KAleep-Net: A Kolmogorov-Arnold Flash Attention Network for Sleep Stage Classification Using Single-Channel EEG with Explainability

Zubair Akbar, *Graduate Student Member, IEEE*, Farhad Hassan, *Graduate Student Member, IEEE*, Jingzhen Li, *Member, IEEE*, Ubaidullah alias Kashif, *Graduate Student Member, IEEE*, Yuhang Liu, *Member, IEEE*, Jia Gu, *Senior Member, IEEE*, Kaixin Zhou, and Zedong Nie, *Senior Member, IEEE*

***Abstract*— Sleep monitoring is essential for assessing sleep quality and understanding its broader implications for overall health. Although electroencephalography (EEG) remains the gold standard for sleep analysis, multichannel techniques are often cumbersome and impractical for real-world application. As a more feasible alternative, single-channel EEG offers greater practicality but still faces several persistent challenges, including reduced spatial resolution, feature instability, and limited clinical interpretability. To address these limitations, we propose *KAleep-Net* (Kolmogorov-Arnold based Sleep Network) for sleep stage classification. It employs a Multispectral Feature Pipeline to extract both fine-grained and coarse-grained features from single-channel EEG signals. It integrates a Temporal Sequencing Network with Flash Attention to capture rich and stable features effectively. The proposed approach achieved an accuracy of 86.5%, an F1-score of 79.6%, and a Cohen's κ of 79.9% on the Sleep-EDF-20 dataset, along with a 41.7% improvement in training speed. For the Sleep-EDF-78 dataset, it attained 85.0% accuracy, 77.0% F1-score, 78.0% κ, and a 67.5% gain in training efficiency. On the SHHS dataset, the model achieved 86.4% accuracy, an F1-score of 0.79, and a κ of 0.81, with an 8.18% improvement in training speed. For interpretability, an integrated gradient technique was adopted to enhance decision transparency and promote clinical adoption. The framework offers an efficient solution for sleep staging in resource-constrained environments with clinically trusted insights for single-channel EEG-based sleep monitoring.**

This work was supported by the Noncommunicable Chronic Diseases–National Science and Technology Major Project (Grant Nos. 2024ZD0532000 and 2024ZD0532002); the National Natural Science Foundation of China (Grant No. 62173318); the Science and Technology Service Network Plan of CAS–Huangpu Special Project (Grant No. STS-HP-202203); the Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences (Grant No. 2022A8017729013) (*Corresponding author: Zedong Nie*)

Zubair Akbar, Farhad Hassan, Jingzhen Li, Ubaidullah alias Kashif, Yuhang Liu, and Zedong Nie are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: {zubair, lijz, zd.nie, farhad, kashif, yh.liu2}@siat.ac.cn).

Zubair Akbar, Farhad Hassan, and Ubaidullah alias Kashif are also with the University of Chinese Academy of Sciences, Beijing, China.

Jia Gu is with the Faculty of Data Science, City University of Macau, Macau, China (e-mail: gujia1981@hotmail.com).

Kaixin Zhou is with the School of Public Health, Guangzhou Medical University, and Guangzhou National Laboratory, Guangzhou, China (e-mail: zhoukx@ucas.ac.cn).

Source code is available at: `https://github.com/zobi-logs/KAleep-Net`

***Index Terms*— Kolmogorov-Arnold Network (KAN), Single-Channel EEG, Temporal Characteristics, Flash Attention, Sleep Stage Classification**

I. Introduction

SLEEP is a critical aspect of human well-being as it covers essential physiological processes, including cellular repair, memory consolidation, and hormone regulation [1]. It is also crucial for maintaining cognitive functions, emotional stability, and reducing the risk of chronic conditions such as anxiety, depression, hypertension and diabetes [2]. Sleep disruptions can lead to various health issues impacting mental and physical well-being [3]. The significant variations in sleep stage proportions can contribute to sleep disorders, including insomnia and sleep apnea [2]. Therefore, effective sleep stage monitoring is crucial in healthcare. Sleep studies involve systematic observation, recording, and analysis of various physiological parameters and behavioral patterns during sleep. These are classified according to standards set by the American Academy of Sleep Medicine (AASM), which defines sleep stages as: Wakefulness (W), Rapid Eye Movement (REM), non-Rapid Eye Movement Stage 1 (N1), Stage 2 (N2), and Stage 3 (N3) [4]. PSG is widely considered the gold standard for sleep monitoring and stage classification. PSG includes recordings of EEG, Electromyography (EMG), Electrooculography (EOG), and Electrocardiography (ECG) [5]. Among these, EEG is pivotal in capturing brain-wave patterns critical for sleep stage monitoring and analysis [6]. Despite its effectiveness, traditional EEG in PSG involves multi-channel [7]–[9] which can be cumbersome and inconvenient for patients and clinicians due to the extensive wiring and setup required [5]. As an alternative, single-channel EEG presents a more streamlined and user-friendly approach that simplifies signal-capturing. Making it a practical and effective method for monitoring sleep in the home environment [6]. Traditionally, sleep stages are labeled by physicians through the visual inspection of PSG or EEG recordings, which are divided into 30-second segments. This process of manually annotating sleep stages is both time-consuming and labor-intensive [1]. Developing an efficient, reliable, and automated sleep stage monitoring using single channel EEG could enhance the accuracy and efficiency of

sleep assessments, providing support to clinicians, and ease for users.

Numerous studies have been conducted using traditional machine learning algorithms based on single-channel EEG. These studies devised manual features from raw EEG signals for automatic sleep stage classification. Zapata et al. [10] extracted time-frequency features from EEG signals and employed a support vector machine with quadratic equation (SVM-Q) for classification. Alickovic [11] applied discrete wavelet transformation (DWT), extracted statistical values from sub-band signals, and used a rotational support vector machine (RotSVM) ensemble algorithm. These methods outperform earlier approaches but often struggle due to manual feature extraction, which is time-consuming and requires extensive domain knowledge to extract relevant features. Therefore, deep learning has emerged as a powerful multi-objective approach that automatically extracts features from raw EEG signals, overcoming the limitations of traditional machine learning methods [12].

Among the pioneering deep learning models, SleepEEGNet [13] integrates convolutional neural networks (CNNs) for feature extraction with bi-directional residual neural networks (BiRNNs) and attention mechanisms, achieving an accuracy of 84.26%. Similarly, TinyUStaging [14] incorporates multiple attention mechanisms to analyze EEG and EOG signals, reaching an accuracy of 84.62%. The Sleep Hybrid Neural Network (SHNN) [15] combines CNNs with bi-directional gated recurrent units (Bi-GRUs) to capture temporal dependencies in single-channel EEG data, reporting an accuracy of 78.62% and a micro-F1 score of 71.0% under a semi-supervised learning paradigm. Another noteworthy approach is the use of multi-scale convolutional filters by Phan et al. [16], who converted sleep epochs into time-frequency images for classification. More recently, AttnSleep [17] has incorporated multi-resolution CNNs with multi-head attention mechanisms, achieving 83.5% and 86.2% accuracy on the Sleep-EDF and Sleep Heart Health Study (SHHS) datasets, respectively. Furthermore, Zhao et al. [18] developed a multi-task learning model utilizing a U-Net architecture with one-dimensional channel attention, reporting accuracy scores of 85.6% on Sleep-EDF and 83.4% on Sleep-EDFx. Meanwhile, Siddiqa et al. [19] introduced a multi-branch CNN for single-channel EEG, achieving an accuracy of 74.27%, which improved to 75.36% with multi-channel EEG inputs.

Despite advancements in deep learning-based sleep staging, existing models have several limitations. Although single-channel EEG methods are more practical than multi-channel setups for real-world applications, they suffer from reduced spatial coverage owing to the limited number of electrodes, resulting in suboptimal feature extraction that complicates the differentiation of sleep stages with similar characteristics. Furthermore, the features extracted from EEG signals are often unstable due to noise, variations in electrode placement, and individual differences in EEG patterns, which undermines the robustness of the models and leads to inconsistent performance across datasets and subjects. In addition, the lack of interpretability arising from the absence of mechanisms to highlight key EEG patterns that inform their predictions poses a significant barrier to the clinical adoption of these models, as transparency is essential for their effective use in medical practice. While recent models such as MaskSleepNet [20], ASTGSleep [21], and FlexibleSleepNet [22] have reported accuracies close to or exceeding 86% on public datasets like SleepEDF-20, performance on these benchmarks appears to be approaching saturation. In this context, pushing for marginal accuracy improvements may offer limited practical value. Instead, there is a growing need to develop models that provide enhanced interpretability, adaptability, and robustness qualities that are crucial for real-world deployment.

