JYX
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Xu, Xu; Cong, Fengyu; Chen, Yongyong; Chen, Junxin

Title: Sleep Stage Classification With Multi-Modal Fusion and Denoising Diffusion Model

Year: 2024

Version: Accepted version (Final draft)

Copyright: © IEEE

Rights: In Copyright

Rights url: http://rightsstatements.org/page/InC/1.0/?language=en

Please cite the original version:

Xu, X., Cong, F., Chen, Y., & Chen, J. (2024). Sleep Stage Classification With Multi-Modal Fusion and Denoising Diffusion Model. IEEE Journal of Biomedical and Health Informatics, Early Access. https://doi.org/10.1109/jbhi.2024.3422472

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

Sleep Stage Classification with Multi-Modal Fusion and Denoising Diffusion Model

Xu Xu, *Graduate Student Member, IEEE*, Fengyu Cong, Yongyong Chen, *Member, IEEE*, and Junxin Chen, *Senior Member, IEEE*

***Abstract*— Sleep stage classification plays a crucial role in sleep quality assessment and sleep disorder prevention. Nowadays, many studies have developed algorithms for this purpose, but they still face two challenges. The first is noise in physiological signals from various devices. The second challenge is that most studies simply concatenate multi-modal features without considering their correlations. To this end, we propose a framework, namely Diff-SleepNet, to efficiently classify sleep stages from multi-modal input. This framework begins with a diffusion model with peak signal-to-noise ratio (PNSR) loss function that adaptively filters noise. The filtered signals are then transformed into a multi-view spectrum through data pre-processing. These spectra are processed by a transformer-based backbone to extract multi-modal features. The production is fed into the following multi-scale attention module for robust feature fusion. The sleep stage category is finally determined by a fully connected layer. Our framework is trained and validated on three typical datasets, i.e., SHHS, Sleep-EDF-SC, and Sleep-EDF-X. Experimental results demonstrate that it is effective and has advantages over other peer methods.**

***Index Terms*— Sleep Stage Classification, Deep Learning, Multi-Modal Fusion, Diffusion Model, Attention Mechanism**

I. Introduction

SLEEP is an essential physiological phenomenon associated with human health [1]. According to a population-based survey, about 30% of adults have sleep disorders [2]. Therefore, sleep-related illness has become a prominent issue threatening global public health. Recent research has determined that sleep stage classification is an effective approach for diagnosing sleep disorders [3]. Normally, sleep physicians manually classify sleep stages as five stages, i.e., wake (W), non-rapid eye movement REM (NREM) (N1, N2, N3), and REM [4]. However, manual classification is time-consuming and highly subjective. Recently, many studies have focused on automatic sleep stage classification for reducing diagnosis time and improving classification performance [5], [6]. They can be divided into two categories.

The first category generally extracts features using signal processing algorithms and then utilizes machine learning for decision-making. For instance, Heyat *et al.* [5] extracted power spectral density features from electroencephalogram (EEG), after which these features were input into a decision tree for sleep staging. Fraiwan *et al.* [7] extracted time-frequency entropy features and used linear discriminant analysis to classify sleep stages. Differently, Csen *et al.* [8] combined feature selection algorithm with machine learning-based classification algorithm to filter irrelevant and redundant features from EEG signals, which could reduce the time consumed for sleep staging. An *et al.* [9] improved the classification performance of stage N1 and REM by fusing multiply support vector machine (SVM) models. However, traditional methods rely on prior knowledge and have limited performance improvement.

The other type is mainly through deep learning to automatically extract features and classify sleep stages. It mainly includes single-modal models and multi-modal ones. Among them, EEG, electrooculogram (EOG), and electromyography (EMG) are considered as inputs for the first network [10], [11]. Typically, Michielli *et al.* [6] proposed a cascaded framework based on the long short-term memory (LSTM) blocks with EEG signals, which could improve the classification performance of stage N1. Ebrahimi *et al.* [12] combined temporal convolutional neural network (TCNN) with conditional random field (CRF) to extract low-level feature vectors from signal-modal EEG signal, which was able to minimize engineering bias. On the other hand, multi-modal input is able to provide more information for the model. For instance, Abdulla *et al.* [13] proposed a multi-channel model with a binary pattern algorithm and texture color analysis for sleep staging. In addition, Dai *et al.* [14] developed a network with multi-modal input and a Transformer-based backbone to extract and fuse features for sleep stage classification.

Previous works have demonstrated that multi-modal input can improve the performance of sleep staging [13], [14].

This work is funded by the National Natural Science Foundation of China (No. 62171114), the Fundamental Research Funds for the Central Universities (No. DUT22RC(3)099), and Xiaomi Young Talents Program.

Xu Xu is with the School of Software, Dalian University of Technology, Dalian 116621, China; also with the School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China.

Fengyu Cong is with the School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian, 116024, China; also with the Faculty of Information Technology, University of Jyväskylä, Jyväskylä 40014, Finland, the School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, and the Key Laboratory of Integrated Circuit and Biomedical Electronic System, Dalian University of Technology, Dalian 116024, China.

Yongyong Chen is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China.

Junxin Chen is with the School of Software, Dalian University of Technology, Dalian 116621, China.

Junxin Chen is the corresponding author (junxinchen@ieee.org).

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

However, there are two challenges, i.e., data noise from acquisition equipment and multi-modal fusion without feature correlation. Recently, the development of denoising model and attention-based multi-modal fusion have brought new ideas to solve above challenges. First, denoising models have been applied to various physiological signal with good results. For example, Singh *et al.* [15] developed a denoising model based on the generative adversarial networks (GAN) for EEG signals. Zeng *et al.* [16] proposed a network to remove noise and extract spatial features from EEG signals. Meanwhile, peak signal-to-noise ratio (PSNR) is a typical metric used in signal denoising. Recent research has determined that it improves model performance better than mean square error (MSE) [17]. In addition, many attention mechanisms are designed to simulate the brain for multi-modal feature fusion. Typically, Lebal *et al.* [18] proposed an attention-based model to fuse multi-modal features for epilepsy detection. Li *et al.* [19] found that neurons in the same region have different receptive field sizes, which can collect multi-scale information.

Motivated by the above discussion, we develop a framework with denoising modal and multi-modal fusion, namely Diff-SleepNet, for automatic sleep stage classification. Specifically, an improved diffusion model is designed to reduce noise interference. Its general loss, i.e., the MSE loss, is replaced by the PSNR loss during the training stage. On the other hand, we develop a module based on multi-scale convolution and self-attention to efficiently fuse multi-modal features. First, the multi-modal physiological signals, i.e., EOG and EEG, are denoised by the improved diffusion model. A short-time Fourier transform (STFT) is then introduced to transfer these signals into spectrums. Next, these spectrograms are inputted into the Transformer-based backbone to extract high-level features, which are then fused by an attention mechanism. Finally, the sleep stage is generated by the classification branch. To evaluate our Diff-SleepNet, we select three high-quality datasets, i.e., SHHS [20], Sleep-EDF-SC [21], and Sleep-EDF-X [21]. In addition, we choose two typical denoising datasets, i.e., EEGDenoise [22] and MIT-BIH [23] to train and validate our denoising model. The results well demonstrate the effectiveness and robustness of our Diff-SleepNet, which also has advantages over state-of-the-art (SOTA) algorithms. Our contributions are summarized as follows.

1) A novel approach is developed with a denoising diffusion model and multi-modal input for automatic sleep stage classification. Experimental results demonstrate that it is effective and has advantages over peer SOTA algorithms. Our codes are available at GitHub[1].
2) A diffusion model with transfer learning and PSNR loss has been found effective in denoising multi-modal physiological signals and further improving the classification performance.
3) A multi-scale attention mechanism has been validated that is able to effectively fuse the multi-modal features and well empower sleep stage classification.

This paper is structured as follows. First, Section II reviews related works about sleep stage classification methods. Our modules and whole framework architecture are introduced in Section III, and detailed information about implementation is in Section IV. Then, performance evaluation and discussions are presented in Section V. Finally, the study and potential work are summarized in Section VI.

[1]https://github.com/Odysseus0816/Diff-SleepNet

II. Related Work

According to the input form, existing automated sleep staging studies can be divided into two categories, i.e., single-view input and multi-view input. These typical works are summarized and listed in Table I, which will be introduced in the following paragraphs.

A. Single-View Input

The original 1D signal and the 2D time-frequency image represent different distributions of the same data. Researchers have developed various models for analyzing data representations. Recently, they have used short-context inputs, such as adjacent epochs, to predict the sleep stage corresponding to the target epoch. However, the short context input fails to capture long-term dependencies between sleep data epochs. To address this limitation, Supratak *et al.* [40] proposed DeepSleepNet based on a two-stage training approach. A multi-scale CNN was first used to extract temporal and spectral features. Then, a LSTM was employed to learn the transition rules between adjacent epochs of sleep stages. Afterward, Supratak *et al.* [29] proposed TinySleepNet, which used a unidirectional LSTM and did not require a large buffer of epochs for backward processing. As a result, it reduced the computational resources needed for sequence learning by approximately half. Zhu *et al.* [41] introduced a multi-branch CNN model called MaskSleepNet, which considers multi-modal signals and incorporates a self-attention mechanism for model regularization through channel fusion. Furthermore, Phan *et al.* [32] proposed SleepTransformer based on the self-attention mechanism, eliminating the need for RNN and CNN.

