A Novel Sleep Staging Method Based on EEG and ECG Multimodal Features Combination

Juntong Lyu, Wenbin Shi, *Member, IEEE*, Chuting Zhang, *Student Member, IEEE*, and Chien-Hung Yeh, *Senior Member, IEEE*

***Abstract*— Accurate sleep staging evaluates the quality of sleep, supporting the clinical diagnosis and intervention of sleep disorders and related diseases. Although previous attempts to classify sleep stages have achieved high classification performance, little attention has been paid to integrating the rich information in brain and heart dynamics during sleep for sleep staging. In this study, we propose a generalized EEG and ECG multimodal feature combination to classify sleep stages with high efficiency and accuracy. Briefly, a hybrid features combination in terms of multiscale entropy and intrinsic mode function are used to reflect nonlinear dynamics in multichannel EEGs, along with heart rate variability measures over time/frequency domains, and sample entropy across scales are applied for ECGs. For both the max-relevance and min-redundancy method and principal component analysis were used for dimensionality reduction. The selected features were classified by four traditional machine learning classifiers. Macro-F1 score, macro-geometric mean, and Cohen kappa value are adopted to evaluate the classification performance of each class in an imbalanced dataset. Experimental results show that EEG features contribute more to wake stage classification while ECG features contribute more to deep sleep stages. The proposed combination achieves the highest accuracy of 84.3% and the highest kappa value of 0.794 on the support vector machine in the ISRUC-S3 dataset, suggesting the proposed multimodal features combination is promising in accuracy and efficiency compared to other state-of-the-art methods.**

***Index Terms*— Sleep stage classification, multiscale entropy analysis, empirical mode decomposition, multimodal, heart rate variability.**

I. Introduction

SLEEP takes up the majority of our lifetime. Well-organized sleep at night ensures the well-function of the body and matters in maintaining physical and mental health [1], [2], [3]. Inversely, insufficient/ineffective sleep deteriorates cognition, learning, and memory [4]. As the population suffering from sleep disorders is on the rise, developing multimodal sleep staging techniques (e.g., light sleep and deep sleep) are of special importance to monitoring health in the immune system, memory, metabolism, etc. [5], [6], [7].

Polysomnography (PSG) is commonly manually scored by sleep experts and has multiple types of physiological signals including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and electrocardiogram (ECG) [8]. Typically, PSG recordings are divided into 30-s epochs, and each epoch is manually scored by sleep experts with five stages being classified, i.e., awake (W), rapid eye movement (REM), and non-REM stages (N1, N2, and N3) [9]. Such a manual scoring process is quite time-consuming and highly depends on personal clinical experience. Thus, to address the above problems, it is an urgent need to develop automatic methods for sleep classification tasks.

II. Related Works and Preliminary

A. Related Works

Automatic sleep staging methods or systems can greatly improve the efficiency and accuracy of traditional sleep scoring and monitoring, meanwhile to further support the diagnosis of sleep disorders. Many studies have attempted to develop automated sleep stage classification systems or methods based on cutting-edge machine learning technologies, which extract features from physiological signals, especially the EEG signal. Recently, deep learning methods have drawn much attention to the field owing to their ability in extracting abstract features automatically. For instance, Jia et al. [10] designed a multi-view spatial-temporal graph convolutional network (MSTGCN) with domain generalization for sleep stage classification, utilizing time-varying spatial and temporal features from multichannel brain signals, and achieved an accuracy of 89.5% and 82.1% in the MASS-SS3 and ISRUC-S3

Manuscript received 4 April 2023; revised 3 August 2023; accepted 17 September 2023. Date of publication 11 October 2023; date of current version 20 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171028 and Grant 62001026, in part by the Open Project of Key Laboratory of Medical Electronics and Digital Health of Zhejiang Province under Grant MEDH202204 and Grant MEDC202303, and in part by the Beijing Institute of Technology High-Level Fellow Research Fund Program under Grant 3050012222022. *(Juntong Lyu and Wenbin Shi contributed equally to this work.) (Corresponding author: Chien-Hung Yeh.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Juntong Lyu, Wenbin Shi, and Chuting Zhang are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China.

Chien-Hung Yeh is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, and also with the Center for Dynamical Biomarkers, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA 02215 USA (e-mail: nzdiw1120@gmail.com).

Digital Object Identifier 10.1109/TNSRE.2023.3323892

datasets respectively. Eldele et al. [11] proposed an attention-based two-module architecture—AttnSleep to finely extract features in single-channel EEG signals. A multi-resolution convolution neural network (CNN) and adaptive feature recalibration as the first module, then a temporal context encoder that leverages a multi-head attention mechanism is the second. AttnSleep achieved an accuracy of 84.4%, 81.3%, and 84.2% in the Sleep-EDF-20, Sleep-EDF-78, and SHHS datasets respectively, outperforming state-of-the-art models (e.g., DeepSleepNet [12], SleepEEGNet [13], and ResnetLSTM [14], etc.). Other pieces of literature also employed sequential structure [15] or hybrid CNN [16] in sleep classification, which also presented a fair performance in sleep staging. Although these studies demonstrated the potential of deep-learning methods to sleep classification, their heavy training time cost, complicated architecture, interpretability, and massive datasets for training prohibit deep-learning methods from being direct application in clinical scenarios.

On the other hand, developing reliable features with a precise physiological inference or statistical significance using physiological signals could leave out complicated architecture design meanwhile reduce the training procedure with much less data. Many studies have proposed single-mode analysis methods that rely on EEG or ECG signals; briefly, extracting statistical, temporal, and spectral features, then adopting conventional machine-learning methods, such as random forest (RF) [17], support vector machine (SVM) [18], [19], and linear discriminant analysis (LDA) [20], etc., for sleep staging. For example, Hassan et al. [21] decomposed EEG segments using Ensemble Empirical Mode Decomposition (EEMD) and extracted various statistical moment-based features, by using random under-sampling boosting (RUSBoost), reaching an accuracy of 88.07%. Long et al. [22] calculated the spectral features of RR intervals from ECG signals, improving the discrimination ability of heart rate variability (HRV) spectral features by increasing the spectral boundary resolution, thus achieving higher accuracy on sleep classification. Some studies even achieved an accuracy above 90% in the sleep classification task. In [23], the tunable-Q factor wavelet transform (TQWT) was applied to decompose sleep-EEG signal segments into TQWT sub-bands. Normal inverse Gaussian (NIG) distribution modeling was then used in feature extraction along with an adaptive boosting (Adaboost) technique as a classifier to obtain an accuracy rate of 90.01%. Hassan et al. [24] employed complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and bootstrap aggregating (Bagging) methods to extract statistical features from a single-channel EEG signal, which achieved an accuracy of 90.69% in a 5-class sleep stage classification task on the Sleep-EDF dataset.

Although existing methods achieved high accuracy for sleep stage classification by using traditional machine-learning techniques, these methods have not solved the following points: 1) Previous studies mainly focused on using a single-mode signal like EEG signal only, while other electrophysiological signals also reveal distinctive characteristics during sleep. For example, an ECG signal can retrieve autonomic nerve system (ANS) activity, which reflects cardiovascular activities during sleep. Combining features from other physiological signals may further help to improve the performance of sleep staging. 2) Physiological signals contribute to sleep staging from various aspects, however, little attention has been paid to elucidating the contribution of different modalities. 3) Nonlinear dynamics of these physiological signals across frequencies and time scales during sleep still matter.

To address these points, we propose a novel multimodal features combination by integrating time-frequency and nonlinear features derived from the brain and heart during sleep. The multimodal features combination can provide better insight into sleep from both the central nervous system and the autonomic nervous system, hence improving the performance of the sleep stage classification scheme. The contribution of this study is threefold: Firstly, we characterize crucial biomarkers from brain/heart oscillatory dynamics during sleep using time-frequency analysis and nonlinear approaches, in which the multiscale entropy method is applied to extract nonlinear features of EEG/ECG at different frequencies or scale ranges. Secondly, we propose an optimized generalized multimodal features combination to drive classifiers for sleep staging. Thirdly, we investigate the contribution of the extracted features from each modality. We also interpret our results in terms of the physiological meanings of the extracted features.

