Discover Artificial Intelligence

RESEARCH Open Access

A hybrid deep learning framework for sleep stage classification using single channel EEG signals

Suren Kumar Sahu[1], Santosh Kumar Satapathy[2], Sudhir Kumar Mohapatra[3*], Jukka Heikkonen[3], Rajeev Kanth[4] and Tapan Kumar Das[5]

*Correspondence:
Sudhir Kumar Mohapatra
skmoha@utu.fi
[1]Faculty of Emerging Technology, Sri Sri University, Cuttack, Odisha, India
[2]Pandit Deendayal Energy University, Gandhinagar, Gujarat, India
[3]University of Turku, Turku, Finland
[4]Savonia University of Applied Sciences, Kuopio, Finland
[5]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India

Abstract

Background For the diagnosis of sleep disorders and analysis of sleep habits, the precise identification of sleep stages is crucial. Although manual scoring methods are common, the work process can be laborious, and it inherently involves variability. To get over these limitations, this research introduces AdvancedSleepClassifier, a hybrid deep learning system that is capable of classifying sleep stages from a single channel of electroencephalogram (EEG). The introduced model seeks to exploit temporal and spectral aspects of EEG information by the use of dilated Convolutional Neural Network (CNN) integration, squeeze-and-excitation blocks, bidirectional long short-term memory (LSTM) layers, and multi-head self-attention. The approach guarantees the system complies with the traditional EEG analysis techniques and, at the same time, facilitates the automated processes and results.

Methods The evaluation was done by drawing on the Institute for Systems and Robotics University of Coimbra (ISRUC-Sleep) and Sleep European Data Format (SleepEDF) data that are freely available. EEG recordings were segmented into segments of 30 s each, which were then filtered, normalised, and the power spectral density (PSD) was computed within five frequency bands. By doing spectral-temporal feature extraction simultaneously, the architecture enables the use of both raw EEG data and PSD features. The team applied weighted cross-entropy loss and data augmentation to address the equilibrium of the class counting problem. The model was iterated 10-fold cross-validation without subject bias and tested for two-class, three-class, four-class, and five-class tasks, using accuracy, F1-score and Area Under the ROC Curve (AUC-ROC) metrics.

Results The AdvancedSleepClassifier was capable of delivering precise and trustworthy classification results on different levels of class detail. The maximum accuracy was derived at 97.2% for binary classification (Wake vs. Sleep), 91.3% for three-class (Wake, NREM, REM), 89.4% for four-class and 85.07% for five-class classification on SleepEDF. It obtained significantly high F1 scores in clinically significant stages, particularly Wake. The classifier gave an F1 score of 0.94 for Wake and 0.87 for N3; with the support of robust class separation, the ROC-AUC scores exceeded 0.98 in most situations. The use of Grad-CAM (Gradient-weighted Class Activation Mapping) and

t-Distributed Stochastic Neighbour Embedding N(t-SNE) provided additional support for the physiological relevance of the features that were found by the model.

Conclusion The model shows great promise for the provision of interpretable, useful solutions to the automatic sleep staging task utilizing only EEG data from a single channel. It has good generalization on different datasets having different class structures, while maintaining clinical significance along with high reliability. Future work may involve extending the model to multimodal signals and optimizing it for real-time deployment in portable health-monitoring devices.

Keywords Sleep stage classification, EEG, Deep learning, CNN, LSTM, Attention mechanism, Automatic sleep scoring, ISRUC, SleepEDF, Biomedical AI

1 Introduction

As a person sleeps, the different stages serve particular purposes in the body. When N1 occurs, the brain is approaching sleep, having light and easily interrupted stages with fewer alpha waves. At N2, distinguishing signs like sleep spindles and K-complexes help the brain maintain sleep and improve its memory functions. Having n3 restores and supports the immune system, gives the body energy, and recovers stamina. It mainly contains delta waves with high power, which make it most difficult to be roused from. REM sleep is related to helping manage emotions and joining together memories. While the body is completely still during REM sleep, brain activity is highly similar to when we are awake. They go through these stages again after roughly 90 min throughout the night. To detect these phases in sleep, the system processes EEG, EOG, and EMG data accurately. To successfully apply deep learning models in the medical field, one must first fully comprehend each process and its key factors.

Sleep represents an essential biological process that helps preserve cognitive performance, together with keeping metabolic functions stable while managing emotional responses [1]. A sleep cycle of humans follows sequential patterns between Wakefulness (W), Non-Rapid Eye Movement (NREM) and Rapid Eye Movement (REM) stages. During NREM sleep, three substages, N1, N2 and N3, advance from lighter to deeper stages of sleep [2]. Physicians need precise identification of sleep stages known as sleep staging for diagnosing medical conditions, including insomnia, narcolepsy and sleep apnea [3]. Staging sleep with traditional methods entails expert annotation of polysomnography (PSG) signals containing electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) data, while facing challenges related to extensive work efforts along with substantial time requirements and inconsistency between different scorers [4].

Researchers have started using deep learning frameworks that focus on EEG signals for automated sleep staging because current techniques show various limitations [5]. CNNs demonstrate top performance for local temporal feature extraction, yet sequential dependencies and Recurrent Neural Network (RNN) and their subtypes, LSTM and Gated Recurrent Unit (GRU), operate effectively for capturing these dependencies [6]. Attention techniques utilising Transformer frameworks have appeared in recent times to deal with long-range temporal interactions while improving interpretability [7].

This study presents AdvancedSleepClassifier, which functions as a hybrid deep learning system that uses single-channel EEG recordings to extract multi-scale temporal and spectral characteristics for automated sleep stage classification. The model uses dilated

CNNs for capturing multi-resolution temporal dynamics, besides squeeze-and-excitation residual blocks for channel feature adaptation and multi-head self-attention to focus on important EEG segments. Each sequential dependency functions using bidirectional Long Short-Term Memory layers to collect data, and the spectral feature fusion operates with delta, theta, alpha, beta, and gamma frequency band integration to imitate clinical EEG review practices.

The model achieves performance evaluation on both ISRUC-Sleep and Sleep-EDF Physionet datasets to show its ability to work across different settings. A class-weighted cross-entropy loss optimisation drives the training process and domain-specific data enhancement methods tackle the uneven stage distributions. Our developed system reaches a top accuracy of 71.38% and demonstrates exceptional ability to detect N3 (F1 = 0.85) and Wake (F1 = 0.78) stages. The analysis incorporates Grad-CAM interpretability tools to display EEG segments that discriminate between different sleep stages. This research offers a complete and understandable sleep staging method by fusing spectral-temporal learning processes with new hardware features that can become applicable in medical settings. Table 1 summarizes the standard EEG frequency bands and their corresponding physiological relevance, which are widely used in sleep stage characterization and neurophysiological signal analysis.

2 Literature review

Automating the process of sleep staging is now essential for biomedical signal processing because manually scoring is not ideal, and more people are suffering from sleep-related illnesses. Traditionally, artificial intelligence (AI) used algorithms based on rules or machine learning with features that experts made from polysomnography records. Still, these approaches didn't always work for different situations and needed experts in the field. Deep learning gives the ability to develop end-to-end systems that can process raw data from EEG, EOG or EMG. Many models like CNNs, RNNs and transformers have been designed to improve how stages are classified and make the approach independent from the features chosen by experts. Even so, certain issues exist: each physician may interpret in a unique way, there are overall few subjects compared to others in the research, and each database has its own structure. Additionally, most data in studies are either small or homogeneous, which limits their usefulness in real life. That is why recent studies have examined data augmentation, transfer learning and self-supervised solutions. Overall, more progress could be made in easy-to-understand models, using them for several datasets and running them efficiently on energy-saving hardware. Significant progress has been noted, yet some ongoing problems mean more work is needed to develop systems that businesses, healthcare and science can depend on.

Tsinalis et al. worked to establish a sleep stage classification system using single-channel scalp EEG recordings, which did not need any manually generated signal features.

Table 1 Frequency bands of EEG signals

Frequency band	Range (Hz)	Description
Delta	0.5–4 Hz	Associated with deep sleep and restorative processes.
Theta	4–8 Hz	Linked to light sleep, drowsiness, and relaxation.
Alpha	8–13 Hz	Related to relaxed and calm states, often with eyes closed.
Beta	13–30 Hz	Associated with active thinking, focus, and problem-solving.
Gamma	> 30 Hz	Linked to higher mental activity, perception, and memory.

The research used the Sleep-EDF dataset that contained EEG recordings from twenty healthy young adults. The proposed method incorporated a CNN structure with class-balanced stochastic gradient descent for the correction of class imbalances. The proposed model attained an accuracy rate of 74% with an F1-score average of 81%, which matched existing leading methods in the field. One major shortcoming of the model was its diminished capability to recognize N1 sleep stages, along with a few other stages. The conclusion from this research showed CNNs effectively acquire sleep stage discriminative features directly from raw EEG data, which presents opportunities for automated sleep diagnostics [8].

Phan et al. created a combined classification and prediction system through CNNs for automatic sleep staging. The researchers based their study on the Sleep-EDF Expanded and Montreal Archive of Sleep Studies datasets. Through their CNN model, they developed a system that performed both current epoch classification and near-epoch prediction by using temporal information. The CNN model delivered an overall achievement rate of 82.3% on Sleep-EDF Expanded and 83.6% on Montreal Archive of Sleep Studies. The framework showed restricted performance because it depended on using static input segments. Joint classification and prediction methods, when integrated together, result in better sleep staging performance according to the study authors [9].

Luo et al. developed automatic sleep staging through a method that integrated CNNs with BiLSTM networks for signal processing. Research used the Sleep-EDF dataset for all experimental procedures. The research method featured CNNs extracting spatial signal features from EEG signals before Bidirectional Long Short-Term Memory (BiLSTM) networks analysed the temporal dependencies. The authors implemented an improved Modified Synthetic Minority Over-sampling Technique (MSMOTE) algorithm for data oversampling as a solution to handle class imbalance. The tested model reached 92.21% classification success. The main drawback of this work was its potential overfitting caused by oversampling techniques. The combined CNN and BiLSTM approach succeeded in gathering spatial data features along with temporal elements that boosted the accuracy rates of sleep stage analysis [10].

Guillot and Thorey created RobustSleepNet, which serves as a deep learning model for scalable sleep staging across different datasets. The model training took place along with evaluation tests using eight heterogeneous sleep staging datasets. The model employed transfer learning functionality, which managed different formats of polysomnography data and different patient demographics. RobustSleepNet obtained 97% of the F1-score above other models that received single-dataset training. The model showed decreased performance when processing completely new data sets unless it received fine-tuning. RobustSleepNet presents a solution for automated sleep staging that provides general applicability across multiple clinical environments, according to the authors [11].

XSleepNet represents a multi-view sequential model for automatic sleep staging, which Phan et al. presented. Researchers analysed five different sleep datasets, which contained varied-sized information. XSleepNet merged EEG signals with their frequency domain components into a sequence-to-sequence learning system that adjusted learning parameters separately for each signal type. The proposed model proved better than both single-view methods as well as older multi-view approaches in all cases. The advanced model design presents difficulties when trying to operate in environments with restricted resources. Multi-view learning systems improve both the accuracy and

the robustness of sleep stage identification according to the authors' research conclusions [12].

LWSleepNet represents a lightweight attention-based deep learning model that stages sleep utilizing single-channel EEG data according to Yang et al. Researchers used the Sleep-EDF dataset for their evaluation purposes. This framework contained Light-MRCNN features along with Temporal Feature Extraction (TFE) blocks that detected temporal patterns through depthwise separable convolutions and bottleneck blocks for efficient operation. LWSleepNet underwent cross-validation tests, which showed it maintained high performance levels while achieving a simplified model architecture. The main drawback of our model design was the accuracy-performance trade-off between simplicity and accuracy within these parameters. As per the study's findings, LWSleepNet reveals aptness for portable devices by achieving performance optimization via computational efficiency standards [13].

The Multi-Scale Dual Attention Network (MSDAN), which Wang et al. developed, concentrates on automatic sleep staging through single-channel EEG information. Regulation of evaluation required the researchers to use both Sleep-EDF and Sleep-EDFx datasets. This approach leverages multi-scale convolutional layers that analyse various EEG waveform patterns while employing channel and spatial attention methods to enhance relevant data segments, and employs soft thresholding for reducing repetitive information. Nutritional fly model demonstrated Sleep EDF evaluation yielding 91.74% and Sleep EDFx evaluation producing 90.35%. Structural improvements in N1 stage recognition appeared throughout both evaluations. The model poses complications for its implementation in resource-limited operational settings. The research findings confirmed that MSDAN extracted essential sleep stage features which lead to better detection performance, especially during difficult N1 sleep stages [14].

