3DSleepNet: A Multi-Channel Bio-Signal Based Sleep Stages Classification Method Using Deep Learning

Xiaopeng Ji, Yan Li, and Peng Wen

***Abstract*— A novel multi-channel-based 3D convolutional neural network (3D-CNN) is proposed in this paper to classify sleep stages. Time domain features, frequency domain features, and time-frequency domain features are extracted from electroencephalography (EEG), electromyogram (EMG), and electrooculogram (EOG) channels and fed into the 3D-CNN model to classify sleep stages. Intrinsic connections among different bio-signals and different frequency bands in time series and time-frequency are learned by 3D convolutional layers, while the frequency relations are learned by 2D convolutional layers. Partial dot-product attention layers help this model find the most important channels and frequency bands in different sleep stages. A long short-term memory unit is added to learn the transition rules among neighboring epochs. Classification experiments were conducted using both ISRUC-S3 datasets and ISRUC-S1, sleep-disorder datasets. The experimental results showed that the overall accuracy achieved 0.832 and the F1-score and Cohen's kappa reached 0.814 and 0.783, respectively, on ISRUC-S3, which are a competitive classification performance with the state-of-the-art baselines. The overall accuracy, F1-score, and Cohen's kappa on ISRUC-S1 achieved 0.820, 0.797, and 0.768, respectively, which also demonstrate its generality on unhealthy subjects. Further experiments were conducted on ISRUC-S3 subset to evaluate its training time. The training time on 10 subjects from ISRUC-S3 with 8549 epochs is 4493s, which indicates its highest calculation speed compared with the existing high-performance graph convolutional networks and U^2−Net architecture algorithms.**

***Index Terms*— Deep learning, 3D convolutional networks, sleep stages classification.**

I. Introduction

SLEEP disorders, including insomnia, apnea, and circadian rhythm sleep disorders, are widespread in most populations. Sleep stages classification is the first step for sleep research and sleep disorder diagnosis. Polysomnograms (PSGs) are physiological signals, which are collected and analyzed by experts to explore brain activities during humans action [1], like measuring the depth of anesthesia [2], [3], identifying the motor imagery [4], seizure prediction [5], etc. Rechtschaffen and Kales sleep staging rules (R&K rules) [6] and American Academy of Sleep Medicine (AASM) standards [7] are golden criteria for experts to evaluate sleep quality through observing their bio-signals. The R&K rules divide sleep into six stages, namely awake (W), rapid eye movement (REM), and four non-REM stages (N1, N2, N3, and N4), but the AASM standards merged N3 and N4 into N3, called slow wave sleep (SWS). The exhaustive manual classification work is not only time-consuming and labor-consuming but also subjective depending on trained experts.

Various prior machine learning methods have been attempted to classify sleep stages automatically and efficiently. Traditional machine learning methods, including random forests, and support vector machines, have been investigated in sleep stages classification for decades, and many other classifiers also demonstrated their high performance in this task. Despite their success in sleep stages classification, shallow learning algorithms normally extract features according to experts' knowledge, which means that the classification effectiveness is limited by feature engineering and feature selections. Even though experts have extracted features from time domain [8], [9], frequency domain [10], [11], and time-frequency domain [12], [13], it is still a huge challenge to find new and effective features to improve the classification performance of traditional classifiers. Deep learning algorithms, including convolutional neural networks (CNNs) [14], recurrent neural networks (RNNs) [15], [16], deep belief networks (DBNs) [17], and graph convolutional networks (GCN) [18] have been proposed for this shortcoming.

Based on different data representations, the inputs of CNNs are normally one-dimensional signals [14] or two-dimensional signals [19], and the convolutional operations are with one-dimension and two-dimension data, respectively. These two types of models aggregate temporal information in time series, while they ignore the intrinsic connections among different bio-signals. GCN models have thus been proposed for the shortcoming. Unlike CNNs and RNNs that require Euclidean data, the inputs to GCNs are non-Euclidean structures, which means that the functional connections among

Manuscript received 22 December 2022; revised 23 May 2023; accepted 17 August 2023. Date of publication 28 August 2023; date of current version 7 September 2023. *(Corresponding author: Xiaopeng Ji.)*

This work involved human subjects from publicly available databases. The authors confirm that all human subject research procedures and protocols are exempted from review board approval.

Xiaopeng Ji and Yan Li are with the School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia (e-mail: xiaopeng.ji@usq.edu.au; yan.li@usq.edu.au).

Peng Wen is with the School of Engineering, University of Southern Queensland, Toowoomba, QLD 4350, Australia (e-mail: paul.wen@usq.edu.au).

Digital Object Identifier 10.1109/TNSRE.2023.3309542

different channels and spatial correlations can be explored by GCN models automatically. However, due to their low computing efficiency, it would still have a long way to eventually apply GCNs for clinical diagnosis. Compared with CNNs and GCNs, 3D-CNNs not only can extract temporal features from raw data but also have the capacity to aggregate spatial information with less computing complexity. Transitional rules may also impact the final classification results, which means that they need to be considered during sleep stages identification to improve classification performance. However, few algorithms pay much attention to it, and sometimes this important factor is even ignored.

To tackle the challenges above, a 3D-CNN model based on a backbone from the spatial-spectral-temporal based attention 3D dense network [20] is proposed for automatic sleep staging. To explore relations among signals and frequency bands deeply, temporal features, frequency features, and temporal-frequency features are extracted and fed into this proposed model.

The main contributions of this paper are summarized as follows:

- A 3D-CNN and 2D-CNN mixed deep learning model named 3DSleepNet is proposed to classify sleep stages automatically. 3D convolutional operations are used to extract spatial-temporal features and spatial-spectral-temporal features from temporal inputs and temporal-frequency inputs, respectively. 2D convolutional operations are also utilized in the proposed model to extract spatial-spectral features from frequency inputs.
- A novel partial dot-product attention mechanism is designed for 3D convolutional operations to efficiently capture the most relevant information. A spatial-spectral attention mechanism is designed for 2D convolutional operations to capture the most relevant spatial-spectral information.
- To evaluate the classification performance on healthy and unhealthy subjects, the classification experiments were performed on ISRUC-S3 and 50 random subjects from ISRUC-S1 (https://sleeptight.isr.uc.pt/). The accuracy, F1-score, and Cohen's kappa on ISRUC-S3 are 0.832, 0.814, and 0.783, respectively, which indicates that the proposed model achieves a state-of-the-art performance. The overall accuracy, F1-score, and Cohen kappa on ISRUC-S1 (the datasets with sleep-disorder patients) achieved 0.820, 0.797, and 0.768, respectively, which also demonstrates its generality on unhealthy subjects. The training speed experiments on ISRUC-S3 show that the proposed model outperforms other GCN models and U^2-Net architecture models in terms of the model training time.
- The impact of the ratio of unhealthy and healthy subjects in the training set is explored using a set of mixed training data from ISRUC-S1 (unhealthy datasets) and ISRUC-S3 (healthy datasets). The experimental results show that the classification performance on unhealthy patients achieved the best when the training set consists of 100% abnormal patients.
- Incremental experiments are conducted on the ISRUC-S3 dataset to explore the effects of different model variants. The experimental results show that the proposed 3DSleepNet model achieves its best performance when the attention layers and a long short-term memory layer (LSTM) are added with all three input branches.

