
WAVE_SLEEPNET: AN INTERPRETABLE NETWORK FOR EXPERT-LIKE SLEEP STAGING

A PREPRINT

✉ Yan Pei

College of Biomedical Engineering
Zhejiang University
summer_bae@zju.edu.cn

Wei Luo*

College of Biomedical Engineering
Zhejiang University
luo.wei@zju.edu.cn

April 25, 2024

ABSTRACT

Although deep learning algorithms have proven their efficiency in automatic sleep staging, the widespread skepticism about their "black-box" nature has limited its clinical acceptance. In this study, we propose WaveSleepNet, an interpretable neural network for sleep staging that reasons in a similar way to sleep experts. In this network, we utilize the latent space representations generated during training to identify characteristic wave prototypes corresponding to different sleep stages. The feature representation of an input signal is segmented into patches within the latent space, each of which is compared against the learned wave prototypes. The proximity between these patches and the wave prototypes is quantified through scores, indicating the prototypes' presence and relative proportion within the signal. The scores are served as the decision-making criteria for final sleep staging. During training, an ensemble of loss functions is employed for the prototypes' diversity and robustness. Furthermore, the learned wave prototypes are visualized by analysing occlusion sensitivity. The efficacy of WaveSleepNet is validated across three public datasets, achieving sleep staging performance that are on par with the state-of-the-art models when several WaveSleepNets are combined into a larger network. A detailed case study examined the decision-making process of the WaveSleepNet which aligns closely with American Academy of Sleep Medicine (AASM) manual guidelines. Another case study systematically explained the misidentified reason behind each sleep stage. WaveSleepNet's transparent process provides specialists with direct access to the physiological significance of its criteria, allowing for future adaptation or enrichment by sleep experts.

Keywords Automatic sleep staging, interpretability, deep learning

1 Introduction

The pervasive issue of sleep deprivation and the prevalence of sleep disorders represent a significant public health concern, with profound implications for morbidity and mortality. Sleep disturbances, which are closely associated with various psychiatric, neurodegenerative, and cardiovascular conditions, underscore the necessity for precise sleep analysis [Tobaldini et al. [2018]]. Polysomnography (PSG) is the cornerstone of sleep evaluation, holding a pivotal role in research and clinical practice by detecting unexpected events and thus revealing sleep abnormalities [Berry et al. [2017], Malhotra and Loscalzo [2009]].

The process of sleep staging, traditionally conducted by trained experts, involves extensive manual annotation, often necessitating several hours per patient. This labor-intensive approach incurs significant costs in terms of human resources, which can become a bottleneck for large-scale clinical research [Rosenberg and Van Hout [2013]]. Additionally, manual scoring is subject to inter-rater and intra-rater variability, posing challenges in maintaining consistency with scoring rules. Consequently, a succession of automatic sleep staging methodologies has been developed to support sleep specialists [Muto and Berthomier [2023]]. These automated methods aim to reduce the subjectivity and resource demands of manual annotations while striving to match or surpass the accuracy of human experts.

The recent integration of deep learning (DL) methods into automatic sleep staging has demonstrated superior performance compared to traditional machine learning (ML) models, largely due to DL’s ability to operate without extensive prior knowledge of physiological processes. Notably, Convolutional Neural Networks (CNNs) Memar and Faradji [2018]Sors et al. [2018], Recurrent Neural Networks (RNNs) and their hybrids Sokolovsky et al. [2020]Phan et al. [2019] have been effectively employed in this field. However, the protracted training times associated with RNNs, due to their sequential data processing nature, and the critical need for contextual information in sleep signal analysis, have led to the adoption of transformer networksQu et al. [2020]Phan et al. [2022] and attention mechanismsEldele et al. [2021]. These networks excel in capturing temporal dependencies. Among these developments, single-channel sleep staging has gained prominence for its user-friendliness and compatibility with home-based sleep monitoring.

Despite considerable advancements, the clinical adoption of automatic sleep staging algorithms remains limited. Discussions with leading sleep experts reveal a prevailing skepticism regarding the black box characteristic of deep learning models, a common concern within artificial intelligence in healthcare and medicine. Both ML and DL require the extraction of representative features, which are then used to inform classification models. ML methodsHassan et al. [2015]Saastamoinen et al. [2006] offer greater interpretability but are limited by the labor-intensive process of manual feature engineering and often result in inferior performance in sleep stage classification when compared to DL methods. Conversely, DL algorithms have the advantage of autonomously learning features from raw data, without requiring prior knowledge of physiology. However, this advantage is a double-edged sword, as it obscures which specific sleep features are encoded. The consequent lack of explainability limits the adoption of deep learning models in clinical practice because clinicians often need to understand the reasons behind each classification to avoid data noises and unexpected biasesAl-Hussaini et al. [2019]. These issues highlight the necessity for advancements in DL-based sleep stage methodologies that prioritize interpretability to make the sleep stage process more explainable, by revealing more details of their inner workings.

Recent advancements have been made towards explainable sleep stage models. Dutt et. alDutt et al. [2022] developed SleepXAI, an integrated CNN-CRF model that employs a modified Grad-CAMSelvaraju et al. [2017] for explainable multi-class sleep stage classification from single channel EEG signals. This model highlights the signal segments most influential in stage prediction, enhancing explainability. Yet, the method falls short in explaining the reasoning process of how a network actually makes its decisions and clarifying the physiological relevance of its criteria. Phan et. alPhan et al. [2022] and Pradeepkumar et. alPradeepkumar et al. [2022] both introduced transformer-based interpretable sleep stage models, a sequence-to-sequence model underpinned by transformer architecture, which interprets EEG data through self-attention scores visualized as heat maps to highlight sleep-relevant features captured from the input EEG signal. They mainly define the areas of focus without clarifying the relationship between the signal segments and their waveform characteristics, which can lead to ambiguity. Furthermore, these models concentrate on specific waveforms but neglect the significance of waveform prevalence in sleep staging task. For instance, alpha waves and theta waves are present in both Wake stage and N1 stage. However, the predominance of alpha waves typically indicates the Wake stage, while a higher proportion of theta waves is characteristic of the N1 stage—an important distinction that is not sufficiently captured by these models.

To address the above problem, we propose a waveform-sensing-based explainable neural network named WaveSleepNet. It enables network models to classify in a manner similar to human cognitive processes and explains their decision-making in a manner that agrees with the way sleep experts describe their own thinking in sleep staging tasks. How do human experts describe their process of sleep staging? According to AASM Berry et al. [2017] standards, which is briefly summarized in Table 1, sleep stages are classified by experts through the systematic analysis of characteristic waveforms (such as spindles, alpha waves, etc.). This process involves the examination of the presence and relative proportions of the specific waveforms within the input signals. To acquire such a sleep staging model that imitates this way of thinking and explains its reasoning process in a human-understandable way, we propose WaveSleepNet based on prototype learningBien and Tibshirani [2011]Priebe et al. [2003]Chen et al. [2019a] which is a classical form of case-based reasoningKolodner [1992]. In the proposed model, a fully-convolutional neural network is utilized as a local feature extraction network to project raw EEG signals into a low-dimensional features. These deep features are segmented into patches, which are then compared with wave prototypes learned by the network within a latent feature space to calculate corresponding distances. These distances are fed into a subsequent module and transformed into scores which respectively reflect the presence of a wave pattern within the signal and the relative proportion of the wave pattern. These scores are then weighted and combined to derive the final sleep staging prediction. The wave prototypes are key elements of the WaveSleepNet model, signifying the latent representations of characteristic waveforms learned by the network that are capable of distinguishing between different sleep stages. To visualize the wave prototypes embedded in the network as signal waveforms, we employed occlusion sensitivityZeiler and Fergus [2014] analysis to equate each pattern with a signal segment derived from the training dataset. Furthermore, to better balance the accuracy and interpretability of the network, the proposed model incorporates an ensemble of loss functions, enabling the model

to learn more diverse wave prototypes during the training process. In this way, our model is interpretable, in the sense that it has a transparent reasoning process when making predictions.