B. Preliminaries

1) Multiscale Entropy Method: Sample Entropy (SampEn) is a well-known measure of entropy rate and is useful for short-time series analysis in particular [25]. As per Grassberger and Procaccia's definition [26], SampEn is fundamentally a regularity statistic, an improvement of the approximate entropy method, which is known to be biased [27], [28]. Two important parameters to determine sample entropy are the embedding dimension m and a threshold value r [29], [30]. Computation of $SampEn\,(m, r, N)$ given a data sequence $x\,(n) = \{x\,(1), x\,(2), \ldots, x\,(N)\}$ of length N is as follows: Firstly, reconstruct the m-dimension vector $x_m\,(i)\,(i = 1, 2, \ldots, N - m + 1)$, where

$$x_m\,(i) = [x\,(i), x\,(i+1), x\,(i+2), \ldots, x\,(i+m-1)] \quad (1)$$

Next, determine the distance between $x_m\,(i)$ and $x_m\,(j)$, which is defined as the maximum distance between each of the elements in the two vectors, i.e.,

$$d\left[x_m\,(i), x_m\,(j)\right] = max\left|x_{i+k} - x_{j+k}\right|, \quad 0 \le k \le m-1 \quad (2)$$

For each vector $x_m\,(i)$, calculate the probability that any vector $x_m\,(j)$ is close to it, i.e.,

$$B_i^m\,(r) = \frac{B_i}{N-m-1} \quad (3)$$

where B_i represents the number of vectors that satisfy

$$d\left[x_m\,(i), x_m\,(j)\right] \le r, \quad i \ne j \quad (4)$$

and the density is

$$B^m\,(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m\,(r) \quad (5)$$

Similarly, $A_i^m(r) = \frac{A_i}{N-m-1}$ and $A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r)$ represent the probability and density for $x_{m+1}(i)$. The total number of template matches in a m-dimensional ($(m+1)$-dimensional) state space within a tolerance r can be calculated as

$$B(r) = \frac{1}{2}(N-m-1)(N-m)B^m(r) \tag{6}$$

And

$$A(r) = \frac{1}{2}(N-m-1)(N-m)A^m(r) \tag{7}$$

Finally, the sample entropy is mathematically stated as

$$SampEn(m, r, N) = -\log\left(\frac{A(r)}{B(r)}\right) \tag{8}$$

Classical multiscale entropy (MSE) is an extension of SampEn to a multiscale fashion. In terms of the time series defined above, scaled versions of $x(n)$ are obtained by the coarse-grain procedure. The coarse-grained time series are denoted as $\{y^\tau\}$ at scale τ. Then, SampEn is calculated for sequentially scaled time series, resulting in a curve of entropy versus scale as

$$y^\tau(j) = \frac{1}{\tau}\sum_{i=(j-1)\tau+1}^{j\tau} x(i), \quad 1 \le j \le [N/\tau] \tag{9}$$

2) Intrinsic Mode Function: Physiological signals comprise multiple frequency modulations. Decomposed elementary components provide finer insight into the key characteristics of these frequency dynamics [31], [32]. Thus, appropriate signal decomposition method is carefully considered in this study.

Several advanced adaptive signal decomposition methods, including empirical mode decomposition (EMD) [33], empirical wavelet transform (EWT) [34], and the Fourier decomposition method (FDM) [35], [36], were developed and used for medical signal processing. Among them, the EWT and FDM use either wavelet or Fourier basis to construct adaptive filter banks and decompose a signal as per its spectral characteristics. These two methods do perform competently in nonlinear and nonstationary physiological signal analyses, especially for FDM which invalidate the perception that the Fourier theory fails to be used for non-stationary signal analysis. FDM obtains decompositions by two frequency-scan techniques, i.e., FDM-LTH and FDM-HTL, however, this requires a relatively larger computational cost [37]. For EWT, its empirical wavelet function may cause mode mixing problems with the number of predefined components being difficult to determine.

The EMD is an empirical algorithm that extracts non-sine waves by the so-called "sifting process". It simply requires local extrema to generate the upper and lower envelopes for the "sifting process". In this study, competitive performances among EMD, EWT, and FDM methods in signal decomposing were carefully pre-validated. We decomposed the signal using these methods to extract delta and alpha waves that were further proceeded for feature extractions and performed sleep classification tasks using these features. Our results show that EMD achieves a better result in overall performance and is easier to implement in terms of computational time (presented in the "Experimental Results and Discussions" section). Considering constructing an efficient multimodal features combination in practical use, we use the EMD method for feature extraction in the following analyses.

EMD unveils the nonlinear dynamics of a signal by decomposing it into elementary components referred to as intrinsic mode functions (IMFs) [33]. Every IMF satisfies two properties: (1) the number of extrema and the number of zero crossings are either equal or differ by one; (2) the mean value of the envelope defined by the local maxima and the envelope defined by the local maxima is constant zero. The steps of an EMD algorithm for a given input $x(t)$ are as follows:

(1) Generating local mean curve: the algorithm begins with identifying all the local maxima and minima. The upper envelope $u(t)$ is generated by connecting all the local maxima using a cubic spline curve. Likewise, all the local minima are connected to obtain the lower envelope $v(t)$. Then, the mean $m_1(t)$ of these two envelopes is constructed as

$$m_1(t) = \frac{[u(t) + v(t)]}{2} \tag{10}$$

(2) Sifting process: the first step of a sifting process is to calculate the difference $h_1(t)$ as

$$h_1(t) = x(t) - m_1(t) \tag{11}$$

However, given $h_1(t)$ rarely directly satisfies the two IMF properties to serve as the first IMF of the input. The sifting process requires repeated on $h_1(t)$ until it meets the two IMF properties. After recursively applying the above step on $h_1(t)$, the sifting process terminated once the shortest-period component $c_1(t)$ is obtained. Noted the sifting process terminated when the sum of the difference is zero. Then, $c_1(t)$ is removed from the input $x(t)$ to obtain the first residue $r_1(t)$, i.e.,

$$r_1(t) = x(t) - c_1(t) \tag{12}$$

(3) Generating all IMFs: If the residue $r_1(t)$ still contains information in greater scales, it is treated as a new input to the next sifting process. Such a process is repeated on all IMFs $c_i(t)$ with the subsequent residues. The whole procedure is terminated once the residue $r(t)$ is either a constant or monotonic slope or a function with only one extremum. When the decomposition procedure is to the end, the input signal can be expressed as:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{13}$$

where $c_1(t), c_2(t), \ldots, c_n(t)$ correspond to the IMFs in order, and $r_n(t)$ is a negligible residue.