B. Multi-View Input

Compared to single-view input, utilizing information from multiple views enhances the recognition performance of the model. A multi-view input model comprises multiple base networks, each dedicated to processing a specific view. It is important for the network structures of these base networks to exhibit diversity, allowing for tailored handling of each view. In the epoch-level, Phan *et al.* [30] designed fully convolutional network and bi-directional LSTM to extract feature from 1D signals and 2D spectrograms, respectively. Furthermore, they combined features from both base networks using a feature-weighted fusion approach to maximize integration and improve staging accuracy. Similarly, Siddharth *et al.* [31] proposed SpectroTemporalNet based on TemporalNet and SpectralNet. Among them, TemporalNet extracted features from 1D signals, while SpectralNet handled spectrum image feature extraction. These multi-model features were conditionally weighted to achieve better predictive performance. Li *et al.* [33] developed a multi-view fusion model, i.e., MVF-SleepNet. MVF-SleepNet was based on multi-modal signals

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

TABLE I: An overview of typical methods for sleep stage classification

Author	Year	Abbreviation	Signal	Technology	Dataset
Perslev *et al.* [24]	2019	U-Time	EEG	U-Net, Time Series Segmentation	Sleep-EDF [21], Physionet 2018 [25], ISRUC [26], CAP [27], SVUH-UCD [28], DCSM**
Supratak *et al.* [29]	2020	TinySleepNet	EEG	Data Augmentation, Lightweight	Sleep-EDF [21], MASS*
Phan *et al.* [30]	2021	XSleepNet	EEG	Multi-view, Gradient Blending	SHHS [20], Sleep-EDF [21], MASS*
Sanghavi *et al.* [31]	2021	STemporalNet	EEG	WaveNet, Frequency Spectrum	Sleep-EDF [21]
Phan *et al.* [32]	2022	SleepTransformer	EEG	Transformer, Entropy	SHHS [20], Sleep-EDF [21]
Li *et al.* [33]	2022	MVF-SleepNet	EOG, EEG, EMG	Time-frequency, Graph-learned graphs	ISRUC-S1 [26], ISRUC-S3 [26]
Yao *et al.* [34]	2023	VSTTN	EOG, EEG	Spatial-Temporal, Transfer Learning	PSG-35 [35], ISRUC-S1 [26], ISRUC-S3 [26]
Ji *et al.* [36]	2023	3DSleepNet	EOG, EEG, EMG	3D-CNN, 2D-CNN, Attention	ISRUC-S1 [26], ISRUC-S3 [26]
Kong *et al.* [37]	2023	N/A	EEG	Neural Architecture Search	SHHS [20], Sleep-EDF [21]
Dai *et al.* [14]	2023	MCSleepNet	EOG, EEG, EMG	Transformer, Multi-channel	SHHS [20], Sleep-EDF [21]
Ji *et al.* [38]	2024	MixSleepNet	EOG, EEG, EMG	3D Convolution, Graph Convolution	ISRUC-S1 [26], ISRUC-S3 [26]
Chao *et al.* [39]	2024	N/A	EOG, EEG, EMG	Directed Transfer Function, Residual Attention	Sleep-EDF [21]

* http://ceams-carsm.ca/mass/
** https://www.rigshospitalet.dk/english/research/units-and-groups/Pages/danish-center-for-sleep-medicine-.aspx

(i.e., EEG, EOG, and EMG) and multi-view input (i.e., time-frequency images and graph-learned graphs). They utilized VGG-16 and GRU to learn spectro-temporal information from sequentially timed images, respectively. In addition, Ji *et al.* [36] developed 3D-CNN and 2D-CNN with attention to extract spatial-temporal and spatial-spectral features, respectively.

III. Proposed Method

A. Framework Architecture

Formally, the multi-channel input is expressed as $X_i, 1 \leqslant i \leqslant n$, where n is the number of channels. This work designs a framework F to classify sleep stages C based on these inputs, which can be described as $C = F(X_1, \ldots, X_n)$. Its architecture is illustrated in Fig. 1, which mainly includes two parts, i.e., data pre-processing and classification model.

1) First, the multi-channel signals are denoised by the diffusion model.
2) The STFT is then introduced to transfer and normalize these signals into spectrograms.
3) Next, these spectrograms are inputted into the Transformer-based backbone to extract high-level features, whose products are fused by a custom attention.
4) Finally, the sleep stage is predicted through the classification branch.

These components are described in the following paragraphs.

B. Data Pre-Processing

1) Denoising Model: During physiological signal acquisition, it is often affected by environmental noise, which can interfere with classification results. Therefore, a denoising model is developed based on DDPM, which is inspired by non-equilibrium thermodynamics [42]. Its backbone is U-Net and consists of two stages, i.e., forward process (adding noise) and reverse process (filtering noise). For the first stage, the $x_0 \sim q(x_0)$ are sampled from the input data. Then, the Gaussian distribution noise is added to these samples. It can be described as

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I\right),$$

where T, I, and $\beta_t \in (0,1)$ are the diffusion steps, unit matrix, and the level variance of noise, respectively. In addition, $N(x; \mu, \delta)$ represents a normal distribution that generates x with the mean μ and covariance δ. The samples at each step are related only to the samples at time $t-1$. Thus, when t follows a uniform distribution yields

$$(x_t|x_0) = N\left(x_t; \sqrt{\widehat{\beta}_t} \cdot x_0, \left(1-\widehat{\beta}_t\right) \cdot I\right),$$

where $\widehat{\beta}_t = \prod_{i=1}^{t}(1-\beta_t)$. x_t is introduced through the standard inverse transformation as

$$x_t = \sqrt{\widehat{\beta}_t} \cdot x_0 + \sqrt{1-\widehat{\beta}_t} \cdot z_t,$$

where $z_t \sim N(0, I)$. On the other hand, reverse iteration of T time steps for data denoising can be described as Eq. (1), where P_θ and θ are the filtering signals and model parameters, respectively. The output of this model is $F_{Di} = X_i - P_{\theta i}$.

$$P_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t, t), \sum\nolimits_\theta (x_t, t)\right). \quad (1)$$

2) Data Transfer: This part mainly includes four steps, i.e., data normalization, hamming window, STFT, and spectrum gain. First, different channel signals are normalized into a range of 0 to 1, which can accelerate the model convergence. This process can be expressed as

$$N(x) = \frac{x - \min(x)}{\max(x) - \min(x)},$$

where x and $N(x)$ are the original signal and normalized signal functions, respectively. Then, normalization signals are