This paper is organized as follows: Section II introduces related works about sleep stages classification and 3D convolutional neural networks. In Section III, the details of the proposed models are illustrated. Experimental data, experimental setting, experimental results, and discussions are presented in Section IV. Finally, conclusions are drawn in Section V.

II. Related Work

A. Sleep Stages Classification

Automatic sleep stages classification algorithms have been studied for decades and many shallow machine learning classifiers have been reported in sleep scoring. Support vector machine (SVM) is one of the most widely used classifiers in many classification tasks and has demonstrated its high performance in identifying sleep stages. Zhu et al. [21] extracted graph domain features through a visibility graph similarity method to perform a five-state classification based on single-channel EEG. Naive Bayes [22], random forest (RF) [23], [24], complex networks [25], [26], and ensemble learning-based classifiers [27], [28], [29] also gave acceptable classification results. However, all these research methods are based on hand-crafted features, the classification performance and efficiency depend on feature engineering and researchers' understanding of data.

Compared to other shallow machine learning algorithms, deep learning methods not only allow experts to input original/raw data but also can extract higher-level features. Motivated by their breakthroughs in many fields, such as image recognition [30], [31] and natural language processing [32], many researchers have applied deep learning algorithms to process bio-signals. DeepSleepNet [14] is a two-step training CNN model with an BiLSTM layer for sleep stages classification. A representation learning module helps this model to capture both temporal information and frequency information through two CNNs with small and large filter sizes for the first layers. A sequence residual learning module is used to learn stage transition rules. Ji et al. [18] proposed a multi-channel-based graph convolutional network to perform five-stage classification, which achieved the best performance on ISRUC-S3 and ISRUC-S1. U-Nets and U^2-Nets are novel complex models with a multi-scale extraction module, which also gave acceptable results in sleep scoring tasks [33], [34], [35].

B. 3D Convolutional Neural Networks

2D-CNNs are normally utilized to recognize images [36], [37], [38] through 2D convolutional operations, which means that spatial features can be extracted efficiently. Compared with 2D-CNNs, 3D-CNNs can capture spatial features and temporal features simultaneously. Motivated by this characteristic, many researchers turn to build 3D-CNNs from time series data, such as human action recognition [39] and pose estimation [40]. In the bio-signal analysis area, some 3D-CNN

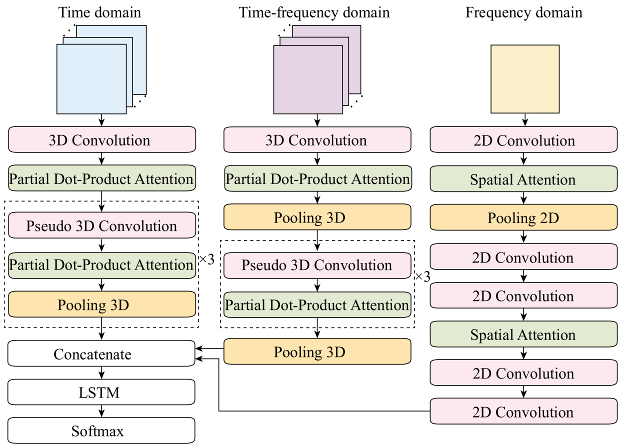

Fig. 1. The overall architecture of the 3DsleepNet. Time domain features, time-frequency domain features and frequency domain features are inputted into this model after feature extraction. Time domain features are down-sampled from raw/original signals. Time-frequency domain features come from short-term differential entropies and the frequency domain features come from power spectral densities.

models are also reported in emotion recognition [41], [42], [43], epileptic seizure prediction [44], and motor imagery analysis tasks [45], but there is still little study to apply 3D-CNNs for classifying sleep stages.

III. Methodology

Fig. 1 shows the overall architecture of the proposed model in this paper. The proposed model consists of one spatial-temporal stream, one spatial-spectral-temporal stream, and one spatial-spectral stream. The inputs of the spatial-temporal stream and spatial-spectral-temporal stream are 3D representations of raw/original signals in the time domain and time-frequency domain. The inputs of the spatial-spectral stream are 2D representations of raw signals in the frequency domain. In terms of the spatial-temporal stream and spatial-spectral-temporal stream, a partial dot-product attention mechanism is designed to help the proposed model to pay more attention to valuable information in the time series of each frequency band from each channel. A long-short term memory layer is added to learn transition rules among neighboring epochs for classification.

There are four key components in this proposed 3DSleepNet model: 1) A three-stream 3D-CNN model is designed for automatic sleep staging through time space, time-frequency space, and frequency space after extracting features from multi-channel bio-signals. 2) For each 3D-CNN stream, partial dot-product attention layers are added to help the proposed model to focus on more valuable information. 3) Pseudo 3D-CNN modules are used to decrease computing complexity. 4) A LSTM layer is added to learn transitional rules among neighboring epochs.

A. Feature Extraction and 3D Representation

Fig. 2 shows the procedure of feature extraction, where time domain features and time-frequency domain features are spatial-temporal and spatial-spectral-temporal 3D representations of bio-signals, respectively, and frequency domain features are 2D representations. The bio-signal of Nchannels are defined as $S = (s_1, s_2, \ldots, s_N) \in \mathbb{R}^{N\times L}$, where $s_i \in \mathbb{R}^{L}(i \in \{1, 2, \ldots, N\})$ is a channel of EEG, ECG or EMG signal with L data points. Before feature extraction, all channels to be used will be filtered by M bandpass filters and the filtered signals are defined as $S' = (s'_1, s'_2, \ldots, s'_N) \in \mathbb{R}^{N\times M\times L}$, where the $i-$th channel with M frequency band waves of L data points is represented by $s'_i \in \mathbb{R}^{M\times L}(i \in \{1, 2, \ldots, N\})$. In sleep stages classification tasks, $L-$length bio-signals are usually segmented into epochs of 30 seconds (s), and the filtered multi-channel signals in each epoch can be represented as $E = (e_1, e_2, \ldots, e_N) \in \mathbb{R}^{N\times M\times T}$, where e_i is the $i-$th channel with M frequency band waves of T data points in that epoch. To extract temporal features, the filtered signals E are down-sampled, and the length of signals changes from T to τ. The temporal features can be represented as $\chi_t = (x^1, x^2, \ldots, x^\tau) \in \mathbb{R}^{N\times M\times \tau}$, each $x^j \in \mathbb{R}^{N\times M}(j \in \{1, 2, \ldots, \tau\})$ is a 2D feature map of time step j. In terms of 3D representations of time-frequency features, the short-term differential entropy is calculated based on the filtered signals E. The spatial-spectral-temporal 3D representation of bio-signals is defined as $\chi_{tf} = (\hat{x}^1, \hat{x}^2, \ldots, \hat{x}^{\hat{\tau}}) \in \mathbb{R}^{N\times M\times \hat{\tau}}$, where $\hat{x}^k \in \mathbb{R}^{N\times M}(k \in \{1, 2, \ldots, \hat{\tau}\})$ is a 2D differential entropy feature map of time step k. The 2D representation of frequency domain features are defined by the power spectral density of M frequency band waves from all Nchannels of that epoch.