3) Time-Domain HRV Measures: Time-domain HRV measures, which analyze the variation of RR intervals through statistical methods, are the simplest and most intuitive measures to characterize HRV. The most global HRV measure could be the standard deviation of all NN intervals (SDNN), which evaluates general HRV and is relatively insensitive to small errors in scanning. Other time-domain HRV measures characterize short-term variations in heart rate. The degree to which HRV changes on a beat-to-beat basis is reflected in the average change in the interbeat interval between beats

TABLE I
FEATURES EXTRACTED FROM THE EEG AND ECG SIGNALS

	No.	Feature	No.	Feature	No.	Feature	No.	Feature
EEG Features	1	MSE_{δ} at F3 channel	7	MSE_{α} at F3 channel	13	IMF_{δ} at F3 channel	19	IMF_{α} at F3 channel
	2	MSE_{δ} at F4 channel	8	MSE_{α} at F4 channel	14	IMF_{δ} at F4 channel	20	IMF_{α} at F4 channel
	3	MSE_{δ} at C3 channel	9	MSE_{α} at C3 channel	15	IMF_{δ} at C3 channel	21	IMF_{α} at C3 channel
	4	MSE_{δ} at C4 channel	10	MSE_{α} at C4 channel	16	IMF_{δ} at C4 channel	22	IMF_{α} at C4 channel
	5	MSE_{δ} at O1 channel	11	MSE_{α} at O1 channel	17	IMF_{δ} at O1 channel	23	IMF_{α} at O1 channel
	6	MSE_{δ} at O2 channel	12	MSE_{α} at O2 channel	18	IMF_{δ} at O2 channel	24	IMF_{α} at O2 channel
ECG Features	25	$MSE_{1\sim5}$	30	*HR* (Heart rate)	35	*RMSSD* (Root mean square of successive differences of NN intervals)	40	*LF* / *HF* (Low-to-high-frequency power ratio)
	26	$MSE_{6\sim10}$	31	*SD* (Semi-minor axis length of Poincare plot of RR intervals)	36	*ULF* (Ultra-low-frequency power)		
	27	$MSE_{11\sim15}$	32	*SDNN* (Standard deviation of NN intervals)	37	*VLF* (Very-low-frequency power)		
	28	$MSE_{16\sim20}$	33	*SDRatio* (Semi-minor to -major axis length ratio of Poincare plot)	38	*LF* (Low-frequency power)		
	29	$MSE_{1\sim20}$	34	*mRR* (Mean RR intervals)	39	*HF* (High-frequency power)		

(RMSSD). Heart rate also reflects the varying parasympathetic nerve system (PNS) and sympathetic nerve system (SNS) activities, associates respiratory frequency, and provides information on cardiac activity during sleep [38].

4) Frequency-Domain HRV Measures: Frequency-domain HRV features are critical indicators to reflect the activity of ANS. Largescale underlying periodicities of every 5 min to 24 hr in the (heart rate) HR signal are reflected by ultra-low-frequency power (ULF), while very-low-frequency power (VLF) corresponds to a shorter scale of every 25 sec to every 5 min (i.e., 0.0033 to 0.04 Hz). In addition, VLF could imply the underlying frequency of most sleep-disordered breathing and periodic limb movements. Faster underlying periodicities in HR patterns are captured by low frequency (LF) power (0.04-0.15 Hz) and high frequency (HF) power (0.15-0.4 Hz). The power in LF (0.04-0.15 Hz) and HF (0.15-0.4 Hz) bands were related to the regulation of SNS and PNS nervous systems, respectively [39]. LF/HF ratio is used to assess changes in autonomic function between sleep stages, which generally increases with increased SNS activity during the transition from non-rapid eye movement (NREM) to rapid eye movement (REM) sleep [40], thus capturing subtle changes in different sleep stages in cardiac dynamics.

III. MULTIMODAL FEATURES COMBINATION

A. Feature Extraction

A brief description of the features used in this work is given in Table I. A total of 40 features were extracted from each 30-s segment, including 16 ECG features, 24 EEG features based on two types of signal processing techniques as well as two sets of measures associated with cardiac dynamics.

1) EEG Features: Since EEG signals behave irregularly, entropy measures were considered to quantify the amount of roughness captured inside a signal [41], [42], [43]. In this study, the multiscale entropy analysis was applied to the pre-processed EEG signals, generating a curve of sample entropy versus scale factors. Sample entropy parameters were set to $m = 2$ and $r = 15\%$ of the signal standard deviation. Then, the areas in the delta and alpha frequency bands under the curve were calculated, serving as the delta-entropy and alpha-entropy features for classification. The range of scale factors under the alpha and delta bands were calculated according to the formula as follows [44]:

$$f_{N,\tau_{SF}} = \frac{f_{S,1}}{2 \times \tau_{SF}} \tag{14}$$

where $f_{N,\tau_{SF}}$ is the frequency corresponds to the scale factor τ_{SF}, $f_{S,1}$ is the sampling frequency at time scale 1 or the original time series.

As the frequencies of EEG components vary across different sleep stages, grouping these components can help better sleep staging from a spectral perspective. Therefore, the IMFs' peaking frequency bands in delta (0.5-4 Hz) and alpha (8-13 Hz), were extracted using the EMD method. Then, the average envelope power of the delta- and alpha-peaking IMFs were estimated using the Welch method. The envelope power of delta- and alpha-peaking IMFs then served as features for sleep staging.

2) ECG Features: The fractal and entropy methods are commonly used in the analysis of complex systems. The multiscale entropy method evaluates the nonlinear dynamics of HRV, giving an insight into its complexity across time scales

TABLE II
DETAILED DESCRIPTION OF DATASETS USED IN OUR EXPERIMENT

Datasets	Subjects	Channels	Sampling Rate	Number of sleep phases					Total Samples
				W	N1	N2	N3	REM	
MASS-SS3	62	20 EEG channels 1 ECG channel	512 Hz	6442 *10.9%*	4839 *8.2%*	29802 *50.2%*	7653 *12.9%*	10581 *17.8%*	59317 *100%*
ISRUC-S3	10	6 EEG channels 1 ECG channel	200 Hz	1651 *19.3%*	1215 *14.2%*	2609 *30.5%*	2014 *23.6%*	1060 *12.4%*	8549 *100%*

during sleep. Thus, MSE was also applied for ECG feature extraction in this experiment.

RR intervals were first interpolated at a sampling frequency of 4 Hz with the cubic spline function, to fulfill the requirement in data length [29], [30]. The MSE of RR intervals at scale 1-20 was calculated, which generated a curve of entropy versus scale. Both short (1-5) and long (6-20) time scales of the MSE curve for RR intervals were assessed. Then, we calculated the area under the curve in different time scale ranges of interest, including 1-5, 6-10, 11-15, 16-20, and 1-20 as features. Sample Entropy parameters were set to $m = 2$, and $r = 25\%$ of the standard deviation of an input signal.

Apart from the cross-scaled entropies, six time-domain HRV measures and five frequency-domain HRV measures were included as features for sleep classification. To meet the requirement of data length for calculating HRV time-domain and frequency-domain measures, the whole RR sequence was segmented into sections in 5 min centered around every 30-s sleep epoch. With a 30-s step size, every overlapping 5-min RR sequence section belongs to a specific sleep stage, which is the RR sequence to be analyzed for HRV measures calculation.

B. Feature Selection

The min-redundancy and max-relevance (mRMR) feature selection algorithm was employed in a k-fold manner to analyze the importance of features in the sleep classification task. The mRMR feature selection is a method that selects the features with the highest relevance to the target classes while it minimizes the redundancy among the selected features [45]. Top-ranking features were selected and integrated based on the weights and rankings obtained from the mRMR feature selection method.

IV. EXPERIMENTAL PREPARATIONS

A. Dataset Acquisition

Two publicly available datasets were employed in this experiment: 1) ISRUC-S3 dataset [46] contains 10 healthy subjects (gender: 9 male and 1 female; age: 40±10 years). Each recording includes 6 EEG channels, 2 EOG channels, 3 EMG channels, and 1 ECG channel. All EEG, EOG, and chin EMG signals were sampled at 200 Hz. 2) MASS-SS3 [47] dataset contains 62 healthy subjects (28 male and 34 female). Each recording contains 20 EEG channels, 2EOG channels, 3 EMG channels, and 1 ECG channel. All recordings of the two datasets were segmented into 30-s epochs, and visually scored into five sleep stages: awake, NREM (N1, N2, and N3), and REM sleep, by sleep experts according to the guideline of the American Academy of Sleep Medicine (AASM) [46]. Details of the datasets used in our experiments are summarized in Table II. Informed consents of all subjects were obtained in both datasets.