To address these challenges, we propose a Multi-Spectral Feature Pipeline that integrates parallel-stacked Kolmogorov–Arnold Network (KAN) convolutional filters. This architecture improves feature representation [23] and enables the simultaneous extraction of both high- and low-frequency EEG components, thereby enhancing the model's ability to capture the diverse spectral characteristics inherent in sleep data. In addition, a Temporal Sequencing Network combined with Flash Attention improves the modeling of long-range contextual dependencies while maintaining computational efficiency [24], [25]. This design strengthens feature representation and helps mitigate the effects of spatial limitations and feature instability. Furthermore, the model employs the Integrated Gradients technique to provide physiologically meaningful attributions by highlighting key EEG patterns, thereby enhancing interpretability and supporting clinical adoption [26]. The primary contributions of this work are as follows.

1) We introduce a Multi-Spectral Feature Pipeline with a Temporal Attention Block to enhance feature space and alleviate the challenges posed by spatial limitations and feature inconsistency.
2) We incorporate Integrated Gradients (IG) to provide physiologically meaningful insights into model predictions, aligning the model's outputs with established clinical sleep markers.
3) KAleep-Net outperforms existing models in accuracy and robustness across benchmark datasets, with statistical significance confirmed via a paired t-test.

The remainder of this paper is organized as follows: Section II presents our proposed model with a detailed methodology. Section III covers the experimental framework, including datasets, evaluation metrics, implementation details, and comparative analysis of results. Section IV further Includes Results and a discussion of the model. Finally, Section V concludes the paper with key findings and future directions.

II. Methodology

The proposed KAleep-Net processes a 30-second EEG signal $\mathbf{X} \in \mathbb{R}^{3000\times 1}$ using a multi-spectral feature extraction pipeline. This pipeline consists of parallel Fine-Scale Feature (FSF) and Coarse-Scale Feature (CSF) blocks designed to capture spectral information at multiple resolutions. The FSF block applies KAN-augmented convolutional filters with small kernel sizes to extract high-frequency components, denoted as $\mathbf{F}_{\text{fine}}$. In parallel, the CSF block employs wider convolutional kernels to capture low-frequency features, denoted as $\mathbf{F}_{\text{coarse}}$.

The outputs of these two branches are concatenated to form a unified feature representation: $\mathbf{F}_{\text{cat}} = \mathbf{F}_{\text{fine}} \oplus \mathbf{F}_{\text{coarse}}$. This concatenated feature map is then passed into a temporal attention network, which models temporal dependencies across the 30-second signal. It generates bidirectional hidden states $[\mathbf{h}_t^{\rightarrow}, \mathbf{h}_t^{\leftarrow}]$ at each time step t, resulting in a sequence representation $\mathbf{H} \in \mathbb{R}^{3000\times 2d}$, where d denotes the hidden size of the LSTM layers. To enhance temporal focus, a Flash Attention mechanism is integrated, computing attention weights and producing an attended output $\mathbf{F}_a \in \mathbb{R}^{2d}$. Finally, this attention-weighted representation is passed through a softmax classification layer, yielding the predicted sleep stage distribution $\hat{\mathbf{y}} \in \mathbb{R}^5$ across five predefined sleep stages, as illustrated in as illustrated in Fig. 1. The detailed explanation of KAleep-Net components are as follows:

A. Multi-Spectral Feature Pipeline(MSFP)

The Multi-Spectral Feature Extraction component is designed to capture a wide range of spectral information from an EEG signal by employing two specialized blocks:

1) Fine Scale Feature (FSF) Block: The Fine-Scale Feature Block extracts detailed features from the EEG input signal $\mathbf{X} \in \mathbb{R}^{3000\times 1}$ using three CNN-KAN pairs. Each pair, denoted as (C_i, K_i) for $i = 1, 2, 3$, combines a convolutional layer C_i with a corresponding KAN layer K_i, allowing progressive refinement of the feature representation. The first convolutional layer C_1 applies one-dimensional convolution with $f = 64$ filters, a kernel size $k = 3$, and same padding. After processing through the first pair (C_1, K_1), max pooling m_p and dropout D_p are applied to reduce dimensionality and prevent overfitting. Each of the FSF branches applies two MaxPooling layers with a pooling size of 2. This successively reduces the temporal resolution from 3000 to 1500, and then to 750 time steps, compressing the input signal temporally while retaining key local features.

The output of the dropout layer is passed to the second pair (C_2, K_2), which applies convolution with 128 filters (kernel size 3), and the KAN layer further transforms the feature space by capturing more complex patterns. This is followed by the third CNN-KAN pair (C_3, K_3), after which a final max pooling m_p and flattening operation f prepare the features for subsequent processing. The overall transformation in the Fine-Scale Feature Block is mathematically expressed as:

$$\mathbf{F}_{\text{fine}} = f\left(m_p\left(K_3\left(C_3\left(K_2\left(C_2\left(D_{\text{fine}}\right)\right)\right)\right)\right)\right) \tag{1}$$

$$D_{\text{fine}} = D_p\left(m_p\left(Z_{\text{fine},K_1}\right)\right) \tag{2}$$

and the output of each convolutional layer C_i is given by $Z_{\text{fine},C_i} = w_{\text{fine},C_i} * l_i + b_{\text{fine},C_i}$ with $w_{\text{fine},C_i} \in \mathbb{R}^{k\times f\times o}$ as the weight tensor, $b_{\text{fine},C_i} \in \mathbb{R}^f$ as the bias, $*$ denoting convolution, and l_i being the input to the layer. The KAN transformation for FSF is defined as:

$$K_i\left(Z_{\text{fine},C_{i,j}}\right) = \sum_{j=1}^{d}\sum_{k=1}^{K} w_{ijk}\cdot\sigma\left(h_{ijk}\cdot Z_{\text{fine},C_{i,j}} + b_{ijk}\right) \tag{3}$$

where K represents the number of basis functions, w_{ijk} are the learnable weights of K_i, h_{ijk} are the scaling factors, b_{ijk} are the bias terms, and σ denotes the non-linear kernel activation function, which is implicitly defined and adaptively learned rather than fixed. This block enables the extraction of high-frequency, detailed features $\mathbf{F}_{\text{fine}}$, extracted using small convolutional kernels, which are particularly well-suited for capturing subtle patterns essential for differentiation.

2) Coarse Scale Feature (CSF Block): The Coarse Scale Feature Block (CSF) employs a tiered architecture comprising three sequential convolutional-KAN pairs (C_a, K_a), where $a \in \{1, 2, 3\}$. Each pair progressively extracts and refines low-frequency temporal patterns from the input EEG signal $\mathbf{X} \in \mathbb{R}^{3000\times 1}$ using larger kernel sizes to capture broader temporal contexts. The first convolutional layer C_1 uses 32 channels, a kernel size $k = 5$, and "same" padding to perform a one-dimensional convolution. This is followed by a KAN layer K_1 that models nonlinear interactions while preserving spatial fidelity. To further reduce dimensionality and prevent overfitting, a max pooling layer m_p and dropout layer D_p are applied. Each of the CSF branches applies two MaxPooling layers with a pooling size of 2. This successively reduces the temporal resolution from 3000 to 1500, and then to 750 time steps, compressing the input signal temporally while retaining key local features. The output of the dropout layer is then passed into the second CNN-KAN pair (C_2, K_2), which increases the feature depth to 64 channels to enhance discriminative capacity while maintaining temporal resolution. The final CNN-KAN pair (C_3, K_3) further refines the representation using the same number of channels and adds additional nonlinear transformation. A max pooling layer m_p is then applied to compress the feature map, followed by a flattening operation f to prepare the features for subsequent processing. The overall operation of the Coarse Scale Feature Block is defined as:

$$\mathbf{F}_{\text{coarse}} = f\left(m_p\left(K_3\left(C_3\left(K_2\left(C_2\left(D_{\text{coarse}}\right)\right)\right)\right)\right)\right) \tag{4}$$

where $D_{\text{coarse}} = D_p\left(m_p\left(Z_{\text{coarse},K_1}\right)\right)$ and the output of each convolutional layer C_i is given by $Z_{\text{coarse},C_i} = w_{\text{coarse},C_i} * l_i + b_{\text{coarse},C_i}$ with $w_{\text{coarse},C_i} \in \mathbb{R}^{k\times c\times o}$ representing the weight tensor, $b_{\text{coarse},C_i} \in \mathbb{R}^c$ the bias vector, $*$ denoting the convolution operation, and l_i the input to the layer. The KAN transformation of CSF is:

$$K_i\left(Z_{\text{coarse},C_{i,j}}\right) = \sum_{j=1}^{d}\sum_{k=1}^{K} w_{ijk}\cdot\sigma\left(h_{ijk}\cdot Z_{\text{coarse},C_{i,j}} + b_{ijk}\right) \tag{5}$$

where K is the number of basis functions, and w_{ijk}, h_{ijk}, and b_{ijk} are learnable parameters representing the weights, scaling factors, and bias terms, respectively. σ denotes the non-linear kernel activation function, which is implicitly defined and adaptively learned rather than fixed.

