syst* **2017**, *30*.
42. Seo, H.; Back, S.; Lee, S.; Park, D.; Kim, T.; Lee, K. 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.* **2020**, *61*, 102037.
43. Almutairi, H.; Hassan, G.M.; Datta, A. Detection of obstructive sleep apnoea by ecg signals using deep learning architectures. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 24–28 August 2021; pp. 1382–1386.
44. Malafeev, A.; Laptev, D.; Bauer, S.; Omlin, X.; Wierzbicka, A.; Wichniak, A.; Jernajczyk, W.; Riener, R.; Buhmann, J.; Achermann, P. Automatic human sleep stage scoring using deep neural networks. *Front. Neurosci.* **2018**, *12*, 781. [[CrossRef](#)] [[PubMed](#)]
45. American Academy of Sleep Medicine. *International Classification of Sleep Disorders—Third Edition*, 3rd ed.; Darien, I.L., Ed.; American Academy of Sleep Medicine: Darien, IL, USA, 2014.
46. CAP—Sleep Database. Available online: <https://physionet.org/content/capslpdb/1.0.0/> (accessed on 30 May 2025).
47. Terzano, M.G.; Parrino, L.; Sherieri, A.; Chervin, R.; Chokroverty, S.; Guilleminault, C.; Walters, A. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Med.* **2001**, *2*, 537–554. [[CrossRef](#)]
48. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220.
49. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L.A. *Feature Extraction: Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 207.
50. Karadeli, M.I.; Cansiz, B.; Aydin, N.; Serbes, G. Enhanced Embolic Signal Analysis through Ensemble Deep Learning Techniques Utilizing Transfer Learning and Layer Freezing Strategies. *Adv. Intell. Syst.* **2025**, *7*, 2500129. [[CrossRef](#)]

51. Mutlu, E.; Serbes, G. Bridging Auscultation and Tiny Machine Learning: A Digital Stethoscope Leveraging Convolutional Neural Networks on an Embedded Device for Organ Sound Analysis. *Trait. Du Signal* **2024**, *41*, 2471. [[CrossRef](#)]
52. Fernandez-Blanco, E.; Rivero, D.; Pazos, A. Convolutional neural networks for sleep stage scoring on a two-channel EEG signal. *Soft Comput.* **2020**, *24*, 4067–4079. [[CrossRef](#)]
53. Şen, B.; Peker, M.; Çavuşoğlu, A.; Çelebi, F.V. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. *J. Med. Syst.* **2014**, *38*, 18. [[CrossRef](#)]
54. Kim, J.; Lee, J.; Shin, M. Sleep stage classification based on noise-reduced fractal property of heart rate variability. *Procedia Comput. Sci.* **2017**, *116*, 435–440. [[CrossRef](#)]
55. Widasari, E.R.; Tanno, K.; Tamura, H. Automatic sleep disorders classification using ensemble of bagged tree based on sleep quality features. *Electronics* **2020**, *9*, 512. [[CrossRef](#)]
56. Stenwig, H.; Soler, A.; Furuki, J.; Suzuki, Y.; Abe, T.; Molinas, M. Automatic sleep stage classification with optimized selection of eeg channels. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; pp. 1708–1715.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.