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

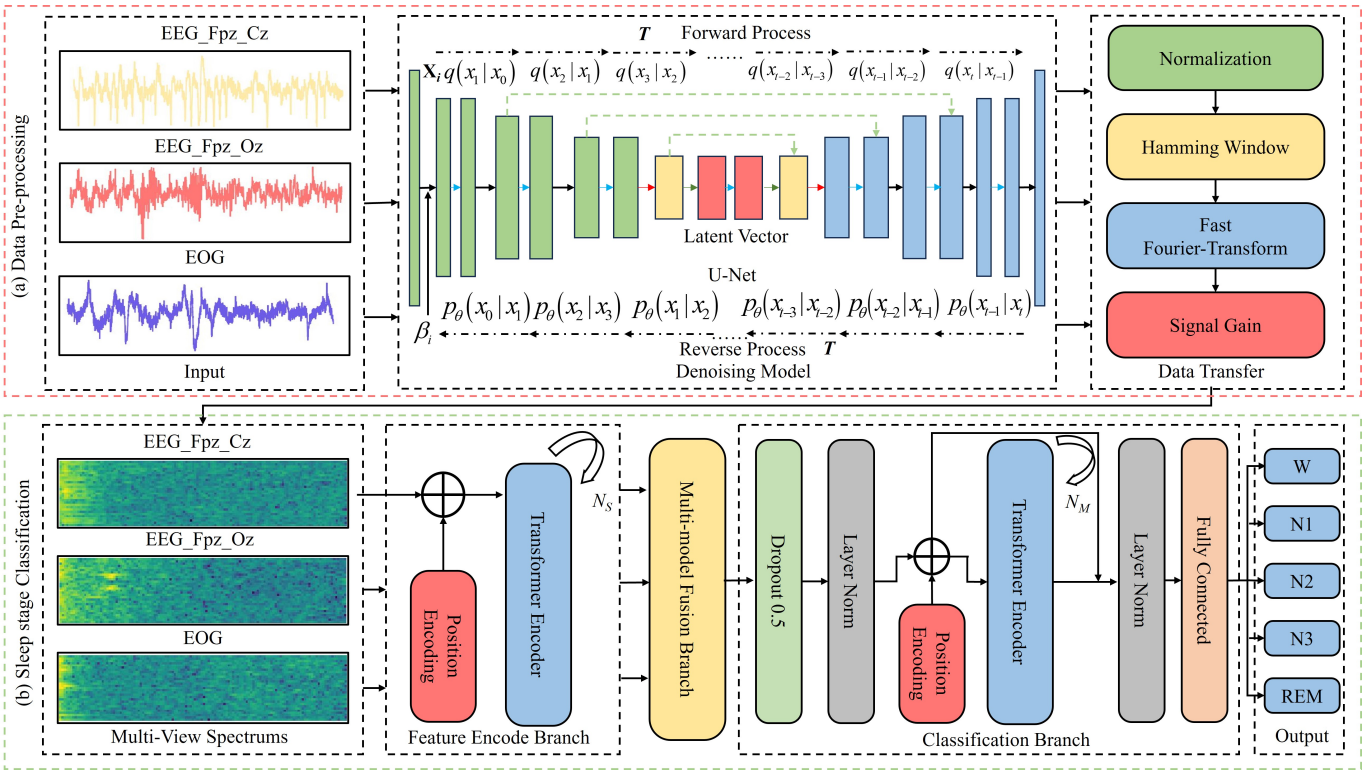


Fig. 1: The architecture of our proposed approach. (a) Data pre-processing. (b) Classification network. Notably, the C4-A1 EEG, C3-A2 EEG, LOC EOG, and ROC EOG channels are adopted for the SHHS dataset.

encoded by the Hamming window function $H(x)$, which is able to suppress the spectral leakage as

$$H(n) = \frac{1}{2}\left[1 + \cos\left(2\pi \frac{n}{N-1}\right)\right],$$

where N is the length of window. Next, the STFT transfers windowed signal to frequency spectrum, and can be expressed as Eq. (2), where $f(t)$ and $h(t)$ are time domain and frequency domain functions, respectively.

$$STFT(t,x) = \int_{-\infty}^{\infty} f(t) H(\tau - t)\, e^{-j2\pi x\tau} dt, \qquad (2)$$

Finally, the spectrum are gained as $G(x) = 20 \cdot \log_{10}(|x|)$, where $G(x)$ is gained spectrum function. Overall, the output spectrum S_i of data pre-processing is generated by denoising model DN and data transfer DT as $S_i = DT(DN(x_i))$.

C. Sleep Stage Classification

1) Feature Encode Branch: This branch is developed to encode the spectrum as high-level features, which mainly includes two parts, i.e., position encoder and Transformer encoder N_S. Position encoder is illustrated in Fig. 2(a), and is able to capture the dependencies of the physiological signals. In specific, different frequency spectrums are parallel inputted into sine function $PE(pos, 2i)$ and cosine function $PE(pos, 2i+1)$. This processing can be expressed as Eq. (3), where pos and i are position information and spatial information, respectively. Then, the output of position encoder is generated through adding and dropping as $PE_{out} = Drop(PE(pos, 2i) + PE(pos, 2i+1))$. It is added to the original input to introduce temporal information.

$$\begin{aligned} PE(pos, 2i) &= sin\left(pos \Big/ 10000^{2i/d}\right), \\ PE(pos, 2i+1) &= cos\left(pos \Big/ 10000^{2i/d}\right). \end{aligned} \qquad (3)$$

On the other hand, the Transformer encoder is shown in Fig. 2(b), which mainly includes three parts, i.e., normalization, feed forward network (FFN), and multi-attention head. Normalization helps to normalize the attention weights across different tokens, ensuring that the network attends to relevant information. In addition, the FFN consists of two linear transformations with a non-linear activation function in between. It is able to capture complex relationships and learn more expressive representations through non-linear transformation. Figure 2(c) illustrates the architecture of a multi-attention head, which consists of multiple attention sub-layers, typically denoted as i. Each attention head has three learnable parameters, including query (Q), key (K), and value (V) weight matrices. Its output F_i can be calculated as Eq. (4), where δ is softmax activation.

$$F_i(Q, K, V) = \delta\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (4)$$

The attention weight matrix is divided by d_k to eliminate variance too large for back propagation. This approach is able to stabilize the model optimization process. The final output of multi-attention head F_M is obtained by concatenating and dropout F_i as $F_M = Drop(FC(Concat[F_1, ...F_i]))$.

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

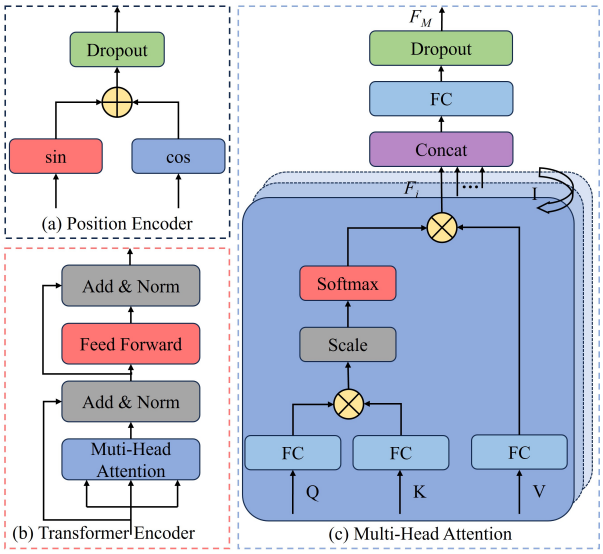


Fig. 2: The structure of feature encode branch. (a) Position encoder. (b) Transformer encoder. (c) Multi-Head attention.

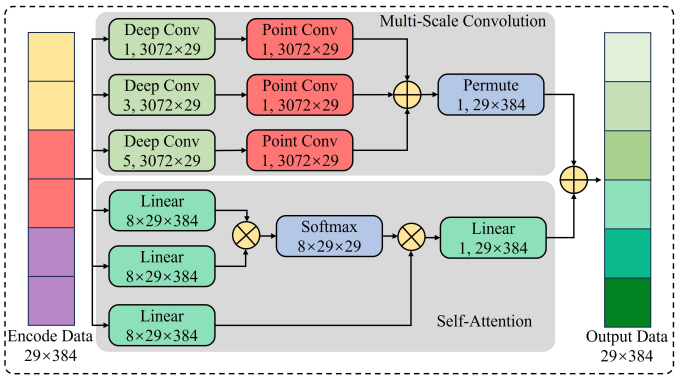


Fig. 3: The architecture of multi-model fusion branch.

2) Multi-Model Fusion Branch: This branch is designed to efficiently fuse features from different physical modes. Self-attention is well known and widely used for its ability to recognize global interdependencies between input features. Nevertheless, it often lead to the under-representation of data in longer sequences [19]. Recent research has identified that localized convolution can effectively address this problem. Inspired by these, a multi-scale attention is developed to fuse multi-modal information. It differs from earlier approaches that were primarily based on non-learnable techniques like concatenation or pooling. The architecture of this attention is illustrated in Fig. 3, which mainly includes three components, i.e., self-attention, multi-scale depth-separable convolution, and concatenation. First, the MSA fuses the depth-wise ($Deep$) and position-wise convolutions ($Point$) as Eq. (5), where i is the size of the convolution kernel.

$$C\left(x\right) = \sum_{i=1,3,5} Point_i\left(Deep_i\left(x\right)\right), \tag{5}$$

This approach can offer a balance between computational efficiency and the capacity to learn complex [43]. It efficiently captures both channel-specific characteristics and spatial relationships within the data, leading to a more powerful and computationally optimized feature representation. Next, self-attention is introduced to capture global features between sequences, which can be described as Eq. (4). Finally, the output of MSA can be described by $MSA(x) = C(x) + A(x)$.

3) Classification Branch: This branch aims to interpret the fused features and reliably categorize sleep stages. First, the features are processed through a dropout layer to reduce redundancy in various modalities. Then, they are normalized using layer normalization to speed up training. We refine the combined features using a series of position and transformer encoders N_M, connected by residuals. This method maintains each channel's distinct characteristics while integrating similarities and differences among them. Next, the encoded features are fed into two fully connected layers, which are key in selecting and merging features identified in earlier stages. The first layer includes a ReLU activation function and a dropout layer. Lastly, a softmax function is applied to calculate output probabilities for each category.

IV. Experiment Configuration

A. Datasets

Regrading the denoising model training and testing, we use two typical datasets, i.e., EEGDenoise [22] and MIT-BIH [23].

