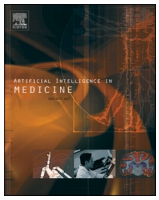
Artificial Intelligence In Medicine 127 (2022) 102279

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Automatic sleep stage classification: A light and efficient deep neural network model based on time, frequency and fractional Fourier transform domain features

Yuyang You [a], Xuyang Zhong [a,c], Guozheng Liu [a], Zhihong Yang [b,*]

[a] Beijing Institute of Technology, Beijing, China
[b] Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[c] Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
Sleep stage classification
Fractional Fourier transform
Bidirectional LSTM

ABSTRACT

This work proposed a novel method for automatic sleep stage classification based on the time, frequency, and fractional Fourier transform (FRFT) domain features extracted from a single-channel electroencephalogram (EEG). Bidirectional long short-term memory was applied to the proposed model to train it to learn the sleep stage transition rules according to the American Academy of Sleep Medicine's manual for automatic sleep stage classification. Results indicated that the features extracted from the fractional Fourier-transformed single-channel EEG may improve the performance of sleep stage classification. For the Fpz-Cz EEG of Sleep-EDF with 30 s epochs, the overall accuracy of the model increased by circa 1% with the help of the FRFT domain features and even reached 81.6%. This work thus made the application of FRFT to automatic sleep stage classification possible. The parameters of the proposed model measured 0.31 MB, which are 5% of those of DeepSleepNet, but its performance is similar to that of DeepSleepNet. Hence, the proposed model is a light and efficient model based on deep neural networks, which also has a prospect for on-device machine learning.

1. Introduction

Sleep is an active and regulated process with an essential restorative function for physical and mental health [1]. Monitoring and evaluating the quality of sleep exerts an important effect on people's health and the diagnosis of sleep-related diseases.

Sleep experts determine the quality of sleep using electrical activity recorded from sensors attached to different parts of the body. A set of signals from these sensors is reflected in a polysomnogram (PSG), which consists of electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and electrocardiogram (ECG) [2]. According to several sleep manuals, such as those of Rechtschaffen and Kales (R&K) [3] and the American Academy of Sleep Medicine (AASM) [4], human sleep processes can be classified into different stages. In the AASM manual, the sleep stages are categorised as awake (W), non-rapid eye movement (NREM) and rapid eye movement (REM). NREM sleep is further divided into three stages: the N1, N2 and N3 stages. For the longest time, sleep stage classification has relied on a manual and time-consuming approach.

Several recent studies have attempted to develop machine learning methods on the basis of the channels of PSG signals to help classify sleep stages automatically. These studies have two main directions. One direction aims to utilise a number of algorithms to extract features from PSG signals, and some manual feature engineering based algorithms represented by EMD [5], FBSE-EWT [6], DESA [7], FBDM [8] have achieved very good results. The method proposed by [9] used tunable-Q wavelet transform (TQWT) based filter-bank to decompose the segment of ECG signal into several constant bandwidth sub-band signals, centred correntropies are computed from the various sub-band signals. In [7], EEG signals are decomposed using iterative filtering method, and the discrete energy separation algorithm (DESA) was applied to the modes to determine amplitude envelope and instantaneous frequency functions, which was used to compute features. Ref. [10] decomposed both HBI and EDR signals into modes using the sliding mode singular spectrum analysis (SM-SSA), and extract features from each mode to detect sleep apnea. The smoothed pseudo Wigner–Ville distribution (SPWVD) [11] and Fourier-Bessel decomposition method (FBDM) [8] are used to extract time-frequency representation from EEG signals. These studies

* Corresponding author.
E-mail address: zhyang@implad.ac.cn (Z. Yang).

https://doi.org/10.1016/j.artmed.2022.102279
Received 10 April 2021; Received in revised form 28 February 2022; Accepted 7 March 2022
Available online 9 March 2022

then use the extracted features as input to train classifiers by using models such as random forest [12] and ensemble support vector machine (SVM) [13], all of which [14–17] are based on statistical learning. The other direction aims to construct models based on deep neural networks (DNNs) which can be trained to extract useful features from raw data automatically. Such methods are based on deep learning. The classification performance of DNNs has been proved effective in previous studies, most of them [2,18,20,21] are based on CNN and RNN; Ref [22,23] introduced attention mechanism into the sleep staging; a latest research [24] also proposed a graph-temporal fused CNN model. We can perceive that more and more DNNs with different structures were successfully applied in the field of sleep staging. However, most of these neural networks can only be run on high-performance devices, such as GPUs and TPUs, due to their large parameters. Besides, slight performance gains in models often come at the cost of ever-increasing number of parameters.

Fourier transform is generally utilised to analyse signals in the frequency domain, but extracting significant features from nonstationary signals is difficult. A solution to this problem is the fractional Fourier transform (FRFT) [25]. FRFT can be used to analyse nonstationary signals in a time–frequency domain, on which the information of signals is concentrated. Information that is difficult to extract in the time or frequency domain, can be captured by FRFT. FRFT has been widely applied to many fields, including medical image registration [26], moving target detection [27], speech enhancement [28] and so on. However, this technique has yet to be used in sleep stage classification.

In the current work, we proposed a novel method for automatic sleep stage classification that is based on the time, frequency and FRFT domain features extracted from a single-channel EEG in two different datasets. The results indicated that the features extracted from fractional Fourier-transformed single-channel EEG signals improve the performance of sleep stage classification, and the DNN-based classifier provides better generalization.

We applied bidirectional long short-term memory (Bi-LSTM) [29,30] as the classifier. A previous study [2] demonstrated that Bi-LSTM can be trained to learn temporal information, such as sleep stage transition rules [4], from the time, frequency and FRFT domain features. Such information is used by sleep experts to determine the next possible sleep stages on the basis of previous stages. We also applied the batch normalisation and dropout layers to our model, and the L2 weight decay and discrete-descending learning rate to the training algorithm to alleviate overfitting. The gradient clipping technique was used to avoid gradient explosion caused by Bi-LSTM.