Table 1: Characteristics of each sleep stage according to the AASM rule

Sleep Stage	Characteristics
Wake	• alpha rhythm (8-13 Hz)
	• Eye blinks
	• Rapid eye movements (REM)
	• Slow eye movements (SEM)
N1	• More than 50% of low-amplitude, mixed-frequency (LAMF) EEG activity (4-7 Hz)
	• Slow eye movements (SEM)
N2	• Vertex sharp waves (V waves)
	• K-Complex
N3	• Sleep spindle (12-14 Hz)
	• More than 20% of Delta waves
REM	• Sleep spindle (12-14 Hz)
	• Rapid eye movements (REM)
	• Sawtooth waves (2-6Hz)

The main contributions of this work are as follows:

- 1) The proposed WaveSleepNet is a case-based model which automatically sleep scoring in a manner similar to the cognitive processes of sleep experts. To the best of our knowledge, this is the first application of case-based learning to the task of sleep staging. The effectiveness of WaveSleepNet model for sleep staging is validated on three public datasets.
- 2) Our model, WaveSleepNet, provides a level of interpretability characterized by a transparent reasoning process. It determines the presence and proportion of characteristic waveforms within a signal based on their proximity to the learned waveform prototypes, thereby classifying the sleep stage. Our experimental case studies also corroborate the alignment of our model’s sleep staging logic with that of the AASM.
- 3) We further analyzed the misclassification errors of the model which were not explored in prior studies. By systematically examining the reasons behind the incorrect classification of signals in various stages, we have provided direction for further improvements to the model to enhance the accuracy of sleep staging.

2 METHOD

The proposed model is designed to classify an input sequence of single-channel EEG epochs $\mathbf{x}^{(L)} = (x_1, x_2, \dots, x_L)$ into the sleep stage of the last input EEG epoch $\hat{\mathbf{y}}^{(L)} = (\hat{y}_L)$ (see Fig. 1). The input single-channel EEG signal is denoted as $\mathbf{x}^{(L)} \in \mathbb{R}^{E \cdot F \cdot L}$, where E is the duration of an EEG epoch in seconds and F is the sampling frequency. The predicted sleep stage for the L -th EEG epoch is presented as $\hat{\mathbf{y}}^{(L)} \in \{Wake, N1, N2, N3, REM\}$. It follows the five-stage sleep classification in the AASM rule Berry et al. [2017].

Fig. 1 illustrates the overall framework of our WaveSleepNet model. It comprises three major components: Feature Extraction Network, WaveSensing Network and Decision Network. Firstly, the Feature Extraction Network derives deep features from the raw input signals, effectively capturing information across various temporal scales and channel dimensions. Subsequently, within the WaveSensing Network, the convolutional filters can be regarded as a series of specific waveform prototypes, which are constrained to be identical to some latent training patch. The process involves the comparison of extracted features against these learned waveform prototypes by calculating the L^2 distance for each waveform feature. These distances are subsequently fed into the Decision Network, which comprises two decision-making estimators. The two estimators converted the distances into scores to assess the presence and the relative proportion of the wave prototypes within the input signals based on the proximity to the wave prototypes. The scores are weighted combined to get the final prediction.

2.1 Feature Extraction Network

The role of the backbone network is to learn the feature map, $\mathcal{F}(\mathbf{x}^{(L)}) : \mathbf{x}^{(L)} \in \mathbb{R}^{E \cdot F \cdot L} \mapsto \mathbf{s} \in \mathbb{R}^{K \times C}$, in order to transform an input EEG epoch $\mathbf{x}^{(L)}$ into a high-level feature vector $\mathbf{s}$ for representation. Herein, K and C denotes the dimensionality and the channel number of the extracted feature, respectively.

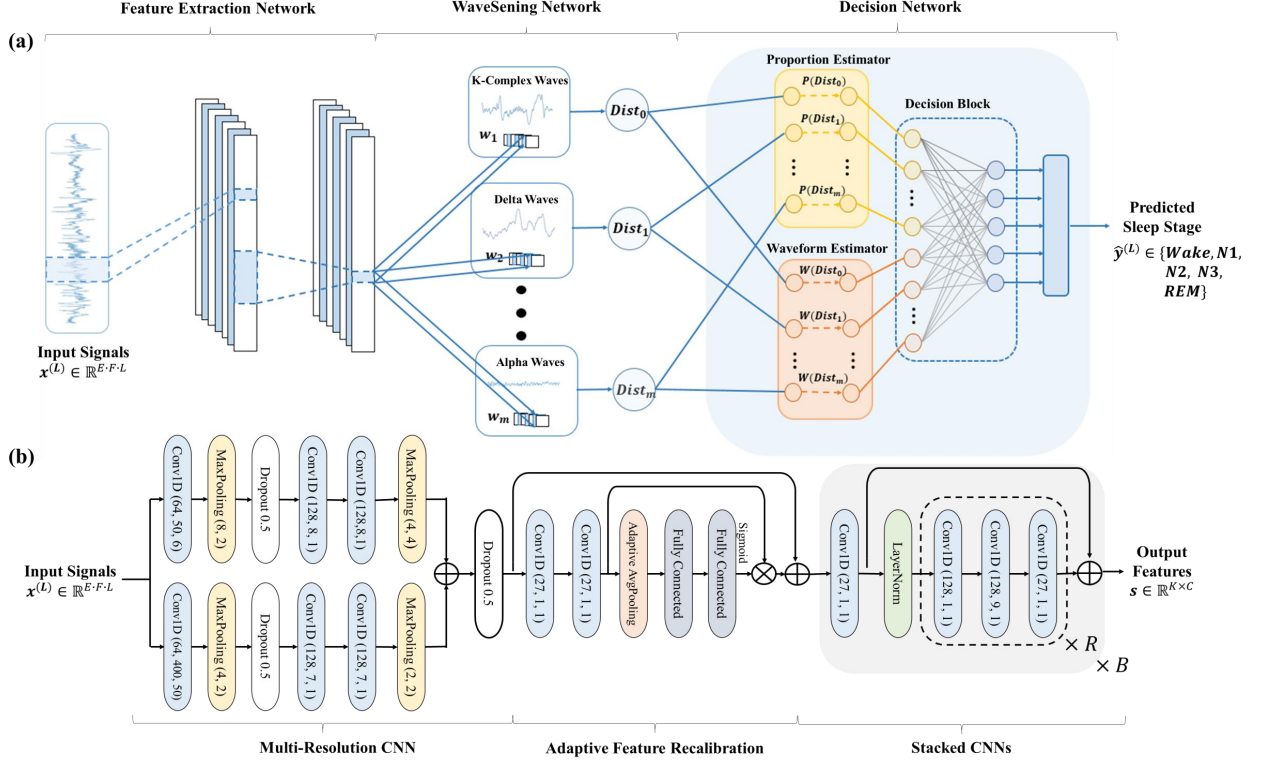


Figure 1: The architecture of the proposed WaveSleepNet. (a) The overall model architecture. (b) The structure of the Feature Extraction Network.

The architecture of the network is shown in Fig. 1(b), comprising modules of Multi-Resolution CNN (MRCNN), Adaptive Feature Recalibration (AFR) and stacked convolutional networks (stacked CNNs). The MRCNN and AFR are adopted from previous research Supratak et al. [2017] Eldele et al. [2021] Zhu et al. [2023].

Given that different sleep stages are characterized by different frequency ranges, we adopt the MRCNN to extract features in the signal across different scales. The MRCNN has two branches of CNNs with different kernel sizes, which are determined by the characteristics of the EEG signal and the signal sampling rate. In particular, Conv1D(64, 50, 6) in Fig. 1 represents using a 1D convolution layer with 64 filters, a kernel size of 50 and a stride of 6. Similarly, MaxPooling(4, 4) refers to a MaxPooling layer with a kernel size of 4 and a stride of 4.

AFR is employed to model the inter-dependencies among features, enabling the selective emphasis on the most discriminative features via a residual Squeeze-and-Excitation (SE) block Hu et al. [2017]. The features learned by the MRCNN are fed into two 1×1 Conv1D convolutional layers, which integrate information across different channels and reduce the dimensionality of the features. Subsequently, a Squeeze-and-Excitation (SE) block is utilized to recalibrate the feature maps using “squeeze” operation of adaptive average pooling, followed by an “excitation” operation that uses two fully connected layers.

To extract complex features in EEG signals, we adopted stacked CNNs that consists of B stacked 1-D convolutional blocks, as shown in Fig. 1. Each convolutional block is formed by a convolutional layer, a residual connection, a 1-D Layer-Norm layer and R unit convolutional layers which is formed by a sequence of 1-D convolutional layers. The residual path of a block serves as the input to the next block. LayerNorm refers to applying layer normalization Ba et al. [2016], stabilizing the hidden state dynamics in deep networks. In each convolutional unit, there’s a standard convolutional layer to increase feature dimensions, and depthwise separable convolution Chollet [2017] which decouples the standard convolution operation into depthwise convolution and pointwise convolution to decrease the parameter numbers.

2.2 WaveSensing Network

The WaveSensing Network architecture is inspired by ProtoPNet Chen et al. [2019b], which dissects the image by finding prototypical parts and combines evidence from the prototypes to make a final classification. The WaveSensing

Network capitalizes on trainable wave prototypes to transform the features learned by the Feature Extraction Network into a set of distance values between the features and the corresponding wave prototypes.