- *EEGDenoise*[2]*:* This dataset comprises 4514 EEG, 3400 EOG, and 5598 EMG segments. Each segment, resampled at 256 Hz, maintains a duration of 2s. To create the semi-simulated dataset, we merge the original EEG segments with the EMG segments. In addition, to generate EEG contaminated with artifacts, we utilize 3400 EEG-EOG pairs, 4514 EEG-EMG pairs, and 3400 EEG pairs combined with mixed artifacts (EOG and EMG).
- *MIT-BIH*[3]*:* ECG data is sourced from the MIT-BIH Arrhythmia dataset, where each recording has a duration of 30 minutes and a sampling frequency of 360 Hz. We process ECG data through a low-pass filter set at a 45 Hz cutoff frequency and then resample it to 256 Hz. Segmenting all data into 2s intervals, we obtain a total of 3600 ECG segments. This work assesses the ability to eliminate ECG artifacts by integrating clean ECG segments with EEG segments.

On the other hand, this work utilizes three high-quality open-access datasets, i.e., SleepEDF-SC, SleepEDF-X, and Sleep Heart Health Study (SHHS). These datasets are summarized and introduced as follows.

- *SleepEDF-SC*[4]*:* This subset is from the Sleep-EDF dataset at 2013. It encompassed 20 individuals aged 25 to 34 years, aiming to explore the age-sleep relationship in healthy subjects. SleepEDF-SC features two-day-and-night PSG recordings for each participant. Sleep stages in this dataset are categorized into five types for each 30s epoch, based on R&K rules, through visual analysis of signal characteristics. To perform sleep staging, we

[2]https://github.com/ncclabsustech/EEGdenoiseNet
[3]https://www.physionet.org/content/mitdb/1.0.0/
[4]https://physionet.org/content/sleep-edf/1.0.0/

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

utilize channels such as Fpz-Cz EEG, Pz-Oz EEG, and ROC-LOC EOG (horizontal), all sampled at 100 Hz. A common practice is to focus on a 60 minutes window centered on the in-bed portion of a recording [40].

- *SleepEDF-X*[5]*:* This dataset is part of the expanded Sleep-EDF dataset from 2018, including 78 subjects aged between 25 and 101 years and featuring 153 whole-night PSG sleep recordings. Similar to the SleepEDF-SC, each participant underwent PSG recordings for two consecutive nights. The manual scoring standards and 30s epoch labeling in this dataset mirror those of SleepEDF-SC. This work choose the Fpz-Cz EEG, Pz-Oz EEG, and ROC-LOC EOG (horizontal) channels as inputs.
- *SHHS*[6]*:* This dataset is part of a larger multi-center cohort study. It focuses on the impact of sleep-disordered breathing on cardiovascular and other health aspects. Participants in the study included individuals with various conditions, such as pulmonary and coronary heart diseases. Each participant provided a night's recording, containing two EEG channels (C3-A2, C4-A1) and other physiological signals, sampled at 125 Hz. Sleep experts manually annotate these patient recordings following R&K guidelines. To reduce the influence of diseases on sleep patterns, we select individuals deemed to have normal sleep, and a total of 329 individuals met these criteria. Notably, this work adopts the C4-A1 EEG, C3-A2 EEG, LOC EOG, and ROC EOG channels.

B. Implementation

In Section III-B.2, it is explained that PSG signals within a 30s epoch undergo transformation into a time-frequency image, achieved through a 256-point STFT. Regrading this process, we use a 2s Hamming window, applying a 50% overlap rate. As a result, the generated image comprises 128 frequency bins ($F = 128$) and 29 time points ($T = 29$).

In the Transformer encoder, this work employs $I = 8$ and 1024 hidden units for the multi-attention heads and the FFN, respectively. A dropout rate of 0.1 is used for each sub-layer of Transformer encoder. For each channel, $N_S = 16$ is used to extract features in the Transformer encoder. In the classification branch, $N_M = 4$ transformer encoders with the same parameters are used, and a dropout rate is 0.5. In addition, 1024 hidden units and a dropout rate of 0.5 are used in the last fully connected layers.

The hardware setup for our experiment is an NVIDIA RTX-3090, and the software environment uses PyTorch 1.11. Following the study [14], we use multi-fold cross-validation to evaluate our model. In specific, the subjects in SleepEDF-20, SleepEDF-78, and SHHS are divided into 20, 10, and 20 folds, respectively. In addition, we performed on the typical dataset, i.e., SleepEDF-X, for the ablation experiments. The AdamW optimizer with a learning rate of 5e-6 is used to minimize it. Its weight decay and batch size are set to 1e-2 and 64, respectively. We introduce the early stopping strategy, and the training process is stopped if the validation accuracy does not increase after 20 epochs.

On the other hand, we use the same training setup as study [42] for our denoising model. More specifically, the training step is 1000 , with the forward process variances linearly increasing from $\beta_1 = 1e-4$ to $\beta_{1000} = 2e-2$. Regrading the reverse process, a U-Net backbone [45] is employed with group normalization across all layers. In addition, we use self-attention at the 16×16 feature map resolution.

The loss function of denoising model is the PSNR loss, which can be formulated as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2,$$
$$Loss = -20 \times \log_{10}\left(\frac{MAX_y}{\sqrt{MSE}}\right),$$

where y, $\widehat{y}_i$, and N are the original signal, reconstructed signal, and a signal of length, respectively. Furthermore, MAX_y represents the maximum possible value of the signal. On the other hand, a cross-entropy loss with weight is designed for classification network. It is able to reduce the over-fitting due to unbalanced distribution and can be described as Eq. (6), where L_i represents the count of label.

$$\begin{aligned} Loss &= \sum_{i=0}^{4} \lambda_i \cdot \log\left(1/q_i\right), \\ \lambda_i &= L_i \Big/ \sum_{i=0}^{4} L_i. \end{aligned} \tag{6}$$

C. Evaluation Metric

To evaluate our denoising model, this work follows the study [46] to use root mean squared error (RMSE), correlation coefficient (CC), and signal to noise ratio (SNR), along with their weighted counterparts, weighted CC (WCC), and weighted SNR (WSNR). Among them, the initial three metrics are used to evaluate overall performance, while WCC and WSNR focus on local sub-band distortions. On the other hand, regrading the performance of sleep stage classification, we follow the study [14] to utilize precision (Pre), recall (Rec), F1-score (F1), accuracy (ACC), macro-averaged F1-score (MF1), and Cohen's kappa (k). Notably, MF1 calculates the average F1-score across the five categories. Furthermore, floating-point operations (FLOPs), parameters (Param.), and frames per second (FPS) represent the model's time complexity, spatial complexity, and inference speed, respectively. These are utilized to evaluate the computational cost.

V. Results and Discussions

A. Sleep Classification Results

Three typical datasets, i.e., SleepEDF-SC, SleepEDF-X, and SHHS, are trained and tested to evaluate our proposed Diff-SleepNet. The performance is evaluated by key metrics and listed in Table II, where the optimal and sub-optimal performance are highlighted in bold and underline, respectively. The left of Table II lists the confusion matrix of Diff-SleepNet. Meanwhile, the right of Table II lists the key metrics of Diff-SleepNet and other baseline models, i.e., MCSleepNet [14]

[5]https://physionet.org/content/sleep-edfx/1.0.0/
[6]https://sleepdata.org/datasets/shhs

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

TABLE II: Performance of sleep stage classification on three datasets. The optimal and sub-optimal performance are highlighted in **Bold** and Underline, respectively

Sleep Stage	Predicted Confusion Matrix					Diff-SleepNet			MCSleepNet [14]			SleepContextNet [44]		
	W	N1	N2	N3	REM	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
SleepEDF-SC														
W	**8453**	144	87	28	186	**0.941**	**0.951**	**0.947**	0.934	0.921	0.928	0.912	0.880	0.896
N1	379	**1376**	532	6	511	**0.698**	0.491	**0.576**	0.550	0.444	0.491	0.469	**0.547**	0.505
N2	64	245	**16570**	453	467	**0.909**	**0.931**	**0.920**	0.895	0.903	0.900	0.880	0.887	0.884
N3	5	5	489	**5201**	3	**0.913**	**0.912**	**0.913**	0.891	0.895	0.893	0.905	0.867	0.885
REM	87	201	544	7	**6878**	0.855	**0.891**	**0.928**	**0.873**	0.875	0.848	0.825	0.816	0.820
SleepEDF-X														
W	**63326**	2573	307	48	568	**0.964**	**0.948**	**0.955**	0.950	0.931	0.940	0.930	0.925	0.928
N1	1548	**11688**	6257	26	1825	**0.689**	**0.578**	**0.611**	0.581	0.487	0.530	0.500	0.480	0.490
N2	276	1432	**65145**	1012	1267	**0.867**	**0.942**	**0.903**	0.840	0.900	0.869	0.843	0.854	0.848
N3	22	20	1565	**11428**	4	**0.912**	**0.876**	**0.894**	0.831	0.807	0.818	0.802	0.810	0.806
REM	546	1261	1880	21	**22127**	**0.858**	**0.856**	**0.857**	0.820	0.831	0.826	0.786	0.791	0.789
SHHS														
W	**43716**	672	845	259	827	**0.935**	**0.944**	**0.940**	0.917	0.927	0.922	0.888	0.896	0.892
N1	988	**4723**	1745	5	2843	**0.600**	0.458	**0.520**	0.484	0.348	0.405	0.514	**0.527**	**0.520**
N2	1148	1921	**126199**	4773	8084	**0.912**	**0.888**	**0.900**	0.896	0.876	0.886	0.870	0.883	0.876
N3	35	5	7586	**52489**	38	**0.912**	**0.873**	**0.892**	0.905	0.860	0.882	0.866	0.821	0.843
REM	847	548	1998	55	**62505**	0.841	**0.948**	0.891	0.827	0.934	0.877	**0.890**	0.896	**0.893**

and SleepContextNet [44]. As can be observed, our Diff-SleepNet yields F1 of 94.7%, 57.6%, 92.0%, 91.3%, and 92.8% at SleepEDF-SC on the five sleep stages, i.e., W, N1, N2, N3, and REM, respectively. Furthermore, it outperforms the other frameworks with F1 of 95.5%, 61.1%, 90.3%, 89.4%, and 85.7% at SleepEDF-X on the five sleep stages. Diff-SleepNet also has better performance on other frameworks on the W, N1, N2, and N3, respectively. However, it performs slightly worse than SleepContextNet [44] for REM. This may be because SleepContextNet adds more context.