B. 3D Convolution

The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple temporal-contiguous 2D feature maps together [39]. As a result, comparing with 1D temporal convolution and 2D spatial convolution, 3D temporal-spatial convolution is more advantageous in both representing brain connections and their activities. The convolutional value at (x, y, z) on the $j-$th feature map in the $i-$th layer is given by [39]:

$$v_{ij}^{xyz} = \sigma\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

where P_i, Q_i and R_i are the size of the 3D kernels along the three dimensions. w_{ijm}^{pqr} is the $(p, q, r)-$th value of the kernel connected to the $m-$th feature map in the previous layer, and σ is an activation function. Fig. 3 shows an example of the 3D convolution on the extracted 3D features. The kernel size along the time dimension in the current layer is 3, which means that each convolutional value is decided by the 3D kernel and 3 contiguous 2D feature maps. This characteristics of aggregating spatial information and temporal information help the proposed model to capture valuable features comprehensively.

C. Pseudo-3D Convolution

Even though 3D convolutions are more advantageous in exploring temporal features and spatial features, the requirements of a large amount of computing resources limit its

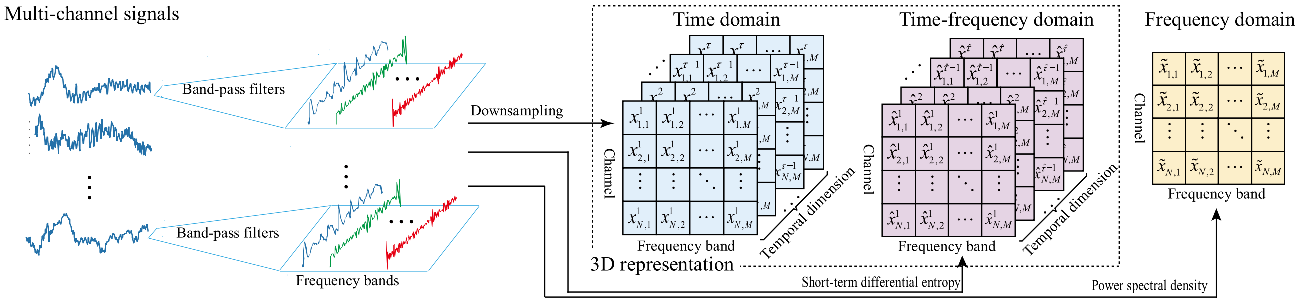

Fig. 2. Each channel is filtered by m filters to band waves. 3D temporal features are extracted through down-sampling from filtered band waves of n channel bio-signals. 3D temporal-frequency features are extracted by calculating the short-term differential entropy of each band waves from each channel. The frequency domain features are extracted by calculating the power spectral density of m band waves of n channels.

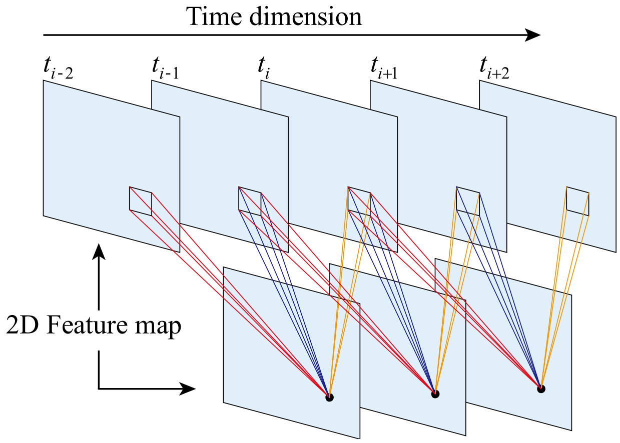

Fig. 3. An example of 3D convolution on the extracted 3D features. The kernel size is three in the time dimension.

applications. To tackle this problem, Pseudo-3D convolutions [46] are adopted in the proposed model to reduce the computational complexity. The kernel of the standard 3D convolutions are (P, Q, R), where P and Q can be seen as the kernel size of 2D spatial convolutions and R is the kernel size along the time dimension. In Pseudo-3D convolution, the kernel (P, Q, R) are decoupled into $P \times Q \times 1$ and $1 \times 1 \times R$, where $P \times Q \times 1$ represents convolutional filters equivalent to 2D CNN on spatial domain and $1 \times 1 \times R$ convolutional filters like 1D CNN tailored to the temporal domain. Hence, the output of a Pseudo-3D convolution module l can be defined as:

$$\chi^l = \Phi^{1\times 1\times R}(\Phi^{P\times Q\times 1}(\chi^{l-1})) \tag{2}$$

where χ^{l-1} is the output of $l-1$th layer, $\Phi^{P\times Q\times 1}$ and $\Phi^{1\times 1\times R}$ denote the 2D convolution on spatial domain with a kernel of $P \times Q$ and 1D convolution on temporal domain with a kernel of R, respectively.

D. Partial Dot-Product Attention

The attention mechanism is often utilized to automatically extract the most relevant information. In the proposed model, a simple but effective partial dot-product attention is designed to quantify the importance of input features, where higher weights are assigned to the most relevant information and lower weights are assigned to the less relevant information. For a given input $\chi \in \mathbb{R}^{N\times M\times T}$, the partial dot-product attention is computed as:

$$Att = \chi \otimes \sigma((\chi \cdot M_1) \cdot M_2 + b) \tag{3}$$

where $M_1 \in \mathbb{R}^{T\times M}$, $M_2 \in \mathbb{R}^{M\times T}$, $b \in \mathbb{R}^{N\times M\times T}$ are learnable parameters, $\cdot$ denotes dot-product, $\otimes$ refers to the point-wise multiplication, and σ is a softmax function.

IV. Experiments

A. Experimental Data and Experimental Settings

ISRUC-Sleep [47] is an open-source database, which consists of three subsets, namely, ISRUC-S1, ISRUC-S2, and ISRUC-S3. All signals were collected according to the international 10–20 standard electrode placement. The detailed information of subjects and the distribution of sleep stages are listed in TABLE I. Each recording consists of 19 channels, including, EOG, EEG, EMG, ECG, snore, body position, etc. The sampling rate of EOG, EEG, and EMG are 200Hz.

Classification experiments are conducted on ISRUC-S1 and ISRUC-S3 to evaluate the classification performance for both healthy cases and sleep-disorder patients and the evaluation measures, including accuracy (ACC), precision (PR), recall (RE), F1-score (F1), and Cohen's kappa (κ) are defined as below:

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

$$precision = \frac{TP}{TP+FP}\% \tag{5}$$

$$recall = \frac{TP}{TP+FN}\% \tag{6}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives and false negatives, respectively.

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{7}$$

where $precision$ and $recall$ are defined as in equation (5) and equation (6).

$$\kappa = 1 - \frac{1 - Accuracy}{1 - p_e} \tag{8}$$

where $Accuracy$ is defined as question (4), and p_e is the hypothetical probability of chance agreement calculated:

$$p_e = \frac{1}{N^2}\sum_k n_{k1}n_{k2} \tag{9}$$

TABLE I
THE COMPREHENSIVE INFORMATION OF ISRUC DATASETS

Subset	Subject Number	Age	Sex ratio (Male: Female)	Sleep health conditions	Distribution of each sleep stage					
					Wake	N1	N2	N3	REM	Total
ISRUC-S1	100	51±16	55:45	Unhealthy	20098	11062	27511	17251	11265	87187
ISRUC-S2	8	46.8±18.8	6:2	Unhealthy	2282	2211	5042	2609	2063	14207
ISRUC-S3	10	40±10	9:1	Healthy	1651	1215	2609	2014	1060	8549

TABLE II
DETAILED INFORMATION OF CHANNELS USED IN EXPERIMENTS

Signal type	Label	Description
EOG	LOC-A2	Left eyes movements.
	ROC-A1	Right eyes movements.
EEG	F3-A2	A1 and A2 are placed in the left and right ear-lobes.
	C3-A2	
	O1-A2	
	F4-A1	
	C4-A1	
	O2-A1	
Chin EMG	X1	Placed between the chin and the lower lip.

where N is the total number of samples, and k is the number of categories, and n_{ki} is the number of times rater i predicted category k.