The kernel size selection for the fine-scale features (kernel size $= 3$) and coarse-scale features (kernel size $= 5$) blocks is grounded in EEG signal characteristics. This dual-scale approach ensures comprehensive spectral coverage, facilitating robust and interpretable feature extraction.

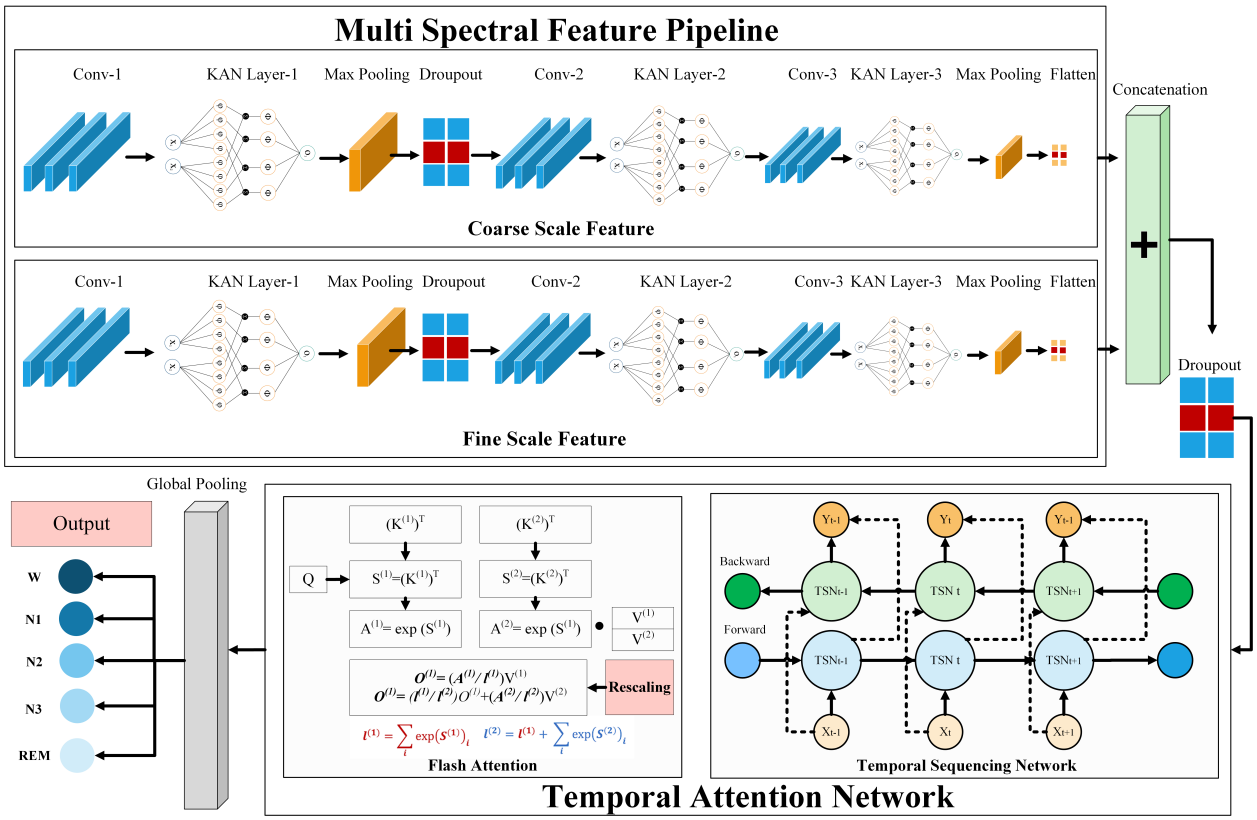

Fig. 1. Architecture of KAleep Net: comprises two key components: (1) Multi-Spectral Feature Pipeline, which extracts coarse- and fine-scale temporal features using convolutional layers followed by Kolmogorov–Arnold Network (KAN) layers; and (2) Temporal Attention Network, which combines Temporal Sequencing Network (TSN) to capture stage transitions and Flash Attention for efficient wieghtage on temporal dependency and predicts sleep stages (Wake, N1, N2, N3, and REM.)

B. Concatenation of Features

After processing through both the FSF and CSF blocks, each branch yields a temporally downsampled sequence of shape $\mathbb{R}^{750\times d}$, where 750 denotes the reduced number of time steps after two MaxPooling operations, and d is the feature dimensionality. These outputs, $\mathbf{F}_{\text{fine}}$ and $\mathbf{F}_{\text{coarse}}$, are concatenated along the feature/channel dimension to form a unified feature sequence:

$$\mathbf{F}_{\text{cat}} = \mathbf{F}_{\text{fine}} \oplus \mathbf{F}_{\text{coarse}} \in \mathbb{R}^{750\times D} \tag{6}$$

where D represents the total number of concatenated feature channels from both branches. This sequence $\mathbf{F}_{\text{cat}}$ is then passed to the temporal sequencing network for further modeling.

C. Temporal Attention Network (TAN)

The Temporal Attention Network (TAN) consists of two key sub-blocks: the Temporal Sequencing Network and the Flash Attention Block. It captures temporal patterns by analyzing both forward and backward directions, dynamically highlighting critical time steps in the sequence through the attention mechanism.

1) Temporal Sequencing Network Block: The Temporal Sequencing Network (TSN) utilizes a bidirectional Long Short-Term Memory (BiLSTM) network to process the concatenated feature sequence $\mathbf{F}_{\text{cat}} \in \mathbb{R}^{750\times D}$, which represents the temporally downsampled EEG features after the FSF and CSF blocks. The BiLSTM processes this sequence in both forward ($\mathbf{h}_t^{\rightarrow}$) and backward ($\mathbf{h}_t^{\leftarrow}$) directions, capturing bidirectional temporal dependencies across the 750 time steps. This dual perspective enables a richer contextual representation of temporal patterns within the EEG signal, an essential factor for accurate sleep stage classification, where stage transitions and subtle dynamics vary across time. For each time step $t \in \{1, 2, \ldots, 750\}$, the hidden states are computed as:

$$\mathbf{h}_t^{\rightarrow} = \text{LSTM}_{\text{fwd}}(\mathbf{F}_{\text{cat},t}, \mathbf{h}_{t-1}^{\rightarrow}) \tag{7}$$

$$\mathbf{h}_t^{\leftarrow} = \text{LSTM}_{\text{bwd}}(\mathbf{F}_{\text{cat},t}, \mathbf{h}_{t+1}^{\leftarrow}) \tag{8}$$

The final representation at each time step is formed by concatenating the forward and backward hidden states:

$$\mathbf{H}_t = [\mathbf{h}_t^{\rightarrow}; \mathbf{h}_t^{\leftarrow}] \in \mathbb{R}^{2d} \tag{9}$$

Thus, the TSN produces a sequence of bidirectional hidden states:

$$\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_{750}] \in \mathbb{R}^{750\times 2d} \tag{10}$$

2) Flash Attention Block (FAB): The Flash Attention Block (FAB) efficiently models long-range temporal dependencies in the BiLSTM output sequence $\mathbf{H} \in \mathbb{R}^{750\times 2d}$ using a block-wise

scaled dot-product attention mechanism. The input is linearly projected into query, key, and value matrices, and attention scores are computed in memory-efficient blocks using:

$$S_{ij} = \frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sqrt{d_a}}, \quad \mathbf{F}_{a,i} = \frac{1}{Z_i} \sum_j \exp(S_{ij} - m_i)\mathbf{V}_j \tag{11}$$

where Z_i and m_i apply the log-sum-exp trick for numerical stability. The attended output $\mathbf{F}_a \in \mathbb{R}^{750 \times d_a}$ is pooled across time to form a fixed-size representation $\mathbf{O}_{\text{pooled}} \in \mathbb{R}^{d_a}$, which is used for final classification via a softmax layer over five sleep stages.