Figure 4 illustrates the visualized performance of subject SC4001E0 from the SleepEDF-X. The upper part of the figure illustrates the actual ground truth, while the lower section depicts the classification outcomes obtained from Diff-SleepNet. Notably, there is a high degree of similarity between the ground truth and the predicted hypnogram by Diff-SleepNet. Diff-SleepNet achieves an accuracy of 91.4% and an F1-score of 83.7% in this subject. Visually, the hypnogram generated by Diff-SleepNet might display less smoothness compared to the ground truth, possibly exhibiting unusual sleep transitions. This characteristic stems from the fact that in our framework, each epoch is shuffled prior to input, and the temporal sequence of epochs for an individual subject is not taken into account. Despite this limitation, such an approach significantly simplifies the model's complexity and enhances its capability in sleep stage classification. In conclusion, these quantitative and qualitative results determined that our model has advantages over other peer studies.

B. Comparison with Counterparts

To evaluate the advantages of our proposal, the following studies [14], [24], [30], [32], [40], [44], [47]–[50] are introduced for peer comparison on SleepEDF-SC, SleepEDF-X, and SHHS. They are trained and validated using the same strategy as the study [14] to ensure a fair comparison. Experimental results are listed in Table III, where the optimal and sub-optimal performance are highlighted in bold and underline, respectively. The left and right of Table III are

TABLE III: Performance of our method and counterparts for sleep stage classification. The optimal and sub-optimal performance are showed in **Bold** and Underline, respectively

Method	Overall Metrics			Per-class F1 Score				
	Acc	k	MF1	W	N1	N2	N3	REM
SleepEDF-SC								
Supratak *et al.* [40]	0.820	0.76	0.769	N/A	N/A	N/A	N/A	N/A
Mousavi *et al.* [47]	0.843	0.79	0.797	0.892	0.522	0.898	0.851	0.850
Phan *et al.* [48]	0.860	0.81	0.797	N/A	N/A	N/A	N/A	N/A
Eldele *et al.* [49]	0.844	0.79	0.781	0.897	0.426	0.888	0.902	0.790
Phan *et al.* [30]	0.864	0.81	0.809	N/A	N/A	N/A	N/A	N/A
Dai *et al.* [14]	0.872	0.82	0.812	0.928	0.491	0.900	0.893	0.848
Ours	**0.897**	**0.87**	**0.846**	**0.942**	**0.698**	**0.909**	**0.913**	**0.855**
SleepEDF-X								
Mousavi *et al.* [47]	0.800	0.73	0.736	N/A	N/A	N/A	N/A	N/A
Perslev *et al.* [24]	N/A	N/A	0.760	N/A	N/A	N/A	N/A	N/A
Phan *et al.* [48]	0.838	0.78	0.782	N/A	N/A	N/A	N/A	N/A
Zhu *et al.* [50]	0.828	N/A	0.778	0.903	0.471	0.860	0.821	0.832
Eldele *et al.* [49]	0.813	0.74	0.753	0.920	0.420	0.850	0.821	0.741
Phan *et al.* [30]	0.840	0.78	0.787	N/A	N/A	N/A	N/A	N/A
Zhao *et al.* [44]	0.827	0.76	0.772	0.928	0.490	0.848	0.806	0.789
Phan *et al.* [32]	0.814	0.74	0.743	0.917	0.404	0.843	0.779	0.772
Dai *et al.* [14]	0.850	0.79	0.796	0.928	0.530	0.869	0.818	0.826
Ours	**0.886**	**0.84**	**0.824**	**0.964**	**0.689**	**0.877**	**0.902**	**0.857**
SHHS								
Eldele *et al.* [49]	0.842	0.78	0.753	0.867	0.332	0.871	0.871	0.821
Zhao *et al.* [44]	0.864	0.81	0.805	0.892	0.520	0.876	0.843	**0.893**
Dai *et al.* [14]	0.875	0.82	0.793	0.922	0.405	0.886	0.882	0.877
Ours	**0.892**	**0.85**	**0.828**	**0.935**	**0.600**	**0.912**	**0.912**	0.841

performance records of overall metrics (i.e., accuracy, Cohen's kappa, and MF1) and five stages, respectively.

From Table III, our model yields the accuracy, Cohen's kappa, and MF1 of 89.7%, 0.87, and 84.6% on the SleepEDF-SC, respectively. In addition, it also achieves the F1 score of 94.2%, 69.8%, 90.9%, 91.3%, and 85.5% on five stages with SleepEDF-SC, respectively. Comparing with the sub-optimal results, Diff-SleepNet improves the accuracy, Cohen's kappa, and MF1 by 2.5%, 0.05, and 3.4%, respectively. The per-class F1 scores also achieve better performance than other models. Regrading the SleepEDF-X, our Diff-SleepNet improves the accuracy, Cohen's kappa, and MF1 by 3.6%, 0.05, and 2.8%,

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

TABLE IV: Performance of ablation study for denoising model on the SleepEDF-X

Method	Overall Metrics			Per-class F1 Score				
	Acc	k	$MF1$	W	N1	N2	N3	REM
w/o DDPM	0.855	0.79	0.805	0.942	0.673	0.857	0.881	0.837
w/ MSE	0.873	0.82	0.813	0.950	0.679	0.864	0.889	0.844
w/ PNSR	**0.886**	**0.84**	**0.824**	**0.964**	**0.689**	**0.877**	**0.902**	**0.857**

TABLE V: Performance of physical signal denoising on the EEGDenoise and MIT-BIH

Method	RRMSE (↓)	CC (↑)	SNR (↑)	WSNR (↑)	WCC (↑)
EEGDenoise					
Sun *et al.* [51]	0.302	0.931	12.746	16.897	0.906
Zhang *et al.* [52]	0.344	0.904	12.121	15.094	0.888
Pu *et al.* [53]	0.301	0.945	13.251	17.584	0.924
Gao *et al.* [54]	0.315	0.919	13.236	16.374	0.902
Yin *et al.* [46]	0.262	0.948	14.534	18.702	0.926
Ours	**0.254**	**0.934**	**14.775**	**18.942**	**0.934**
MIT-BIH					
Sun *et al.* [51]	0.326	0.935	11.114	14.075	0.936
Zhang *et al.* [52]	0.302	0.933	12.912	15.698	0.938
Pu *et al.* [53]	0.323	0.939	11.204	13.767	0.941
Gao *et al.* [54]	0.283	0.946	13.428	16.259	0.950
Yin *et al.* [46]	0.254	0.958	13.975	17.422	0.958
Ours	**0.249**	**0.964**	**14.126**	**17.512**	**0.967**

respectively, compared to the sub-optimal results. It also has higher performance than other methods on overall metrics. However, Zhao *et al.* [44] developed SleepContextNet, which yields the best performance on the F1 score of REM. The reasons have been analyzed in Section V-A.

C. Ablation Study

1) Contribution of Denoising Model: This work develops a denoising model with PNSR loss to filter out the interference from various noise and ocular artifacts, which can cause greater difficulties in the subsequent analysis process. To evaluate the contribution of the denoising model, we analysis the performance records of model with and without DDPM on the SleepEDF-X. In addition, we compare the classification results with different loss functions, i.e., MSE loss and PNSR loss. These results are listed in Table IV. When removing DDPM from our framework, we observe a decrease in the accuracy, Cohen's kappa, and MF1 of 3.1%, 0.05, and 1.9%, respectively. Besides, when using MSE loss from the DDPM, we observe a decrease in the accuracy, Cohen's kappa, and MF1 of 1.3%, 0.02, and 1.1%, respectively. These results indicated that our denoising model is well suited for sleep stage classification.