Incremental experiments and executing efficiency experiments are also carried out on ISRUC-S3 to explore the contribution of each module in the proposed model and computing complexity, respectively. Moreover, the ISRUC-S1 data are also used to test the generality of the proposed method on unhealthy patients. The code will be uploaded on Github (https://github.com/XiaopengJi-USQ/3DSleepNet) once the paper is published.

All these experiments are carried out on a computer with an Intel I9-12900KF CPU, 128 GB memory, and an Nvidia 3090 GPU.

B. Preprocessing and Feature Extraction

All PSGs were pre-processed by the data provider: 1) A notch filter was used to eliminate the 50 Hz electrical noise; 2) For EEG and EOG data, a bandpass Butterworth filter was utilized to obtain waves from 0.3 Hz to 35 Hz. The EMG data were filtered by a lower cutoff frequency of 10 Hz and a higher cutoff frequency of 70 Hz. 3) The last 30 epochs of each subject were removed due to noise.

The detailed information of channels used to extract features in all experiments is listed in TABLE II

In terms of time domain feature extraction, the features are obtained through down-sampling from 200 Hz to 10 Hz for each selected channel. According to our experiments, the excessive down-sampling in time series has little negative effects on the classification results, but the training time is substantially reduced. To extract time-frequency domain features, the differential entropy (STDE) of 9 crossed frequency bands: 0.5-4 Hz, 2-6 Hz, 4-8 Hz, 6-11 Hz, 8-14 Hz, 11-22 Hz, 14-31 Hz, 22-40 Hz and 31-49 Hz are calculated for each channel. These time-frequency features are obtained from filtered bio-signals with a 200 Hz sampling rate and the window size of the short-term differential entropy is set to 3s, which means that there are 10 feature maps in each epoch. The frequency domain features are obtained by computing the power spectral density of each epoch with 200 Hz sampling.

C. Comparison With the State-of-the-Art Methods

We compare the proposed model with traditional machine learning methods, including SVM [48], Random Forest [24], and Multilayer Perceptron neural network [16] on both ISRUC-S1 and ISRUC-S3 datasets. Ensemble classifiers, like Bootstrap aggregating (Bagging) [27], boosting [28], eXtreme Gradient Boosting (XGBoost) [29] are also evaluated on these two subsets. Multi-types deep learning algorithms with different architectures, such as CNNs [14], [49], U^2-Net [35], and GCNs [18], [50], [51] are included for comparisons purpose as well. For a fair comparison, all models were reproduced based on our hardware computational environments, except the JK-STGCN model, which we have all results in our previous work [18] and subject-independent cross-validation strategy were adopted on both of the subsets.

TABLE III shows the comparison results on ISRUC-S3 and random selected 50 unhealthy subjects from ISRUC-S1 subsets. 10-fold cross-validation and 25-fold cross-validation are carried out on ISRUC-S3 and 50 unhealthy subjects from ISRUC-S1, respectively, to test the classification performance on healthy subjects and unhealthy cases.

The classification performances of the SVM model and the RF model are the lowest among all these methodologies, even more features are extracted and selected by researchers. There are two reasons leading to this phenomenon. On the one hand, the inefficient extracted features cannot represent the original data comprehensively. On the other hand, the classification results are limited by the performance of the classifier itself, which are improved by ensemble methods.

Compared to shallow classifiers, 1D-dimension convolution models, including TinySleepNet [49] and DeepSleepNet [14], not only can input original data without feature engineering but also improve the classification performance. There are two reasons to explain this improvement. CNNs can extract high-level features and aggregate temporal information in continuous time series, which means that the correlation between previous and next data points can be learned comprehensively. As a result, the classification accuracy increases. Moreover, LSTM layers are added to learn transition rules, which help these models to further improve classification performance. However, the 1D-dimension convolution only can focus on one-dimension data, the spatial connections among

TABLE III
COMPARISON AMONG 3DSLEEPNET AND OTHER METHODS ON ISRUC S3 AND 50 RANDOM SUBJECTS FROM ISRUC S1 SUBGROUP

Subset		Method	Overall Metrics			Per class F1 score (F1)				
			ACC	F1	κ	W	N1	N2	N3	REM
ISRUC-S3	Alickovic et al.	SVM	0.714	0.672	0.626	0.824	0.428	0.724	0.815	0.569
	Memar et al.	RF	0.702	0.685	0.616	0.838	0.470	0.671	0.763	0.684
	Dong et al.	MLP+LSTM	0.751	0.708	0.675	0.852	0.378	0.758	0.872	0.682
	Hassan & Subasi	Bagging	0.740	0.706	0.662	0.847	0.465	0.751	0.843	0.625
	Hassan & Bhuiyan	Boosting	0.714	0.681	0.625	0.821	0.458	0.725	0.825	0.576
	Liu et al.	XGBoost	0.749	0.722	0.676	0.866	0.481	0.751	0.848	0.666
	Supratak et al.	CNN+BiLSTM	0.719	0.696	0.643	0.831	0.463	0.742	0.851	0.595
	Supratak & Guo	CNN+LSTM	0.753	0.737	0.682	0.809	0.533	0.758	0.851	0.734
	Jia et al.	U^2-Net	0.807	0.791	0.751	0.867	0.581	0.808	0.895	0.805
	Jia et al.	STGCN	0.786	0.770	0.724	0.864	0.540	0.782	0.869	0.793
	Jia et al.	MSTGCN	0.818	0.803	0.765	0.898	0.581	0.808	0.880	**0.848**
	Ji et al.	JK-STGCN	0.831	0.814	0.782	**0.900**	**0.598**	0.826	0.901	0.845
	This study	3D-CNN	**0.832**	**0.814**	**0.783**	0.896	0.596	**0.832**	**0.909**	0.838
ISRUC-S1	Alickovic et al	SVM	0.684	0.608	0.583	0.793	0.242	0.708	0.808	0.490
	Memar et al.	RF	0.699	0.649	0.607	0.841	0.307	0.705	0.750	0.640
	Dong et al.	MLP+LSTM	0.703	0.648	0.614	0.807	0.301	0.724	0.817	0.591
	Hassan & Subasi	Bagging	0.693	0.621	0.595	0.813	0.248	0.715	0.798	0.529
	Hassan & Bhuiyan	Boosting	0.663	0.614	0.555	0.789	0.325	0.687	0.766	0.504
	Liu et al.	XGBoost	0.736	0.691	0.657	0.866	0.372	0.742	0.835	0.638
	Supratak et al.	CNN+BiLSTM	0.730	0.691	0.654	0.850	0.385	0.739	0.830	0.648
	Supratak & Guo	CNN+LSTM	0.764	0.745	0.695	0.846	0.548	0.729	0.830	0.794
	Jia et al.	U^2-Net	0.816	**0.800**	0.764	0.903	**0.577**	0.801	**0.886**	0.832
	Jia et al.	STGCN	0.780	0.751	0.715	0.889	0.463	0.763	0.825	0.813
	Jia et al.	MSTGCN	0.808	0.787	0.752	0.885	0.539	0.799	0.876	0.838
	Ji et al.	JK-STGCN	0.820	0.798	0.767	0.895	0.550	**0.811**	0.883	0.850
	This study	3D-CNN	**0.820**	0.797	**0.768**	**0.908**	0.534	0.808	0.880	**0.855**