ProductGraphSleepNet represents a model that demonstrates the ability to detect both space and time patterns in EEG data during sleep staging, according to Einizade. The research used two EEG datasets from the Montreal Archive of Sleep Studies (MASS), SS3 and Sleep-EDF. The researchers implemented adaptive product graph learning together with bidirectional gated recurrent units and modified graph attention networks to represent spatial relationships and temporal patterns. The model processed data from MASS, achieving 86.7% accuracy while delivering 83.8% accuracy on Sleep-EDF data and 81.8% and 77.4% F1-scores, respectively. Adaptation graph models complicated the system to the point of affecting its operational timeliness as an applied tool. Research studies demonstrated the success of implementing spatio-temporal graph structures because they increase the model's effectiveness in tracking complex sleep behaviors [15].

Jia et al. designed SalientSleepNet as a multimodal network that uses salient wave detection in EEG data to enhance sleep stage classification. The sleep stage detection model received evaluation through tests performed on two publicly available datasets containing sleep data. The research method employed a temporally convoluted network derived from U^2-Net that contained dual streams to extract dimensional information and multi-scale extraction units to detect transition criteria, in addition to multimodal attention units, which directed analysis toward essential details. SalientSleepNet achieved better performance than previous models while requiring fewer model parameters for its operation. Several input modalities in the proposed method reduce its use in cases that employ only single-channel EEG systems. The research investigation established

that sleep stage identification becomes more accurate when methods focus simultaneously on meaningful wave detection across different file sources [16].

Using single-channel EEG input, Mousavi et al. presented SleepEEGNet as a sequence-to-sequence deep learning model that performs automatic sleep stage scoring functions. The Sleep-EDF datasets became the foundation for training and evaluation of this model. The system used deep CNNs for extracting both time-stable and frequency-based patterns while implementing a sequence-to-sequence model to identify long-range patterns between sleep stages. Training included the application of new loss functions to handle the class imbalance problem. SleepEEGNet scored 84.26% accuracy and reached 79.66% macro F1-score together with 0.79 Cohen's Kappa coefficient. The main weakness of this model design arises from its subpar results when detecting rare sleep stages. CNN sequence model integration produces effective results in capturing complicated sleep pattern characteristics, which enhances automatic sleep staging [17]. Table 2 presents a comparative review of existing automated sleep stage classification

Table 2 Comparative literature review of automated sleep stage classification methods and their key challenges

References	Model / approach	Dataset(s)	Key contribution	Key challenges / limitations
Tsinalis et al. [8]	CNN-based end-to-end model	Sleep-EDF	Demonstrated that CNNs can learn discriminative features directly from raw single-channel EEG without handcrafted features.	Poor recognition of N1 sleep stage; limited dataset size (20 healthy subjects); restricted generalization.
Phan et al. [9]	CNN with joint classification and prediction	Sleep-EDF Expanded, MASS	Integrated current-epoch classification with near-epoch prediction using temporal context.	Dependence on static input segments; limited ability to model long-range temporal dependencies.
Luo et al. [10]	CNN + BiLSTM with MSMOTE	Sleep-EDF	Combined spatial feature extraction and temporal dependency modeling; addressed class imbalance via oversampling.	Risk of overfitting due to synthetic data generation; limited robustness on unseen data.
Guillot & Thorey [11]	RobustSleepNet (Transfer Learning)	8 heterogeneous datasets	Achieved scalable sleep staging across datasets using transfer learning.	Performance degrades on completely unseen datasets without fine-tuning; added training complexity.
Phan et al. [12]	XSleepNet (Multi-view sequential model)	5 public datasets	Leveraged time-domain and frequency-domain EEG jointly for improved robustness.	High computational complexity; difficult deployment in resource-constrained environments.
Yang et al. [13]	LWSleepNet (Lightweight attention model)	Sleep-EDF	Designed an efficient architecture suitable for portable and wearable devices.	Accuracy–efficiency trade-off; simplified architecture may limit peak performance.
Wang et al. [14]	MSDAN (Multi-scale Dual Attention Network)	Sleep-EDF, Sleep-EDFx	Improved N1 stage detection using multi-scale and attention mechanisms.	Computationally intensive; challenges in real-time and low-power deployment.
Einizade [15]	Product-GraphSleepNet (Graph-based DL)	MASS SS3, Sleep-EDF	Modeled spatio-temporal EEG relationships using graph attention mechanisms.	Increased architectural complexity; reduced operational efficiency for real-time use.
Jia et al. [16]	SalientSleepNet (Multimodal attention network)	Public multimodal datasets	Focused on salient wave detection and multimodal feature fusion.	Requires multiple input modalities; unsuitable for single-channel EEG systems.
Mousavi et al. [17]	SleepEEGNet (Seq2Seq CNN model)	Sleep-EDF	Captured long-range sleep-stage transitions using sequence-to-sequence learning.	Weak performance for rare sleep stages; class imbalance remains challenging.

methods, highlighting their datasets, methodologies, performance, and key limitations, thereby motivating the need for the proposed approach.

The main contributions of the proposed method in this article can be summarised as follows.

- We introduce our proposal, AdvancedSleepClassifier, a hybrid deep learning framework – a parallel-stream framing architecture formed by dilated CNNs, squeeze-and-excitation, Bi-LSTM and multi-head self-attention that combines raw EEG signal processing and spectral feature analysis. This design successfully encapsulates both temporal and spectral dependences of single-channel recordings of EEG.
- The model combines clinical EEG priors by taking the PSD features of five canonical EEG bands (delta, theta, alpha, beta, gamma) and using them to obtain improved physiological interpretability and classification robustness.
- To overcome severe class imbalance (especially in transitional sleep states such as N1, N3), the training pipeline integrates class-weighted cross-entropy loss and targeted data augmentations, including random Gaussian noise injection and signal cutout, which contribute to the effectiveness of the recognition of the minority states.
- Detailed analysis on ISRUC-Sleep and Sleep-EDF datasets in a binary, three-class, four-class, and five-class classifications scheme shows the model possesses a strong generalisation capability.
- A multiple model initialization ensemble averaging strategy additionally improves the stability of classification by 3.2%, indicating the robustness and reproducibility of the suggested system.

3 Methodology

The model we have designed, AdvancedSleepClassifier, is a mixture of convolutional, SE and recurrent units specifically designed for accurately detecting sleep stages [18]. The SE blocks adjust the responses of channels in the features, helping the network identify changes across sleep stages [19]. Its goal is to increase performance for every class and keep the model understandable and adjustable by applying an organized attention and module framework [20].

The proposed study presents AdvancedSleepClassifier, which represents a customised hybrid deep learning system for single-channel EEG-based automatic sleep stage detection [21]. AdvancedSleepClassifier employs two parallel processing streams, which leverage temporal characteristics in combination with spectral information to boost both accuracy rates and interpretability during sleep stage classification operations [22]. This section outlines how the research dataset was specified, describes how data was prepared, explains the model design, details the training process and presents evaluation methodologies.

3.1 Dataset description

The research utilized Publicly available EEG datasets containing PSG recordings from two sources: ISRUC-Sleep (Level 1) and Sleep-EDF PhysioNet (SC subset) [23]. The ISRUC dataset includes EEG signals from a total of 100 people, comprising healthy participants and patients with sleep issues, along with 200 Hz sampling rate information. The C3-A2 EEG channel served for all experiments since this measurement has both

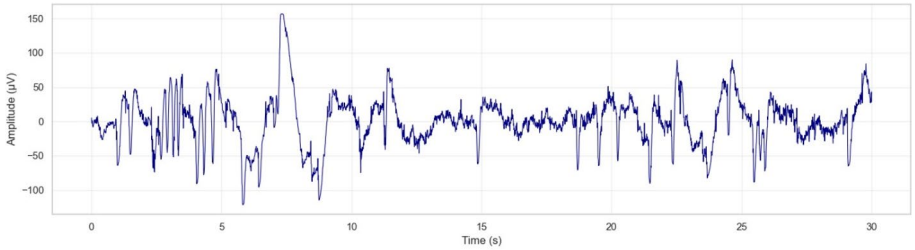

Fig. 1 Sleep EDF EEG wave sample diagram

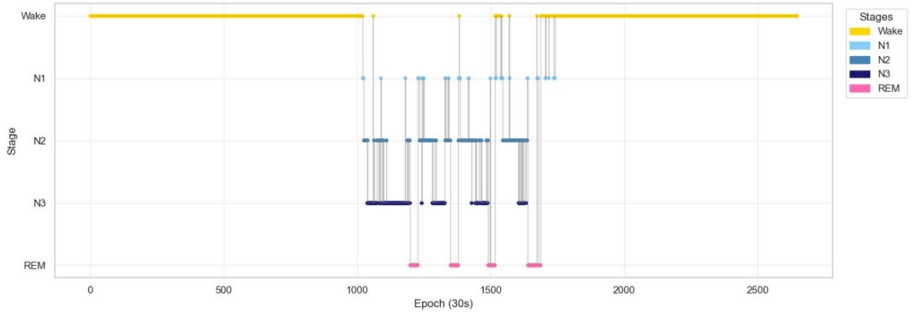

Fig. 2 Sleep EDF hypnogram sample diagram

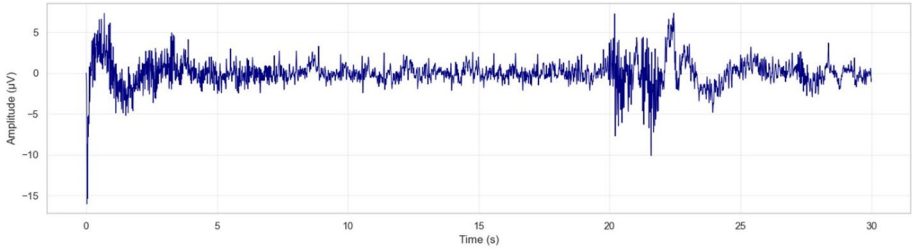

Fig. 3 ISRUC EEG wave sample diagram

clinical importance and data compatibility across datasets. The C3–A2 channel was chosen due to its widespread clinical use in polysomnography and its proven effectiveness in capturing sleep-specific EEG features, as adopted in several benchmark sleep staging studies. The Sleep-EDF PhysioNet dataset contains 20 healthy adult recordings, which were sampled at 100 Hz and underwent manual evaluation through Rechtschaffen and Kales scoring standards. Each EEG recorded section received 30-second segmenting and received one of these five sleep stage labels: Wake, N1, N2, N3 and REM [24]. Both database collections serve as standard benchmarks that enable generalization examination across demographic groups and physiological aspects [25]. Figures 1, 2, 3 and 4 illustrate representative EEG waveforms and corresponding hypnograms from the Sleep-EDF and ISRUC-Sleep datasets, highlighting typical signal characteristics and sleep stage transitions used for automated sleep stage classification.

3.2 Preprocessing

The robust preprocessing pipeline applied to raw EEG recordings filtered noise out while maintaining important signal temporal features and frequency characteristics [26]. The first step involved filtering signals between 0.5 and 35 Hz to eliminate muscle artifacts

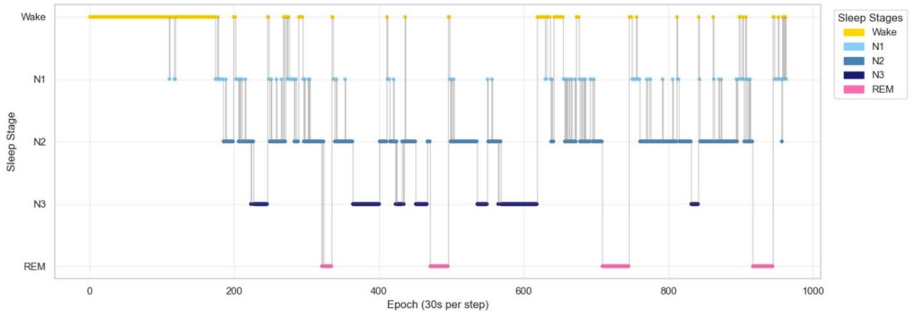

Fig. 4 ISRUC hypnogram sample diagram

along with low-frequency drift [27]. Standard EEG recording procedures divided data into 30-second intervals, which did not overlap with one another. The z-score normalization technique normalized every epoch before data sharing among participants within each recording session. The researchers used hypnograms to extract sleep stages, which they assigned to the AASM five-class scheme. Power spectral density (PSD) estimates obtained through Welch's technique analysed delta, theta, alpha, beta and gamma bands from data using 4-second Hanning windows with 50% window overlap [28]. Welch's method was implemented using a 4-second window as a compromise between frequency resolution and statistical stability of the PSD estimates within each 30-second EEG epoch. A shorter window enables multiple overlapping segments to be averaged, thereby reducing variance and improving robustness to transient artifacts commonly present in EEG signals. Although longer windows provide higher resolution at very low frequencies, the selected 4-second window was sufficient to capture clinically relevant sleep-related frequency bands (delta, theta, alpha, and sigma) while maintaining temporal consistency across epochs. Preliminary experiments with longer window lengths (e.g., 6–8 s) showed no significant improvement in classification performance, whereas they increased sensitivity to non-stationarities. Therefore, a 4-second window was adopted in line with common practice in sleep EEG spectral analysis. The PSD values underwent log transformation before they served as input data for the parallel spectral model stream. Equation (1) represents the power spectral density (PSD) estimation, where the spectrum is obtained by averaging the squared magnitude of the Fourier transforms across signal segments.

$$P_{xx}(f) = \frac{1}{L}\sum_{k=1}^{L} |X_k(f)|^2 \tag{1}$$

Where $X_f(f)$ is the DFT of the k^{th} windowed segment, and L is the number of segments. Used to compute frequency band power for EEG.