III. Experimental Setup

A. Dataset

This study utilized the publicly available Sleep-EDF-78 Sleep-EDF-20 [26], and Sleep Heart Health Study (SHHS) [27], [28] datasets [29], which are widely recognized as benchmark datasets for sleep stage classification. These datasets were selected due to their credibility, standardized annotation protocols, and reproducibility. Sleep-EDF datasets contain recordings from two consecutive nights per subject, acquired using EEG (Fpz-Cz, Pz-Oz), EOG, and EMG channels, sampled at 100 Hz. Sleep stages were initially annotated based on the 1968 Rechtschaffen and Kales (R&K) scoring system [30]. For consistency with current standards, these annotations were subsequently converted to the AASM classification, comprising Wake (W), N1, N2, N3 (merged from R&K stages 3 and 4), and REM stages. The SHHS dataset is one of the largest publicly available collections of overnight polysomnographic recordings. The SHHS dataset includes 6,441 subjects, from which 329 were selected based on an Apnea-Hypopnea Index (AHI) criterion of AHI < 5. The recordings in SHHS were acquired using EEG (C4-A1), EOG, EMG, ECG, respiratory effort, airflow, oxygen saturation (SpO_2), and body position, sampled at 125 Hz. Sleep stages were annotated according to the AASM classification system, and N3 and N4 stages were merged into a single N3 stage. The dataset includes 30-minute periods of wake data before and after the sleep period and excludes any periods marked as UNKNOWN stages. The SHHS dataset aims to explore the cardiovascular consequences of sleep-disordered breathing in individuals aged 40 years and older, offering a diverse cohort with real-world variability in signal quality. This makes it an ideal dataset for developing robust sleep classification models. Detailed demographic characteristics are summarized in Table I.

B. Data Preparation

To mitigate noise and artifacts present in the raw EEG signals, denoted as X_{raw}, which were acquired at a sampling frequency of $f_s = 100$ Hz, and to subsequently normalize these signals, the following preprocessing steps were employed:

1) Segmentation: The continuous raw EEG signal X_{raw} was divided into non-overlapping segments, each with a duration of $T_e = 30$ seconds. Given the sampling frequency f_s, the number of samples per segment is computed as $S = f_s * T_e$. This segmentation facilitates the localization of EEG-related events and improves memory efficiency, especially when handling large datasets. The 30-second window length is consistent with standard practices in EEG-based research [31], [32]. Formally, the i^{th} segment is defined as:

$$X_{\text{seg}}^{(i)} = X_{\text{raw}}[i \cdot S : (i+1) \cdot S - 1] \tag{12}$$

for i = 1, 2, …, N and $N = \lfloor \frac{T}{S} \rfloor$ denotes the total number of full segments. Each resulting segment $X_{\text{seg}}^{(i)} \in \mathbb{R}^{S \times 1}$. Specifically, with $f_s = 100$ Hz and $T_e = 30$ seconds, we have $S = 3000$ samples and $X_{\text{seg}}^{(i)} \in \mathbb{R}^{3000 \times 1}$.

2) Noise Removal: A fourth-order Butterworth high-pass filter with a cutoff frequency of 0.5 Hz and a low-pass filter with a cutoff frequency of 40 Hz are applied to each segment to attenuate low-frequency trends [33]. The implementation is as follows:

$$X_f^{(i)}[n] = \sum_{k=0}^{M} b_k \cdot X_{\text{seg}}^{(i)}[n-k] - \sum_{k=1}^{N} a_k \cdot X_f^{(i)}[n-k] \tag{13}$$

Here, b_k and a_k are the feedforward and feedback coefficients of the filter, respectively, derived from the design of a fourth-order Butterworth filter ($M = N = 4$). This corresponds to the digital transfer function $H(z)$ of the high-pass system. The output $X_f^{(i)}[n] \in \mathbb{R}^{S \times 1}$ preserves the original signal length while attenuating low-frequency noise such as electrodermal drift and slow movement artifacts.

3) Artifact Reduction: Amplitude-based thresholding was applied to eliminate segments containing transient artifacts such as eye blinks, muscle activity, or electrode discharges. Segments were retained only if all samples satisfied the physiologically acceptable amplitude constraint:

$$|X_f^{(i)}[n]| \leq 100\,\mu V, \quad \forall n \in \{1, \ldots, S\} \tag{14}$$

The output $X_{\text{clean}}^{(i)} \in \mathbb{R}^{S \times 1}$ is retained only for segments meeting the amplitude constraint.

4) Normalization: To ensure numerical stability and to prevent segments with large amplitudes from dominating the training process, [34] each segment is standardized using z-score normalization:

$$X_{\text{norm}}^{(i)}[n] = \frac{X_{\text{clean}}^{(i)}[n] - \mu^{(i)}}{\sigma^{(i)}} \tag{15}$$

for n = 1, 2, …, S while $\mu^{(i)}$ and $\sigma^{(i)}$ are the mean and standard deviation of the i^{th} segment, respectively, computed as:

$$\mu^{(i)} = \frac{1}{S} \sum_{n=1}^{S} X_{\text{clean}}^{(i)}[n] \tag{16}$$

$$\sigma^{(i)} = \sqrt{\frac{1}{S} \sum_{n=1}^{S} \left(X_{\text{clean}}^{(i)}[n] - \mu^{(i)} \right)^2} \tag{17}$$

Finally, the preprocessed EEG signal is denoted as:

TABLE I
DEMOGRAPHIC SUMMARY OF SLEEP-EDF-20, SLEEP-EDF-78, AND SHHS DATASETS

Parameter	Sleep-EDF-20	Sleep-EDF-78	SHHS
Number of Subjects	20	78	6,441 (SHHS-1), 2,651 (SHHS-2)
Age Range	25–34 years	25–101 years	40 years and older
Gender Distribution	10 Male, 10 Female	38 Male, 40 Female	Approximately equal male/female distribution
Health Status	Healthy adults (no sleep disorders)	Mix of healthy and individuals with sleep disorders	General population with cardiovascular risk factors
Medication Use	None reported	Some on sleep-affecting medications	Varies; not explicitly excluded
Data Collection Period	1989–1994	1989–1997	1995–2006 (SHHS-1: 1995–1998, SHHS-2: 2001–2006)

$$X := X_{\text{norm}}^{(i)} \in \mathbb{R}^{S\times 1} \tag{18}$$

C. Implementation Settings

All experiments were conducted using an NVIDIA RTX A6000 GPU, equipped with 48 GB of GDDR6 memory. The software environment included Python 3.8, PyTorch 1.7.1, CUDA 11.0, and cuDNN 8.0.5, ensuring optimized performance and reproducibility. A subject-out validation strategy was employed to evaluate generalization across unseen subjects. Optimal hyperparameters were identified through a grid search over key parameters such as learning rate and weight decay. The Adam optimizer was used with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} to promote stable convergence and regularization. A `ReduceLROnPlateau` scheduler dynamically adjusted the learning rate based on validation loss plateaus. The training was performed with a batch size of 64 to balance gradient stability and GPU memory constraints. An early stopping criterion of 10 epochs was implemented to terminate training if validation performance did not improve, thereby minimizing overfitting and reducing computational overhead. Furthermore, the detailed hyperparameters and kernel sizes used in our KAleep-Net model are given in Table II to ensure reproducibility.

TABLE II
HYPERPARAMETERS AND KERNEL SIZES USED IN KALEEP-NET

Hyperparameter	Value
Segment Length	30 sec (3000 samples @ 100 Hz)
FSF Convolutional Layers	3
FSF Kernel Size	3
FSF Filters	$64 \rightarrow 128 \rightarrow 128$
FSF Pooling	MaxPooling (pool size = 2)
FSF Dropout	0.25
CSF Convolutional Layers	3
CSF Kernel Size	5
CSF Filters	$32 \rightarrow 64 \rightarrow 64$
CSF Pooling	MaxPooling (pool size = 2)
CSF Dropout	0.25
Merge Method	Concatenate
BiLSTM Layers	2 (Bidirectional)
BiLSTM Hidden Units	128 per direction (256 total per layer)
Attention Type	Explicit Block-wise Flash Attention
Attention Block Size	5
Attention Dimension (d_a)	256
Output Classes	5
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	64
Epochs	100 (early stopping = 10 epochs)
Filtering	Bandpass (Butterworth 0.5–40 Hz)
Artifact Threshold	$\pm 100\ \mu$V
Normalization	Z-score normalization

D. Evaluation Metrics

To assess KAleep-Net, we employed the following three established evaluation metrics:

- **Accuracy (A)** [35], defined as:

$$A = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{19}$$

- **Macro F1-Score** (F_1) [36], calculated as:

$$F_1 = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \cdot \text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{20}$$

  where K represents the number of classes.
- **Cohen's Kappa Coefficient** (κ) [37], computed as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{21}$$

  where P_o is the observed agreement and P_e is the expected agreement by chance.

Furthermore, confusion matrices, along with plots of loss and accuracy curves, were employed to analyze the agreement between model predictions and ground-truth labels [38], [39].