Given the deep features $s \in \mathbb{R}^{K \times C}$ extracted by the Feature Extraction Network, the network learns M wave prototypes $W = \{w_j\}_{j=1}^M$, whose shape is $K_1 \times C$ with $K_1 \leq K$. In our experiments, we set $K_1 = 1$. Since the channel (C) of each wave pattern is the same as that of the deep features but the length of each wave pattern is smaller than those of the deep features, each wave pattern will be used to represent some prototypical activation pattern in a patch of the deep features, which in turn will correspond to some prototypical waveform in the original EEG signals. Consequently, each wave pattern w_j can be regarded as the latent representation of distinctive waveforms that are instrumental in differentiating sleep stages. As a schematic illustration, the wave prototypes in Fig. 1 corresponding to K-complex wave, Delta wave and Alpha wave.

The deep features s are segmented into K one-point-overlapped patches which have the same shape as w_j . The distance $Dist_j \in \mathbb{R}^K$ between the j -th wave pattern w_j and all patches of s is calculated as (1). The result is an activation map of distance scores whose value indicates the proximity between the wave pattern and the corresponding feature patches.

$$Dist_j = Stack\{\| \tilde{s} - w_j \|_2^2 \text{ for } \tilde{s} \in patches(s)\} \quad (1)$$

Every wave prototype w_j is the latent representation of some training signal segments, which naturally and faithfully becomes the prototype's visualization. In this way, we can interpret the wave prototypes as visualizable prototypical signal segments. To determine which segment of the EEG signals from the training set X_{Train} corresponds to wave pattern w_j , we initiate by selecting the EEG epoch x_n that contains a feature patch $\tilde{s}_k$ with the minimum distance $Dist_j$ to w_j , as shown in Eq. (2). Subsequently, to pinpoint the relevant signal segment within X_n , we systematically occlude varying segments of the input signal with arrays of zeros, and monitoring the changes in $Dist_j$. The occlusion that results in a substantial increase in the distance $Dist_j$ identifies the signal segment that we then designate as the visual representation of w_j .

$$\begin{aligned} \min\{Dist_j\} &= \min_{s \in \mathcal{S}} \{\|s - w_j\|_2^2\}, \text{ where } \mathcal{S} = \\ &\quad \{\tilde{s} : \tilde{s} \in patches(\mathcal{F}(x_i)) \text{ for } x_i \in x_{train}\} \\ &= \|\tilde{s}_k - w_j\|_2^2, \text{ where } \tilde{s}_k \in patches(\mathcal{F}(x_n)) \end{aligned} \quad (2)$$

2.3 Decision Network

Considering the fundamental scoring rationales for sleep staging, as outlined in Table 1, we propose a decision network composed of Waveform Estimator, Proportion Estimator, and Decision Block. The Waveform Estimator and the Proportion Estimator transform $Dist$, derived from WaveSensing Network, into scores that reflect the presence of a wave pattern within the signal and the relative proportion of the wave pattern, respectively. The Decision Block integrates these two scores to predict the final sleep stage accurately.

2.3.1 Waveform Estimator

The distance $Dist_j$ between deep features s and the j -th wave pattern w_j is converted into a similarity score as delineated by Eq. (3) Chen et al. [2019b]. The similarity function employed is monotonically decreasing with respect to $Dist_j$. The resultant activation map, which comprises similarity scores produced by wave pattern w_j , undergoes dimensional reduction via global max pooling to yield a single similarity score. This score quantitatively reflects the presence of a specific waveform within the input EEG epoch, as mathematically represented by Eq. (4). Consequently, a large output of $WScore_j$, indicates that a deep feature patch closely aligns with the j -th wave pattern in terms of the L2-norm. This, in turn, suggests that a corresponding segment of the EEG signals has a similar concept to the characteristic waveform corresponding to the j -th wave pattern.

$$Similarity_j = \log((Dist_j + 1)/(Dist_j + \epsilon)) \quad (3)$$

$$WScore_j = ReLU(BatchNorm(GlobalMaxPooling(Similarity_j))) \quad (4)$$

2.3.2 Proportion Estimator

Similar to Waveform Estimator, the $Dist_j$ derived from WaveSensing Network is converted into a similarity score $Similarity_j$. Subsequently, this similarity score is transformed into a quantifiable measure that indicates the prevalence of the characteristic waveform within the input signals, which is shown in Eq. (5). In this score, we employ GlobalAveragePooling in place of the GlobalMaxPooling utilized in the Waveform Estimator. Thus, a large output of $PScore_j$ signifies not only the presence of a waveform in the input signal that closely resembles the j -th wave pattern but also indicates that such a waveform constitutes a significant proportion of the input signal.

$$PScore_j = ReLU(BatchNorm(GlobalAveragePooling(Similarity_j))) \quad (5)$$

2.3.3 Decision Block

In the Decision Block, the outputs from Waveform Estimator ($WScore \in \mathbb{R}^M$), and the Proportion Estimator ($PScore \in \mathbb{R}^M$), are concatenated to yield a composite score, denoted as $Score \in \mathbb{R}^{2M}$. Next, a fully connected (FC) layer is utilized to assign suitable weights to the concatenated scores. The outputs of this layer are then fed into the sigmoid function to obtain the predicted sleep stage of the L -th epoch, denoted as $\hat{y}^{(L)}$. This operation can be represented as shown in Eq. (6).

$$\hat{y}^{(L)} = Sigmoid(FC(Concat([WScore, PScore]))) \quad (6)$$

2.4 Loss Function Ensemble

The loss function of the network is carefully designed to optimize classification accuracy and enhance interpretability. In addition to the classification loss, there is a penalty on learned wave prototypes to promote their diversity and coverage of the data in latent representations. There are two error terms to reinforce the alignment of learned wave prototypes with distinct feature patches in the latent space. Furthermore, an L1 regularization term has been adopted to facilitate the wave prototypes selection for each sleep stage, which effectively improves interpretability.

We employ the standard cross-entropy loss to penalize the sleep stage misclassification. The cross-entropy loss is denoted by L_{class} , and is given by:

$$L_{class} = \frac{1}{n} \sum_{i=1}^n \sum_{c=1}^{N_C} -1[y_i = c] \log(P(\hat{y} = c)) \quad (7)$$

where y_i is the label of the i -th EEG epoch and $P(\hat{y} = c)$ denotes the probability of the model predicting sleep stage class c .

We use the average minimum squared L_2 distance between any two wave prototypes, w_i, w_j to penalize the wave prototypes' diversity error. The diversity loss, denoted by L_{dist} , is given by Eq. (8). The logarithm function tapers large distances so that the penalty does not diminish or vanish, while the inclusion of the ϵ term enhances numerical stability. By computing the inverse of the logarithm of the distances between wave prototypes, we penalize wave prototypes that are close in distance while making sure the minimum distance between wave prototypes does not become excessively large.

$$L_{dist} = \frac{1}{\log(\frac{1}{m} \sum_{j=1}^m \min_{i>j \in [1,m]} \|w_i - w_j\|_2^2) + \epsilon} \quad (8)$$

There are two loss terms that relate the distance of the feature patches to the wave prototypes in latent space:

$$L_{R1} = \frac{1}{m} \sum_{j=1}^m \min_{i \in [1,n]} \|w_j - s_i\|_2^2 \quad (9)$$

$$L_{R2} = \frac{1}{n} \sum_{i=1}^n \min_{j \in [1,m]} \|s_i - w_j\|_2^2 \quad (10)$$

Here both terms represent the mean of the minimum squared distances. The minimization of L_{R1} would necessitate each wave pattern to be as close as possible to at least one instance within the training set in the latent space, while the minimization of L_{R2} would require every encoded training signal to be as close as possible to one of the wave prototypes. This balance ensures significant pixel-to-pixel correspondences between the wave prototypes and the training signals, thereby enhancing the representational fidelity of the model.