3.3 Feature extraction and model architecture

Using a dual-input stream architecture, AdvancedSleepClassifier processes derived spectral features and raw EEG signals concurrently. A one-dimensional Convolutional Neural Network (1D-CNN) backbone is used to extract hierarchical temporal features from the raw signal stream. To capture macro-scale patterns, the first layer uses a large-kernel convolution (kernel size = 50, stride = 3). To exponentially increase the receptive field, subsequent layers incorporate residual convolutional blocks with dilated convolutions

(dilation rates of 1, 2, and 4). Squeeze-and-excitation (SE) modules for adaptive channel recalibration are incorporated into each residual block to improve attention to informative signal regions. Concurrently, the spectral stream employs fully connected layers with batch normalization and ReLU activation to process the five-band PSD features. Before classification, the temporal stream outputs are concatenated with these dense representations. A multi-head self-attention module comes after the CNN backbone to extract temporal dependencies between epochs [29]. Each of the four attention heads projects input independently into query, key, and value vectors, calculating scaled dot-product attention. Positional encodings are included to preserve temporal order and enhance the model's capacity to identify time-bound sleep events like K-complexes or spindles [30]. Equation (2) defines the scaled dot-product attention mechanism, where the similarity between queries and keys is normalized and used to weight the value vectors through a softmax operation.

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

A core component of the self-attention mechanism used to weigh the relevance of different time steps. Its outputs are then fed to a Bidirectional LSTM layer of 64 units, which continues to enhance temporal dependencies by processing sequences in both forward and backward directions. Dropout regularization (rate = 30%) is used to avoid overfitting. Figure 5 illustrates the overall architecture of the proposed AdvancedSleepClassifier, highlighting its main processing stages and feature learning components.

Figure 6 presents the workflow of the proposed AdvancedSleepClassifier, including preprocessing, feature extraction, temporal modeling, and classification stages.

3.4 Classification and training strategy

The last classification head comprises two dense layers (256 and 128 units) with batch normalization, ReLU activation, and dropout (30–50%). The output layer uses softmax

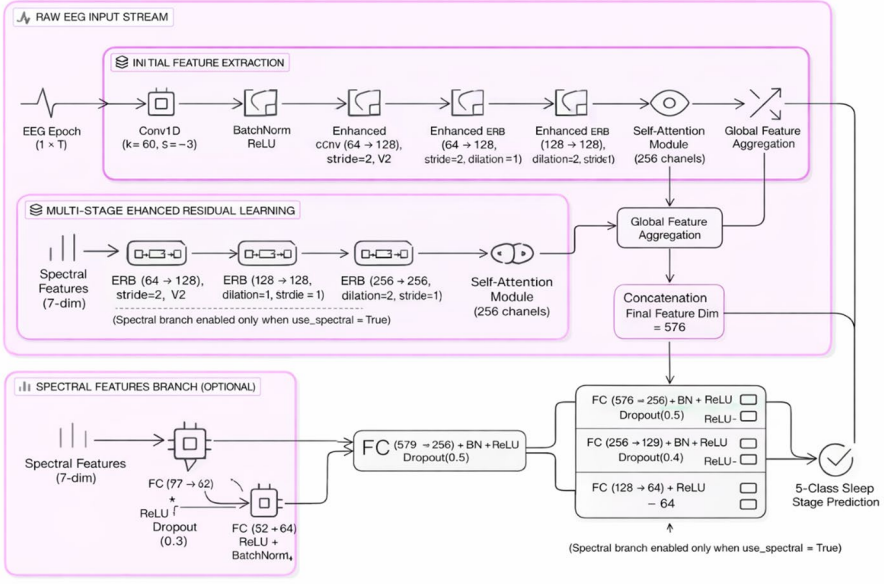

Fig. 5 AdvancedSleepClassifier architecture

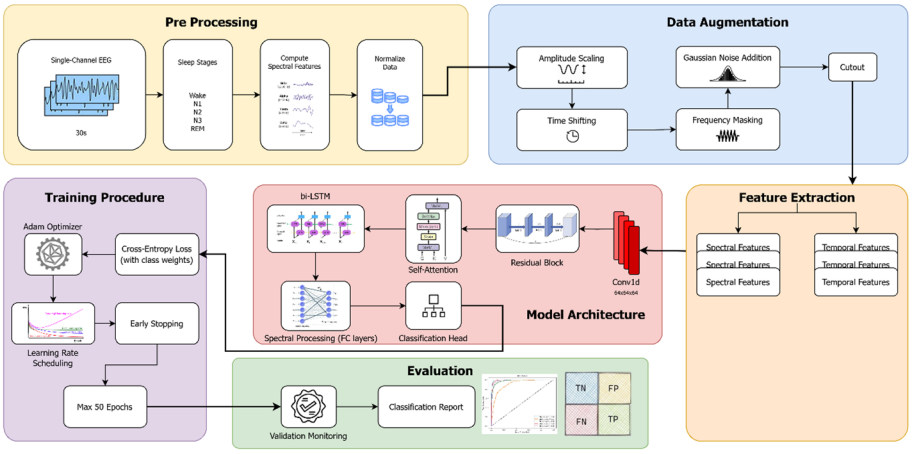

Fig. 6 AdvancedSleepClassifier flowchart

activation to provide class probabilities for five-stage classification. To impose a penalty for overfitting and promote generalization, L2 weight decay (1e-4) is used for all trainable parameters. Class imbalance in the N1 and N3 stages is resolved through class-weighted cross-entropy loss with weights proportional to the inverse of class frequency (e.g., N1: 0.48, N3: 1.65). To mitigate the severe class imbalance inherent in sleep stage datasets, a class-weighted cross-entropy loss was employed. The class weights were computed based on the inverse frequency of each sleep stage in the training set, ensuring that underrepresented stages such as N1 and N3 received higher penalties during optimization. Specifically, for each class c, the weight $w_c$was calculated as $w_c = \frac{N}{K \cdot n_c}$, where $N$is the total number of training samples, $K$is the number of classes, and $n_c$denotes the number of samples belonging to class c. This weighting strategy encourages balanced learning across all sleep stages and reduces bias toward majority classes. A data augmentation technique involving random Gaussian noise (σ = 0.1× signal SD) and random cutout (10% masking of 3-second samples) is also used. Equation (3) defines the weighted cross-entropy loss function, where class-specific weights are incorporated to address class imbalance by penalizing misclassification of minority sleep stages more heavily

$$WeightedCrossEntropyLoss = -\sum_{i=1}^{N} w_{y_i} \log\left(p_{y_i}\right) \tag{3}$$

.

Where w_{yi} is the weight for class y_i and p_{yi} is the predicted probability for the correct class. This helps in addressing class imbalance during training.

3.5 Evaluation metrics and interpretability

Performance is measured in terms of accuracy, F1-score, and AUC-ROC per sleep stage. For better interpretability, Grad-CAM is employed to visually highlight class-informative areas in the EEG signal, roughly indicating slow-wave activity in N3 and sharp transients in N1. t-SNE visualization of embeddings learned by features shows distinct clustering between Wake/REM and NREM stages, confirming the model's latent space discriminative ability. Also, an ensemble of five differently initialized AdvancedSleepClassifier instances is used with a majority vote mechanism to enhance the performance

```
Input:
   x ← EEG data of shape [batch_size, channels=1, time_steps]
   spectral ← Optional spectral features of shape [batch_size, 7]
Hyperparameters:
   num_classes ← 5
   dropout_rates ← Various (0.3, 0.4, 0.5)
   use_spectral ← True or False
Model Architecture:
   Define AdvancedSleepClassifier:
      Initial_Feature_Extraction:
         conv1: Conv1d(1 → 64, kernel_size=64, stride=3)
         batch_norm1: BatchNorm1d(64)
         relu: ReLU()
         max_pool: MaxPool1d(kernel_size=2, stride=2)
         conv1b: Conv1d(64 → 64, kernel_size=5, stride=1, padding=2)
         batch_norm1b: BatchNorm1d(64)
      Residual_Layers:
         layer1: Sequential(
            EnhancedResidualBlock(64 → 64, dilation=1),
            EnhancedResidualBlock(64 → 64, dilation=2)
         )
         layer2: Sequential(
            EnhancedResidualBlock(64 → 128, stride=2, dilation=1),
            EnhancedResidualBlock(128 → 128, dilation=2)
         )
         layer3: Sequential(
            EnhancedResidualBlock(128 → 256, stride=2, dilation=1),
            EnhancedResidualBlock(256 → 256, dilation=2)
         )
      Self_Attention:
         attention: SelfAttention(256)
      Global_Pooling:
         avg_pool: AdaptiveAvgPool1d(output_size=1)
      Sequence_Modeling:
         lstm:    BidirectionalLSTM(input_dim=256,    hidden_dim=256,    dropout=0.3,
batch_first=True)
      Spectral_Branch (if use_spectral):
         spectral_fc: Sequential(
            Linear(7 → 32),
            ReLU(),
            BatchNorm1d(32),
            Dropout(0.3),
            Linear(32 → 64),
            ReLU(),
            BatchNorm1d(64)
         )
         final_feature_dim ← 512 + 64
      else:
         final_feature_dim ← 512
      Classification_Head:
```

Pseudocode:

```
        fc: Sequential(
          Linear(final_feature_dim → 256),
          BatchNorm1d(256),
          ReLU(),
          Dropout(0.5),
          Linear(256 → 128),
          BatchNorm1d(128),
          ReLU(),
          Dropout(0.4),
          Linear(128 → 64),
          ReLU(),
          Linear(64 → num_classes)
        )
Forward Pass:
  Function Forward(x, spectral=None):
    x ← conv1(x)
    x ← batch_norm1(x)
    x ← relu(x)
    x ← max_pool(x)
    x ← conv1b(x)
    x ← batch_norm1b(x)
    x ← layer1(x)
    x ← layer2(x)
    x ← layer3(x)
    x ← attention(x)
    x ← avg_pool(x)
    x ← flatten(x)    # shape: [batch_size, 256]
    x ← reshape(x, [batch_size, 1, -1])  # add sequence dimension for LSTM
    x, _ ← lstm(x)
    x ← reshape(x, [batch_size, -1])     # flatten LSTM output
    if use_spectral and spectral is not None:
      spectral_features ← spectral_fc(spectral)
      x ← concatenate([x, spectral_features], dim=1)
    output ← fc(x)
    return output
Training:
  Initialize model, CrossEntropyLoss, Adam optimizer(lr=0.001)
  For epoch = 1 to epochs:
    Set model to train mode
    For each batch (x_batch, spectral_batch, y_batch):
      optimizer.zero_grad()
      predictions ← model.Forward(x_batch, spectral_batch)
      loss ← CrossEntropyLoss(predictions, y_batch)
      loss.backward()
      optimizer.step()
```

Fig. (continued)

by 3.2%. Stochastic data augmentation variance in training supports robustness with the ensemble.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Measures the proportion of correct predictions out of all predictions. Equation (4) defines classification accuracy as the ratio of correctly predicted instances to the total number of samples.

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{5}$$

This averages the F1 scores for each class, treating all classes equally. Equation (5) defines the Macro-F1 score, computed as the average of class-wise F1-scores, thereby providing an unbiased performance measure for multi-class and imbalanced sleep stage classification.