To address class imbalance, particularly for underrepresented sleep stages such as N1 and REM. We incorporated a class-weighted cross-entropy loss function. The weight w_c assigned to each class c was determined based on its frequency f_c in the training set, using the following formulation:

$$w_c = \frac{1}{\log(1 + f_c)} \tag{22}$$

This weighting scheme penalizes misclassifications of minority classes more heavily, guiding the model to learn more discriminative representations for them and improving overall performance on imbalanced datasets.

IV. RESULTS

A. Ablation Studies

To validate the effectiveness of the proposed approach, we conducted an ablation study to evaluate several architectural

configurations. The baseline configuration (**CASE 1**) employed a simple CNN for coarse feature extraction combined with a Temporal Sequencing Network (TSN), but it struggles to capture comprehensive feature representations. We then introduced a dual-branch architecture (**CASE 2**) to extract both coarse and fine features, integrated with a TSN, which improved the performance but remained suboptimal. Next, we augmented the CNN with a Kolmogorov-Arnold Network (KAN) for coarse feature extraction paired with a TSN (**CASE 3**) to achieve good performance. In addition, we developed a dual-branch setup (**CASE 4**) with a standard CNN for fine features and a CNN with a KAN for coarse features, integrated with a TSN, leveraging complementary feature extraction strategies. Finally, we extended this by incorporating a Flash Attention mechanism (**CASE 5**) to enhance the focus on salient temporal segments, further boosting performance. Each configuration incrementally improved upon its predecessor, with the ablation study confirming that the combination of dual-branch feature extraction, KAN augmentation, and Flash Attention in **CASE 5** delivers the best results as shown in Table IV.

B. Performance Evaluation

To comprehensively evaluate the performance of KAleep-Net, we present the confusion matrices that visualize its classification effectiveness on both the Sleep-EDF-20, Sleep-EDF-78 and SHHS datasets, as illustrated in table IV.

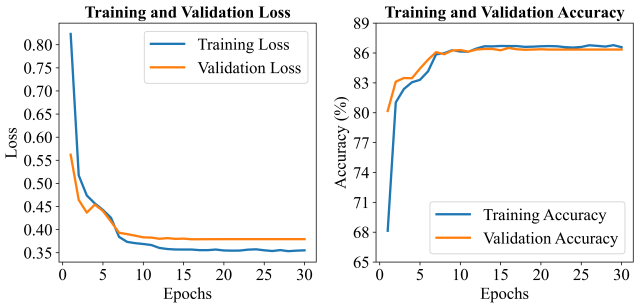

Fig. 2. Loss and Accuracy Curves for train and test on Sleep-EDF-20 dataset.

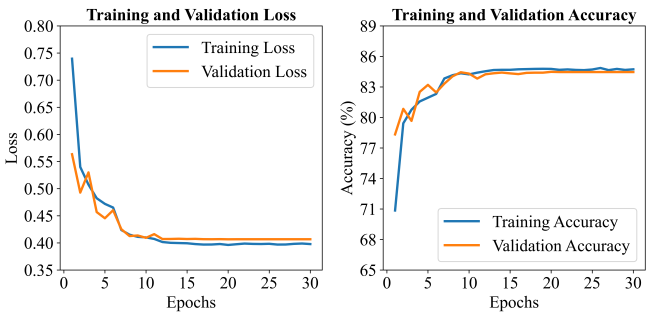

Fig. 3. Loss and Accuracy Curves for train and test on Sleep-EDF-78 dataset.

Furthermore, to analyze the learning behavior of KAleep-Net, we plot training and validation loss and accuracy curves

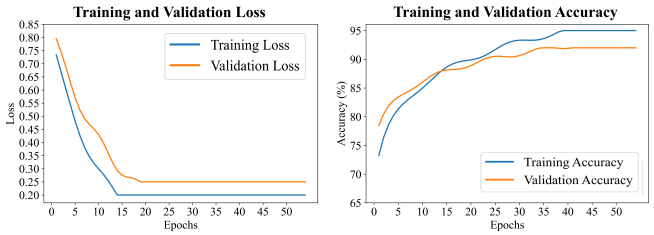

Fig. 4. Loss and Accuracy Curves for train and test on SHHS dataset.

for all the three datasets Fig 2, 3, and 4. All Figures show that model shows a rapid decrease in training and validation loss over the initial epochs, with both curves continuing to converge and remaining close, indicating minimal overfitting.

For robust evaluation, we employed 10-fold cross-validation with subject-wise partitioning, ensuring that each fold contains data from distinct subjects (2 subjects per fold for Sleep-EDF-20, 7-8 subjects per fold for Sleep-EDF-78, and 32-33 subjects per fold for SHHS), as shown in Table V. Stratified sampling was used to preserve the class distribution, and group-based splitting was applied to prevent data leakage by ensuring that all data from a single subject remained within the same fold.

C. Comparison with the state-of-the-art Studies

To evaluate the effectiveness of the proposed KAleep-Net model, we conducted comprehensive comparisons against several state-of-the-art methods across three benchmark datasets: Sleep-EDF-20, Sleep-EDF-78, and SHHS. Model performance was assessed using four primary metrics: Accuracy, Cohen's Kappa, F1 Score, and Training Time per fold (in minutes), as summarized in Table VI.

On the Sleep-EDF-20 dataset, KAleep-Net achieved the highest accuracy of 86.5%, surpassing CausalAttenNet (84.7%) while maintaining a substantially lower training time of 10.5 minutes per fold, compared to 18 to 156 minutes reported for other models. For Sleep-EDF-78, KAleep-Net retained its leading performance with an accuracy of 85.0%, outperforming CausalAttenNet (84.0%) and significantly reducing training time to just 17.9 minutes per fold, in contrast to the 55–432 minutes observed in competing methods.

In terms of other evaluation metrics, KAleep-Net achieved a Cohen's Kappa of 0.79 and an F1 Score of 76.8% on Sleep-EDF-20, and a Kappa of 0.78 and F1 Score of 77.0% on Sleep-EDF-78. On the SHHS dataset, KAleep-Net attained an accuracy of 86.4%, a Kappa of 0.81, and an F1 Score of 79.0%, while reducing training time to 110 minutes per fold. The consistently high classification performance and efficient training profile across datasets highlight KAleep-Net's potential as a scalable and effective solution for automatic sleep stage classification.

D. Statistical Significance

To validate the statistical significance of our accuracy improvement, we conducted a paired t-test comparing KAleep-Net and CausalAttenNet across multiple runs ($n = 5$). The test yielded a t-statistic of 5.28 and a p-value of 0.0062,

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT ARCHITECTURAL CONFIGURATIONS ACROSS SLEEP DATASETS

Architecture	Sleep-EDF-20			Sleep-EDF-78			SHHS		
	Acc.	Macro F1	Kappa	Acc.	Macro F1	Kappa	Acc.	Macro F1	Kappa
Case 1: Simple CNN for Coarse Feature + TSN	80.0	70.0	73.5	78.5	70.5	72.0	83.5	75.0	79.5
Case 2: Dual Architecture Coarse Feature + Fine Feature + TSN	82.5	72.5	76.0	81.0	73.0	74.5	84.0	77.0	82.5
Case 3: CNN with KAN for CSF + TSN	84.0	74.0	77.5	82.5	74.5	76.0	84.9	78.0	79.0
Case 4: CNN with KAN for (CSF + FSF) + TSN	85.5	75.5	78.0	84.0	76.0	77.0	85.5	78.5	78.0
Case 5: CNN with KAN Based (CSF + FSF) + TSN + FA	86.5	76.8	79.9	85.5	77.0	78.0	86.4	79.0	81.0

TABLE IV
CONFUSION MATRICES FOR SLEEP STAGE CLASSIFICATION ACROSS DATASETS. SLEEP STAGES: W = WAKE, N1 = STAGE 1, N2 = STAGE 2, N3 = STAGE 3, REM = RAPID EYE MOVEMENT.

(a) Sleep-EDF20 Dataset						
	Predicted Label					
True Label	W	N1	N2	N3	REM	Total
W	82	20	12	3	26	143
N1	38	85	72	0	97	292
N2	15	12	1610	75	79	1791
N3	2	0	30	532	0	564
REM	8	25	58	0	664	755
Total	145	142	1782	610	866	3545
(b) Sleep-EDF78 Dataset						
	Predicted Label					
True Label	W	N1	N2	N3	REM	Total
W	896	67	27	4	57	1051
N1	84	185	209	3	153	634
N2	13	73	2939	112	159	3296
N3	1	0	67	704	0	772
REM	32	65	94	1	117	309
Total	1026	390	3336	824	486	6062
(c) SHHS Dataset						
	Predicted Label					
True Label	W	N1	N2	N3	REM	Total
W	42004	1598	1988	304	425	46319
N1	856	4947	1587	46	2868	10304
N2	1675	2223	127242	4891	6094	142125
N3	214	19	5051	50695	4174	60153
REM	1405	2258	6143	398	55749	65953
Total	42754	12245	141011	60234	68610	324854

indicating that the observed accuracy improvement is statistically significant ($p < 0.05$) and unlikely due to random variation. Specifically, KAleep-Net achieved an average accuracy of $86.5\% \pm 0.21\%$, compared to $84.7\% \pm 0.35\%$ for CausalAttenNet.The 95% confidence interval for KAleep-Net accuracy is [86.3%, 86.7%], which does not overlap with that of CausalAttenNet [84.4%, 85.0%], further reinforcing the robustness of our model's performance gains. A detailed statistical comparison is presented in Table VII.