Table 2: Experiment settings and dataset statistics

Dataset	No. of subjects	EEG channel	Experimental setting		Class distribution					
			Evaluation scheme	Held-out validation set	Wake	N1	N2	N3	REM	Total
SleepEDF-20	20	Fpz-Cz	20-fold CV	1 subject	8,285 (19.6 %)	2,804 (6.6 %)	17,799 (42.1 %)	5,703 (13.5 %)	7,717 (18.2 %)	42,308
SleepEDF-78	78	Fpz-Cz	10-fold CV	7 subject	69,824 (35.8 %)	21,522 (10.8 %)	69,132 (34.7 %)	13,039 (6.5 %)	25,835 (13.0 %)	199,352
SHHS	5,793	C4-A1	Train/Test: 0.7/0.3	100 subject	1,691,288 (28.8 %)	217,583 (3.7 %)	2,397,460 (40.9 %)	739,403 (12.6 %)	817,473 (13.9 %)	5,863,207

Dataset	SleepEDF-20					SleepEDF-78					SHHS				
AC \ PC	Wake	N1	N2	N3	REM	Wake	N1	N2	N3	REM	Wake	N1	N2	N3	REM
Wake	8936 (91%)	633 (6%)	119 (1%)	31 (0%)	127 (1%)	62,905 (92%)	4,219 (6%)	490 (1%)	65 (0%)	768 (1%)	457,714 (91%)	14,599 (3%)	19,994 (4%)	1,629 (0%)	7,266 (1%)
N1	477 (17%)	1424 (51%)	506 (18%)	12 (0%)	385 (14%)	3,724 (17%)	10,210 (47%)	5,524 (26%)	70 (0%)	2,184 (10%)	10,567 (16%)	32,242 (49%)	17,212 (26%)	30 (0%)	6,236 (9%)
N2	303 (2%)	604 (3%)	15581 (88%)	546 (3%)	765 (4%)	540 (1%)	4,594 (7%)	58,926 (85%)	2,047 (3%)	3,025 (4%)	18,888 (3%)	22,109 (3%)	638,353 (88%)	32,701 (4%)	15,206 (2%)
N3	41 (1%)	8 (0%)	818 (14%)	4832 (85%)	4 (0%)	38 (0%)	11 (0%)	2,646 (20%)	10,320 (79%)	24 (0%)	970 (0%)	16 (0%)	40,156 (19%)	171,773 (81%)	88 (0%)
REM	117 (2%)	479 (6%)	579 (8%)	0 (0%)	6542 (85%)	336 (1%)	1,862 (7%)	2,430 (9%)	8 (0%)	21,199 (82%)	4,748 (2%)	6,375 (3%)	14,340 (6%)	126 (0%)	209,800 (89%)

Figure 2: Confusion matrix of WaveSleepNet for SleepEDF-20, SleepEDF-78, and SHHS dataset. The values in parentheses indicate per-class recall. AC and PC denote Actual Class and Predicted Class, respectively.

Furthermore, we have observed that without the constraints on the fully-connected layer within the Decision Block, the model indiscriminately assigns weights to all wave prototypes, regardless of their relevance to the specific sleep stage. This can lead to confusing model interpretability. To address this issue, we incorporated an $L1$ regularization term into our loss function to select significant wave prototypes in the sleep stage, effectively pushing the weights associated with less important wave prototypes toward zero. The $L1$ regularization term are shown as follows:

$$L1 = \sum_i |W_i|, \text{ for } W_i \in W \quad (11)$$

where $W \in \mathbb{R}^{2M \times 5}$ represents the weight matrices of fully-connected layer in the Decision Block.

Putting everything together, the overall objective function, denoted as L , can be formulated as: (6).

$$L = \lambda_{class} L_{class} + \lambda_{dist} L_{dist} + \lambda_{R1} L_{R1} + \lambda_{R2} L_{R2} + \lambda_{L1} L1 \quad (12)$$

where λ_{class} , λ_{dist} , λ_{R1} , λ_{R2} and λ_{L1} are real-valued hyperparameters that adjust the ratios between terms.

3 EXPERIMENTS AND RESULTS

3.1 Datasets

To assess both the classification performance and the interpretability of our model, we conducted experiments on three public datasets, including SleepEDF-20, SleepEDF-78Kemp et al. [2000] and Sleep Heart Health Study. In this study, we employed a single EEG channel from each dataset for classification. For each dataset, we excluded EEG epochs with annotations unrelated to sleep stages, such as MOVEMENT class. Moreover, to address the five-class problem in datasets annotated with $R\&K$, we merged the N3 and N4 classes into a single N3 class according to the AASM criteria. Table 2 outlines the number of subjects, EEG channels used, evaluation scheme, number of reserved validation subjects, and sample distribution.

3.1.1 SleepEDF-20

This subset of the Sleep-EDF Expanded dataset (2013 version) was obtained from the PhysioBankGoldberger et al. [2000] and included two studies. The Sleep Cassette (SC) study included 20 healthy Caucasian individuals aged from 25

Table 3: Performance comparison between WaveSleepNet and state-of-the-art (SOTA) methods for automatic sleep scoring via deep learning; bold indicate the best model; the results indicated by * indicates the larger combined WaveSleepNet model.

Dataset	Method			Overall Metrics				Perclass F1 Score			
	Model	Interpretable	Channel	ACC	MF1	κ	Wake	N1	N2	N3	REM
SleepEDF-20	XSleepNetPhan et al. [2021]	No	Single-Channel	86.3	80.6	0.81	-	-	-	-	-
	TinySleepNetSupratak and Guo [2020]	No	Single-Channel	85.4	80.5	0.80	90.1	51.4	88.5	88.3	84.3
	SeqSleepNetPhan et al. [2019]	No	Multi-Channel	86.0	79.7	0.81	-	-	-	-	-
	AttnSleepEldele et al. [2021]	No	Single-Channel	84.4	78.1	0.79	89.7	42.6	88.8	90.2	79.0
	DeepSleepNetSupratak et al. [2017]	No	Single-Channel	82.0	76.9	0.76	-	-	-	-	-
	SleepEEGNetMousavi et al. [2019]	No	Single-Channel	84.3	79.7	0.79	89.2	52.2	89.8	85.1	85.0
	SleepContextNetZhao et al. [2022]	No	Single-Channel	84.8	79.8	0.79	89.6	50.5	88.4	88.5	85.0
	Ours	Yes	Single-Channel	85.1	79.5	0.80	90.6	47.8	88.0	86.9	84.2
Ours*	Yes	Single-Channel	86.5	81.0	0.87	92.2	48.7	89.1	88.2	86.7	
SleepEDF-78	XSleepNetPhan et al. [2021]	No	Single-Channel	83.9	77.9	0.78	93.3	49.9	86	78.7	84.2
	Korkalainen et al.Korkalainen et al. [2019]	No	Single-Channel	83.7	-	0.77	-	-	-	-	-
	TinySleepNetSupratak and Guo [2020]	No	Single-Channel	83.1	78.1	0.77	92.8	51.0	85.3	81.1	80.3
	SeqSleepNetPhan et al. [2019]	No	Multi-Channel	82.6	76.4	0.76	-	-	-	-	-
	SleepTransformerPhan et al. [2022]	Yes	Single-Channel	81.4	74.3	0.74	91.7	40.4	84.3	77.9	77.2
	U-TimePerslev et al. [2019]	No	Single-Channel	81.4	74.3	0.74	91.7	44.1	82.5	73.5	76.1
	SleepEEGNetMousavi et al. [2019]	No	Single-Channel	80.0	73.6	0.73	91.7	44.1	82.5	73.5	76.1
	SleepContextNetZhao et al. [2022]	No	Single-Channel	82.7	77.2	0.76	92.8	49.0	84.8	80.6	78.9
Ours	Yes	Single-Channel	82.5	77.1	0.76	92.5	47.5	84.7	80.8	80	
Ours*	Yes	Single-Channel	83.9	78.5	0.77	93.3	49.8	85.9	81.3	82.2	
SHHS	SleepTransformerPhan et al. [2022]	Yes	Single-Channel	87.7	80.1	0.82	92.2	46.1	88.3	85.2	88.6
	XSleepNetPhan et al. [2021]	No	Single-Channel	87.6	80.7	0.82	92.0	49.9	88.3	85.2	88.6
	IITNetSeo et al. [2020]	No	Single-Channel	86.7	79.8	0.81	90.1	48.1	88.4	85.2	87.2
	SeqSleepNetPhan et al. [2019]	No	Multi-Channel	86.5	78.5	0.81	-	-	-	-	-
	Ours	Yes	Single-Channel	86.7	79.1	0.81	92.1	45.5	87.6	81.9	88.5
	Ours*	Yes	Single-Channel	87.6	80.7	0.83	92.4	49.5	88.4	84.6	88.5

to 34 years without any sleep-related medication. The Sleep Telemetry (ST) study included 22 subjects ranging in age from 18 to 79 years, aimed at investigating the effects of Temazepam on sleep. To avoid the influence of drug-related factors and following prior research, the SC subset was adopted in this study. The PSG recordings from the dataset were segmented into 30-second epochs and manually annotated by expert sleep clinicians according to the *R&K* manual. Each 30-second epoch was labeled as one of eight sleep stages (W, S1, S2, S3, S4, REM, MOVEMENT, UNKNOWN) by the sleep experts. Moreover, according to the common setting, we trimmed the PSG recordings to include a 30-minute window before and after the in-bed period.

3.1.2 SleepEDF-78

This dataset is an expanded version of the Sleep-EDF database, containing a total of 79 subjects ranging in age from 25 to 101 years, and comprising 153 complete overnight polysomnography (PSG) sleep recordings. The scoring standard and labeling methodology for the 30-s PSG epochs are the same as for SleepEDF-20. We evaluated our model using the Fpz-Cz EEG channel provided in the PSG recordings, which is sampled at a frequency of 100 Hz.