3.6 Experimental setup

The experimental setup for the sleep stage classification project involves a comprehensive pipeline encompassing data preprocessing, feature extraction, model design, training, and evaluation. EEG recordings are segmented into 30-second epochs and categorized into five sleep stages: Wake, N1, N2, N3, and REM. A subject-independent 10-fold cross-validation strategy was adopted to avoid subject bias. Specifically, all recordings from a given subject (i.e., all 30-second epochs from the same night) were assigned exclusively to a single fold. During each cross-validation iteration, subjects in the test fold were completely excluded from the training and validation folds, ensuring that no EEG segments from the same subject or recording session appeared in both training and testing sets. This protocol prevents data leakage at both the subject level and the epoch level and ensures that the evaluated performance reflects the model's ability to generalise to unseen subjects rather than memorizing subject-specific EEG patterns. Spectral features such as band powers and their ratios are computed, followed by normalization. Data augmentation is applied using various transformations including amplitude scaling, time shifting, frequency masking, Gaussian noise addition, and cutout. The proposed model, AdvancedSleepClassifier, integrates a Conv1D-based feature extractor, enhanced residual blocks with Squeeze-and-Excitation modules, a self-attention mechanism, and a bidirectional LSTM for temporal context processing. Spectral features are processed in a parallel path via fully connected layers and then concatenated with LSTM outputs for classification. The model employs cross-entropy loss with class weights to mitigate imbalance and uses learning rate scheduling and early stopping (patience of 10) during training. While the optimizer is not explicitly specified, Adam is likely used. Training is performed for a maximum of 50 epochs, with validation accuracy guiding the selection of the model.

4 Results

4.1 5-class classification

In this task, all standard sleep stages are classified as Wake, N1, N2, N3 and REM. To accurately monitor sleep, fine detail is essential and is usually applied in clinical work and leading research projects.

4.1.1 ISRUC dataset

Figures 7 and 8 show that the AdvancedSleepClassifier was evaluated on the ISRUC dataset, for classification of five sleep stages, despite providing comparable and good results in different metrics. Training and validation curves demonstrate a steady decrease in loss and increase in accuracy across 50 epochs and validation accuracy exhibiting parallels with training accuracy, indicating effective learning with low risk of overfitting1. In the test set, the classifier showed the overall accuracy of 79.97%, together with the macro-averaged F1-score equal to 0.79 and weighted average equal to 0.80, i.e., the classifier performed equally well for all stages of sleep. Wake stage in particular is a

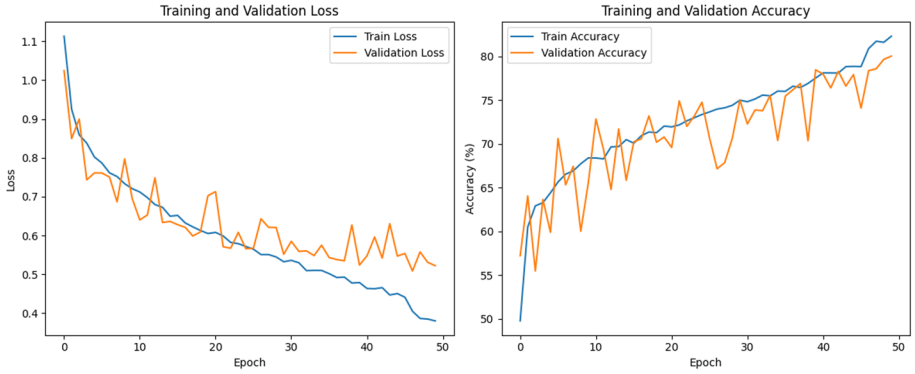

Fig. 7 Training and validation metrics over epochs for 5-class classification with the ISRUC dataset

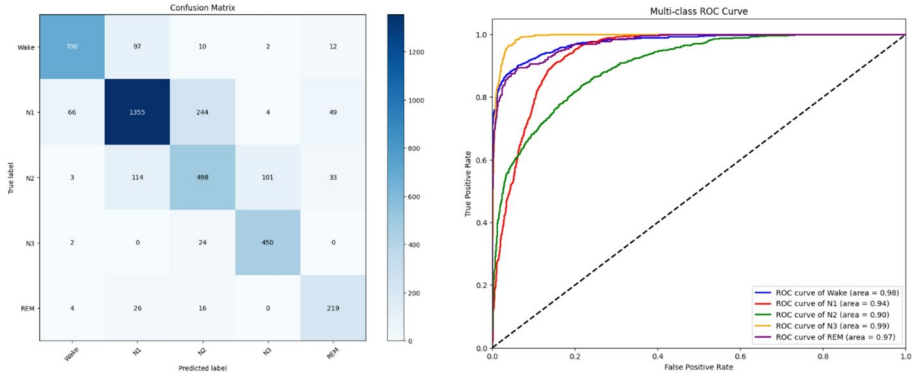

Fig. 8 ROC curve and confusion matrix for 5-class classification with ISRUC dataset

very good performer, as shown by the high level of precision and recall according to the classification report. While results for Wake and N3 stages are as good as 0.90, recall: 0.85, F1: 0.88, performance for N2 and REM stages is lower, but still quite sufficient. The confusion matrix shows that most of the errors occur in the transition from one neighbouring sleep stage to another, which is typical of difficulties with the slow changes in EEG characteristics3. The multi-class ROC analysis serves to reinforce the classifier's effectiveness, because AUC values for all classes exceed 0.90, reaching their inclinations at 0.99 N3 and 0.98 Wake, validating its outstanding ability to discriminate between classes. Taken altogether, these results support the validation of AdvancedSleepClassifier as a potent, high-accuracy means of automatic identification of sleep stages across complex single-channel EEG datasets.

The training and validation loss and accuracy curves (Fig. 7) demonstrate a consistent convergence trend, with steadily decreasing loss and increasing accuracy across epochs, indicating stable optimization behavior. Similarly, the multi-class ROC curves (Fig. 8) show high separability across all sleep stages, with AUC values consistently exceeding 0.90, suggesting strong discriminative capability of the proposed model. In the current version of the manuscript, these curves represent the mean performance across the 10-fold subject-independent cross-validation for visual clarity. Individual fold curves were not plotted to avoid visual clutter and to maintain readability of the figures.

Table 3 Comparative analysis for the ISRUC dataset

Model	Architecture	Accuracy (%)	Cohen's κ
DeepSleepNet [17]	CNN + Bi-LSTM (raw EEG)	73.00	0.65
MetaSleepLearner †[9]	Meta-learning with CNN-RNN	71.00	0.62
DSSNet† [8]	Single-channel CNN	71.39	0.63
Co-ScaleNet† [18]	Multi-scale CNN	84.60	0.80
RUSBoost† [36]	Ensemble learning with class balancing	78.80	0.73
FFTCN (cross-dataset) [37]	FFT-based CNN	77.82	0.71
SFNet [31]	Spectral feature fusion network	78.40	0.72
Proposed AdvancedSleepClassifier	Dilated CNN + SE + Bi-LSTM + Multi-Head Attention	79.97	0.72

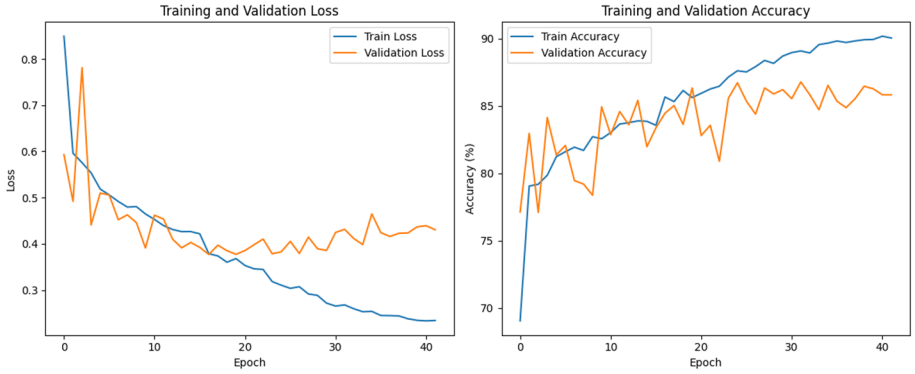

Fig. 9 Training and validation metrics over epochs for 5-class classification with SleepEDF dataset

4.1.1.1 Comparative analysis Table 3 presents a comparative performance analysis of the proposed AdvancedSleepClassifier against several state-of-the-art sleep staging models evaluated on the ISRUC dataset under a 5-class sleep stage classification setting. The compared approaches represent a diverse range of methodological paradigms, including CNN–RNN hybrid architectures (DeepSleepNet), meta-learning frameworks (MetaSleepLearner), single-channel CNN-based models (DSSNet), multi-scale convolutional networks (Co-ScaleNet), ensemble learning with class balancing (RUSBoost), FFT-based CNNs for cross-dataset learning (FFTCN), and spectral feature fusion networks (SFNet). Performance is evaluated using overall accuracy, macro F1-score, and Cohen's kappa, providing a balanced assessment of classification effectiveness and inter-class agreement. While Co-ScaleNet achieves the highest accuracy and F1-score, it relies on a comparatively complex multi-scale architecture. The proposed AdvancedSleepClassifier demonstrates competitive performance with an accuracy of 79.97%, an F1-score of 79.00%, and a Cohen's kappa of 0.72, outperforming several established baselines such as DeepSleepNet, MetaSleepLearner, DSSNet, FFTCN, and SFNet. These results highlight the effectiveness of integrating dilated convolutional feature extraction, residual learning, bidirectional temporal modelling, and attention mechanisms, particularly in handling the inherent variability of the ISRUC dataset.

4.1.2 SleepEDF dataset

Figures 9 and 10 show that a thorough assessment of the AdvancedSleepClassifier on the five-class sleep stage classification on the SleepEDF dataset revealed particularly

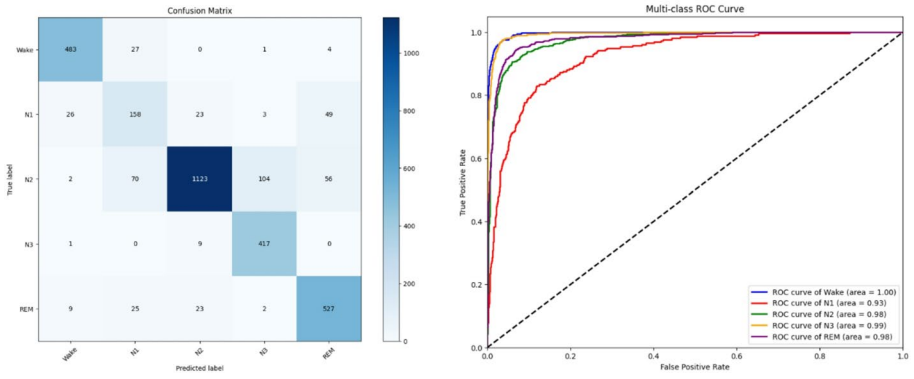

Fig. 10 ROC curve and confusion matrix for 5-class classification with SleepEDF dataset

Table 4 Comparative analysis for the Sleep-EDF dataset

Model	Architecture / Approach	Accuracy (%)	Cohen's κ
Tsinalis et al. (2016) [32]	CNN on raw single-channel EEG	74.0	0.60
DeepSleepNet (2017) [21]	CNN + Bi-LSTM	82.0	0.72
Phan et al. (2019) [33]	CNN with epoch-level prediction	82	0.73
U-Time (2019)[34]	Fully convolutional U-Net-based model	83	0.74
SleepEEGNet (2019)[17]	Seq-to-Seq CNN	84.2	0.75
TinyUStaging (2023) [35]	Lightweight U-Net + focal loss	84.0	0.76
Proposed method	Residual CNN + Attention + BiLSTM + Spectral Fusion	85.07	0.79

fine results. The plotted loss curves show a consistent decline marked with validation loss stabilizing after a period of initial oscillation, whereas accuracy curves continue to improve and comeback at above 85% with training. The classifier did well on the test set with an overall accuracy of 85.07%, macro-averaged F1-score of 0.81 and weighted average F1-score of 0.85, demonstrating the same effectiveness in all sleep stages. The classification report shows astonishingly high precision and recall for Wake: precision = 0.92, recall = 0.95, F1 = 0.94. advanced performance for N2 (precision: 0.92, recall: 0.88, F1: 0.90) and N3 stages, and it also demonstrates robust results for REM. Given such a transitional nature, the N1 stage, which is often hardest to classify, has lower, but still satisfactory scores (precision: 0.52, recall: 0.53, F1: 0.53). The confusion matrix indicates that most misclassifications issue from N1 and adjacent stages, which is widely viewed as a recognized challenge in sleep staging. The hospitality of the multi-class ROC analysis reflects the hard-won discriminative capability of the classifier, showing AUC values at 1.00 for Wake, 0.93 for N1, 0.98 for N2, 0.99 for N3, and 0. These findings prove that AdvancedSleepClassifier provides accurate, trustworthy, and flexible performance in sensitive EEG data sets in the real world, placing it among sufficient tools for automatic detection of sleep stages.