B. Data Preprocessing

Data preprocessing is a prerequisite for feature extraction. Removing artifacts and noises helps improve the quality of the extracted features, thus ensuring the accuracy of classification. For the ISRUC-S3 cohort, by the EEG and ECG preprocessed steps, 6827 epochs were extracted in total, of which 974 epochs were annotated as W stage, 775 epochs as N1 stage, 2298 epochs as N2 stage, 1779 epochs as N3 stage, and 1001 epochs as REM stage. For the MASS-SS3 dataset, 55852 epochs were extracted in total, of which 5844 epochs were labeled as W stage, 4526 epochs as N1 stage, 27946 epochs as N2 stage, 7378 epochs as N3 stage, and 10158 epochs as REM stage.

1) EEG Data Preprocessing: EEG signal preprocessing is crucial for further feature extraction. Typically, Butterworth bandpass filtering is implemented to achieve signal denoising. Recently, several novel signal denoise techniques have been proposed by using adaptive signal decomposition methods like FDM [48], which can effectively remove baseline wander and artifacts that are commonly encountered in physiological signals. Let the FDM signal denoise method as an example, it decomposes a signal into Fourier intrinsic band functions (FIFBs) by frequency scanning. Baseline wandering and power-line interference occur in a very low-frequency range (< 0.7 Hz). By removing corresponding components in FIFBs and reconstructing signals, artifact removal can be effectively carried out.

To assure a satisfactory performance in signal denoising, we compared the denoise results of bandpass filtering with FDM. As illustrated in Fig 1, Butterworth bandpass filtering and FDM can both remove a certain level of baseline wander and artifacts. The power spectra by these two methods showed resemble landscapes, wherein the FDM presented a sharper cutoff frequency in component elimination. Considering a large amount of computation for frequency scanning, plus the similarity in performances of these two approaches for sleep EEGs, noting the very low-frequency range is beyond our interest in EEG analyses during sleep, we implemented the Butterworth bandpass filter for the following signal-denoising.

Fig. 2 (a) shows a representative segment of a raw EEG signal from subject 1, which contained large and prolonged

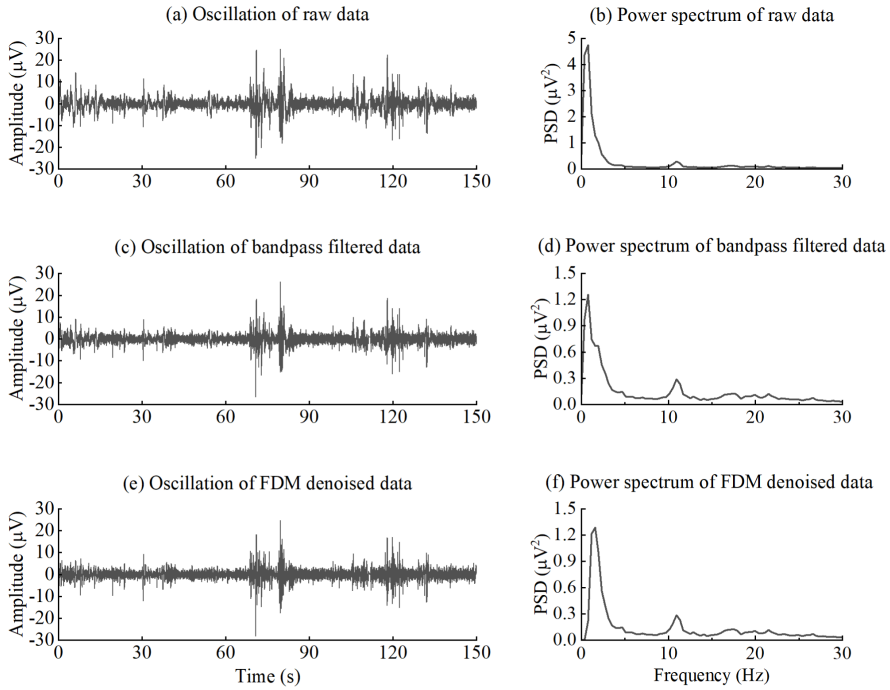

Fig. 1. Performance comparisons of bandpass filtering and FDM in denoising. (a-b) EEG raw data (demonstrated by subject 1) in F3 channel with the corresponding power spectrum. (c-d) The denoised EEG oscillation by 0.5-40 Hz Butterworth bandpass filtering with the corresponding power spectrum. (e-f) The denoised EEG oscillation by Fourier decomposition method with the corresponding power spectrum.

artifacts. These were considered to be the bad segments in a recording that required manual removal. The corresponding power spectrum estimated by the Welch method is shown by the right side of the time series, presenting a higher power in the low-frequency band ($<$ 0.5 Hz). The low-frequency artifacts possibly by limb movement were required to be excluded as the next.

To this end, bad segment rejection was first performed manually according to their amplitude. Bad segments or epochs were interpolated using the cubic spline interpolation method by a predetermined threshold. Then, the data was further denoised by 0.5-40 Hz Butterworth bandpass filtering to remove both low- and high-frequency noise. The time series and frequency spectrum of the preprocessed data is illustrated in Fig. 2(b), where outliers and artifacts were removed, and low-frequency noise ($<$ 0.5 Hz) were eliminated as well.

Then, we compared the sleep stages scoring results evaluated by the two sleep experts. If consistent then the annotation of the epoch remains as it is. Otherwise, the epoch was removed.

2) ECG Data Preprocessing: QRS complex analysis with an order-static filter was used to perform peak detection to extract R waves from ECG signals [49], as shown in Fig. 3(a). Then RR wave intervals were calculated as per the adjacent R peaks differences. RR preprocessing is a prerequisite for extracting reliable HRV features. To remove the outliers, each RR value was compared to the corresponding mean value of RR intervals (mRR) within a 21-points window centered around the tested value. If the RR is either less than 0.5×mRR or larger than 1.5×mRR, it is replaced by mRR, otherwise, it remains as it was. Next, each RR sequence was sectioned. Each RR record was divided into 30-s epochs synchronizing in time with the corresponding sleep staging annotation. RR sequence before and after the preprocessing step is illustrated in Fig. 3(b).

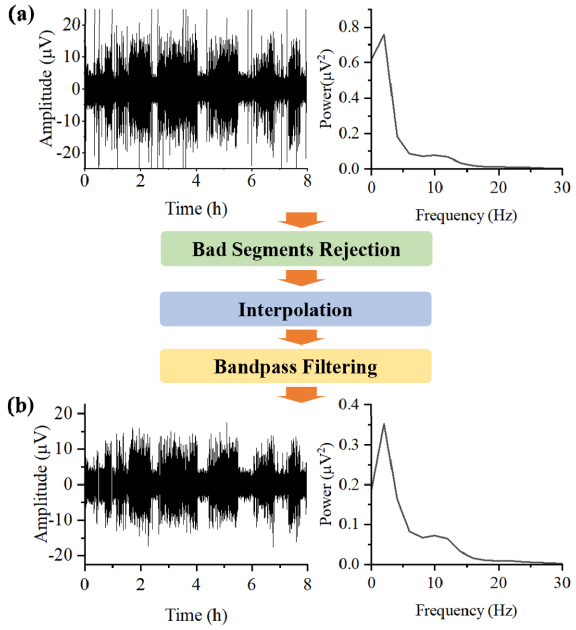

Fig. 2. EEG data preprocessing procedure. (a) EEG raw data (demonstrated by subject 1) in F3 channel with the corresponding power spectrum. Bad segments rejection was performed manually to remove artifacts. Next, single-point outliers were interpolated under a threshold. Then data was denoised by 0.5-40 Hz Butterworth bandpass filtering. (b) The preprocessed data and the corresponding power spectrum.