* W=awake. N1, N2 and N3 are sleep stage 1, 2, 3, separately, and are non-rapid eye movement. REM= rapid eye movement.

different brain regions are ignored consequently, which finally limits their performance. Even though pure 1D-dimension convolution models cannot reach a very high classification accuracy, U^2-Net-based models using 1D-dimension convolution layers still improve the performance slightly through their complex architecture and exponentially higher computational resources. GCN classifiers including a GraphSleepNet classifier, an MSTGCN classifier, and a JK-STGCN classifier, can extract spatial features and temporal features efficiently and this characteristic ensures their high performance in sleep stages classification tasks. Compared with the algorithms above, the 3DSleepNet model not only requires fewer crafted features than the traditional machine learning methods but also can capture spatial-temporal information more effectively than CNN, U^2-Net, and GCN models. The classification results on ISRUC-S3 and random subjects from ISRUC-S1 demonstrate that the 3DSleepNet model not only can classify sleep stages with high accuracy for healthy subjects but also have good generality on unhealthy cases.

Fig. 4 shows the confusion matrix of classifications results obtained from all compared models on ISRUC-S3. For each model, the performance of classifying stages of Wake, N2, N3, and REM have higher results than N1 stage, but some minor samples are still misclassified into other classes. Compared with multi-channel-based models, single-channel-based classifiers have lower REM accuracy. One explanation is that EOG channels help to classify REM stages. As a result, algorithms using single EEG channels without EOG signals fail to classify REM stages correctly. The N1 stage has the lowest classification results, and some samples are incorrectly classified into Wake, N2, and REM. Due to the fact that slow eye movements also make great contribution in classifying N1, all multi-channel-based models have better results for N1 than single-channel-based classifiers. TABLE IV shows the comparison results among several deep learning methods [14], [34], [35], [49], [52], [53], [54] on other public datasets [55], [56], [57]. Compared with these algorithms, the accuracy of the proposed model on ISRUC-S3 is 0.832 which stays similar level, but N1 stages and N3 stages of the 3DSleepNet model outperforms other models.

In practice, when the classification performance of the model is similar, the one taking a shorter training time will have the advantage of capturing the market. According to TABLE III, the 3DSleepNet model and the JK-STGCN model can achieve the top 2 classification accuracy on both ISRUC-S3 and ISRUC-S1, where the MSTGCN model and

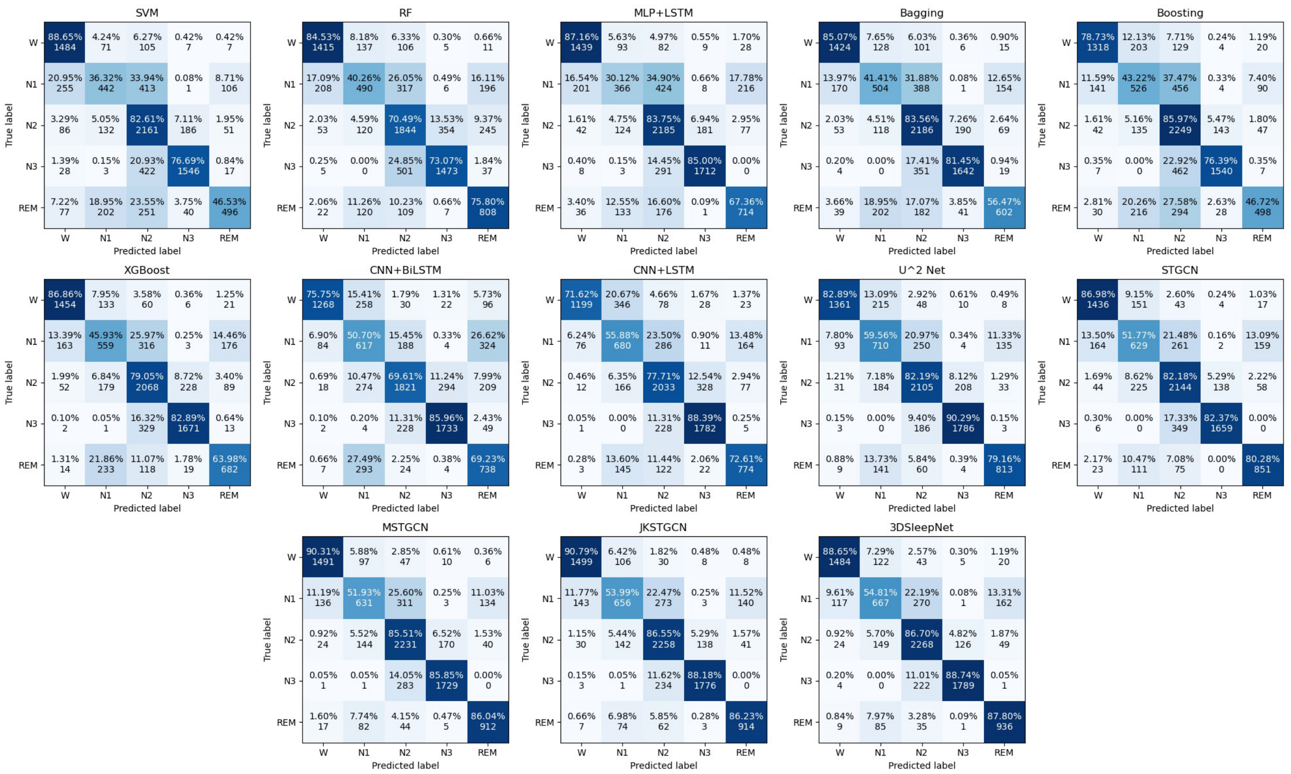

Fig. 4. Confusion matrix from all the compared models on ISRUC-S3 subset.