3.1.3 SHHS

The SHHS dataset is a comprehensive multicenter cohort study designed to investigate the impact of sleep apnea on cardiovascular diseases. Initiated by Quan et al. [1997] and further explored by Zhang et al. [2018], the SHHS dataset encompasses two rounds of PSG recordings, labeled as Visit 1 (SHHS-1) and Visit 2

(SHHS-2). These recordings include a variety of physiological signals such as two-channel EEGs, two-channel EOGs, a single-channel EMG, a single-channel ECG, and two-channel respiratory inductance plethysmography. Sleep stages within each epoch of the recordings were classified according to the *R&K* rule into categories such as Wake (W), REM, N1, N2, N3, N4, MOVEMENT, and UNKNOWN. For the purposes of our study on single-channel EEG classification, we employed the C4-A1 EEG channel data in the experiment.

3.2 Performance Metrics

To evaluate the classification performance of different models, precision (PRE), recall (RE), and per-class F1 score (F1) were adopted as the performance criteria of each sleep stage, and the overall accuracy (ACC), macro F1-score (MF1), and Cohen’s Kappa (κ) were adopted as the overall performance criteria. These metrics are calculated as follows:

$$PRE = \frac{TP}{TP + FP} \quad (13)$$

$$RE = \frac{TP}{TP + FN} \quad (14)$$

$$F1 = \frac{2 \times RE \times PRE}{RE + PRE} \quad (15)$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)$$

$$MF1 = \frac{1}{C} \sum_i^C F1_i \quad (17)$$

$$\kappa = \frac{ACC - P_e}{1 - P_e} = 1 - \frac{1 - ACC}{1 - P_e} \quad (18)$$

where TP , FP , TN , and FN represent the number of true positives, false positives, true negatives and false negatives samples, respectively. $F1_i$ denotes per-class F1-score of the i -th class. P_e represents the hypothetical probability of chance agreement.

3.3 Implementation Details

To evaluate the performance of WaveSleepNet model, we adopted a subject-wise k-fold cross-validation by dividing the subjects in each dataset (except SHHS) into k groups. As listed in Table 2, we set k as 10, 20 for the SleepEDF-20, SleepEDF-78, respectively. For instance, subject-wise 20-fold cross-validation on SleepEDF-20 dataset with 20 subjects is thus leave-one-subject-out (LOSO) cross-validation. For each round, one group of subjects was used for testing and the remaining k-1 groups were divided into training and validation datasets. The held-out validation set was used to track the validation loss for early termination in the training process. Unlike these datasets, the SHHS dataset was randomly divided in a ratio of 0.7 and 0.3 for training and testing, respectively. We further split 100 subjects of the training set into a validation set, as performed in Phan et al. [2021].

We implemented our model using Pytorch 2.1.0, on a server equipped with a NVIDIA GeForce RTX 4090 GPU card. We adopted Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to 5×10^{-4} , and the batch size is set to 64. The early-stop patience of the model training is set to 50. The number of EEG epochs for input signal L is set to 10. The loss hyperparameters λ_{class} , λ_{dist} , λ_{R1} , λ_{R2} , λ_{L1} are set to 50, 8, 9, 18 and 0.3, respectively.

3.4 Sleep Staging Performance

We present the sleep staging performance of WaveSleepNet through confusion matrixes depicted in Fig. 2. The confusion matrix is calculated by adding up all the scoring values of the testing data across all folds in each dataset. In the matrix, each row and column represent the classification results of sleep expert and WaveSleepNet, respectively. Notably, stage N1 achieves the lowest classification performance, where it is often misclassified as Wake and N2. We would further analyze the misclassification results in Section III.F.

Additionally, we compared the performance of WaveSleepNet model against various state-of-the-art approaches in terms of overall accuracy, macro F1-score, cohen kappa on three datasets. As shown in Table 3, the sleep stage performance of WaveSleepNet model is comparable with that of the corresponding baseline (non-interpretable or interpretable) models. However, the accuracy of WaveSleepNet could be further improved by adding the logits of several WaveSleepNet

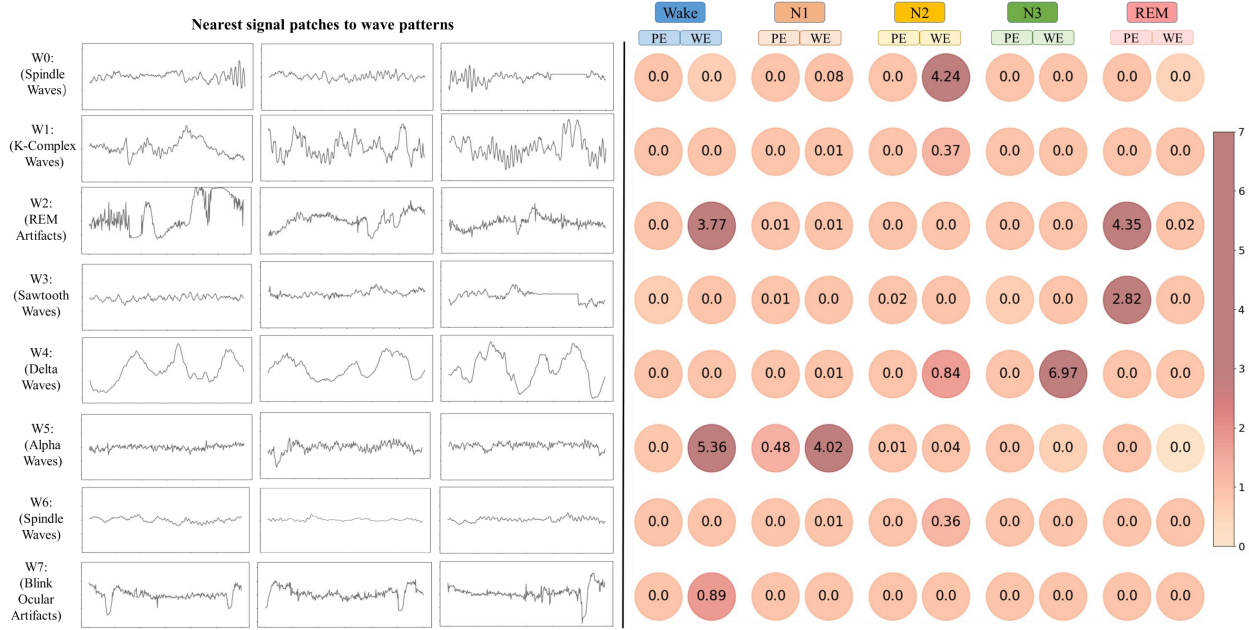


Figure 3: The reasoning process of the trained WaveSleepNet. The left part is the nearest signal patches in the training set to wave prototypes. The right part denotes the contribution of each wave pattern’s proportion Estimator (PE) and waveform Estimator (WE) to various sleep stages. The wave prototypes are learned from the SleepEDF-78 dataset.

models together. Each WaveSleepNet model can be understood as a "scoring sheet" for sleep stages, as depicted in Fig. 3. Therefore, aggregating the logits from several WaveSleepNet models effectively creates the combined scoring sheet. In the combined model, *Scores* with wave prototypes from all these models are all taken into account to get the final predicted sleep stage. The combined model will maintain the same interpretable characteristic, though there will be more wave prototypes for each class. In this experiment, we combined five WaveSleepNet model together to get the final sleep stage performance. The test MF1 of combined WaveSleepNet can reach 81.0, 78.5 and 80.7 for three datasets respectively, which is on par with the best-performing sleep stage model. Moreover, unlike other methods, WaveSleepNet has been simplified primarily to explore the feasibility of the proposed method, which only considers the local information in the input signals without taking into account the contextual information. This indicates that our model possesses potential for further improvement in staging performance, indicating a notable opportunity for advancement.

3.5 Case Study I: Reasoning Process of WaveSleepNet

In this section, we perform a case study to understand how our WaveSleepNet, which was trained on the SleepEDF-78 dataset, classifies sleep stages based on single-channel EEG signals. As shown in Fig. 3, the left part of the figure shows the nearest three signal segments in the entire training set to the wave prototypes encoded in the WaveSleepNet. It’s generally true that the nearest signal segments of a wave pattern all bear the same signal characteristic. After inspection by sleep experts, we identified the sleep characteristic waveforms that these signal segments belong to, that is, the sleep features corresponding to the wave prototypes learned by WaveSleepNet. In this case study, the wave prototypes learned by the network that can effectively stage sleep signals include spindle waves, K-complex waves, rapid eye movement (REM) artifacts, sawtooth waves, delta waves, alpha waves, and blink ocular artifacts. The right section of Fig. 3 illustrates the contribution of each wave pattern relative to various sleep stages. The contribution is calculated by multiplying the *Score* with the weights in the last FC layer. A higher contribution value not only indicates the presence or greater proportion of a sleep feature within the input signal but also carries a greater weight in determining the sleep stage. Notably, we found that while N1 stage is determined by W5 waveforms that are referenced as alpha waves, its *Score* value in W5 was lower than that of Wake stage signals. This is because its characteristic waveform does not exactly match alpha waves, but the LAMF waves that are similar to alpha waves but with a slower frequency of at least 1Hz in comparison.