C. System Performance Evaluation

To estimate the classification performance of all features combination-driven classifier, we adopted four metrics to eval-

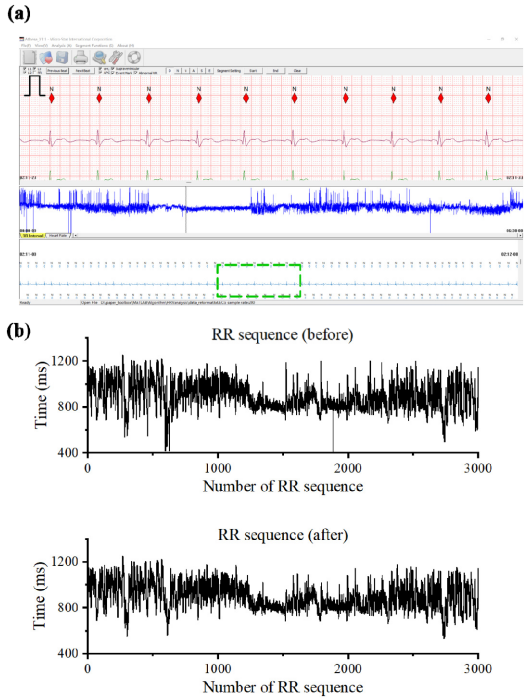

Fig. 3. The performance of ECG preprocessing. (a) QRS complex detection with order-static filter was used to obtain the RR sequence. (b) Comparison of RR sequence before and after preprocessing step.

uate the performances of various sets of features for sleep stage classification, namely, the accuracy (ACC), macro-averaged F1-score (MF1) and Cohen Kappa (κ) [50]. The MF1 metric is commonly used to assess model performance on imbalanced datasets [51]. Given the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) for the i -th class, the overall accuracy ACC, MF1, and MGm are defined as follows:

$$ACC = \frac{\sum_{i=1}^{K} TP_i}{M} \tag{15}$$

$$MF1 = \frac{1}{K}\sum_{i=1}^{K} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{16}$$

where $Precision_i = \frac{TP_i}{TP_i+FP_i}$, $Recall_i = \frac{TP_i}{TP_i+FN_i}$ and $Specificity_i = \frac{TN_i}{TN_i+FP_i}$. M represents the total number of samples, and K stands for the number of classes or sleep stages.

We also employed per-class precision (PR), per-class recall (RE) and per-class F1-score (F1) to evaluate each of our combination-driven classifiers. These metrics were calculated by treating one class as positive and the other four classes as negative in binary classification. The metrics were then calculated by averaging scoring values of the testing data across k folds.

D. Classification

Multiple traditional machine learning classifiers were used for sleep stages classification to investigate the generalization ability of the proposed multi-modal features combination, including support vector machine (SVM), random forest (RF), k-nearest neighbor (kNN), and linear discriminant analysis (LDA). To evaluate the performance of our multimodal features combination, we performed a k-fold cross-validation process in the classification tasks. The datasets were divided randomly into k equal-size subsets. At each iteration, the (k-1) subsets were used as the training and validating data and 1 subset was used for testing. In the experiment using the ISRUC-S3 dataset, 9 subsets were for training and 1 subset was for testing. In the experiment using the MASS-SS3 dataset, 31 subsets were for training and 1 subset was for testing. To further reduce the dimensions of features in the training procedure, principal component analysis was performed after the feature selection step. 99% of the principal components were kept.

V. Experimental Results and Discussions

A. Comparisons With State-of-the-Art Methods

We evaluated the performance of our multi-modal features combination-driven classifiers against various state-of-the-art (SOTA) approaches that have emerged in recent years. We compared their performances in terms of overall accuracy, macro F1-score, and Cohen Kappa value on the ISRUC-S3 dataset as well as the MASS-SS3 dataset.

Table III presents a comparison in the ISRUC-S3 dataset among cutting-edge deep learning methods including [12], [15], [52], [53], [54], [10], as well as other traditional machine learning-based approaches such as [17] and [18]. We observed that our multimodal feature combination-driven classifiers outperformed traditional machine learning approaches in terms of accuracy, F1-score, and kappa value. In particular, SVM and RF models with this combination achieved the optimal and suboptimal macro-F1 score and kappa value respectively among the state-of-the-art methods. Additionally, other classification models like LDA and kNN also outperformed the SOTA traditional machine learning methods [17] and [18] under various evaluation metrics, indicating that the proposed combination is capable of handling imbalanced data as well as model generalization.

As shown in Table IV, in the MASS-SS3 dataset, the performance of our multimodal features combination in most classifiers achieved an accuracy of above 80%. Although the proposed method may not outperform partial deep learning methods, its performance is superior to all the baselines with standard machine learning methods [17] and [18]. Our multimodal combination reached fair performances in both cohorts, indicating its generalization ability.

Although the proposed combination-driven classifiers demonstrated success in overall performance, they failed to meet expectations in the N1 stage classification in both datasets, similar to other baseline models. The reasons may be twofold. Firstly, the N1 stage is a transitional period between W and N2 stages, and the sample size of the N1 stage is relatively small. Secondly, the traditional machine learning methods may be inadequate to learn complex spatial or temporal features, whereas deep learning approaches such as CNN and GCN may favor extracting spatial or temporal features from multichannel EEG signals, N1 stage classification in particular. Nevertheless, the classification performance

TABLE III
THE PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART METHODS ON THE ISRUC-S3 DATASET

	Method	Modalities	Overall results			F1-score for each class				
			Accuracy	F1-score	Kappa	Wake	N1	N2	N3	REM
Alickovic et al. [18]	SVM	EEG	0.733	0.721	0.657	0.868	0.523	0.699	0.786	0.731
Memar et al. [17]	RF	EEG	0.729	0.708	0.648	0.858	0.473	0.704	0.809	0.699
Dong et al. [52]	MLP+LSTM	EEG	0.779	0.758	0.713	0.860	0.469	0.760	0.875	0.828
Supratak et al. [12]	CNN+BiLSTM	EEG	0.788	0.779	0.730	0.887	**0.602**	0.746	0.858	0.802
Chambon et al. [53]	CNN	EEG & EOG	0.781	0.768	0.720	0.870	0.550	0.760	0.851	0.809
Phan et al. [15]	ARNN+RNN	EEG & EOG	0.789	0.763	0.725	0.836	0.439	0.793	0.879	0.867
Jia et al. [54]	STGCN	EEG	0.799	0.787	0.741	0.878	0.574	0.776	0.864	0.841
Jia et al. [10]	MSTGCN	EEG	0.821	0.808	0.769	0.894	0.596	0.806	0.890	0.856
Proposed Method	**SVM**	EEG & ECG	**0.843**	**0.819**	**0.794**	0.904	0.551	**0.830**	**0.933**	**0.876**
Proposed Method	**RF**	EEG & ECG	0.831	0.791	0.775	**0.916**	0.472	0.824	0.924	0.819
Proposed Method	**kNN**	EEG & ECG	0.803	0.768	0.738	0.860	0.463	0.799	0.885	0.832
Proposed Method	**LDA**	EEG & ECG	0.761	0.729	0.683	0.898	0.411	0.751	0.861	0.724

* The bold font indicates the best result and the underlined result is the second best.

TABLE IV
THE PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART METHODS ON THE MASS-SS3 DATASET

	Method	Modalities	Overall results			F1-score for each class				
			Accuracy	F1-score	Kappa	Wake	N1	N2	N3	REM
Alickovic et al. [18]	SVM	EEG	0.779	0.688	0.659	0.801	0.339	0.843	0.645	0.813
Memar et al. [17]	RF	EEG	0.800	0.726	0.697	0.863	0.379	0.858	0.784	0.749
Dong et al. [52]	MLP+LSTM	EEG & EOG	0.859	0.805	-	0.846	0.563	**0.907**	0.848	0.861
Supratak et al. [12]	CNN+BiLSTM	EEG	**0.862**	**0.817**	**0.800**	**0.873**	**0.598**	0.903	0.815	**0.893**
Chambon et al. [53]	CNN	EEG & EOG & EMG	0.739	0.673	0.640	0.730	0.294	0.812	0.765	0.764
Proposed Method	**SVM**	EEG & ECG	0.822	0.754	0.734	0.814	0.414	0.875	0.843	0.823
Proposed Method	**RF**	EEG & ECG	0.833	0.742	0.746	0.832	0.318	0.878	**0.874**	0.803
Proposed Method	**kNN**	EEG & ECG	0.825	0.762	0.742	0.812	0.437	0.878	0.841	0.843
Proposed Method	**LDA**	EEG & ECG	0.743	0.618	0.599	0.714	0.105	0.827	0.818	0.625

* The bold font indicates the best result and the underlined result is the second best.

of our combination-driven classifiers for the N1 stage still outperforms most of the baseline models.