TABLE IV
COMPARISON RESULTS AMONG SEVERAL DEEP LEARNING METHODS BASED ON OTHER PUBLIC DATASETS

Methods	datasets	Overall Metrics			Per class F1 score (F1)				
		ACC	F1	κ	W	N1	N2	N3	REM
DeepSleepNet	Sleep-EDF-v1	0.820	0.769	0.760	0.847	0.466	0.859	0.848	0.824
SleepEEGNet	Sleep-EDF-v1	0.843	0.797	0.790	0.892	0.522	0.868	0.851	0.850
TinySleepNet	Sleep-EDF-v1	0.854	0.805	0.800	0.901	0.514	0.885	0.883	0.843
SeqSleepNet	Sleep-EDF-153	0.838	0.782	-	0.928	0.489	0.854	0.786	0.851
SleepUtime	Sleep-EDF-153	-	0.76	-	0.920	0.510	0.840	0.750	0.800
SalientSleepNet	Sleep-EDF-153	0.841	0.795	-	0.933	0.542	0.858	0.783	0.858
CNN+GRU	SHHS1-700	0.832	-	0.760	0.897	0.311	0.850	0.781	0.808
This study	ISRUC-S3	0.832	0.814	0.783	0.896	0.596	0.832	0.909	0.838

* W=awake. N1, N2 and N3 are sleep stage 1, 2, 3, separately, and are non-rapid eye movement. REM= rapid eye movement.

the SalientSleepNet model also perform very well on ISRUC-S3 or ISRUC-S1 subsets. Therefore, these four models are selected to compare the training efficiency. Considering the architecture and parameters of all these four models are totally different, in order to fairly compare execution efficiency, parameters including batch size, learning epochs, etc. are set the same as those in [18], [35], and [51]. Fig. 5 shows the training time comparison among the top 3 models on ISRUC-S3 or ISRUC-S1. The MSTGCN and JK-STGCN models take the least time to train the classification, but their feature extraction steps take more time, which leads to a higher overall training time compared with the proposed model. Due to its complex architecture, the SalientNet model takes the longest time to complete the training tasks.

D. Model Analysis

According to TABLE III, the proposed 3DsleepNet model can have a good capacity to identify sleep stages for healthy and unhealthy subjects/patients. But it also implies that there are still two important factors, namely the proportion of unhealthy patients and the size of training sets, which can affect the generality on unhealthy patients. To explore the influence of these two factors, two more experiments are carried out on the ISRUC-S3 and ISRUC-S1 datasets.

1) Influence of the Ratio of the Unhealthy Patients in the Training Set: For a fair comparison of the influence of the ratio of unhealthy patients in the training set, all comparison experiments are conducted on a fixed testing set, which consists of ten random subjects from the ISRUC-S1 dataset. The training set initially consists of 10 healthy subjects from the ISRUC-S3 dataset, and the ratio of unhealthy patients is changed by replacing a random healthy subject in it with a random data sample from the remaining 90 patients of the ISRUC-S1 dataset.

It can be seen from Fig. 6 that the classification performance is very low at the beginning when all data in the training set are healthy samples. The accuracy, F1-score and Cohen's kappa only achieve 0.53, 0.47, and 0.38, respectively. However, the classification accuracy increases noticeably and achieves 0.73 when there are 10% unhealthy samples. The F1-score and Cohen's kappa results are also improved a lot and achieve 0.68, and 0.64, respectively. After a steady rising, the classification performance trends keep improving with a slow growth rate, even though there are two decreasing points. Finally, when all data samples are from sleep-disorder cases, the classification performance reaches the highest, with the accuracy, F1-score, and Cohen's kappa are 0.78, 0.73, and 0.69, respectively. Since the distribution of sleep stages from ISRUC-S1 and ISRUC-S3 is totally different, it is reasonable and acceptable that the performance is very low at the beginning, when the model learns normal bio-signals and transitional rules from healthy subjects and has little knowledge about abnormal features. However, it starts to recognize the abnormal features or patterns when an unhealthy subject is added to the training set, and this leads to the improvement of the classification accuracy. The trend of the slow growth of accuracy keeps until all training data are from unhealthy patients.

Under the same size of the training set, the classification results with unhealthy patients are lower than those from healthy subjects. This phenomenon is also caused by the different distribution of sleep stages from ISRUC-S1 and ISRUC-S3, which means that the transitional rules and signal characteristics are quantitatively and complexly higher than those of healthy cases. As a result, models cannot learn features comprehensively with a small dataset, which leads to lower classification performance.

2) Influence of Training Data Size: To explore whether an increased dataset size may increase the classification performance on sleep-disorder cases, an extra experiment is carried out.

The ISRUC-S1 dataset is divided into four disjoint subgroups, and each subgroup contains 10, 20, 30, and 40 patients, respectively. For each subgroup, the Leave-One-Out cross-validation is conducted to validate the influence of the train data size on the proposed model.

Fig. 7 shows the experimental results on the effects of the train set. The classification accuracy, F1-score, and Cohen's kappa on unhealthy patients are 0.79, 0.75, and 0.73, respectively, which are the lowest. However, the classification measurements are improved when the size of the training set increases and finally achieves 0.82, 0.81, and 0.76, respectively.

This experiment also demonstrates that a big training data set size can weaken the negative impacts on classification results from abnormal bio-signals and abnormal transitional rules.

3) Incremental Comparison Experiments: To further investigate the influence of each module used in the proposed model, eight variant models are designed and evaluated using the ISRUC-S3 database. The details are described below:

1. *variant a (basic model)*: The basic model is a one-branch 3D-CNN model of a temporal stream without any attention layer or LSTM layer.

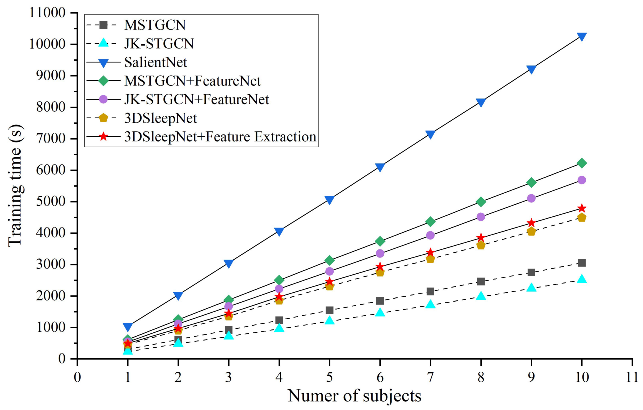

Fig. 5. Training time for the top 3 models on ISRUC-S3 or ISRUC-S1 based on 10-fold training.

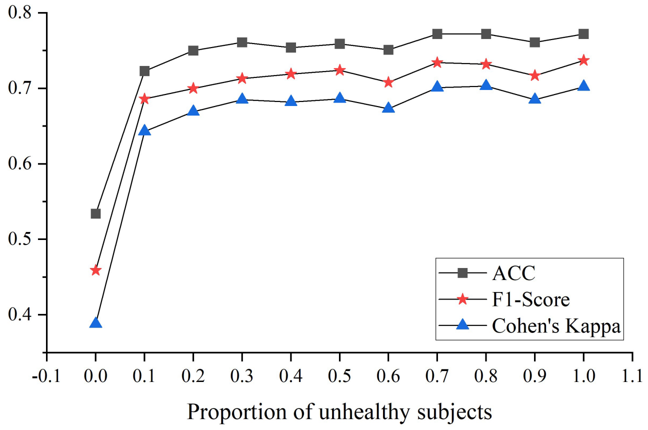

Fig. 6. The trend of classification results with different ratio of unhealthy patients in the training data set.

2. *variant b (basic model + temporal-frequency stream):* A temporal-frequency stream is added to construct a two-stream 3D-CNN model without any attention layer or LSTM layer.

3. *variant c (variant b + frequency stream):* A frequency stream is added to *variant b* to construct a three-stream 3D-CNN model without any attention layer or LSTM layer

4. *variant d (variant c + partial dot-product attention):* A partial dot-product attention is added to quantify the importance of input features.

5. *variant e (variant d + LSTM):* A LSTM layer is added to learn transitional rules among neighboring epochs.