On the other hand, this study compares our denosing model with other peer models [46], [51]–[54]. They are trained and validated using the same strategy as the study [46] to ensure a fair comparison. These results are validated at two typical datasets, i.e., EEGDenoise and MIT-BIH, and listed in Table V. It can be observed that our improved model has advantages over other peer networks.

2) Contribution of STFT: During the data pre-processing stage, STFT is introduced to transfer the PSG to a spectrogram. Different STFT window functions, i.e., Hamming, Hanning, and Blackman, are utilized to evaluate their contribution to sleep stage classification. Furthermore, we analysis performance of the Hamming window with different lengths, i.e., 1s, 2s, and 3s. The performance records with the SleepEDF-X are listed in Table VI.

TABLE VI: Performance of ablation study for STFT with different window functions and lengths on the SleepEDF-X

Method	Overall Metrics			Per-class F1 Score				
	Acc	k	MF1	W	N1	N2	N3	REM
Window Function								
w/ Hanning	0.878	0.84	0.812	0.955	0.683	0.869	0.894	0.849
w/ Blackman	0.875	0.83	0.807	0.952	0.680	0.866	0.891	0.846
Window Length								
w/ 1s	0.866	0.83	0.797	0.942	0.673	0.857	0.882	0.838
w/ 3s	0.864	0.82	0.781	0.940	0.672	0.855	0.880	0.836
w/ Hamming & 2s	**0.886**	**0.84**	**0.824**	**0.964**	**0.689**	**0.877**	**0.902**	**0.857**

As we can observe, when STFT has a Hamming window and its length is 2s, it performs better than other window functions and lengths in sleep stage classification. This indicates that STFT using Hamming window can effectively attenuate the time-domain noise interference in non-smooth PSG and improve the model performance. Moreover, a window length of 2s enhances the temporal resolution of the time-frequency maps and obtains more detailed frequency information to improve sleep staging results.

3) Contribution of Multi-Modal Input: This work introduces multi-modal input to enrich the features for model classification. To evaluate the contribution of multi-modal, we compare the performance records with different combinations on the SleepEDF-X, which are listed in Table VII.

From Table VII, when the model has only Fpz-Cz, Pz-Oz, and EOG as individual input, it yields MF1 as 79.1%, 79.6%, and 77.1%, respectively. In addition, we can observe that multi-modal input can significantly improve overall performance records. When Pz-Oz is combined with Fpz-Cz in the model, it is able to enhance the classification performance in the W and N3 phases, which increases F1 scores by 2.8% and 3.4%, respectively, compared to input Fpz-Cz only. However, adding EOG to the input doesn't significantly improve the performance for stage W. Previous studies have found that the major rhythms of stage W occur in the occipital region. The Pz-Oz channel is closer to the occipital region and therefore captures the rhythm more efficiently. Furthermore, the slow-wave activity is the main factor considered in stage N3. Although it can be detected by the EOG channel, these movements are not important for feature differentiation because there are no distinctive eye movements in stage N3. However, combining the EOG channel in the model is able to improve the performance records of stages N1 and REM. This is because rapid eye movements can be well detected with EOG channel. In comparisons, using Fpz-Cz EEG with EOG performs better than using Fpz-Cz EEG with Pz-Oz EEG, showing the added value of EOG. Both combinations are similarly effective in classifying stage N2, suggesting Pz-Oz and EOG contribute equally to this classification. Nevertheless, integrating both channels leads to further improvement in the F1-score for stage N2, highlighting the potential synergy between these channels. These results align with the physio-

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

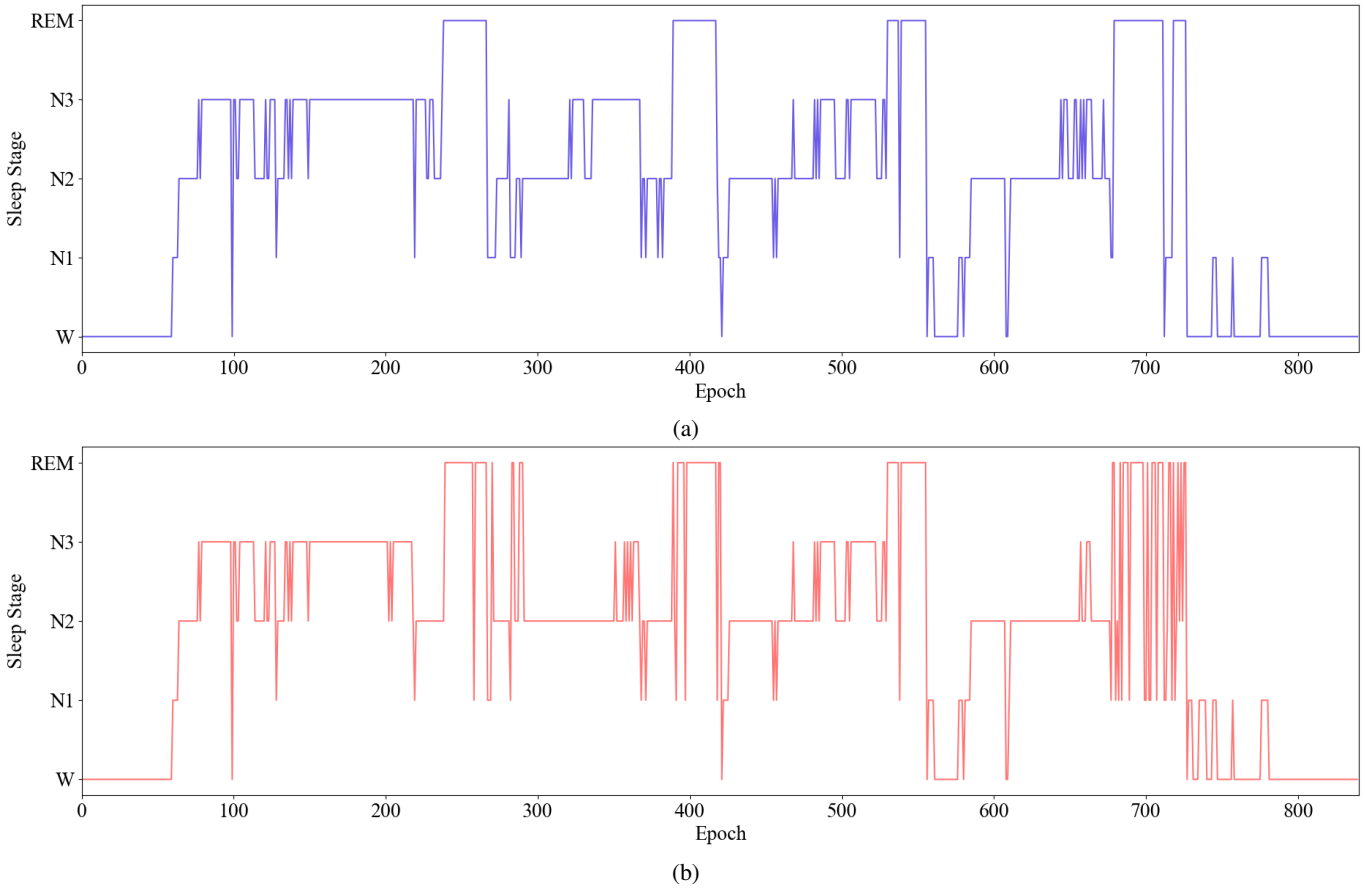


Fig. 4: Visual performance of our framework for sample data. (a) Hypnogram of Ground Truth. (b) Hypnogram of Prediction. Data source: subject SC4001E0 from SleepEDF-X.

TABLE VII: Performance of ablation study for multi-modal input on the SleepEDF-X

Modal			Overall Metrics			Per-class F1 Score				
Fpz-Cz	Pz-Oz	EOG	Acc	k	MF1	W	N1	N2	N3	REM
✓			0.831	0.78	0.791	0.930	0.483	0.854	0.867	0.732
	✓		0.822	0.77	0.768	0.928	0.572	0.848	0.851	0.776
		✓	0.828	0.77	0.771	0.929	0.476	0.845	0.847	0.805
✓	✓		0.865	0.82	0.783	0.958	0.586	0.868	0.901	0.801
✓		✓	0.871	0.83	0.801	0.952	0.630	0.869	0.872	0.832
	✓	✓	0.860	0.82	0.779	0.942	0.627	0.865	0.863	0.825
✓	✓	✓	**0.886**	**0.84**	**0.824**	**0.964**	**0.689**	**0.877**	**0.902**	**0.857**

logical features of the sleep stages, showing that Diff-SleepNet effectively integrates and processes their data.

4) Contribution of Multi-Modal Fusion Branch: This work develops an attention-based branch to efficiently fuse multi-model features. To evaluate the advantages of this branch, we analyze the performance records of our model with concatenation and MFB. Our MFB is designed based on attention mechanisms. Therefore, it is also compared with four typical modules, i.e., squeeze-and-excitation (SE) [55], spatial attention (SA) [56], coordinate attention (CA) [57], and efficient multi-scale attention (EMA) [58]. In addition, we set different numbers of convolutional kernel in our framework to determine the optimal parameters. The performance records with the SleepEDF-X are listed in Table VIII.