B. Effectiveness of EEG/ECG Features Combination

To disclose the key factors potentially contributing to the optimized feature combination, all EEG/ECG and EEG+ECG feature combinations were first widely explored from different perspectives. EEG features combination included all features from the EEG feature set in Table I, and the ECG feature combination performed likewise. EEG+ECG features combination involved all features in Table I. As shown in Table V, EEG features outperformed ECG features in the awake stage classification, which is reflected by the accuracies of EEG (89.4%) and ECG (79.1%) features in the awake staging. ECG features, on the other side, performed superior in deep sleep (N3) staging than EEG features. When integrating all EEG/ECG features as a combination to drive the sleep-stage classifier, the performance of both the awake-sleep and the light-deep sleep classification were much improved (N1 stage: from 11.9% in EEG and 40.6% in ECG to 49.1% in EEG/ECG combination; N3 stage: reached to 91.8%; REM stage: reached to 85.7%), indicating EEG/ECG features may both boost the performance of sleep classifier from different aspects. In addition, a higher kappa value was also achieved (EEG+ECG: 0.7909; EEG: 0.6593; ECG: 0.7890), reflecting a higher consistency with manual scoring.

C. Investigating the Causes of Improvement in the Awake and Sleep Status Classification

Most notably, the proposed multimodal features combination, when applied with traditional machine learning methods, has achieved remarkable improvements in W, N2, and N3 staging performance compared to the state-of-the-art (SOTA) methods in ISRUC-S3 dataset, as illustrated in Table III. Moreover, N2 and N3 stages were better classified among

TABLE V
PERFORMANCE OF SLEEP CLASSIFICATION TASK USING DIFFERENT FEATURES COMBINATIONS AT ALL SLEEP STAGES

Features combinations	Sleep stages	PR (%)	RE (%)	MF1 (%)
EEG Features Combination	Wake	89. 4	82. 6	85. 9
	N1	11. 9	47. 5	19. 0
	N2	82. 6	66. 7	73. 8
	N3	83. 3	84. 6	83. 9
	REM	59. 8	59. 8	59. 8
ECG Features Combination	Wake	79. 1	78. 8	79. 0
	N1	40. 6	53. 6	46. 2
	N2	86. 5	75. 8	80. 8
	N3	91. 6	95. 5	93. 5
	REM	85. 9	90. 5	88. 2
EEG+ECG Features Combination	Wake	**90. 3**	**86. 9**	**88. 6**
	N1	**49. 1**	**60. 1**	**54. 1**
	N2	86. 3	**79. 6**	**82. 8**
	N3	**91. 8**	93. 6	92. 7
	REM	85. 7	**90. 7**	88. 1

Note: PR represents precision, RE represents recall, MF1 represents macro-F1 score. The bold font indicates the optimal performance of the corresponding item among all combinations.

other sleep stages, suggesting that the proposed multimodal features can extract and integrate key factors in both deep sleep and awake stages. Similar results were also shown on the MASS-SS3 dataset, in which performances of N2 and N3 stages outperformed the traditional machine learning baseline models [17] and [18] (Table IV).

Since alpha and delta waves represent the predominant frequency components of EEG during the awake and deep sleep status respectively [55], nonlinear and frequency-domain features derived from oscillations in these two frequency bands may greatly account for the relatively high accuracies in these two statuses. Furthermore, the information from the autonomic nervous system also plays an essential role in distinguishing awake/sleep status with features of heart rate variability, which reflects the dynamics of sympathovagal balance from the parasympathetic predominant drive during NREM sleep to increased sympathetic activity during REM sleep [38]. Hence, integrating information from both central and autonomic nervous systems may ultimately support differentiation between light-deep sleep as well as awake-sleep status.

D. Performance Comparison of Different Classifiers

To investigate the model generalization ability of the proposed multimodal features combination, we utilized four standard machine learning classifiers to perform sleep classification tasks on both ISRUC-S3 and MASS-S3 datasets. Specifically, we investigated the optimal number of sorted features generated by the k-fold mRMR feature selection method. A stage-wise performance comparison was also provided across different classifiers evaluated by F1-score.

TABLE VI
THE AVERAGED TOP 25 SCORED FEATURES OVER K FOLDS ON ISRUC-S3 DATASET

	Features Indices
Top 25 features over k folds	6, 30, 3, 12, 9, 4, 27, 5, 11, 1, 10, 32, 2, 26, 7, 37, 25, 8, 33, 31, 18, 28, 39, 40, 36

Note: The number is correspond to the features' indices listed in Table I. Ranks are listed in order from the highest.

Fig. 4 shows the performance of each classifier with a different number of top-ranking features on both ISRUC-S3 and MASS-SS3 datasets. Generally, the optimal number of features for a small-sample-size cohort (ISRUC-S3) was around 25 to 30. Thus, we set the average of the top 25 scored features over $k$folds (Table VI). While in MASS-SS3, the performances of the four classifiers all showed upward trends in accordance with the number of top-ranking features. Our results indicate that the performances of the proposed multimodal features combination are prone to converge for a smaller cohort, and more features can improve performances further for a larger cohort.

Next, each classifier optimized the number of top-ranking features, as per Fig. 4, to achieve its best performance in sleep staging. Then, the F1-scores were compared among these four classifiers grouped by the five sleep stages (Fig. 5). Overall, the four classifiers all presented satisfactory performances in sleep staging (higher or near 80%), except for the N1 stage. This could be explained by the imbalanced number of epochs, where the epoch number of the N1 stage is much fewer than the rest stages. Data augmentation may help to improve N1 staging further in future works [56].

E. Effectiveness of IMF-Based Features

To further validate the usefulness of IMF-based features by EMD in sleep classification tasks, we then experimented to compare EMD, EWT, FDM, and bandpass filters in sleep classification tasks on the ISRUC-S3 dataset. For signal decompositions, the components with their peaking frequency in delta (0.5-4 Hz) and alpha (8-13 Hz) were extracted by EMD, EWT, and FDM methods, respectively. Then, the average envelope power of the delta- and alpha-peaking IMFs or FIBFs were estimated by the Welch method. The envelope power of delta- and alpha-peaking IMFs and FIBFs then served as features for sleep staging. For the bandpass-filtered approach, signal components at delta (0.5-4 Hz) and alpha bands (8-13 Hz) were extracted using the Butterworth bandpass filter. Likewise, the envelope power of these components served as a feature. These features were then fed into SVM classifier in a k-fold manner resembling our prior setting in the classification section. The average values of accuracy, F1-score, and Kappa value over k folds were used to compare these methods. The experimental result is summarized in Table VII, showing EMD and bandpass filter methods outperform FDM and EWT.