6. *variant f (variant c + spatial-temporal attention + LSTM):* The partial dot-product attention layer is replaced by a spatial-temporal attention layer to indicate the importance of different channels and different sleep epochs. An LSTM layer is added to learn transitional rules among neighboring epochs.

7. *variant g (variant c + self-attention + LSTM):* The partial dot-product attention layer is replaced by a self-attention layer to indicate the interdependence within input features.

8. *variant h (using EEG channels only + the complete model e):* Features are extracted from six EEG signals to test the importance of the complementary channels to the overall performance.

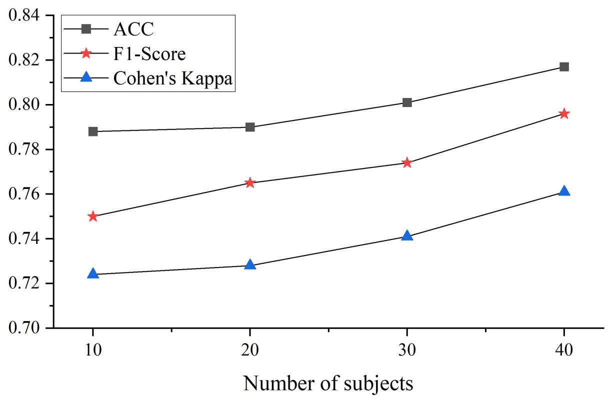

Fig. 7. The classification results of 3DSleepNet on four disjoint groups from ISRUC-S1.

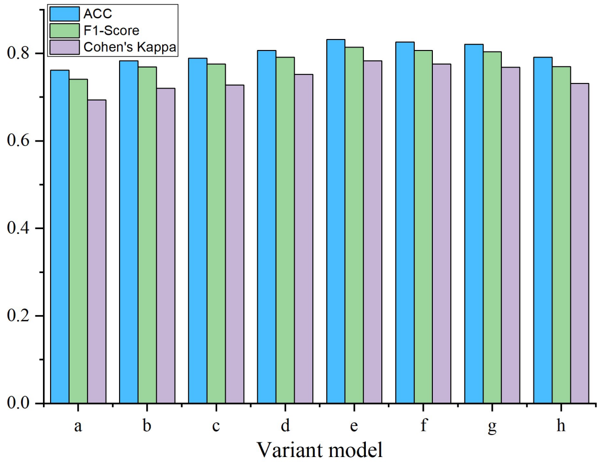

Fig. 8. Comparison of the designed variant models.

Fig. 8 illustrates the classification performance of eight variant models on ISRUC-S3. The basic model which inputs the 3D temporal features has the lowest performance, but all measurements increase when time-frequency features and frequency features are fed into the 3D-CNN model. The reason for this improvement is that the 3D-CNN model only learns the correlation among signals and frequency bands in time series but lacks the knowledge of time-frequency and frequency of signals which plays an important role in sleep stages identification. As a result, the classification accuracy is not very high at the beginning but increases with the added time-frequency features and frequency features. The performance further improves when partial dot-product attention layers and an LSTM layer are added. In terms of the partial dot-product attention layers, they indicate the most important frequency bands in each channel at each epoch and this helps to capture the most relevant information. The LSTM layer helps the proposed model learn the transitional rules among neighboring epochs, which also improves the classification performance. The spatial-temporal attention mechanism shrinks the input features through one dimension which loses more information than the partial dot-product attention mechanism and this leads to the ineffective quantifying of the importance of input features. The self-attention pays more attention role to the correlation among inner elements but the excessive down-sampling in time series weakens the connection among inner elements, which makes it underperform in this sleep stages classification task. To further evaluate the contribution of different channels in the classification performance, the 3DSleepNet model with six EEG channels inputs is tested. Compared to the complete model with all channels, the accuracy, F1-score, and Cohen's kappa of the model only using EEG signals, achieve 0.791, 0.770 and 0.731, respectively, which decrease heavily from 0.832, 0.814 and 0.783, respectively. Lacking EMG and EOG inputs leads to the result that the classifier fails to classify REM and N1 correctly, and these misclassifications also decline the performance of other stages.

V. Conclusion

In this study, a 3D-CNN based sleep stages classification model named as 3DSleepNet is proposed. The 3DSleepNet consists of two 3D-CNN streams and one 2D-CNN stream. The inputs of 3D-CNN streams are time domain features, and time-frequency domain features, and the inputs of the 2D-CNN stream are frequency domain features. For 3D-CNN branches, Pseudo-3D convolution layers are utilized to decrease the computing complexity and partial dot-product attention layers are designed to help the proposed model pay attention to valuable information. After the fusion layer of three streams, an LSTM layer is used to learn the transitional rules among neighboring epochs. Compared with the best results reported by other models, the 3D-CNN model also can achieve competitive performance on both healthy and unhealthy datasets with less computational demand. Based on the classification results, two more factors that may impact the performance are further explored. The experimental results indicate that the poor classification performance on unhealthy cases, which are caused by abnormal bio-signals, can be improved limitedly by increasing the proportion of sleep-disorder patients in the training set or increasing the number of the training data. An incremental experiment is also conducted to identify the contributions of each model variant and several different attention layers are tested to find the best one. A limitation of the proposed model is that the multi-channel-based algorithm requires large storage and memory for computation. However with modern computer hardware technology, this shouldn't be a main problem for its applications and deployment on edge artificial intelligence devices. In the future, we will improve and explore a new 3D representation of a single-channel signal, such as one EEG channel, so that the storage requirements and computing complexity can be further decreased.

References

[1] S. Siuly and Y. Li, "Discriminating the brain activities for brain–computer interface applications through the optimal allocation-based approach," *Neural Comput. Appl.*, vol. 26, no. 4, pp. 799–811, May 2015, doi: 10.1007/s00521-014-1753-3.

[2] T. Nguyen-Ky, P. Wen, and Y. Li, "Consciousness and depth of anesthesia assessment based on Bayesian analysis of EEG signals," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1488–1498, Jun. 2013, doi: 10.1109/TBME.2012.2236649.

[3] T. Nguyen-Ky, P. Wen, Y. Li, and R. Gray, "Measuring and reflecting depth of anesthesia using wavelet and power spectral density," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 4, pp. 630–639, Jul. 2011, doi: 10.1109/TITB.2011.2155081.

[4] N. A. Siuly, Y. Li, and P. Wen, "Identification of motor imagery tasks through CC-LR algorithm in brain computer interface," *Int. J. Bioinf. Res. Appl.*, vol. 9, no. 2, p. 156, 2013, doi: 10.1504/IJBRA.2013.052447.

[5] M. Li, W. Chen, and T. Zhang, "Automatic epilepsy detection using wavelet-based nonlinear analysis and optimized SVM," *Biocybern. Biomed. Eng.*, vol. 36, no. 4, pp. 708–718, 2016, doi: 10.1016/j.bbe.2016.07.004.

[6] E. A. Wolpert, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Arch. Gen. Psychiatry*, vol. 20, no. 2, pp. 246–247, Feb. 1969, doi: 10.1001/archpsyc.1969.01740140118016.

[7] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. V. Vaughn, "The AASM manual for the scoring of sleep and associated events," Rules Terminol. Tech. Specifications, Amer. Acad. Sleep Med., Darien, IL, USA, 2012, vol. 176.