E. Understanding KAleep-Net Decisions: Explainability Evaluation

Clinical trust in sleep staging models hinges on their interpretability. To ensure KAleep-Net's decisions are transparent and clinically grounded, we integrated interpretability directly into the model pipeline using Integrated Gradients (IG), a method that attributes predictions to specific input features. Unlike post-hoc visualizations such as Grad-CAM applied to Vision Transformers [44], our approach operates on raw temporal EEG and provides fine-grained, stage-specific attributions without requiring multimodal data or heavy computation.Figure 5 illustrates representative EEG segments from each sleep stage, where red dots indicate the top 5% of time points with the highest IG attribution, i.e., those EEG regions most influential in the model's classification. For visual context, we also marked general waveform peaks (green circles) using a simple peak-detection algorithm, without assigning them to specific frequency bands. This distinction ensures that attributions reflect the model's decision process, not researcher-imposed labels. The IG markers often align with physiologically meaningful patterns: in Wake stages, with waveforms consistent with alpha rhythms (8–12 Hz); in N2 and N3, with segments resembling sleep spindles, K-complexes, and slow-wave activity. In REM, the attributions are more dispersed, mirroring the mixed-frequency EEG typical of that stage. This correspondence between IG attributions and known neurophysiological features supports the model's interpretability and decision validity.

To further quantify this behavior, Figure 6 shows the average IG attribution over time for each stage across the test set. N2 and N3 display sharp, localized peaks, consistent with the presence of discrete EEG events while Wake and REM show broader, less concentrated patterns. This suggests KAleep-Net does not rely on abrupt signal changes but instead identifies clinically relevant, stage-specific EEG signatures. Together, these results demonstrate that our interpretability framework offers meaningful, transparent insights into model behavior, supporting its use in resource-constrained, real-world sleep analysis systems.

V. DISCUSSION

Sleep is fundamental to both physical and mental health; however, sleep neglect has led to an increase in sleep-related health issues. As manual sleep staging becomes impractical for large-scale analyses, automated methods are a promising solution. This study introduced KAleep-Net, a novel single-channel EEG-based framework that balances accuracy, efficiency, and interpretability, surpassing state-of-the-art approaches while aligning with clinical requirements. The pre-processing pipeline employs a segmentation approach with a 30-second window length. This duration aligns with

TABLE V
10-FOLD CROSS-VALIDATION RESULTS ACROSS SLEEP-EDF-20, SLEEP-EDF-78, AND SHHS DATASETS. ACCURACY, KAPPA, AND MACRO F1 REPRESENT CLASSIFICATION PERFORMANCE METRICS (IN %).

Fold	Sleep-EDF-20			Sleep-EDF-78			SHHS		
	Accuracy	F1 Score	Kappa	Accuracy	F1 Score	Kappa	Accuracy	F1 Score	Kappa
1	86.6	76.7	80.7	84.5	76.4	78.0	86.4	81.0	79.0
2	87.2	77.1	79.2	84.7	77.4	78.6	86.3	80.9	78.9
3	85.7	76.3	79.8	85.4	75.9	79.2	86.6	81.2	79.1
4	86.8	77.0	79.4	85.1	76.2	78.0	86.2	80.8	78.8
5	87.3	77.1	79.4	85.5	76.2	77.4	86.5	81.3	79.3
6	86.7	76.7	80.2	85.5	77.7	78.3	86.4	81.1	79.0
7	87.2	77.4	80.5	84.3	77.7	77.5	86.3	81.0	78.9
8	86.4	76.9	79.8	84.9	77.0	78.2	86.5	81.2	79.2
9	86.5	77.2	80.3	84.4	77.2	78.7	86.6	81.4	79.3
10	86.0	77.0	80.8	84.8	77.3	77.9	86.4	81.0	79.0
Avg.	**86.6**	**76.9**	**80.0**	**84.9**	**76.9**	**78.2**	**86.4**	**81.0**	**79.0**

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY, F1 SCORE (BOTH IN %), AND TRAINING TIME (TT) PER FOLD (IN MINUTES) WITH STATE-OF-THE-ART METHODS ON THE SLEEP-EDF AND SHHS DATASETS. 'N/A' DENOTES VALUES NOT REPORTED IN THE ORIGINAL STUDIES.

Dataset	Model	Accuracy (%)	Kappa	F1 Score (%)	TT (min/fold)
Sleep-EDF-20	DeepSleepNet [40]	81.9	0.76	76.6	150
	SleepEEGNet [13]	84.5	0.79	79.6	90
	DeepResnet [41]	82.5	0.76	73.7	72
	MultitaskCNN [16]	82.1	0.77	75.0	156
	AttnSleep [17]	84.4	0.79	78.1	21
	CausalAttenNet [42]	84.7	0.79	78.1	18
	DistillSleepNet [43]	84.7	N/A	75.8	N/A
	KAleep-Net (Ours)	**86.5**	**0.79**	**76.8**	**10.5**
Sleep-EDF-78	DeepSleepNet [40]	77.8	0.70	71.8	432
	SleepEEGNet [13]	80.0	0.73	73.5	384
	DeepResnet [41]	78.9	0.71	71.4	204
	MultitaskCNN [16]	79.6	0.72	72.8	318
	AttnSleep [17]	81.3	0.74	75.1	102
	CausalAttenNet [42]	84.0	0.74	75.2	55
	DistillSleepNet [43]	83.6	0.73	N/A	N/A
	KAleep-Net (Ours)	**85.0**	**0.78**	**77.0**	**17.9**
SHHS	DeepSleepNet [40]	81.0	0.73	73.0	864
	SleepEEGNet [13]	83.3	0.76	74.0	504
	DeepResnet [41]	84.2	0.78	75.0	438
	MultitaskCNN [16]	83.3	0.74	73.0	492
	AttnSleep [17]	84.2	0.77	75.0	372
	CausalAttenNet [42]	85.0	0.78	77.0	126
	DistillSleepNet [43]	84.2	0.78	78.0	119
	KAleep-Net (Ours)	**86.4**	**0.81**	**79.0**	**110**

TABLE VII
STATISTICAL COMPARISON BETWEEN KALEEP-NET AND CAUSALATTENNET ON THE SLEEP-EDF-20 DATASET. RESULTS ARE REPORTED AS MEAN ± STANDARD DEVIATION ACROSS 10 FOLDS. PAIRED T-TESTS WERE USED TO COMPUTE THE T-STATISTIC AND P-VALUES FOR PERFORMANCE COMPARISON BETWEEN THE TWO MODELS.

Metric	KAleep-Net (Mean ± SD)	CausalAttenNet (Mean ± SD)	t-statistic	p-value
Accuracy (%)	86.5 ± 0.21	84.7 ± 0.35	5.28	0.0062
Cohen's κ	81.0 ± 0.26	79.9 ± 0.34	3.12	0.0357
Macro F1 Score (%)	76.8 ± 0.33	75.2 ± 0.41	3.68	0.0234

the standard sleep staging criteria defined by the American Academy of Sleep Medicine (AASM) and Rechtschaffen & Kales guidelines, which categorize sleep stages into 30-second epochs for clinical consistency [30]. Some prior studies have adopted similar window lengths, another prime reason for this choice is the optimal balance between temporal resolution and spectral fidelity, ensuring sufficient data granularity for feature extraction while minimizing noise from shorter intervals. To enhance the signal quality, a fourth-order Butterworth low-pass filter with a 40 Hz cutoff frequency was applied to