Table 4 presents a comparative evaluation of the proposed method against representative state-of-the-art sleep staging approaches on the Sleep-EDF dataset. Early CNN-based models, such as Tsinalis et al. (2016), report relatively low performance, achieving an accuracy of 74.0% and a Cohen's κ of 0.60, indicating limited agreement beyond chance.Hybrid deep learning architectures, including DeepSleepNet (2017) and the CNN-based epoch prediction framework by Phan et al. (2019), improve temporal modeling and achieve accuracies around 82% with κ values in the range of 0.72–0.73. More recent fully convolutional and sequence-based models, such as U-Time (2019) and

SleepEEGNet (2019), further enhance performance, reaching accuracies of 83–84.2% and κ values up to 0.75, reflecting improved robustness in sleep stage discrimination.The lightweight TinyUStaging (2023) model achieves competitive performance (84.0% accuracy, $\kappa = 0.76$) while emphasizing computational efficiency. In contrast, the proposed method outperforms all compared approaches, achieving the highest accuracy of 85.07% and a Cohen's κ of 0.79, indicating stronger inter-class agreement and more reliable classification. This improvement can be attributed to the integration of residual convolutional feature extraction, self-attention-based feature refinement, bidirectional temporal modeling, and spectral feature fusion, which collectively enhance both discriminative power and temporal consistency.

Table 5 compares the proposed method with existing approaches on the SleepEDF dataset, focusing on class-wise performance across individual sleep stages (W, N1, N2, N3, and REM) in addition to macro F1-score. Unlike Table 1, this comparison emphasizes the impact of different loss functions and learning strategies on sleep stage discrimination, particularly for underrepresented stages such as N1 and N3.

The comparative analysis presented in Table 4 evaluates the performance of existing state-of-the-art sleep staging methods and the proposed approach using the macro F1-score, which is a robust metric for imbalanced multi-class classification problems such as sleep stage classification. Among the existing approaches, SleepGCN (C) achieves the highest F1-score of 85.20%, demonstrating the effectiveness of graph-based modelling combined with an optimised weighted cross-entropy loss. However, this improvement comes at the cost of increased architectural complexity. SleepGCN (A) and SleepGCN (B) show slightly lower F1-scores of 83.42% and 84.77%, respectively, indicating the sensitivity of graph-based models to loss-weight configurations. Traditional deep learning models such as DeepSleepNet, which relies on a CNN–BiLSTM architecture with standard cross-entropy loss, achieve a comparatively lower F1-score of 78.70%, highlighting limitations in handling class imbalance. Fully convolutional architectures like U-Time, optimized using a Dice cost function, improve performance to 80.00% by better modelling temporal continuity, while TinyUStaging, a lightweight U-Net–based model employing focal loss, further improves robustness, achieving an F1-score of 81.10%. The proposed method attains an F1-score of 85.07%, outperforming DeepSleepNet, U-Time, and TinyUStaging, and achieving performance comparable to the best-performing graph-based approach (SleepGCN (C)). This demonstrates that the use of class-weighted cross-entropy loss, combined with an advanced deep learning architecture, effectively mitigates class imbalance while maintaining a favourable balance between performance and model complexity. Overall, the results confirm that the proposed method offers state-of-the-art competitive performance without the overhead associated with graph-based frameworks.

Table 5 Comparative analysis for SleepEDF dataset

Method	Loss function	F1score (%)
SleepGCN (A)[38]	Weighted cross-entropy: A	83.42
SleepGCN (B)[38]	Weighted cross-entropy: B	84.77
SleepGCN (C)[38]	Weighted cross-entropy: C	85.20
DeepSleepNet [21]	Cross-entropy	78.70
U-Time [34]	Dice cost function	80.00
TinyUStaging [35]	Focal loss	81.10
Proposed Method	Class-weighted Cross-Entropy	85.07

4.1.3 Evaluation metrics

The bar graph compares the performance of the AdvancedSleepClassifier for the performance on the 5-class sleep stage dataset on two benchmark datasets. ISRUC and SleepEDF. Four fundamental measures of evaluation-Accuracy, Precision, Recall and F1 Score-are presented along the x-axis, whereas the values along the y-axis depict the same. The blue bars are the results from the ISRUC dataset, and the orange bars are those of the SleepEDF dataset. On all measures, ISRUC is outperformed by AdvancedSleepClassifier on the SleepEDF dataset. For the SleepEDF, the model results in an accuracy of 0.85, precision of 0.87, recall of 0.85 and an F1 score of 0.85. In turn, performance of the ISRUC dataset is slightly lower and quite homogeneous, with 0.80 in accuracy, recall, and F1 score, and a precision of 0.81. Such continuous boosts on SleepEDF indicate that the classifier generalizes well and is highly promising in testing on this dataset, owing to the difference in data quality, subject differences or recording conditions. The increased precision results on SleepEDF represent the model's capacity to decrease the false positive rate, and the balanced recall and F1 values suggest consistent detection at each of the five different sleep stages. In general, the chart brings to light the strength and flexibility of the AdvancedSleepClassifier, and this effectiveness is particularly evident when the system is deployed on disparate real-world EEG records for multi-class sleep stage prediction.

4.2 4-class classification

In the 4-class experiment, we divide the night's sleep into Wake, N1, N2 and a class that includes N3 and REM. For the 4-class classification setting, the N2 and N3 sleep stages were merged into a single class. This decision was motivated by the close physiological relationship between these non-REM sleep stages and their overlapping EEG characteristics, particularly the dominance of slow-wave activity and sleep spindle patterns. Combining N2 and N3 reduces inter-class ambiguity and mitigates class imbalance, thereby improving training stability and robustness in limited-data and subject-independent evaluation scenarios. It is important to note that the primary objective of this reduced-class experiment is to evaluate the robustness of the proposed framework under simplified classification settings rather than to replace clinical sleep staging. The standard 5-class classification results, where N2 and N3 are treated as distinct stages, are therefore reported as the main clinical evaluation (Fig. 11).

4.2.1 ISRUC dataset

Figures 12 and 13 show that the AdvancedSleepClassifier obtains high accuracy and generalises consistently well for four-class sleep stage classification. Both training and validation loss plots show a gradual drop in error with validation loss stabilising, hence indicating successful learning and lack of overfitting up until 50 training epochs (1). In addition to this, the accuracy curves show the incremental growth pattern, validation accuracy being consistent with training accuracy and the latter exceeding 85% during training, which speaks much about the model's stability and credibility.

With an overall accuracy of 85.36% on the test set, the classifier also shows an extension to macro and weighted average F1-scores of 0.84 and 0. This shows equal efficacy in the classification of all four sleep stages. In particular, performance in the classification report for the wake stage is better (precision: 0.90, recall: 0.87, F1: 0.88). Furthermore,

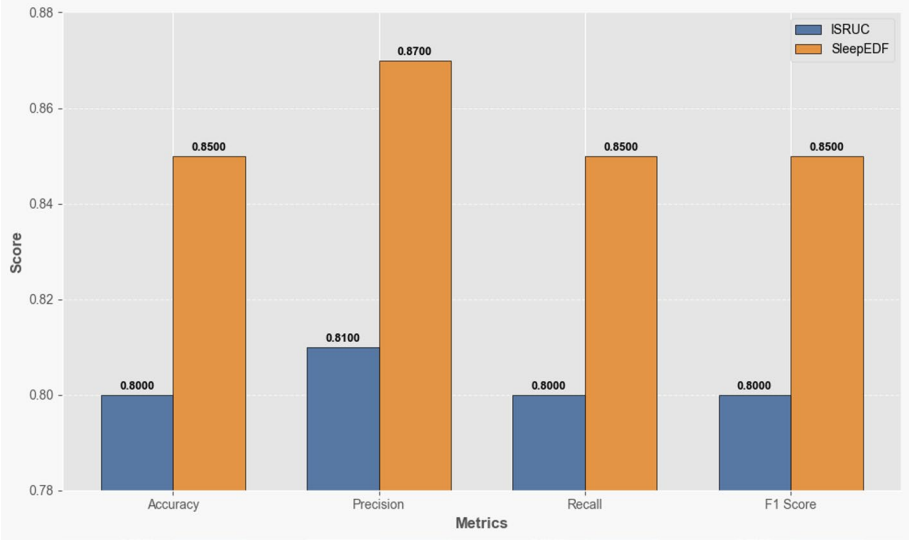

Fig. 11 Evaluation metrics for 5-class classification

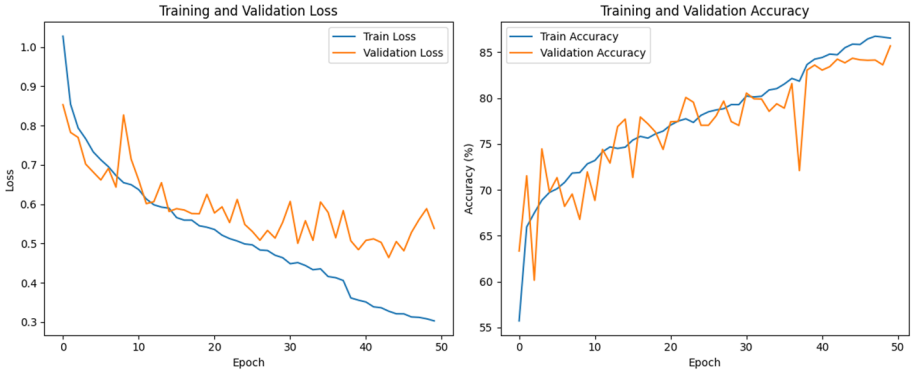

Fig. 12 Training and validation metrics over epochs for 4-class classification with the ISRUC dataset

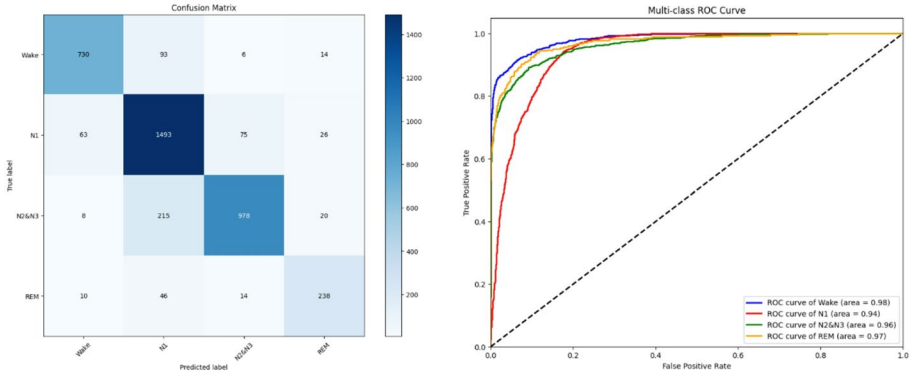

Fig. 13 ROC curve and confusion matrix for 4-class classification with ISRUC dataset

the N1 stage is supported by stable recall (0.90) and strong precision (0.81), evidence that the model's ability to identify this challenging stage is accurate. Per the confusion matrix, mostly errors occur between N1& N2 &N3 in accordance with the physiological characteristics of these stages 3.