EMD is proven to behave as adaptive filter banks [57], which decomposes a signal in a data-driven approach. FDM and EWT also decompose signal adaptively, however, FDM

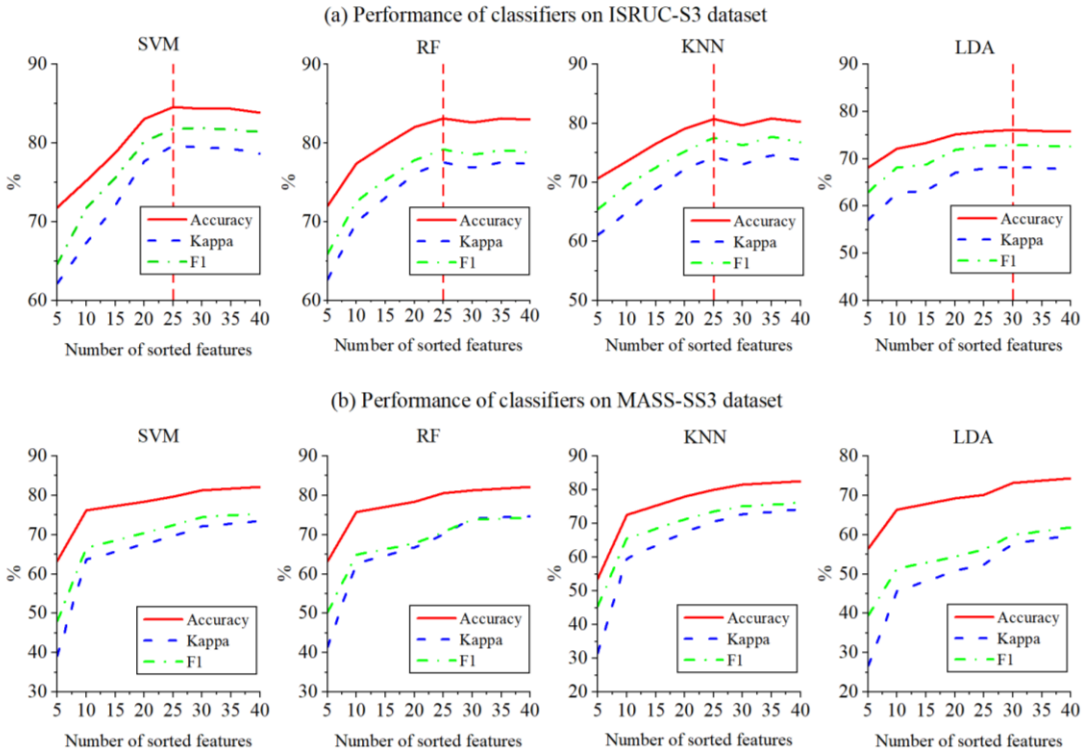

Fig. 4. Performances of four different classifiers using top ranking features from k-fold mRMR feature selection method. (a) Performances of classifiers on ISRUC-S3 dataset. The red vertical dash line annotates the number of sorted features with the best results. (b) Performances of classifiers on MASS-SS3 dataset. The performances of all classifiers show upward trends in accordance with the number of top ranking features.

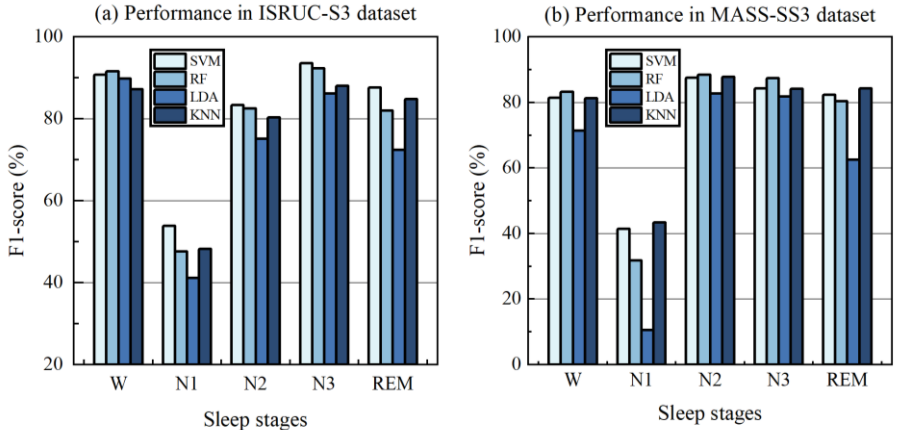

Fig. 5. Performance comparisons of four different classifiers under different sleep stages. Optimized numder of top-ranking features were applied by classifier. (a) Performances of classifiers on ISRUC-S3 dataset. (b) Performances of classifiers on MASS-SS3 dataset.

TABLE VII
PERFORMANCE COMPARISONS AMONG EWT, EMD, FDM AND BANDPASS FILTERING METHODS ON ISRUC-S3 DATASET

Methods	Accuracy (%)	F1-score (%)	Kappa (%)
EMD	71.8	66.0	62.8
Bandpass Filter	71.6	65.0	62.4
FDM	68.3	62.4	58.5
EWT	60.2	50.1	47.0

generates a large number of FIBFs that hinders the reconstruction of the wanted components; whereas EWT requires a predefined number of modes that is difficult to pre-determine. These factors may result in the performances of FDM and EWT failing to achieve expectations. Bandpass filter and EMD share similar performances in classification tasks. The reason for this may be our input has components with strong and concentrated power in delta (0.5-4 Hz) and alpha (8-13Hz) bands (power spectrum in Fig 2(b)), thus the nonlinear waveform distortion by filtering is reduced.

F. Limitations and Future Directions

By extracting and integrating the information in central nervous systems (CNS) and autonomic nervous system (ANS)

derived from EEG and ECG signals, the proposed multimodal features combination exhibited superior performance against various evaluation metrics compared to the SOTA methods, suggesting its effectiveness and efficiency in sleep stage classification task. However, certain limitations should be acknowledged.

Although the multimodal features combination achieved convergent and satisfactory performances in sleep staging on small-sample-size EEG recordings, which is a challenge commonly encountered in clinical scenarios, the performance of a larger dataset requires more candidate features to reach the expectation. Additionally, the EEG features in this study predominantly focused on delta and alpha bands, and the inclusion of features from other frequency bands could further enhance the classification performance. Future research should explore the efficacy of incorporating features from other frequency bands to improve the performance of the proposed multimodal feature combination in classification tasks.

In the present study, we simply extracted and concatenated features from EEG and ECG modalities associated with CNS and ANS functioning, ignoring the interplay between brain and heart that may provide subtle and vital details during sleep. A recent study also revealed that our body is an integrating system that involves interactions from each organ [58]. The dynamic coupling between CNS and ANS may be presented in the electroactivity of the brain and heart. Some studies reported that baroreflex has an essential role in sleep modulation, reflecting a neural pathway associated with cardiovascular and central nervous activities during sleep [59], [60]. Therefore, exploring potential couplings of rhythmicities from these two organs during sleep may provide further insight.

VI. Conclusion

In this study, nonlinear time-frequency analyses and complexity measures were introduced to access the key features derived from the brain/heart signals during sleep. A multimodal feature combination with higher precision and effectiveness was constructed. By using the multi-modal feature combination-driven machine learning classifiers, we achieved a maximum accuracy of 84.3% and a kappa value of 0.794 on the sleep classification task with both the EEG and ECG records on the ISRUC-S3 dataset, indicating the effectiveness and precision of our proposed multimodal features combination. On the MASS-SS3 dataset with a larger sample size, our proposed multimodal features combination still yielded an average level of accuracy (around 80∼83%), further suggesting its generalization ability.

The proposed multimodal feature combination incorporates brain/heart oscillatory features of the most important, thus improving the classification of multiple sleep stages than the single-mode feature combinations (EEG or ECG features). Our proposed multimodal features combination supports classifying between light and deep sleep states, as well as wakeful states, in particular.

The proposed multimodal feature combination was validated to prevail over different classifiers, sharing a high classification accuracy and fair model generalization ability. We expect that the proposed feature combination-driven automatic sleep classifiers to be flexibly implemented in various sleep monitoring systems, either for the healthy or the diseased states (e.g., depression, insomnia, narcolepsy, etc.), in supporting clinical diagnosis and treatment.

References

[1] N. Butkov and T. L. Lee-Chiong, *Fundamentals of Sleep Technology*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2007.