[8] R. Sharma, R. B. Pachori, and A. Upadhyay, "Automatic sleep stages classification based on iterative filtering of electroencephalogram signals," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2959–2978, Oct. 2017, doi: 10.1007/s00521-017-2919-6.

[9] M. Diykh, Y. Li, and P. Wen, "EEG sleep stages classification based on time domain features and structural graph similarity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 11, pp. 1159–1168, Nov. 2016, doi: 10.1109/TNSRE.2016.2552539.

[10] A. Stochholm, K. Mikkelsen, and P. Kidmose, "Automatic sleep stage classification using ear-EEG," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 4751–4754.

[11] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, and F. Chapotot, "Feature selection for sleep/wake stages classification using data driven methods," *Biomed. Signal Process. Control*, vol. 2, no. 3, pp. 171–179, Jul. 2007, doi: 10.1016/j.bspc.2007.05.005.

[12] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Ann. Biomed. Eng.*, vol. 44, no. 5, pp. 1587–1597, May 2016, doi: 10.1007/s10439-015-1444-y.

[13] W. Al-Salman, Y. Li, and P. Wen, "Detection of EEG K-complexes using fractal dimension of time frequency images technique coupled with undirected graph features," *Frontiers Neuroinform.*, vol. 13, p. 45, Jun. 2019.

[14] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017, doi: 10.1109/TNSRE.2017.2721116.

[15] E. Bresch, U. Großekathöfer, and G. Garcia-Molina, "Recurrent deep neural networks for real-time sleep stage classification from single channel EEG," *Frontiers Comput. Neurosci.*, vol. 12, p. 85, Oct. 2018, doi: 10.3389/fncom.2018.00085.

[16] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018, doi: 10.1109/TNSRE.2017.2733220.

[17] I. N. Yulita, M. I. Fanany, and A. M. Arymuthy, "Bi-directional long short-term memory using quantized data of deep belief networks for sleep stage classification," *Proc. Comput. Sci.*, vol. 116, pp. 530–538, Jan. 2017, doi: 10.1016/j.procs.2017.10.042.

[18] X. Ji, Y. Li, and P. Wen, "Jumping knowledge based spatial–temporal graph convolutional networks for automatic sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1464–1472, 2022, doi: 10.1109/TNSRE.2022.3176004.

[19] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018, doi: 10.1109/TNSRE.2018.2813138.

[20] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial–spectral–temporal based attention 3D dense network for EEG emotion recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2909–2917.

[21] G. Zhu, Y. Li, and P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 6, pp. 1813–1821, Nov. 2014, doi: 10.1109/JBHI.2014.2303991.

[22] A. R. Hassan and M. I. H. Bhuiyan, "Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating," *Biomed. Signal Process. Control*, vol. 24, pp. 1–10, Feb. 2016, doi: 10.1016/j.bspc.2015.09.002.

[23] 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," *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 10–19, Oct. 2012, doi: 10.1016/j.cmpb.2011.11.005.

[24] P. Memar and F. Faradji, "A novel multi-class EEG-based sleep stage classification system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 84–95, Jan. 2018, doi: 10.1109/TNSRE.2017.2776149.

[25] M. Diykh, Y. Li, and S. Abdulla, "EEG sleep stages identification based on weighted undirected complex networks," *Comput. Methods Programs Biomed.*, vol. 184, Feb. 2020, Art. no. 105116, doi: 10.1016/j.cmpb.2019.105116.

[26] M. Diykh and Y. Li, "Complex networks approach for EEG signal sleep stages classification," *Expert Syst. Appl.*, vol. 63, pp. 241–248, Nov. 2016, doi: 10.1016/j.eswa.2016.07.004.

[27] A. R. Hassan and A. Subasi, "A decision support system for automated identification of sleep stages from single-channel EEG signals," *Knowl.-Based Syst.*, vol. 128, pp. 115–124, Jul. 2017, doi: 10.1016/j.knosys.2017.05.005.

[28] A. R. Hassan and M. I. H. Bhuiyan, "Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting," *Comput. Methods Programs Biomed.*, vol. 140, pp. 201–210, Mar. 2017, doi: 10.1016/j.cmpb.2016.12.015.

[29] C. Liu et al., "Automatic sleep staging with a single-channel EEG based on ensemble empirical mode decomposition," *Phys. A, Stat. Mech. Appl.*, vol. 567, Apr. 2021, Art. no. 125685, doi: 10.1016/j.physa.2020.125685.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11. Accessed: May 19, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[33] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: Resilient high-frequency sleep staging," *NPJ Digit. Med.*, vol. 4, no. 1, Apr. 2021, Art. no. 72, doi: 10.1038/s41746-021-00440-5.

[34] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," Oct. 2019, *arXiv:1910.11162*.

[35] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," May 2021, *arXiv:2105.13864*.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2014, *arXiv:1409.1556*. Accessed: Jul. 20, 2022.

[37] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2018, pp. 169–175, doi: 10.1109/CCWC.2018.8301729.

[38] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional neural network (CNN) for image detection and recognition," in *Proc. 1st Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, Dec. 2018, pp. 278–282, doi: 10.1109/ICSCCC.2018.8703316.

[39] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[40] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Real-time 3D hand pose estimation with 3D convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 956–970, Apr. 2019, doi: 10.1109/TPAMI.2018.2827052.

[41] Y. Wang, Z. Huang, B. McCane, and P. Neo, "EmotioNet: A 3-D convolutional neural network for EEG-based emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7, doi: 10.1109/IJCNN.2018.8489715.

[42] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby, "EEG-based emotion recognition using 3D convolutional neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 329–337, 2018.

[43] Y. Zhao, J. Yang, J. Lin, D. Yu, and X. Cao, "A 3D convolutional neural network for emotion recognition based on EEG signals," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[44] Z. Wang, J. Yang, and M. Sawan, "A novel multi-scale dilated 3D CNN for epileptic seizure prediction," in *Proc. IEEE 3rd Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2021, pp. 1–4, doi: 10.1109/AICAS51828.2021.9458571.

[45] T. Liu and D. Yang, "A densely connected multi-branch 3D convolutional neural network for motor imagery EEG decoding," *Brain Sci.*, vol. 11, no. 2, p. 197, Feb. 2021, doi: 10.3390/brainsci11020197.

[46] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542, doi: 10.1109/ICCV.2017.590.

[47] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, Feb. 2016, doi: 10.1016/j.cmpb.2015.10.013.

[48] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018, doi: 10.1109/TIM.2018.2799059.

[49] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 641–644, doi: 10.1109/EMBC44109.2020.9176741.

[50] Z. Jia et al., "GraphSleepNet: Adaptive spatial–temporal graph convolutional networks for sleep stage classification," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1324–1330.

[51] Z. Jia et al., "Multi-view spatial–temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1977–1986, 2021.

[52] 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, May 2019, Art. no. e0216456, doi: 10.1371/journal.pone.0216456.

[53] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019, doi: 10.1109/TNSRE.2019.2896659.

[54] X. Shao and C. S. Kim, "A hybrid deep learning scheme for multi-channel sleep stage classification," *Comput., Mater. Continua*, vol. 71, no. 1, pp. 889–905, 2022, doi: 10.32604/cmc.2022.021830.

[55] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.

[56] S. F. Quan et al., "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, Dec. 1997.

[57] G.-Q. Zhang et al., "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018, doi: 10.1093/jamia/ocy064.