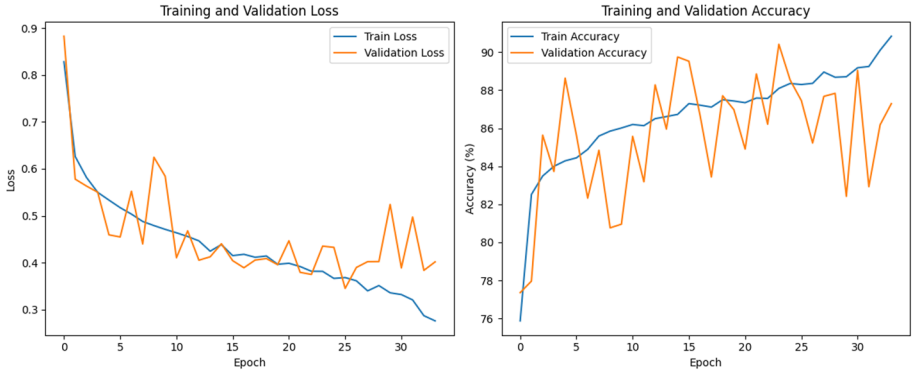

Fig. 14 Training and validation metrics over epochs for 4-class classification with SleepEDF dataset

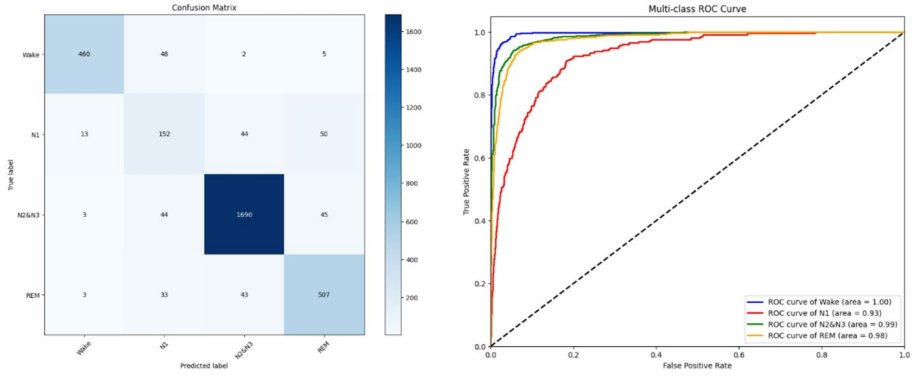

Fig. 15 ROC curve and confusion matrix for 4-class classification with SleepEDF dataset

Multiple-class ROC analysis supports the accuracy of the model, providing AUC scores such as 0.98 for Wake, 0.94 for N1, 0.96 for N2&N3, and 0.97 for REM, representing better discrimination among all stages 4. The results, when combined, show that AdvancedSleepClassifier is capable of detecting all four sleep stages in a single-channel EEG reading, distinct as an accurate and viable solution for practical sleep monitoring solutions.

4.2.2 SleepEDF dataset

Figures 14 and 15 show that the AdvancedSleepClassifier demonstrates robust accuracy and generalization ability when the four-class sleep stage classification task is applied to the SleepEDF dataset. The training/ validation loss graphs demonstrate an unwavering decline in loss, where the validation loss is sometimes erratic but ultimately mirrors the decline of loss of training loss, demonstrating good learning and little overfitting1. The accuracy curves show a constant increase, with validation accuracy close to 90% at the end of training and similar to the training across all epochs. When tested on the test set, the classifier has an excellent overall accuracy of 89.4%, and F1-scores of 0.82, weighted average F1-scores of 90%, which displays uniform performance in all four sleep stages. Interestingly, the Wake stage produces better precision, recall, and F1-score on the classification report. The classifier does an outstanding job on Wake (precision: 0.96, recall: 0.89, F1: 0.93) and N2&N3 (precision: The N1 stage, which is commonly less identifiable, given its transitional properties, shows the stable performance with precision (0.55)

as well as the recall (0.59). 0.55, recall: 0.59, F1: 0.57)2. Errors are commonly observed when N1 is classified against its neighbouring stages, as reflected in the confusion matrix, demonstrating a common problem in sleep staging3. In addition, the multi-class ROC analysis demonstrates the ability of the classifier to separate classes effectively as indicated by the AUC values of 1.00 for Wake, 0.93 for N1, 0.99 for N2 and N3, and 0. All of the outcomes together show the power and usability of AdvancedSleepClassifier as an automated four-class sleep stage analysis tool from one EEG recording.

4.2.3 Evaluation metrics

The bar graph presents comparative performance measures of the AdvancedSleepClassifier performance on the 4-class sleep stage classification function when measuring it with two widely adopted EEG datasets. ISRUC and SleepEDF. The axis of x depicts 4 important metrics-Accuracy, Precision, Recall, F1 Score – whereas the axis of y shows the corresponding score values. Results presented as blue bars represent results on the ISRUC dataset, and orange bars represent results on the SleepEDF dataset.

In all the metrics, the classifier demonstrates significant improvement on the SleepEDF dataset compared to the SleepEDF test set. In particular, on SleepEDF, the model reports R of 0.89, P of 0.90, R of 0.89 and F1-score of 0.90, suggesting very balanced and robust classification in all four stages of sleep. As opposed to this, on the ISRUC-Sleep dataset, scores are slightly lower yet impressive with accuracy, recall and F1 scores of 0.85 each and precision of 0.86 (Fig. 16).

Underlying performance-gap indicate that the AdvancedSleepClassifier generalizes well while achieving a peak performance, with the SleepEDF dataset, maybe as a result of data quality or diverse subjects and recording protocols. The consistently high precision recall and F1 scores for SleepEDF demonstrates the utilities of the model to accurately identify and classify sleep stages, with minimal false positives or negatives. Overall, the results presented in the graph highlight the efficiency and resilience of the classifier for application to four-class sleep staging, especially for high-quality EEG data.

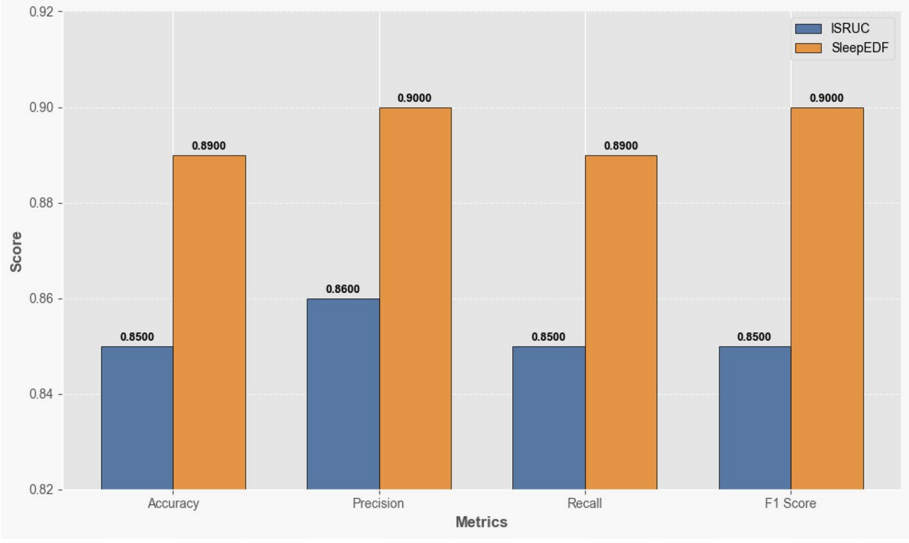

Fig. 16 Evaluation metrics for 4-class classification

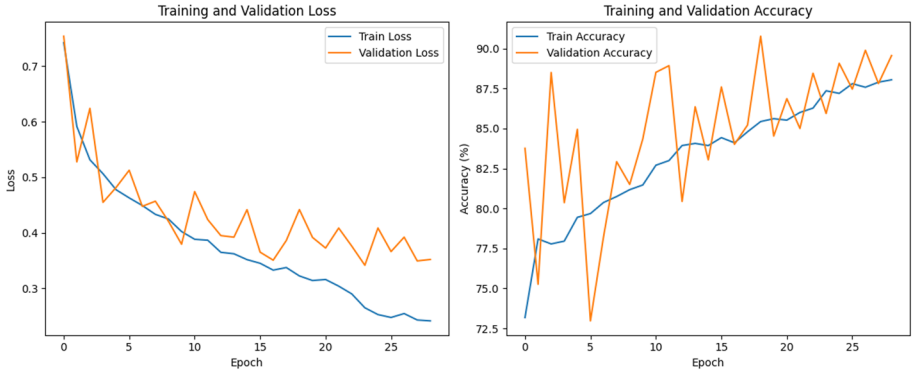

Fig. 17 Training and validation metrics over epochs for 3-class classification with the ISRUC dataset

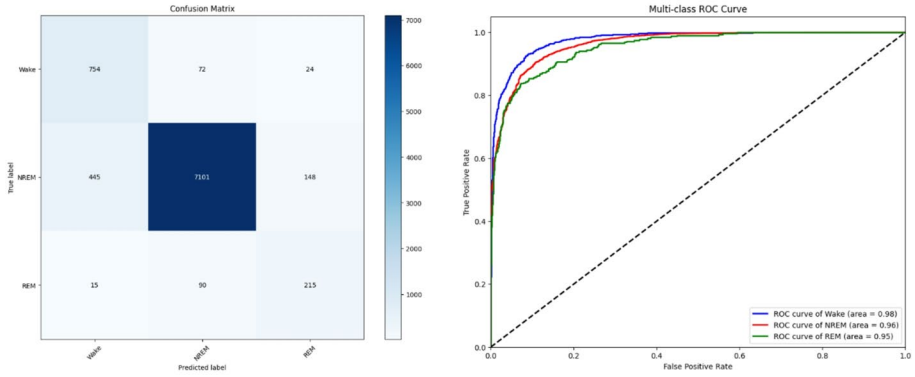

Fig. 18 ROC curve and confusion Matrix for 3-class classification with the ISRUC dataset

4.3 3-class classification

The scheme rules that each EEG segment should be labelled as Wake, NREM (covering all N1-N3 stages) or REM. As a result, the model demonstrates it can detect REM sleep, which is known for being connected to dreaming and key brain functions.

4.3.1 ISRUC dataset

Figures 17 and 18 show that the AdvancedSleepClassifier outperformed in the classification of sleep stages into three classes in the ISRUC dataset. The training and validation loss curves can be observed to follow a gradual decrease trend, with the validation loss closely following the training loss, indicating learning is being effective with limited overfitting1. The accuracy curves show a stable and increasing graph which finally reaches to validation accuracy of 90% at the end of training and thus exhibits the model's stability and generalisation.

In the test set, the classifier achieved an overall accuracy of 91.3%, macro and weighted F1-scores of 0.9. The class Awake performed exceptionally well in terms of precision, as shown in the classification report. For the Wake class, precision, recall, and F1-score all achieve 0.96, 0.95, and 0.95, respectively.

The confusion matrix gives additional evidence of this being the case, as most of the predictions fit correctly on the classes and misclassification is very minimal between the classes. The multi-class ROC analysis presents outstanding discriminative performance, with AUC values of 1.00 for Wake, and Combining the available evidence, it is clear

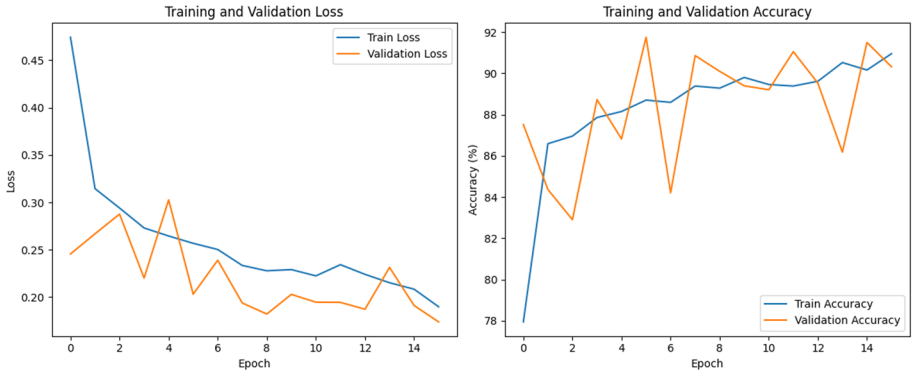

Fig. 19 Training and validation metrics over epochs for 3-class classification with SleepEDF dataset

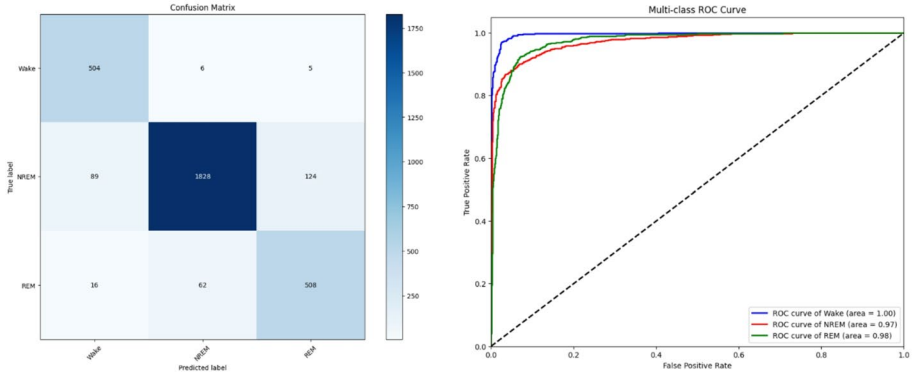

Fig. 20 ROC curve and confusion matrix for 5-class classification with SleepEDF dataset

that AdvancedSleepClassifier is an excellent, generalisable solution for classifying three stages of sleep from single-channel EEG readings with high accuracy.