TABLE VIII: Performance of ablation study for MFB on the SleepEDF-X

Method	Overall Metrics			Per-class F1				
	Acc	k	MF1	W	N1	N2	N3	REM
Fusion Method								
w/ Concatenation	0.852	0.79	0.798	0.945	0.531	0.869	0.821	0.830
w/ SE	0.871	0.82	0.812	0.952	0.612	0.865	0.872	0.847
w/ SA	0.867	0.81	0.806	0.948	0.603	0.870	0.846	0.841
w/ CA	0.867	0.82	0.818	0.956	0.634	0.868	0.880	0.850
w/ EMA	0.879	0.83	0.822	0.958	0.652	**0.877**	0.892	**0.858**
Num. Kernel								
1	0.862	0.81	0.802	0.947	0.605	0.872	0.852	0.840
1&3	0.873	0.82	0.819	0.954	0.657	0.875	0.885	0.852
1&3&5&MFB	**0.886**	**0.84**	**0.824**	**0.964**	**0.689**	**0.877**	**0.902**	0.857

When removing MFB from the model, we observe a decrease in the accuracy, Cohen's kappa, and MF1 by 3.4%, 0.05, and 2.6%, respectively. Furthermore, when other attentions are combined with our framework, we observe a decrease in the MF1 by 1.2%, 1.8%, 0.6%, and 0.1%, respectively. It is worth noting that EMA is similar to our module in some metrics. This is because it is also based on multi-scale. With the increase in the convolutional kernel, the effectiveness of the model predictions also increases. These results determined

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

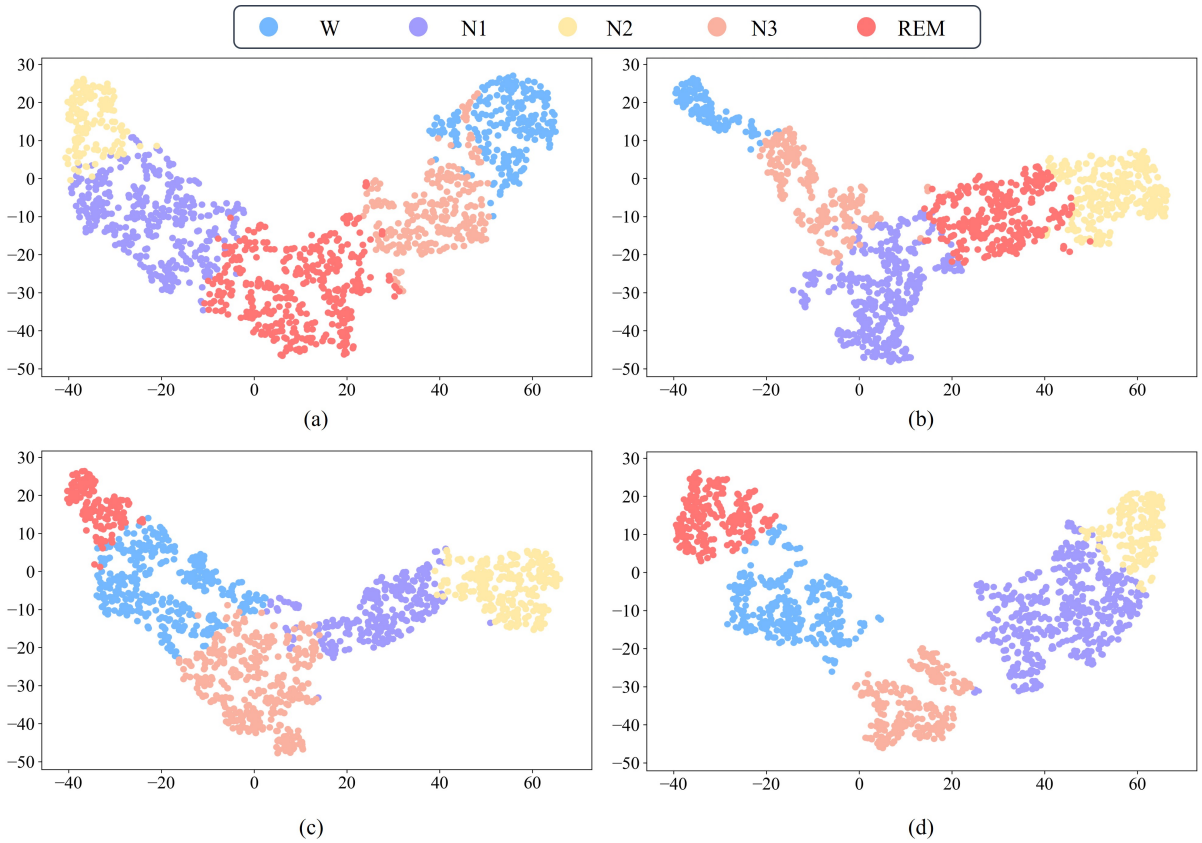


Fig. 5: The feature clustering maps of ablation study for MFB on the SleepEDF-X. (a) Model with concatenation. (b) Model with the SA. (c) Model with the SE. (d) Model with the MFB.

that the MFB can well fuse features from different modalities and enhance the effectiveness of sleep stage classification.

On the other hand, we visualize the decision features with different fusion methods to evaluate our MFB, which can be illustrated in Fig. 5. It can be seen that when the model is with three modules (i.e., concatenation, SE, and SA), clustered features are not clearly distinguishable, which will enhance the error rate of the prediction. In addition, after analyzing the confusion matrix of the SleepEDF-X, it is found that the model is prone to confusing two states, i.e., stage W and N1, as well as stage N2 and N1. It is well known that sleep is a continuous process. From Fig. 5, we can also find the distance of these features is close, which explains the phenomenon at feature level. Therefore, the model tends to confuse the physiological signals of alternating sleep stages.

D. Hyper-Parameter Optimization

This work uses a transformer as backbone network, which plays a key role in feature extraction and integration. Therefore, it is essential to analyze the transformer encoder's structure and parameters. This work focuses on how the number of heads (I) in transformer encoders, the count N_S in the single-channel feature extraction block, and the number N_M in the classification branch impact our model's performance. The performance metrics on the SleepEDF-X with different parameter settings are listed in Tables IX.

From Tables IX, a slight improvement occurs as N_S increases from 4 to 16, due to more network layers and increased complexity. Notably, N_M set to 4 performs better than at 2 or 8. This relates to our model's approach of combining multi-channel joint features with individual single-channel features in the fusion block, where too many transformer encoders could disrupt this balance. The number of heads I has a similar effect to N_M. Model performance initially improves with more heads, but then decreases. More heads expand the model's ability to detect various effective features, but they also reduce the number of features each attention head can capture.

Understanding computational resource usage is crucial for model deployment. To evaluate this, we conducted ablation experiments on key modules, measuring critical metrics like FLOPs, parameters, and FPS. Our model shows a balance between performance and efficiency, with 1419.18 M FLOPs, 35.53 M parameters, and a frame rate of 31.07 FPS. Further analysis reveals that the FLOPs for N_S and N_M are 45.79 M and 45.80 M, respectively. Given the enhancement in detection results, the computational resources required for these branches are reasonable.

E. Discussion

Automatic sleep staging based on PSG recordings has made significant progress over time. The classification problem of sleep stages has been mostly resolved, but there are still challenges to overcome. First, from Tables II and III, we can

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

TABLE IX: Performance of Diff-SleepNet on the SleepEDF-X with different hyper-parameters

N_S	N_M	I	Acc	k	MF1	Flops (M ↓)	Param. (M ↓)	FPS (↑)
4	2	8	0.832	0.76	0.762	**778.10**	**24.45**	**101.30**
8	2	8	0.836	0.77	0.773	961.27	27.62	60.33
16	2	8	0.850	0.79	0.782	1327.60	33.95	32.41
4	4	8	0.833	0.76	0.779	869.69	26.03	89.60
8	4	8	0.858	0.80	0.806	1052.85	29.20	56.21
16	4	8	**0.886**	**0.84**	**0.824**	1419.18	35.53	31.07
4	8	8	0.867	0.81	0.808	1052.85	29.19	73.10
8	8	8	0.875	0.82	0.816	1236.02	32.36	47.60
16	8	8	0.877	0.82	0.818	1602.35	38.69	28.07
16	4	2	0.852	0.80	0.796	1419.18	35.53	30.88
16	4	4	0.873	0.82	0.810	1419.18	35.53	30.35
16	4	16	0.885	**0.84**	0.821	1419.18	35.53	30.74

observe that each model significantly less effective in stage N1 than in the other stages. This is because stage N1 have fewer representative samples than other stages. Further work can increase the quantity and diversity of stage N1 samples through techniques like image augmentation, synthetic data generation, or oversampling. In addition, existing multi-view models have the limitation of slow interference time. The models need to be practically applied in daily life, which may have a delayed impact on real-time sleep monitoring. Therefore, their architecture can be improved in the future by accelerated inference methods such as knowledge distillation. Finally, effective application of automatic sleep staging models in daily life requires consideration of cloud-based processing systems. Real-time collection and transmission of sleep data from wearable monitoring devices to cloud-based models, which can provide feedback to individuals or clinical doctors. However, Data privacy and security are paramount considerations in the cloud-based applications. Therefore, it is crucial to implement strong security measures, including encryption, authentication, and access controls.