[2] K. K. Gulia and V. M. Kumar, "Importance of sleep for health and wellbeing amidst COVID-19 pandemic," *Sleep Vigilance*, vol. 4, no. 1, pp. 49–50, Jun. 2020.

[3] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, Aug. 2016.

[4] M. Torabi-Nami, S. Mehrabi, A. Borhani-Haghighi, and S. Derman, "Withstanding the obstructive sleep apnea syndrome at the expense of arousal instability, altered cerebral autoregulation and neurocognitive decline," *J. Integrative Neurosci.*, vol. 14, no. 2, pp. 169–193, Jun. 2015.

[5] P. Jadhav and S. Mukhopadhyay, "Automated sleep stage scoring using time-frequency spectra convolution neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[6] Z. Wang, J. Zhang, Y. Xia, P. Chen, and B. Wang, "A general and scalable vision framework for functional near-infrared spectroscopy classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1982–1991, 2022.

[7] M. Choi and S.-J. Lee, "Oscillometry-based blood pressure estimation using convolutional neural networks," *IEEE Access*, vol. 10, pp. 56813–56822, 2022.

[8] S. A. Keenan, "An overview of polysomnography," in *Handbook of Clinical Neurophysiology*, vol. 6. Amsterdam, The Netherlands: Elsevier, 2005, pp. 33–50.

[9] R. B. Berry et al., "AASM scoring manual updates for 2017 (version 2.4)," *J. Clinical Sleep Med.*, vol. 13, no. 5, pp. 665–666, 2017.

[10] 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.

[11] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[12] 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.

[13] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

[14] Y. Sun, B. Wang, J. Jin, and X. Wang, "Deep convolutional network method for automatic sleep stage classification based on neurophysiological signals," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–5.

[15] 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.

[16] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, May 2019.

[17] 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.

[18] 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.

[19] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, "A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal," *Exp. Syst. Appl.*, vol. 104, pp. 277–293, Aug. 2018.

[20] T. Sousa, A. Cruz, S. Khalighi, G. Pires, and U. Nunes, "A two-step automatic sleep stage classification method with dubious range detection," *Comput. Biol. Med.*, vol. 59, pp. 42–53, Apr. 2015.

[21] 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.

[22] X. Long, P. Fonseca, R. Haakma, R. M. Aarts, and J. Foussier, "Spectral boundary adaptation on heart rate variability for sleep and wake classification," *Int. J. Artif. Intell. Tools*, vol. 23, no. 3, Jun. 2014, Art. no. 1460002.

[23] A. R. Hassan and M. I. H. Bhuiyan, "An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting," *Neurocomputing*, vol. 219, pp. 76–87, Jan. 2017.

[24] 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.

[25] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 278, no. 6, pp. 2039–2049, Jun. 2000.

[26] P. Grassberger and I. Procaccia, "Estimation of the Kolmogorov entropy from a chaotic signal," *Phys. Rev. A, Gen. Phys.*, vol. 28, no. 4, pp. 2591–2593, Oct. 1983.

[27] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Nat. Acad. Sci. USA*, vol. 88, no. 6, pp. 2297–2301, Mar. 1991.

[28] S. M. Pincus and A. L. Goldberger, "Physiological time-series analysis: What does regularity quantify?" *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 266, no. 4, pp. 1643–1656, Apr. 1994.

[29] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 2, Feb. 2005, Art. no. 021906.

[30] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, no. 6, Jul. 2002, Art. no. 068102.

[31] C. Zhang, C.-H. Yeh, and W. Shi, "Variational phase-amplitude coupling characterizes signatures of anterior cortex under emotional processing," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 4, pp. 1935–1945, Apr. 2023.

[32] C.-H. Yeh, C. Zhang, W. Shi, M.-T. Lo, G. Tinkhauser, and A. Oswal, "Cross-frequency coupling and intelligent neuromodulation," *Cyborg Bionic Syst.*, vol. 4, p. 0034, Jan. 2023.

[33] N. E. Huang et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[34] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, Aug. 2013.

[35] P. Singh, S. D. Joshi, R. K. Patney, and K. Saha, "The Fourier decomposition method for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 473, no. 2199, Mar. 2017, Art. no. 20160871.

[36] P. Singh, "Novel Fourier quadrature transforms and analytic signal representations for nonlinear and non-stationary time-series analysis," *Roy. Soc. Open Sci.*, vol. 5, no. 11, Nov. 2018, Art. no. 181131.

[37] W. Zhou, Z. Feng, Y. F. Xu, X. Wang, and H. Lv, "Empirical Fourier decomposition: An accurate signal decomposition method for nonlinear and non-stationary time series analysis," *Mech. Syst. Signal Process.*, vol. 163, Jan. 2022, Art. no. 108155.

[38] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Med. Rev.*, vol. 16, no. 1, pp. 47–66, Feb. 2012.

[39] R. E. Kleiger, P. K. Stein, and J. T. Bigger, "Heart rate variability: Measurement and clinical utility," *Ann. Noninvasive Electrocardiol.*, vol. 10, no. 1, pp. 88–101, Jan. 2005.

[40] U. J. Scholz, A. M. Bianchi, S. Cerutti, and S. Kubicki, "Vegetative background of sleep: Spectral analysis of the heart rate variability," *Physiol. Behav.*, vol. 62, no. 5, pp. 1037–1043, 1997.

[41] W. Shi, H. Feng, X. Zhang, and C.-H. Yeh, "Amplitude modulation multiscale entropy characterizes complexity and brain states," *Chaos, Solitons Fractals*, vol. 173, Aug. 2023, Art. no. 113646.

[42] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, "Correlated and uncorrelated regions in heart-rate fluctuations during sleep," *Phys. Rev. Lett.*, vol. 85, no. 17, pp. 3736–3739, Oct. 2000.

[43] T. Penzel, J. W. Kantelhardt, C.-C. Lo, K. Voigt, and C. Vogelmeier, "Dynamics of heart rate and sleep stages in normals and patients with sleep apnea," *Neuropsychopharmacology*, vol. 28, pp. 48–53, Jul. 2003.

[44] J. Courtiol et al., "The multiscale entropy: Guidelines for use and interpretation in brain signal analysis," *J. Neurosci. Methods*, vol. 273, pp. 175–190, Nov. 2016.

[45] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[46] 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.

[47] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, Dec. 2014.

[48] A. Singhal, P. Singh, B. Fatimah, and R. B. Pachori, "An efficient removal of power-line interference and baseline wander from ECG signals by employing Fourier decomposition technique," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101741.

[49] C. Lin et al., "Robust fetal heart beat detection via R-peak intervals distribution," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 12, pp. 3310–3319, Dec. 2019.

[50] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[51] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern., B*, vol. 39, no. 1, pp. 281–288, Feb. 2009.

[52] 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.

[53] 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.

[54] Z. Jia et al., "GraphSleepNet: Adaptive spatial–temporal graph convolutional networks for sleep stage classification," in *Proc. IJCAI*, 2020, pp. 1324–1330.

[55] M. A. Carskadon and W. C. Dement, "Normal human sleep: An overview," *Princ. Pract. Sleep Med.*, vol. 4, no. 1, pp. 13–23, 2005.

[56] C.-E. Kuo, G.-T. Chen, and P.-Y. Liao, "An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102981.

[57] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

[58] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, "Network physiology reveals relations between network topology and physiological function," *Nature Commun.*, vol. 3, no. 1, p. 702, Feb. 2012.

[59] Y. Yao et al., "Cardiovascular baroreflex circuit moonlights in sleep control," *Neuron*, vol. 110, no. 23, pp. 3986–3999, Dec. 2022.

[60] M. de Zambotti, J. Trinder, A. Silvani, I. M. Colrain, and F. C. Baker, "Dynamic coupling between the central and autonomic nervous systems during sleep: A review," *Neurosci. Biobehavioral Rev.*, vol. 90, pp. 84–103, Jul. 2018.