4.3.2 SleepEDF dataset

Figures 19 and 20 show that the AdvancedSleepClassifier was thoroughly tested on the SleepEDF dataset for a 3-class Sleep stage classification with Wake, NREM and REM intonations. The test accuracy was 90.39%, showing the models' strong generalization capabilities. The fact that both training and validation accuracy curves remain stable in time, accompanied by a gradual decay in validation loss across epochs, suggests that there is only limited overfitting, and the training does not diverge.

Over the lifetime of the prototype, the classification report gives excellent results with F1-scores of 0.90 for Wake, 0.93 for NREM, and 0.83 for REM. The model demonstrated excellent recall for the Wake stage (0.98); however, precision-recall was stable for the remaining stages. With a macro and weighted F1-scores of 0.89 and 0.91, respectively, the model shows resilience to class imbalance as well as overall competent superiority. From the confusion matrix, the classification was very accurate, particularly with NREM, where 1828 samples, out of the 2041, were correctly classified. Confusion between the NREM and REM stages has been noted, perhaps because the EEG traces for these stages exhibit a high degree of overlap. ROC analysis was also instrumental towards the confirmation of the reliability of the models, showing AUC scores of 1.00 for Wake,

The obtained results corroborate the capacity of the hybrid deep learning construction to encode fine temporal relations in restricted EEG data communication channels. Reported with evidence of high potential, this classifier would quite well support automated sleep stage monitoring for setups that are poorly equipped in terms of EEG channel resources.

4.3.3 Evaluation metrics

The bar chart describes the relative performance of the AdvancedSleepClassifier upon a three-class sleep stage classification task evaluated on two standard EEG corpora. ISRUC and SleepEDF. X axis shows critical evaluation metrics – Accuracy, Precision, Recall and F1 Score, whereas y axis shows respective score values. Results extracted from the ISRUC dataset are illustrated in blue, and from the SleepEDF dataset are illustrated in orange.The classifier shows an excellent performance over both datasets, while the ISRUC results perform slightly better. That is, for the ISRUC dataset, the classifier has an accuracy of 0.91, a precision of 0.93, a recall of 0.91 and an F1 score of 0.92. Conversely, in the SleepEDF dataset, the results are slightly lower. accuracy and recall are 0.90, the precision and F1 score are at 0.91.

These results suggest that the model generalizes well across datasets but shows a marginally superior performance for ISRUC, perhaps attributed to variations in data distribution, preprocessing or among inter-subject variations. The consistently excellent F1 scores of both data providers highlight how reliably the model manages to discriminate the three sleep stages with a sensible compromise of precision and recall. This emphasises the strength and flexibility of the AdvancedSleepClassifier to practical EEG- centred sleep-staging (Fig. 21).

4.4 2-class classification

Two-class classification involves singling out Wake and treating it as separate from sleeping. Together, all non-wake stages, labelled as N1, N2, N3 and REM, are classified as one “Sleep” sleep stage. Because of their simplicity, this division is usually used in light models or in screening tools made for doctors.

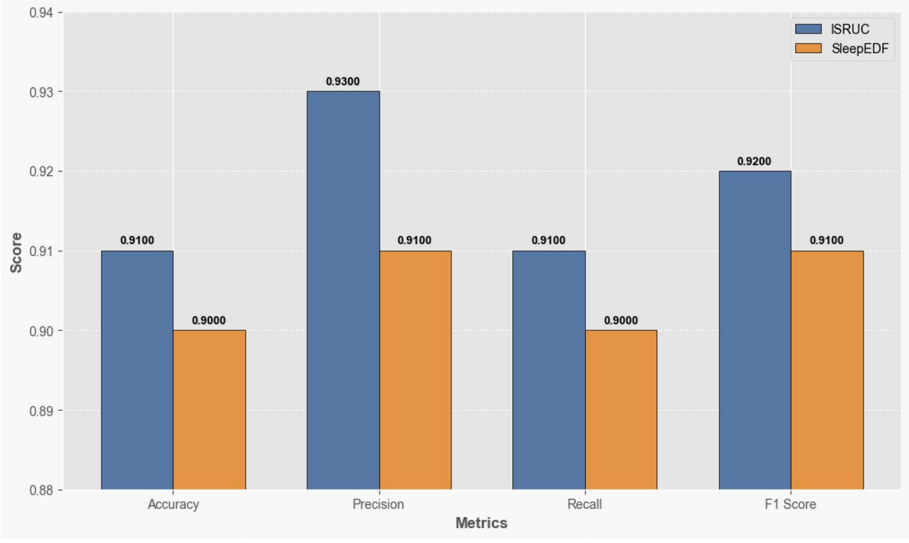

Fig. 21 Evaluation metrics for 3-class classification

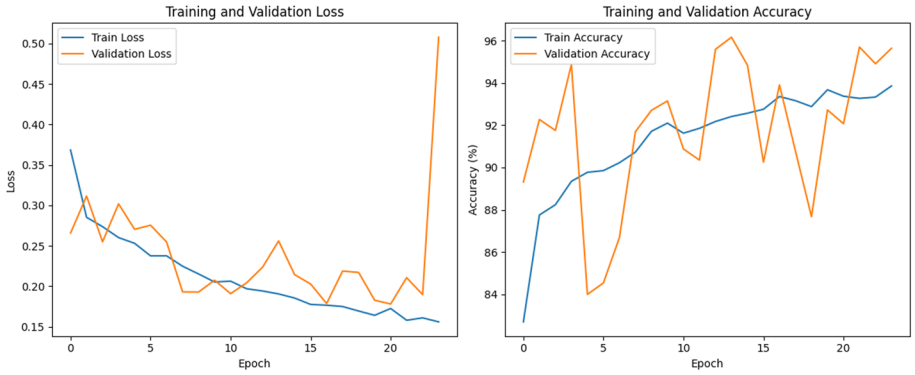

Fig. 22 Training and validation metrics over epochs for 2-class classification with ISRUC dataset

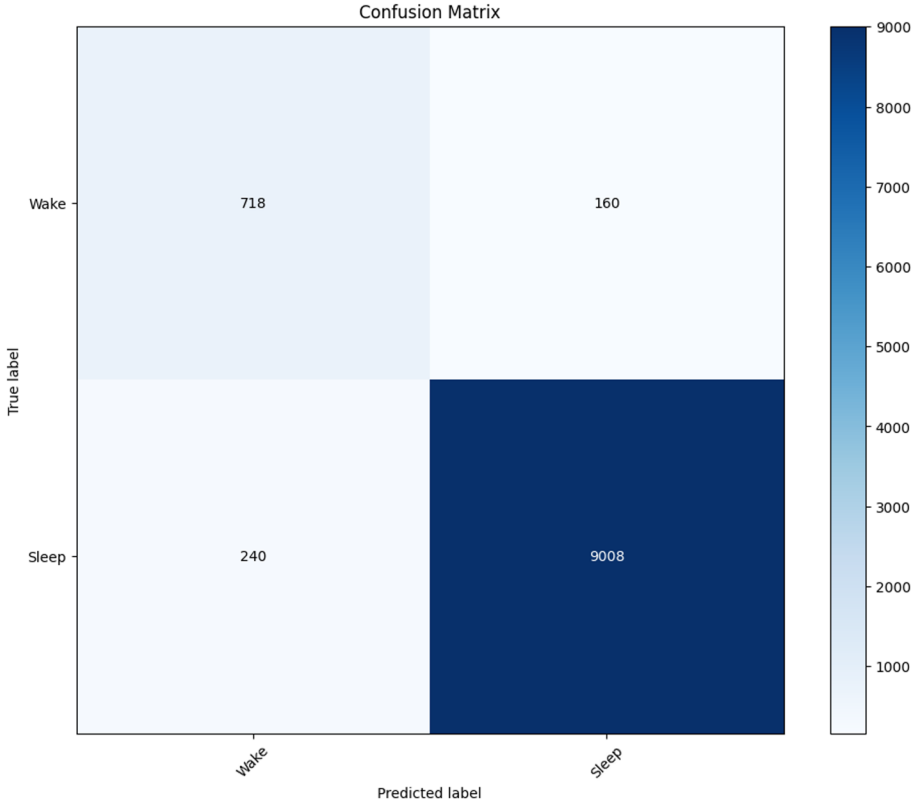

Fig. 23 Confusion matrix for 2-class classification with ISRUC dataset

4.4.1 ISRUC dataset

Figures 22 and 23 show that the sleep staging model was well trained for binary classification-Sleep vs. Wake on the ISRUC dataset and achieved a high test accuracy of 96.05%. According to the training and validation curves results, learning is successful because training loss has continued to drop while accuracy has also grown. However, there are visible oscillations in validation loss, where there is a sudden spike in the final epoch, which may indicate overfitting. The classification report shows amazing performance; the Sleep class shows a precision and recall 97% and an F1-score 98%.

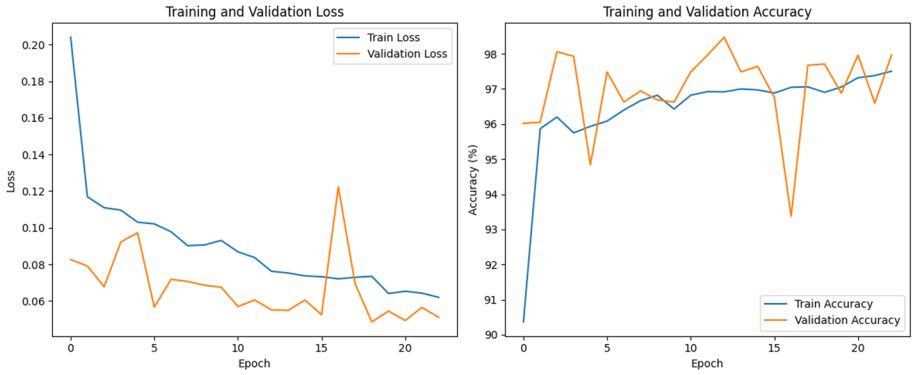

Fig. 24 Training and validation metrics over epochs for 2-class classification with SleepEDF dataset

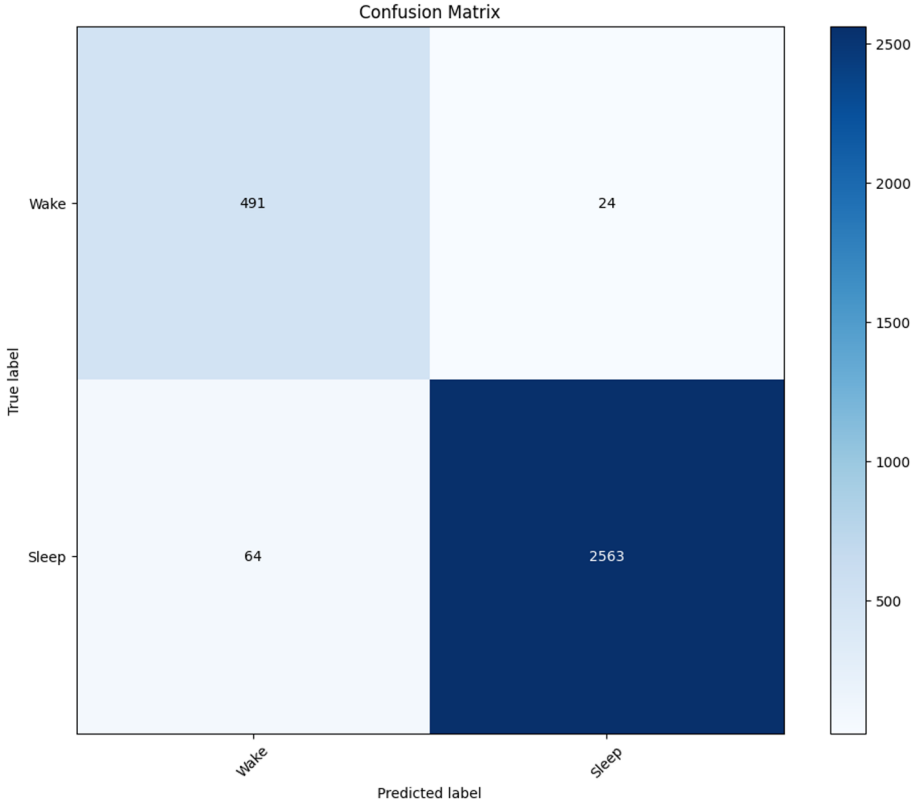

Fig. 25 Confusion matrix for 2-class classification with SleepEDF dataset

4.4.2 SleepEDF dataset

Figures 24 and 25 show that the model displayed extraordinary performances. During training, both training and validation losses gradually improved, with training loss dropping as training took place over 22 epochs. Training and validation accuracies, in turn, also rose steadily, achieving highs of approximately 97.5% and 98.5%, respectively, and with very few indications of overfitting. Eventually, the test accuracy was at 97.20%. Both classes had excellent precision and recall as depicted by the classification report, including: The model showed 0.88 precision and 0.95 recall for wake epochs and 0.99 precision and 0.98 recall for sleep epochs, which in total equates to an F1-score of 0.97. The classification outcomes, which were provided by the confusion matrix, illustrated precise