VI. Conclusion

This work proposes a novel network for efficiently classifying sleep stages from multi-modal data. The core of our approach is a diffusion model that denoises signals, enhanced by a multi-scale attention mechanism for feature fusion. First, the multi-modal physiological signals are denoised by the diffusion model, which is constrained by the PNSR loss. The STFT is then introduced to transfer these signals into spectrums, which are inputted into the Transformer-based backbone to extract high-level features. Next, the production is fused by an attention mechanism. Finally, the sleep stage category is generated through a classification branch. Both quantitative and qualitative evaluations show that our proposed framework is effective and has advantages over other peer methods. In the future, this framework will be optimized for clinical applications using multi-center data.

References

[1] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J. M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, "The future of sleep health: a data-driven revolution in sleep science and medicine," *NPJ Digital Medicine*, vol. 3, no. 1, p. 42, 2020.

[2] M. G. Smith, M. Cordoza, and M. Basner, "Environmental noise and effects on sleep: an update to the WHO systematic review and meta-analysis," *Environmental Health Perspectives*, vol. 130, no. 7, p. 076001, 2022.

[3] W. Al-Salman, Y. Li, A. Y. Oudah, and S. Almaged, "Sleep stage classification in EEG signals using the clustering approach based probability distribution features coupled with classification algorithms," *Neuroscience Research*, vol. 188, pp. 51–67, 2023.

[4] S. C. C. O. T. J. S. O. S. R. S. (JSSR):, T. Hori, Y. Sugita, E. Koga, S. Shirakawa, K. Inoue, S. Uchida, H. Kuwahara, M. Kousaka, T. Kobayashi *et al.*, "Proposed supplements and amendments to 'a manual of standardized terminology, techniques and scoring system for sleep stages of human subjects', the rechtschaffen & kales (1968) standard," *Psychiatry and Clinical Neurosciences*, vol. 55, no. 3, pp. 305–310, 2001.

[5] M. B. B. Heyat, D. Lai, F. I. Khan, and Y. Zhang, "Sleep bruxism detection using decision tree method by the combination of C4-P4 and C4-A1 channels of scalp EEG," *IEEE Access*, vol. 7, pp. 102 542–102 553, 2019.

[6] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Computers in Biology and Medicine*, vol. 106, pp. 71–81, 2019.

[7] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

[8] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms," *Journal of Medical Systems*, vol. 38, pp. 1–21, 2014.

[9] P. An, Z. Yuan, J. Zhao, X. Jiang, and B. Du, "An effective multi-model fusion method for EEG-based sleep stage classification," *Knowledge-Based Systems*, vol. 219, p. 106890, 2021.

[10] A. Balaji, U. Tripathi, V. Chamola, A. Benslimane, and M. Guizani, "Toward safer vehicular transit: implementing deep learning on single channel EEG systems for microsleep detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 1052–1061, 2021.

[11] A. Chougule, J. Shah, V. Chamola, and S. Kanhere, "Enabling safe ITS: EEG-based microsleep detection in vanets," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[12] E. Khalili and B. M. Asl, "Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG," *Computer Methods and Programs in Biomedicine*, vol. 204, p. 106063, 2021.

[13] S. Abdulla, M. Diykh, S. Siuly, and M. Ali, "An intelligent model involving multi-channels spectrum patterns based features for automatic sleep stage classification," *International Journal of Medical Informatics*, vol. 171, p. 105001, 2023.

[14] Y. Dai, X. Li, S. Liang, L. Wang, Q. Duan, H. Yang, C. Zhang, X. Chen, L. Li, X. Li *et al.*, "MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with PSG," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[15] P. Singh and G. Pradhan, "A new ECG denoising framework using generative adversarial network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 2, pp. 759–764, 2020.

[16] H. Zeng, N. Xia, D. Qian, M. Hattori, C. Wang, and W. Kong, "Dm-re2i: A framework based on diffusion model for the reconstruction from eeg to image," *Biomedical Signal Processing and Control*, vol. 86, p. 105125, 2023.

[17] L. Abrahamyan, A. M. Truong, W. Philips, and N. Deligiannis, "Gradient variance loss for structure-enhanced image super-resolution," in *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3219–3223.

[18] A. Lebal, A. Moussaoui, and A. Rezgui, "Epilepsy-net: attention-based 1d-inception network model for epilepsy detection using one-channel and multi-channel eeg signals," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 17 391–17 413, 2023.

[19] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519.

[20] 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, 2018.

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3422472

[21] 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.

[22] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, "EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising," *Journal of Neural Engineering*, vol. 18, no. 5, p. 056057, 2021.

[23] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[24] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] M. M. Ghassemi, B. E. Moody, L.-W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, "You snooze, you win: the physionet/computing in cardiology challenge 2018," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45. IEEE, 2018, pp. 1–4.

[26] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 180–192, 2016.

[27] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa *et al.*, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep," *Sleep Medicine*, vol. 2, no. 6, pp. 537–554, 2001.

[28] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[29] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 641–644.

[30] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2021.

[31] S. Sanghavi, P. Vaid, P. Rathod, and K. Srivastava, "SpectroTemporalNet: Automated sleep stage scoring with stacked generalization," in *Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2021, pp. 1256–1263.

[32] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[33] Y. Li, J. Chen, W. Ma, G. Zhao, and X. Fan, "MVF-sleepnet: Multi-view fusion network for sleep stage classification," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[34] H. Yao, T. Liu, R. Zou, S. Ding, and Y. Xu, "A Spatial-Temporal Transformer architecture using multi-channel signals for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[35] X. Zhang, W. Kou, I. Eric, C. Chang, H. Gao, Y. Fan, and Y. Xu, "Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device," *Computers in Biology and Medicine*, vol. 103, pp. 71–81, 2018.

[36] X. Ji, Y. Li, and P. Wen, "3DSleepNet: A multi-channel bio-signal based sleep stages classification method using deep learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[37] G. Kong, C. Li, H. Peng, Z. Han, and H. Qiao, "EEG-Based sleep stage classification via neural architecture search," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1075–1085, 2023.

[38] X. Ji, Y. Li, P. Wen, P. Barua, and U. R. Acharya, "MixSleepNet: A multi-type convolution combined sleep stage classification model," *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107992, 2024.

[39] D. Cho and B. Lee, "Automatic sleep-stage classification based on residual unit and attention networks using directed transfer function of electroencephalogram signals," *Biomedical Signal Processing and Control*, vol. 88, p. 105679, 2024.

[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] H. Zhu, W. Zhou, C. Fu, Y. Wu, N. Shen, F. Shu, H. Yu, C. Chen, and W. Chen, "MasksleepNet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[44] C. Zhao, J. Li, and Y. Guo, "SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106806, 2022.

[45] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," in *Proceedings of the 5th International Conference on Learning Representations, ICLR*. OpenReview.net, 2017.

[46] J. Yin, A. Liu, C. Li, R. Qian, and X. Chen, "A GAN guided parallel CNN and Transformer network for EEG denoising," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[47] 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.

[48] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. 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*, vol. 27, no. 3, pp. 400–410, 2019.

[49] 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.

[50] T. Zhu, W. Luo, and F. Yu, "Convolution-and attention-based neural network for automated sleep stage classification," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, p. 4152, 2020.

[51] W. Sun, Y. Su, X. Wu, and X. Wu, "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals," *Neurocomputing*, vol. 404, pp. 108–121, 2020.

[52] H. Zhang, C. Wei, M. Zhao, Q. Liu, and H. Wu, "A novel convolutional neural network model to remove muscle artifacts from EEG," in *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1265–1269.

[53] X. Pu, P. Yi, K. Chen, Z. Ma, D. Zhao, and Y. Ren, "EEGDnet: Fusing non-local and local self-similarity for EEG signal denoising with Transformer," *Computers in Biology and Medicine*, vol. 151, p. 106248, 2022.

[54] T. Gao, D. Chen, Y. Tang, Z. Ming, and X. Li, "EEG reconstruction with a dual-scale CNN-LSTM model for deep artifact removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1283–1294, 2022.

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[56] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6688–6697.

[57] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 713–13 722.

[58] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

Authorized licensed use limited to: Northeastern University. Downloaded on September 07,2024 at 08:38:41 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.