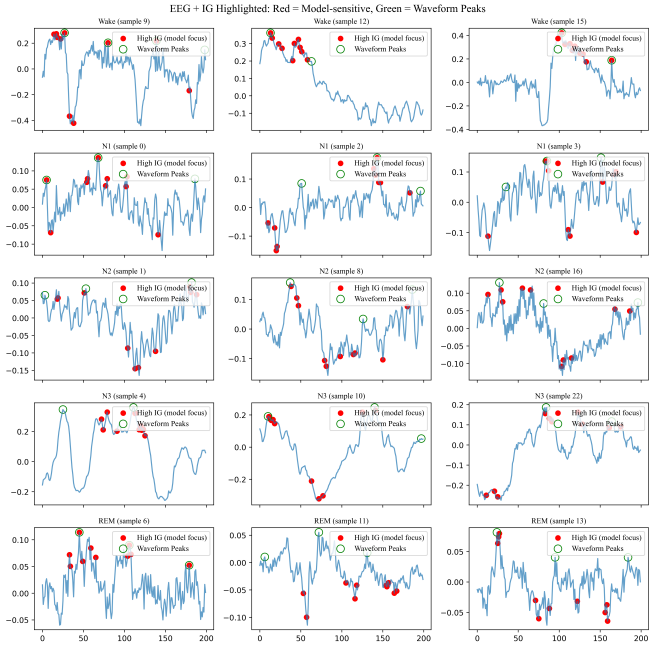

Fig. 5. Example EEG segments from each sleep stage, with red dots indicating the top 5% of high-attribution time points computed using Integrated Gradients (IG). These highlight the regions the model found most influential for classification. Green circles denote prominent waveform peaks identified by a peak-detection algorithm and are provided for visualization purposes only; they are not labeled as specific frequency bands. This figure illustrates model attention patterns without assuming explicit clinical markers.

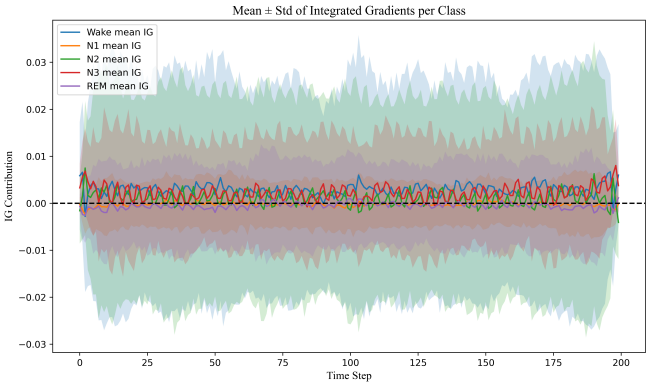

Fig. 6. Mean and standard deviation of Integrated Gradients (IG) contributions across all test samples, grouped by sleep stage. Each curve represents the average model sensitivity to EEG time points for a specific class, with shaded areas indicating variability across samples. The observed trends reflect stage-specific attribution patterns, such as localized focus in N2 and N3 and broader distributions in Wake and REM.

attenuate high-frequency artifacts (e.g., muscle activity and environmental interference), while a high-pass filter (0.5 Hz cutoff) suppressed baseline drift and electrode impedance fluctuations. This dual-band filtering preserves physiologically relevant EEG components (e.g., delta, theta, and spindle waves) that are critical for sleep-stage differentiation. Transient artifacts from muscle contractions, electrode pops, and eye blinks were further mitigated using a threshold-based rejection method (±100 μV) to ensure robust data quality for downstream analysis. For the model evaluation, subject-wise data partitioning was implemented to avoid data leakage, which is a common pitfall in biomedical signal analysis. The datasets (EDF-20, EDF-78, and SHHS) were split into 80% training, 10% validation, and 10% test sets, ensuring no overlap of subject-specific patterns across splits. The higher accuracy reported in [45], [46] may have been influenced by intersubject correlations in the training and test sets. Therefore, we adopted a subject-wise strategy to prioritize clinical realism, ensuring that our approach reflects real-world scenarios where models must generalize to unseen individuals.

To the best of our knowledge, this is the first approach to integrate Kolmogorov-Arnold Network (KAN)-based convolutional filters into sleep staging, constituting the core innovation of KAleep-Net. Compared to traditional signal decomposition methods such as Wavelet Transform (WT) and Empirical Mode Decomposition (EMD), KAN offers a unique set of advantages. The KAN functions as a learnable neural operator, adapting nonlinear basis functions during training, whereas WT relies on fixed basis transforms and EMD employs a hand-crafted, data-driven pipeline. This adaptability allows KANs to be seamlessly integrated and trained end-to-end with deep learning models, eliminating the need for manual parameter tuning or external preprocessing steps. Furthermore, KANs are computationally efficient, GPU-optimized, and suitable for real-time applications, unlike EMD and WT which often involve high computational cost or limited integration flexibility. The architecture leverages a multispectral feature pipeline to simultaneously capture coarse and fine spectral features from sleep EEG signals, enabled by KAN's adaptive approximation properties of KAN. This is complemented by a Temporal Sequencing Network to model long-range contextual dependencies and a Flash Attention mechanism, which optimizes computational efficiency by dynamically prioritizing salient sequence segments. Collectively, this design enhances feature representation while mitigating the spatial limitations and spectral instability inherent in raw EEG data, thereby enabling robust generalization. Collectively, this design enhances feature representation while mitigating the spatial limitations and spectral instability inherent in raw EEG data, thereby enabling robust generalization. Despite the relatively long input length (3000 samples per 30-second epoch), the fine-scale (kernel $=$ 3) and coarse-scale (kernel $=$ 5) filters capture distinct spectral features due to the diverse nature of EEG signals. Temporal Sequencing Networks and Flash Attention further reduce redundancy and enhance feature relevance. Ablation studies confirm the complementary effectiveness of this dual-scale approach in improving KAleep-Net's performance. The proposed method achieved state-of-the-art performance across the three benchmark datasets. On Sleep-EDF-20, KAleep-Net outperformed the existing methods by 2.1% accuracy, 0.2% F1-score, and 0.9% Cohen's kappa. Improvements were more pronounced for Sleep-EDF-78, with gains of 3.4% accuracy, 1.8% F1-score, and 4.0% kappa. On SHHS, it gained an accuracy of 86.4%, kappa of 0.81, and F1 Score of 79.0%, with training time of 110 minutes.Notably, KAleep-Net reduced the training time by 41.7% (EDF-20), 67.5% (EDF-78), and 8.18% (SHHS) compared with resource-intensive alternatives,

demonstrating the viability of resource-constrained clinical settings. Statistical significance was confirmed using a paired t-test against CausalAttenNet (T-statistic = 5.28, p = 0.0062, n = 5 runs), with results significant at ($p < 0.01$). These advancements underscore KAleep-Net's dual strength: maintaining diagnostic-grade performance while drastically reducing computational overhead, which is a critical step toward real-world deployment. A critical challenge in deploying deep learning for sleep staging is model explainability. While existing studies [13], [16], [17], [40]–[42] often treat architectures as "black boxes," the lack of interpretability limits their clinical utility because clinicians require transparency to trust automated decisions. To address this, we prioritize explainability through Integrated Gradients (IG), a post-hoc attribution method that quantifies feature contributions to predictions across all sleep stages (Wake, N1, N2, N3, REM). The IG analysis of representative EEG epochs revealed physiologically plausible decision-making patterns. Highlighted segments (marked by high IG scores) aligned with established biomarkers: alpha rhythms (8–12 Hz) during wakefulness, sleep spindles (11–16 Hz), K-complexes in N2, and slow-wave oscillations ($< 4Hz$) in N3. For example, wake-stage predictions strongly correlated with alpha-like spectral peaks, reflecting the model's reliance on relaxed wakefulness signatures. Similarly, the N2 and N3 attributions were localized to transient spindle bursts and slow-wave amplitudes, mirroring the manual scoring criteria. The statistical aggregation of IG scores across the test set further validates these trends. The mean attribution magnitudes peaked in N2 (0.72 ± 0.15 SD) and N3 (0.68 ± 0.18 SD), indicating reliance on discrete stage-specific waveforms. In contrast, Wake (0.41 ± 0.21 SD) and REM (0.38 ± 0.19 SD) exhibited lower magnitudes with broader variability, suggesting recognition of distributed, transient patterns (e.g., rapid eye movements or mixed-frequency activity). This alignment with neurophysiological hallmarks, quantified through rigorous attribution mapping, enhances the clinical credibility of KAleep-Net, bridging the gap between automated staging and actionable diagnostic insights.

Despite its promising results, this study had several limitations. First, although training efficiency has been enhanced, deploying the framework on embedded systems remains challenging due to hardware constraints. Real-time inference on low-power edge devices (e.g., wearables) necessitates further optimizations, such as structured model pruning or 8-bit quantization, to reduce memory and computational overhead. Second, the generalizability of the framework to diverse demographic groups, such as pediatric, elderly, or patient cohorts with sleep disorders, remains unvalidated, as the sleep-EDF dataset primarily includes recordings from healthy adults. Third, the reliance on 30-second epochs aligns with clinical standards for offline staging but limits applicability to continuous long-term monitoring scenarios (e.g., home-based wearables), where dynamic sleep transitions and multi-scale temporal dependencies require finer resolution.