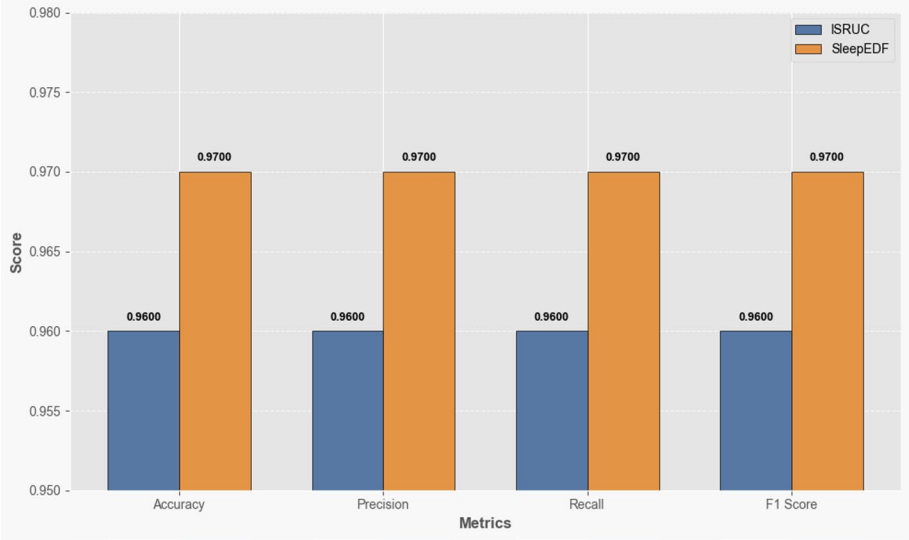

Fig. 26 Evaluation metrics for 2-class classification

detection of 491 out of the 515 wake epochs as well as 2563 out of the 2627 sleep epochs. The macro and weighted averages of the model on the precision, recall, and F1-score were all over 0.94, thus making it robust and effective for varying classes. The model exhibits a superior capability to discriminate sleep-wake states, thus validating its applicability for binary sleep stage classification in the SleepEDF dataset.

4.4.3 Evaluation metrics

The bar chart illustrates the comparison of performance advanced by the AdvancedSleepClassifier on a binary (2-class) sleep stage classification task on the two most popular EEG datasets – the Sleep Darwin dataset and the dimension dating back to emission frequency dataset. ISRUC and SleepEDF. The x-axis represents four evaluation metrics (Accuracy, Precision, Recall, F1 Score), and the y-axis represents scores of performances. The blue represents the ISRUC dataset, and the yellow bars specify the SleepEDF dataset.The classifier demonstrates promising and stable performance on both datasets, SleepEDF – providing slightly higher results on all four measures. Particularly, for the SleepEDF dataset, the classifier yields a score of 0.97 in all the metrics used – accuracy, precision, recall and F1 score, indicating near-perfect classifying ability. On the ISRUC dataset, every metric manages a robust 0.96, which indicates strong, but somewhat decreased performance.

The tight performance margin indicates high generalisation power of the model in binary classification, balanced precision and recall at all times, limiting false positives and false negatives. The marginal advantage visible in the SleepEDF outcomes may be explained by such a factor as the difference in data specificity – signal clarity, subjects' homogeneity, annotation quality. Comparing the data, overall, the chart shows the strength and flexibility of the AdvancedSleepClassifier, which qualifies it well for real-world purposes that demand stable two-class identification of sleeping stages (Fig. 26).

5 Conclusion

In this study, we presented AdvancedSleepClassifier, a new hybrid deep learning paradigm that is oriented towards automated assignment of sleep stages based on single-channel EEG data. Using a combination of raw EEG signal processing and spectrum-based techniques, the design is compatible with clinical knowledge through incorporating the details related to the spectral of bands, resulting in better accuracy and easier interpretation. AdvancedSleepClassifier was tested on two benchmark datasets, ISRUC-Sleep and SleepEDF, and demonstrated strong generalization capabilities over different levels of sleep stage granularity. categories of five, four, and three stages, as well as binary Wake versus Sleep categorizations. It achieved such levels of performance with test accuracies of 85.07% (five-class SleepEDF), 89.4% (four-class SleepEDF), 91.3% (three-class ISRUC) and 97.2% (two-class SleepEDF). The model functioned quite well when classifying the Wake and N3 stages: either in an obvious or medically helpful way, and maintained satisfactory precision and recall for the tougher N1 stage. Through the advancement of EEG-based sleep analysis, AdvancedSleepClassifier can serve as a milestone for accurate, transparent, and accessible diagnostics for the user, which will streamline efficient and technology-supported solutions for everyday clinical care of sleep assessment. The proposed model still faces challenges in accurately classifying the N1 stage due to its transitional characteristics. Furthermore, model complexity and limited evaluation on real-world clinical data remain constraints. Future work will focus on improving minority-stage discrimination, reducing computational overhead, and validating the framework on larger and more diverse datasets.

Author contributions
All authors contributed to conception, analysis, and manuscript preparation.

Funding
No specific funding received.

Data availability
Data available on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent to publication
Not applicable.

Human and animals
Not applicable.

Informed consent
Not applicable.

Competing interests
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.

Received: 24 September 2025 / Accepted: 2 March 2026
Published online: 27 March 2026

References

1. Satapathy SK, Loganathan D. Prognosis of automated sleep staging based on a two-layer ensemble learning stacking model using single-channel EEG signals. Soft Comput. 2021;25:15445–62. https://doi.org/10.1007/s00500-021-06218-x.
2. Iber I, Ancoli-Israel S, Chesson A, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine; 2007.
3. Satapathy SK, Loganathan D. Automated classification of multi-class sleep stages using polysomnography signals: a nine-layer 1D convolutional neural network approach. Multimedia Tools Appl. 2022. https://doi.org/10.1007/s1042022-13195-2.

4. Lee JK, et al. Automated sleep scoring based on polysomnographic data using deep learning. IEEE Trans Neural Syst Rehabil Eng. 2020;28:1381–90.
5. Chambon F, Galtier M, Arnal J, Wainrib E, Gramfort A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Trans Neural Syst Rehabil Eng. Apr. 2018;26(4):758–69.
6. Supratak Y, Dong H, Wu C, Guo Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG, *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.
7. Zhang ZX et al. Apr., Sleep transformer: automatic sleep stage classification with interpretability and temporal context, *IEEE J. Biomed. Health Inform.*, vol. 26, no. 4, pp. 1715–1725, 2022.
8. Cui Y, Pan Z, Wang J, Wang H. DSSNet: A deep sleep staging network based on single-channel EEG. IEEE Access. 2020;8:122957–66. https://doi.org/10.1109/ACCESS.2020.3007391.
9. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. Meta-learning for automatic sleep stage classification, *IEEE Trans. Biomed. Eng.*, vol. 67, no. 10, pp. 2732–2743, Oct. 2020, https://doi.org/10.1109/TBME.2020.2975180
10. Luo J, Hao J, Pan L. An automatic sleep staging method based on CNN–BiLSTM, Trans. Beijing Inst Technol., vol. 40, no. 7, pp. 746–752, Jul. 2020.
11. Guillot A, Thorey V. RobustSleepNet: transfer learning for automated sleep staging at scale, *arXiv preprint arXiv:2101.02452*, 2021.
12. Phan H et al. XSleepNet: multi-view sequential model for automatic sleep staging, *arXiv preprint arXiv:2007.05492*, 2020.
13. Yang C et al. LWSleepNet: a lightweight attention-based deep learning model for sleep staging with single-channel EEG. Digit Health, 9, 2023.
14. Li Y, Wang S, Tang W, Sun J, Chen H. Co-ScaleNet: A multi-scale deep learning network for sleep stage classification using single-channel EEG. IEEE Trans Instrum Meas. 2022;71:Art2504412. https://doi.org/10.1109/TIM.2022.3140275.
15. Einizade A et al. ProductGraphSleepNet: sleep staging using product spatio-temporal graph learning with attentive temporal aggregation, *arXiv preprint arXiv:2212.04881*, 2022.
16. Jia Z et al. SalientSleepNet: multimodal salient wave detection network for sleep staging, *arXiv preprint arXiv:2105.13864*, 2021.
17. Mousavi S, Afghah F, Acharya UR. SleepEEGNet: automated sleep stage scoring using sequence-to-sequence deep learning approach. IEEE J Biomed Health Inf. May 2019;23(3):1271–9. https://doi.org/10.1109/JBHI.2018.2879483.
18. Chambon A, Galtier M, Arnal 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(4):758–69.
19. Hu J, Shen L, Sun G. Squeeze-and-excitation networks, In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
20. Vaswani A et al. Attention is all you need. In Adv Neural Inf Process Syst (NeurIPS), 2017, pp. 5998–6008.
21. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG, IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, Nov. 2017, https://doi.org/10.1109/TNSRE.2017.2721116
22. Phan X, Andreassen O, Salvanes A. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng. 2019;27(3):400–10.
23. Biswal PC et al. The Sleep-EDF database expanded with age and sex matched healthy controls. PhysioNet, 2019.
24. Dement W, Rechtschaffen A. Manual of standardized terminology, techniques and scoring System for sleep stages of human subjects. Los Angeles: UCLA Brain Information Service; 1968.
25. Iber C, Ancoli-Israel S, Chesson AL, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine; 2007.
26. Eldele A, et al. An attention-based deep learning approach for sleep stage classification with single-channel EEG. IEEE Trans Neural Syst Rehabil Eng. 2021;29:809–18.
27. Berry M, et al. The AASM inter-scorer reliability program: Sleep stage scoring. J Clin Sleep Med. 2015;11(1):111–7.
28. Welch P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust. 1967;15(2):70–3.
29. Zhao Z, Zhang M, Zhou D. DeepSleepTransformer: a multi-channel self-attention transformer network for sleep stage classification. IEEE J Biomed Health Inf. 2022;26(4):1322–33.
30. Sun S, Zhang Z, Wu M. Sleep EEG signal classification using transformer-based neural networks, In: *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2021, pp. 1–8.
31. Chen J, Fan X, Ge R, Xiao J, Wang R, Ma W, Li Y. Towards interpretable sleep stage classification with a multi-stream fusion network, *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, Art. no. 164, Apr. 2025, https://doi.org/10.1186/s12911-025-02995-9
32. Tsinalis O, Matthews P, Guo Y. Automatic sleep stage scoring using convolutional neural networks. ann Biomed Eng. May 2016;44(5):1587–97. https://doi.org/10.1007/s10439-015-1446-1.
33. Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE Trans Biomed Eng. May 2019;66(5):1285–96. https://doi.org/10.1109/TBME.2018.2872652.
34. Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum P, Igel C. U-time: a fully convolutional network for time series segmentation applied to sleep staging. Adv Neural Inf Process Syst (NeurIPS), pp. 4415–26, 2019.
35. Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A lightweight deep learning architecture for automatic sleep staging, *Comput. Biol. Med.*, vol. 156, Art. no. 107127, 2023, https://doi.org/10.1016/j.compbiomed.2023.107127
36. Seera M, Lim CP. A hybrid intelligent system for medical data classification using RUSBoost. Appl Soft Comput. Jan. 2013;13(1):353–61. .1016/ j.asoc. 2012.08.040.
37. Zhang Y, Chen Z, Wang S, Zhang D. Cross-dataset automatic sleep stage classification using FFT-based convolutional neural networks, Biomed. Signal Process. Control, vol. 62, Art. no. 102078, 2020, https://doi.org/10.1016/j.bspc.2020.102078
38. Zhang X, et al. Graph convolutional networks for automatic sleep staging. IEEE Trans Neural Syst Rehabil Eng. 2023;31:1–12.

Publisher's note