Moreover, the reasoning process of WaveSleepNet for sleep staging can be outlined in Fig. 4. The results in Fig. 4 show that WaveSleepNet produces promising interpretable outcomes. Its decision-making process significantly overlaps with that of human manual annotators Berry et al. [2017]. Notably, we observed that WaveSleepNet seems to consider Delta waves (W4) when predicting the sleep N2 stage, a deviation from the staging criteria outlined in the AASM

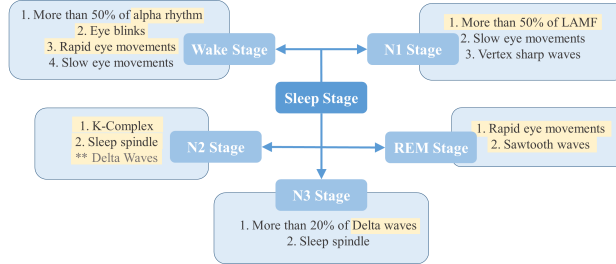


Figure 4: Comparison of the Classification Logic in WaveSleepNet Model versus the AASM Manual Guidelines for Sleep Staging. The numbered annotations indicate the classification logic according to the AASM Manual, light yellow boxes highlight the classification logic implemented by the WaveSleepNet model. ** denotes the classification rules unique to WaveSleepNet and not illustrated in AASM guidelines.

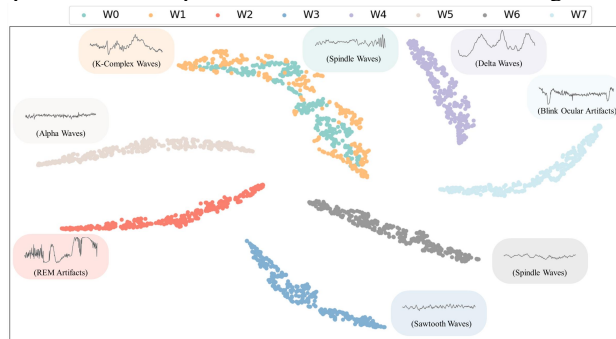


Figure 5: Learned wave prototypes of WaveSleepNet trained in the SleepEDF-78 dataset

manual. This is because K-complex waves are considered Delta waves if they fit the definition of slow activity, which is mentioned in the AASM.

Additionally, we examine *Scores*, as the latent features of each wave pattern, via t-SNE projections where proximity in 2-D space suggests that points are “close” in distance in the original latent space, shown as Fig. 5. The *Scores* of each wave pattern are extracted from a series of EEG signals which form the closest patch-wave pattern pair in the latent space with the corresponding wave pattern. This shows that the learned wave prototypes are easily separated and well-structured. However, it was observed that W0 (Spindle Waves) and W1 (K-Complex Waves) are proximal yet not entirely congruent within the 2-D projection space. This phenomenon can be attributed to the consistent interaction between K-Complex waves and sleep spindles, as spindles frequently coincide with K-Complex waves KOKKINOS and KOSTOPOULOS [2011]. The two wave prototypes frequently exhibit overlapping characteristics, leading to their relative proximity within the feature space. Furthermore, we found that despite W0 and W6 are all represent spindle characteristic waves, they do not exhibit proximity in the 2-D space. This indicates that, although they share similar frequencies, they represent distinct categories of spindle waves. Notably, W6, in comparison to W0, demonstrated a lower amplitude, suggesting that W6 might be the spindle wave interferences from other EEG channels.

3.6 Case Study II: Misclassification Analysis of WaveSleepNet

As an explainable model, WaveSleepNet not only provides a reliable reasoning process for accurate predictions, but it also explains its misclassifications. In this section, we examine the reasons behind WaveSleepNet’s mispredictions across various sleep stages. We achieve this by extracting the *Score* values of wave prototypes. As illustrated in Fig. 6, we conducted a comparison between the *Score* values of accurately classified instances and that of mistakenly classified instances. From Fig. 6, we can see that WaveSleepNet determines sleep stages by identifying specific waveforms in EEG signals. Therefore, if the characteristic waveforms of a certain sleep stage signals weaken or are not significant enough, they may be incorrectly classified into other stages. We can analyze each stage one by one:

1) Wake Stage: The waveforms used to determine the Wake stage are Blink Ocular Artifacts (W7), REM artifacts (W2), and alpha waves (W5), while the waveforms used to determine the N1 stage are alpha waves (W5). When the Blink ocular artifacts and REM artifacts in the input signals of the Wake stage are not pronounced, the signal may be misidentified as N1 stage, as indicated by the results of the confusion matrix.

2) N1 Stage: The classification of N1 sleep stage depends on the *Score* value of alpha waves (W5), while the classification of N2 sleep stage mainly depends on the *Score* value of spindle waves (W0). When the frequency of

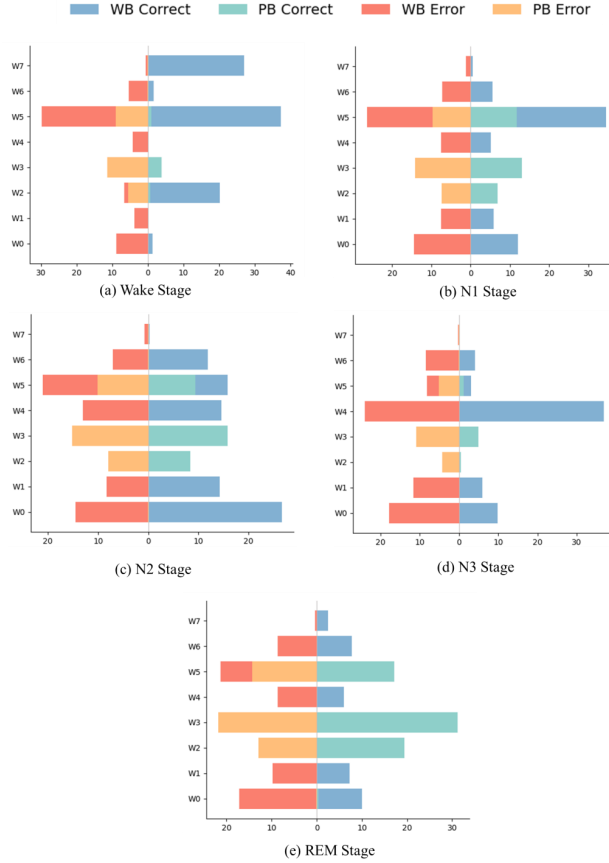


Figure 6: The comparison of *Score* values for accurately classified instances and mistakenly classified instances. *WE Correct* and *WE Error* represent the *WScore* for correctly and incorrectly classified instances, respectively, while *PE Correct* and *PE Error* denote the *PScore* for accurately and mistakenly classified instances, respectively.

alpha waves appearing in stage N1 signals is similar to that of spindle waves, the W0 *Score* value will increase and thus the signal will be mistakenly classified as stage N2.

3) N2 Stage: Spindle waves (12-14 Hz) and alpha waves (8-13 Hz) are similar in frequency, leading to some spindle waves being incorrectly identified as alpha waves. If there is no K-complex wave (W1) present in the signal at this time, this may result in the input signal being erroneously classified as N1 sleep stage. In the N2 stage, V-waves are similar to REM artifacts (sharply contoured waves with a duration <0.5 seconds), thus some V-waves may be mistakenly identified as REM artifacts (W2). Furthermore, both N2 and REM stages have sawtooth waves (W3). These lead to some signals being misclassified as REM stage.

4) N3 Stage: The waveforms used to determine the N3 stage is Delta Waves (W4). When there are prominent spindle waves (W0, W6) in the input signal and Delta waves (W4) are not significant, the signal may be mistakenly identified as N2 stage.

5) REM Stage: The similarity between low-frequency alpha waves during REM stage and alpha waves (W5) causes a high W5 *Score* value, making it easy for the input signal to be misclassified as N1 sleep. Additionally, the low-frequency alpha waves are similar to spindle waves (W0, W6), and if sawtooth waves (W3) are present in the signal simultaneously, it may be misclassified as N2 stage.