VI. Conclusion

This study presents KAleep-Net, for automatic sleep staging that addresses the critical demand for efficient, accurate, and interpretable analyses. The architecture integrates a Multi-Spectral Feature Pipeline based on parallel-stacked Kolmogorov–Arnold Network (KAN) convolutional filters, a Temporal Sequencing Network, and Flash Attention, enabling the extraction of both fine- and coarse-grained spectral features while modeling long-range temporal dependencies with high computational efficiency. Evaluations on the Sleep-EDF-20, Sleep-EDF-78, and SHHS datasets showed that KAleep-Net outperformed existing methods, achieving accuracy gains of 2.1%, 3.4%, 2.61% F1-score improvements of 0.2%, 1.8%, 3.85%, and Cohen's kappa increases of 0.9%, 4.0%, and 1.28% respectively. Moreover, training time reductions of 41.7% 67.5%, and 8.18% emphasize the suitability of the proposed model for deployment in resource-constrained environments. Integrated Gradients (IG) analysis revealed an alignment between the model's predictions and known physiological EEG signatures, such as alpha rhythms in wakefulness, sleep spindles, and K-complexes in N2, and slow-wave activity in N3, reinforcing clinical trust. Quantitative IG aggregation further confirmed the stage-specific feature reliance, substantiating its translational potential in clinical settings.

Acknowledgement

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

[1] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, "The global problem of insufficient sleep and its serious public health implications," in *Healthcare*, vol. 7, no. 1. MDPI, 2018, p. 1.

[2] D. Zhao, Y. Wang, Q. Wang, and X. Wang, "Comparative analysis of different characteristics of automatic sleep stages," *Computer methods and programs in biomedicine*, vol. 175, pp. 53–72, 2019.

[3] L. F. Agnati, P. W. Barlow, F. Baluška, P. Tonin, M. Guescini, G. Leo, and K. Fuxe, "A new theoretical approach to the functional meaning of sleep and dreaming in humans based on the maintenance of 'predictive psychic homeostasis'," *Communicative & Integrative Biology*, vol. 4, no. 6, pp. 640–654, 2011.

[4] D. Moser, P. Anderer, G. Gruber, S. Parapatics, E. Loretz, M. Boeck, G. Kloesch, E. Heller, A. Schmidt, H. Danker-Hopfe *et al.*, "Sleep classification according to aasm and rechtschaffen & kales: effects on sleep scoring parameters," *Sleep*, vol. 32, no. 2, pp. 139–149, 2009.

[5] P. Chriskos, C. A. Frantzidis, C. M. Nday, P. T. Gkivogkli, P. D. Bamidis, and C. Kourtidou-Papadeli, "A review on current trends in automatic sleep staging through bio-signal recordings and future challenges," *Sleep medicine reviews*, vol. 55, p. 101377, 2021.

[6] I. G. Campbell, "Eeg recording and analysis for sleep research," *Current protocols in neuroscience*, vol. 49, no. 1, pp. 10–2, 2009.

[7] H. Liu, H. Zhang, B. Li, X. Yu, Y. Zhang, and T. Penzel, "Msleepnet: a semi-supervision-based multiview hybrid neural network for simultaneous sleep arousal and sleep stage detection," *IEEE transactions on instrumentation and measurement*, vol. 73, pp. 1–9, 2024.

[8] D. Zhou, Q. Xu, J. Wang, H. Xu, L. Kettunen, Z. Chang, and F. Cong, "Alleviating class imbalance problem in automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[9] P. Jadhav and S. Mukhopadhyay, "Automated sleep stage scoring using time-frequency spectra convolution neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–9, 2022.

[10] I. A. Zapata, Y. Li, and P. Wen, "Rules-based and svm-q methods with multitapers and convolution for sleep eeg stages classification," *IEEE Access*, vol. 10, pp. 71 299–71 310, 2022.

[11] E. Alickovic and A. Subasi, "Ensemble svm method for automatic sleep stage classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1258–1265, 2018.

[12] J. Van Der Donckt, J. Van Der Donckt, E. Deprost, N. Vandenbussche, M. Rademaker, G. Vandewiele, and S. Van Hoecke, "Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring," *Biomedical Signal Processing and Control*, vol. 81, p. 104429, 2023.

[13] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, p. e0216456, 2019.

[14] J. Lu, C. Yan, J. Li, and C. Liu, "Sleep staging based on single-channel eeg and eog with tiny u-net," *Computers in Biology and Medicine*, vol. 163, p. 107127, 2023.

[15] Y. Zhang, W. Cao, L. Feng, M. Wang, T. Geng, J. Zhou, and D. Gao, "Shnn: A single-channel eeg sleep staging model based on semi-supervised learning," *Expert Systems with Applications*, vol. 213, p. 119288, 2023.

[16] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.

[17] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[18] C. Zhao, J. Li, and Y. Guo, "Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel eeg," *Biomedical Signal Processing and Control*, vol. 88, p. 105615, 2024.

[19] H. A. Siddiqa, Z. Tang, Y. Xu, L. Wang, M. Irfan, S. F. Abbasi, A. Nawaz, C. Chen, and W. Chen, "Single-channel eeg data analysis using a multi-branch cnn for neonatal sleep staging," *IEEE Access*, vol. 12, pp. 29 910–29 925, 2024.

[20] H. Zhu, W. Zhou, C. Fu, Y. Wu, N. Shen, F. Shu, H. Yu, W. Chen, and C. Chen, "Masksleepnet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2353–2364, 2023.

[21] X. Chen, Y. Zhang, Q. Chen, L. Zhou, H. Chen, H. Wu, Y. Xu, K. Chen, B. Yin, W. Chen *et al.*, "Astgsleep: Attention based spatial-temporal graph network for sleep staging," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[22] Z. Ren, J. Ma, and Y. Ding, "Flexiblesleepnet: A model for automatic sleep stage classification based on multi-channel polysomnography," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[23] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[24] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based bilstm-crf approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[25] M. Pagliardini, D. Paliotta, M. Jaggi, and F. Fleuret, "Faster causal attention over large sequences through sparse flash attention," *arXiv preprint arXiv:2306.01160*, 2023.

[26] Y. Zhuo and Z. Ge, "Ig 2: Integrated gradient on iterative gradient path for feature attribution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[27] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet *et al.*, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[28] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 05 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy064

[29] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[30] R. K. Malhotra and A. Y. Avidan, "Sleep stages and scoring technique," *Atlas of sleep medicine*, pp. 77–99, 2013.

[31] C. Wan, M. C. Nnamdi, W. Shi, B. Smith, C. Purnell, and M. D. Wang, "Advancing sleep disorder diagnostics: A transformer-based eeg model for sleep stage classification and osa prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 878–886, 2025.

[32] J. Huang, L. Ren, X. Zhou, and K. Yan, "An improved neural network based on senet for sleep stage classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4948–4956, 2022.

[33] R. S. Mulyana, "A low voltage, low power 4th order continuous-time butterworth filter for electroencephalography signal recognition," Master's thesis, The Ohio State University, 2010.

[34] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, "On the effects of data normalization for domain adaptation on eeg data," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106205, 2023.

[35] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[36] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.

[37] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.

[38] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1806–1819, 2017.

[39] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[40] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[41] Y. Sun, B. Wang, J. Jin, and X. Wang, "Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–5.

[42] J. Pan, Y. Feng, P. Zhao, X. Zou, A. Hou, and X. Che, "Causalattennet: A fast and long-term-temporal network for automatic sleep staging with single-channel eeg," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[43] Z. Jia, H. Liang, Y. Liu, H. Wang, and T. Jiang, "Distillsleepnet: Heterogeneous multi-level knowledge distillation via teacher assistant for sleep staging," *IEEE Transactions on Big Data*, vol. 11, no. 3, pp. 1273–1284, 2025.

[44] H. Lee, Y. R. Choi, H. K. Lee, J. Jeong, J. Hong, H.-W. Shin, and H.-S. Kim, "Explainable vision transformer for automatic visual sleep staging on multimodal psg signals," *npj Digital Medicine*, vol. 8, no. 1, p. 55, 2025.

[45] S. K. Satapathy, B. Brahma, B. Panda, P. Barsocchi, and A. K. Bhoi, "Machine learning-empowered sleep staging classification using multi-modality signals," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 119, 2024.

[46] D. Jiang, M. Yu, W. Yuanyuan *et al.*, "Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement," *Expert Systems with Applications*, vol. 121, pp. 188–203, 2019.