4 Discussion

In this paper, we propose an interpretable sleep staging model, WaveSleepNet, that explains decision-making process in a manner that aligns with the cognitive processes of sleep experts. By conducting experiments on different datasets, we demonstrated its comparable sleep staging performance with the existing state-of-the-art methods. Additionally, a case study was conducted to discuss the reasoning process of WaveSleepNet in the task of sleep staging. The results

demonstrate that it yields highly interpretable outcomes, which align closely with the logic used by human manual annotators. Beyond this, the study also delves into the reasons behind classification errors made by WaveSleepNet during the sleep staging process. This investigation offers valuable insights for future improvements to sleep staging methods.

From Section III.E, we discovered that ocular artifacts play a significant role in the model’s classification process. This finding is not limited to the specific case study, it has been observed across all datasets. This indicates that, in the task of sleep staging using single-channel EEG signals, ocular movement features are still encapsulated within the single-channel EEG signals in the form of artifacts. This might explain why many single-channel sleep staging models perform comparably to multi-channel modelsPhan et al. [2021]Phan et al. [2022]Zhao et al. [2022]Eldele et al. [2021]Wei et al. [2024]. It also clarifies why models utilizing the Fpz-Cz channel often achieve better staging outcomes than those using Pz-Oz channelEldele et al. [2021]Dutt et al. [2022]Huang et al. [2022], as the Fpz-Cz channel is more susceptible to ocular interference. Nevertheless, ocular artifacts can also disrupt the staging process. As illustrated in section III.F, the similarity between V-waves and REM ocular artifacts can lead to the misclassification of N2 stage signals as REM stage. Furthermore, accurate identification of EOG features typically requires the collaborative judgment of signals from two EOG channels, and relying solely on single-channel ocular artifacts could result in misinterpretation of these features. Therefore, future work should aim to accurately leverage ocular artifacts in single-channel EEG sleep staging without introducing interference.

Through the analysis of misclassifications by WaveSleepNet in section III.F, we noted that the most significant issue arises from the confusion between the alpha waves (8-13Hz) during the Wake (W) stage, the attenuated alpha rhythm waves in the N1 stage, the spindle waves in the N2 stage (11-16Hz), and the low-frequency alpha waves during the REM stage. This confusion is primarily due to the overlapping frequencies of these waves, making it challenging to accurately distinguish them within the signal. Misidentification can lead to incorrect sleep staging. This problem is not unique to WaveSleepNet; it is a common challenge faced by all neural network models when processing single-channel EEG signals. In future work, enhancing the separation of these waveforms through advanced signal processing or detection techniques could potentially improve the accuracy of sleep stage classification models.

The annotators designated by sleep experts should not be regarded as the absolute gold standard, as there is a substantial inter-expert variability of 65-85% Norman et al. [2000]. Therefore, investigating the reasoning process of the model in sleep staging may be more valuable than determining the accuracy of the staging itself. The international consensus criteria are the experiences or theories derived from sleep researchers’ extensive work over the years, and they are constantly being updated and revised. We don’t have to make them an absolute ground truth to be essentialized Muto and Berthomier [2023]. Our WaveSleepNet model, as a data-driven method, identifies discernible features and rules from a large number of sleep datasets, facilitating sleep stage differentiation. This, in turn, could be extremely interesting in revealing new sleep signatures, but their relevance still need to be validated by human experts.

Our model has been simplified primarily to mainly explore the feasibility of the proposed method. Therefore, further adjustments to the base model could be explored in future work to enhance the model’s interpretability and sleep staging performance.

Firstly, extending the ability of the network would unquestionably enhance the classification performance. In this study, a convolutional model was utilized for local feature extraction from input signals, but the contextual relationships between epochs were not considered. Future research should explore integrating memory-augmented neural networks like RNNs and LSTMs into the WaveSleepNet to maintain interpretability while learning epoch contextual relationships, thereby enhancing sleep staging accuracy. Additionally, our Feature Extraction Network in WaveSleepNet adapted a simplified CNN model inspired by AttnsleepEldele et al. [2021]. Future efforts should aim to investigate more advanced network architectures for improved extraction of sleep waveform features.

Second, the WaveSleepNet achieves sleep staging by combining waveform and proportion Estimators, mirroring the methodology of sleep experts. However, the decision-making analysis of the trained model reveals discrepancies with the AASM guidelines in the proportion-based reasoning process. The divergence may be attributed to the neural network’s extensive feature space, making it challenging for the network to identify the global optimum solution that aligns perfectly with human expert annotators. To make the staging process more aligned with human expert logic, future research could consider further exploration of integrating expert knowledge into the model.

5 Conclusion

Interpretability and accuracy represent a trade-off in machine learning modeling. Especially in the age of deep learning, many accurate models are black-box models that lack transparency in explaining their predictions, limiting their clinical acceptance. In this paper, we present a deep learning model WaveSleepNet that provides accurate predictions as well as

offers interpretability by revealing more details of its inner workings. The model was developed and evaluated using the SleepEDF-20, SleepEDF-78, and SHHS datasets, demonstrating performance in sleep staging that is comparable to state-of-the-art models. Moreover, a combined larger WaveSleepNet model outperforms state-of-the-art methods under various evaluation metrics. A qualitative case study was conducted to provide explanations through an interpretable decision-making process that closely align with AASM. Furthermore, we innovatively conducted an analysis of the model’s misclassifications, systematically explaining the reasons for each misidentified stage, which helps to better understand and address classification degradation. The added interpretability in our model allows the experts to have direct access to the physiological meaning of the criteria within the model, indicating a possibility for human-machine collaboration for sleep stage classification.

References

- Eleonora Tobaldini, Elisa M. Fiorelli, Monica Solbiati, Giorgio Costantino, Lino Nobili, and Nicola Montano. Short sleep duration and cardiometabolic risk: from pathophysiology to clinical evidence. *Nature Reviews Cardiology*, 16(4):213–224, November 2018. doi:10.1038/s41569-018-0109-6. URL <https://doi.org/10.1038/s41569-018-0109-6>.
- Richard B. Berry, Rita Brooks, Charlene Gamaldo, Susan M. Harding, Robin M. Lloyd, Stuart F. Quan, Matthew T. Troester, and Bradley V. Vaughn. Aasm scoring manual updates for 2017 (version 2.4). *Journal of Clinical Sleep Medicine*, 13(05):665–666, May 2017. ISSN 1550-9397. doi:10.5664/jcsm.6576. URL <http://dx.doi.org/10.5664/jcsm.6576>.
- Atul Malhotra and Joseph Loscalzo. Sleep and cardiovascular disease: An overview. *Progress in Cardiovascular Diseases*, 51(4):279–284, January 2009. ISSN 0033-0620. doi:10.1016/j.pcad.2008.10.004. URL <http://dx.doi.org/10.1016/j.pcad.2008.10.004>.
- Richard S. Rosenberg and Steven Van Hout. The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring. *Journal of Clinical Sleep Medicine*, 09(01):81–87, January 2013. ISSN 1550-9397. doi:10.5664/jcsm.2350. URL <http://dx.doi.org/10.5664/jcsm.2350>.
- Vincenzo Muto and Christian Berthomier. Looking for a balance between visual and automatic sleep scoring. *npj Digital Medicine*, 6(1), September 2023. doi:10.1038/s41746-023-00915-7. URL <https://doi.org/10.1038/s41746-023-00915-7>.
- Pejman Memar and Farhad Faradji. A novel multi-class eeg-based sleep stage classification system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(1):84–95, January 2018. ISSN 1558-0210. doi:10.1109/tnsre.2017.2776149. URL <http://dx.doi.org/10.1109/tnsre.2017.2776149>.
- Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, 42:107–114, April 2018. ISSN 1746-8094. doi:10.1016/j.bspc.2017.12.001. URL <http://dx.doi.org/10.1016/j.bspc.2017.12.001>.
- Junming Zhang and Yan Wu. Competition convolutional neural network for sleep stage classification. *Biomedical Signal Processing and Control*, 64:102318, February 2021. ISSN 1746-8094. doi:10.1016/j.bspc.2020.102318. URL <http://dx.doi.org/10.1016/j.bspc.2020.102318>.
- Pankaj Jadhav and Siddhartha Mukhopadhyay. Automated sleep stage scoring using time-frequency spectra convolution neural network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–9, 2022. ISSN 1557-9662. doi:10.1109/tim.2022.3177747. URL <http://dx.doi.org/10.1109/TIM.2022.3177747>.
- Michael Sokolovsky, Francisco Guerrero, Sarun Paisarnsrisomsuk, Carolina Ruiz, and Sergio A. Alvarez. Deep learning for automated feature discovery and classification of sleep stages. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):1835–1845, November 2020. ISSN 2374-0043. doi:10.1109/tcbb.2019.2912955. URL <http://dx.doi.org/10.1109/TCBB.2019.2912955>.
- Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chen, and Maarten De Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, March 2019. ISSN 1558-0210. doi:10.1109/tnsre.2019.2896659. URL <http://dx.doi.org/10.1109/tnsre.2019.2896659>.
- Chenglu Sun, Chen Chen, Wei Li, Jiahao Fan, and Wei Chen. A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1351–1366, May 2020. ISSN 2168-2208. doi:10.1109/jbhi.2019.2937558. URL <http://dx.doi.org/10.1109/jbhi.2019.2937558>.

- Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, Nov 2017. ISSN 1534-4320. doi:10.1109/TNSRE.2017.2721116.
- Wei Qu, Zhiyong Wang, Hong Hong, Zheru Chi, David Dagan Feng, Ron Grunstein, and Christopher Gordon. A residual based attention model for eeg based sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2833–2843, October 2020. ISSN 2168-2208. doi:10.1109/jbhi.2020.2978004. URL <http://dx.doi.org/10.1109/JBHI.2020.2978004>.
- Huy Phan, Kaare Mikkelsen, Oliver Y. Chen, Philipp Koch, Alfred Mertins, and Maarten De Vos. SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, August 2022. doi:10.1109/tbme.2022.3147187. URL <https://doi.org/10.1109/tbme.2022.3147187>.
- Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021. ISSN 1558-0210. doi:10.1109/tnsre.2021.3076234. URL <http://dx.doi.org/10.1109/TNSRE.2021.3076234>.
- Ahnaf Rashik Hassan, Syed Khairul Bashar, and Mohammed Imamul Hassan Bhuiyan. Automatic classification of sleep stages from single-channel electroencephalogram. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6, 2015. doi:10.1109/INDICON.2015.7443756.
- Antti Saastamoinen, Eero Huupponen, Alpo Värri, Joel Hasan, and Sari-Leena Himanen. Computer program for automated sleep depth estimation. *Computer Methods and Programs in Biomedicine*, 82(1):58–66, April 2006. ISSN 0169-2607. doi:10.1016/j.cmpb.2006.02.003. URL <http://dx.doi.org/10.1016/j.cmpb.2006.02.003>.
- Irfan Al-Hussaini, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Sleeper: interpretable sleep staging via prototypes from expert rules. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 721–739. PMLR, 09–10 Aug 2019. URL <https://proceedings.mlr.press/v106/al-hussaini19a.html>.
- Micheal Dutt, Surender Redhu, Morten Goodwin, and Christian W. Omlin. SleepXAI: An explainable deep learning approach for multi-class sleep stage identification. *Applied Intelligence*, 53(13):16830–16843, December 2022. doi:10.1007/s10489-022-04357-8. URL <https://doi.org/10.1007/s10489-022-04357-8>.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017. doi:10.1109/iccv.2017.74. URL <https://doi.org/10.1109/2Ficcv.2017.74>.
- Jathurshan Pradeepkumar, Mithunja Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L. Kappel, Anjula C. De Silva, and Chamira U. S. Edussooriya. Towards interpretable sleep stage classification using cross-modal transformers, 2022. URL <https://arxiv.org/abs/2208.06991>.
- Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4), December 2011. ISSN 1932-6157. doi:10.1214/11-aos495. URL <http://dx.doi.org/10.1214/11-AOS495>.
- Carey E. Priebe, David J. Marchette, Jason G. DeVinney, and Diego A. Socolinsky. Classification using class cover catch digraphs. *Journal of Classification*, 20(1):3–23, May 2003. ISSN 1432-1343. doi:10.1007/s00357-003-0003-7. URL <http://dx.doi.org/10.1007/s00357-003-0003-7>.
- Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *This looks like that: deep learning for interpretable image recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019a.
- Janet L. Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1):3–34, 1992. ISSN 1573-7462. doi:10.1007/bf00155578. URL <http://dx.doi.org/10.1007/BF00155578>.
- Hangyu Zhu, Wei Zhou, Cong Fu, Yonglin Wu, Ning Shen, Feng Shu, Huan Yu, Wei Chen, and Chen Chen. Masksleepnet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2353–2364, May 2023. ISSN 2168-2208. doi:10.1109/jbhi.2023.3253728. URL <http://dx.doi.org/10.1109/JBHI.2023.3253728>.
- Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017. URL <https://arxiv.org/abs/1709.01507>.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL <https://arxiv.org/abs/1607.06450>.

- François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. doi:10.1109/CVPR.2017.195.
- Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. *This Looks like That: Deep Learning for Interpretable Image Recognition*. Curran Associates Inc., Red Hook, NY, USA, 2019b.
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), June 2000. ISSN 1524-4539. doi:10.1161/01.cir.101.23.e215. URL <http://dx.doi.org/10.1161/01.cir.101.23.e215>.
- B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Obery. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9): 1185–1194, 2000. doi:10.1109/10.867928. URL <https://doi.org/10.1109/10.867928>.
- Stuart F. Quan, Barbara V. Howard, Conrad Iber, James P. Kiley, F. Javier Nieto, George T. O'Connor, David M. Rapoport, Susan Redline, John Robbins, Jonathan M. Samet, and Patricia W. Wahl. The Sleep Heart Health Study: Design, Rationale, and Methods. *Sleep*, 20(12):1077–1085, 12 1997. ISSN 0161-8105. doi:10.1093/sleep/20.12.1077. URL <https://doi.org/10.1093/sleep/20.12.1077>.
- Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 05 2018. ISSN 1527-974X. doi:10.1093/jamia/ocy064. URL <https://doi.org/10.1093/jamia/ocy064>.
- Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.
- Akara Supratak and Yike Guo. Tinsleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2020. doi:10.1109/embc44109.2020.9176741. URL <http://dx.doi.org/10.1109/EMBC44109.2020.9176741>.
- Sajad Mousavi, Fatemeh Afghah, and U. Rajendra Acharya. Sleeppegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS ONE*, 14(5):e0216456, May 2019. ISSN 1932-6203. doi:10.1371/journal.pone.0216456. URL <http://dx.doi.org/10.1371/journal.pone.0216456>.
- Caihong Zhao, Jinbao Li, and Yahong Guo. Sleepcontextnet: A temporal context network for automatic sleep staging based single-channel eeg. *Computer Methods and Programs in Biomedicine*, 220:106806, June 2022. ISSN 0169-2607. doi:10.1016/j.cmpb.2022.106806. URL <http://dx.doi.org/10.1016/j.cmpb.2022.106806>.
- Henri Korkalainen, Timo Leppanen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac Oluwaseun Afara, Sami Myllymaa, and Juha Toyra. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE Journal of Biomedical and Health Informatics*, page 1–1, 2019. ISSN 2168-2208. doi:10.1109/jbhi.2019.2951346. URL <http://dx.doi.org/10.1109/JBHI.2019.2951346>.
- Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Hogeon Seo, Seunghyeok Back, Seongju Lee, Deokhwan Park, Tae Kim, and Kyoobin Lee. Intra- and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. *Biomedical Signal Processing and Control*, 61:102037, 2020. ISSN 1746-8094. doi:<https://doi.org/10.1016/j.bspc.2020.102037>. URL <https://www.sciencedirect.com/science/article/pii/S1746809420301932>.
- Yu Wei, Yongpeng Zhu, Yihan Zhou, Xiaokang Yu, and Yuxi Luo. Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel eeg. *IEEE Journal of Biomedical and Health Informatics*, 28(2):801–811, February 2024. ISSN 2168-2208. doi:10.1109/jbhi.2023.3332503. URL <http://dx.doi.org/10.1109/JBHI.2023.3332503>.
- Jing Huang, Lifeng Ren, Xiaokang Zhou, and Ke Yan. An improved neural network based on senet for sleep stage classification. *IEEE Journal of Biomedical and Health Informatics*, 26(10):4948–4956, October 2022. ISSN 2168-2208. doi:10.1109/jbhi.2022.3157262. URL <http://dx.doi.org/10.1109/JBHI.2022.3157262>.

Robert G Norman, Ivan Pal, Chip Stewart, Joyce A. Walsleben, and David M Rapoport. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, 23(7):1–8, October 2000. ISSN 0161-8105. doi:10.1093/sleep/23.7.1e. URL <http://dx.doi.org/10.1093/sleep/23.7.1e>.

VASILEIOS KOKKINOS and GEORGE K. KOSTOPOULOS. Human non-rapid eye movement stage ii sleep spindles are blocked upon spontaneous k-complex coincidence and resume as higher frequency spindles afterwards. *Journal of Sleep Research*, 20(1pt1):57–72, February 2011. ISSN 1365-2869. doi:10.1111/j.1365-2869.2010.00830.x. URL <http://dx.doi.org/10.1111/j.1365-2869.2010.00830.x>.

Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*, page 818–833. Springer International Publishing, 2014. ISBN 9783319105901. doi:10.1007/978-3-319-10590-1_53. URL http://dx.doi.org/10.1007/978-3-319-10590-1_53.