The classification performance of the proposed model is similar to that of the SOTA model, but our model is much smaller. After feature extraction, its running efficiency in training and inference is much higher. We can infer that the proposed model has a bright prospect for on-device machine learning.

2. Materials and methods

2.1. Materials

The experimental data were obtained from two public datasets: Sleep-EDF [31,32] and the Montreal Archive of Sleep Studies (MASS) [33].

2.1.1. Sleep-EDF

Two sets of subjects were established from two studies, namely, the study on the effect of age on healthy subjects (SC) and the effects of temazepam on sleep (ST). Each recording contained two scalp-EEG signals from the Fpz-Cz and Pz-Oz channels, 1 EOG (horizontal), 1 EMG and 1 oro-nasal respiration [2]. We used the EEG recordings of the Fpz-Cz and Pz-Oz channels from 20 SC subjects (aged 28.7 $\pm$ 2.9 years) as our training and validation data, respectively. The sampling rate was set to 100 Hz. The recordings were segmented into 30 s epochs, and each epoch was manually classified into one of the eight classes (W, N1, N2, N3, N4, REM, MOVEMENT and UNKNOWN) by sleep experts according to the R&K standard. We utilised the AASM standard as the sleep stage classification standard in this work. Thus, we merged the N3 and N4 stages into the single-stage N3. As we are only interested in sleep periods, we included 30 min of awake periods just before and after the sleep periods. We also excluded the MOVEMENT and UNKNOWN stages, as they do not belong to the five sleep stages [4].

Table 1 summarises the number and percentage of the 30 s epochs corresponding to each sleep stage and the demographic information of the Sleep-EDF dataset.

2.1.2. MASS

MASS cohort 1 comprised five subsets of recordings (SS1-SS5) which were organised according to their research and acquisition protocols. We used data from SS3, which contained PSG recordings from 62 healthy subjects (aged 42.5 $\pm$ 18.9 years). Each recording contained 20 scalp-EEG, 2 EOG (left and right), 3 EMG and 1 ECG channels. The EEG electrodes were positioned according to the international 10–20 system, and the EOG electrodes were positioned diagonally on the outer edges of the eyes. All EEG and EOG recordings had the same sampling rate of 256 Hz. As we downsampled the data to 100 Hz before feature extraction, we did not need to adjust the parameters of the feature extraction algorithms. The recordings were manually classified into one of the five sleep stages (W, N1, N2, N3 and REM) by a sleep expert according to the AASM standard. The movement artefacts at the beginning and end of each subject's recordings were labelled as UNKNOWN. We evaluated our model using the F4-EOG (left) channel, which was obtained via montage reformatting [34].

Table 2 summarises the number and percentage of the 30 s epochs corresponding to each sleep stage and the demographic information of the MASS dataset.

2.2. Methods

2.2.1. Pre-processing

EEG signals are mainly composed of alpha (α) (8–13 Hz), beta (β) (12–30 Hz), theta (θ) (4–8 Hz), delta(δ) (0.5–2 Hz) characteristic waves [35]. To eliminate noise and undesired eye movement, we filtered the recordings by using a bandpass Butterworth filter (order of 1) with a low cut-off of 0.5 Hz and a high cut-off of 35 Hz. Fig. 1 shows a raw (top) and a filtered (bottom) 30 s EEG epoch, respectively.

Table 1
Number and percentage of 30 s epochs corresponding to each sleep stage and demographic information of Sleep-EDF.

Number of epochs						
	W	N1	N2	N3	REM	Total
Number	7420	2529	16,025	5158	6868	38,000
Percentage	19.53%	6.66%	42.17%	13.57%	18.07%	100%

Sex distribution			
	Male	Female	Total
Subjects	10	10	20

Age characteristics			
	Mean	Std	Range
Total	28.7	2.9	25–34
Men	29.1	3.5	25–34
Women	44.2	2.3	26–32

Table 2
Number and percentage of 30 s epochs corresponding to each sleep stage and demographic information of MASS.

Number of epochs						
	W	N1	N2	N3	REM	Total
Number	5435	4032	25,287	6571	8925	50,250
Percentage	10.82%	8.02%	50.32%	13.08%	17.76	100%

Sex distribution			
	Male	Female	Total
Subjects	28	34	62

Age characteristics			
	Mean	Std	Range
Total	42.5	18.9	20–69
Men	40.4	19.4	20–69
Women	44.2	18.6	20–69

2.2.2. Fractional Fourier transform (FRFT)

FRFT is a general form of Fourier transform. FRFT reflects the information of a signal in the time and frequency domains and is suitable for processing nonstationary signals. Its discrete algorithm is mature and fast [25]. As the EEG signal is a type of nonstationary signal, Fourier transform cannot extract its significant characteristics well. However, with the help of FRFT, the property of EEG signals in the time-frequency domain where the information is concentrated, can be analysed. In other words, FRFT rotates a signal with an angle in the time-frequency plane. If the angle is π/2, FRFT is equivalent to Fourier transform. Therefore, some information in the time-frequency domains can be captured by extracting features from the fractional Fourier-transformed EEG signals.

We applied the algorithm, which Ozakats proposed in Ref. [36], to calculate the discrete FRFT of the EEG signals. p-order FRFT is defined as

$$X_p(u) = \begin{cases} \sqrt{\frac{(1-jcot\alpha)}{2\pi}} e^{j\frac{u^2}{2}cot\alpha} \int_{-\infty}^{+\infty} x(t) e^{j\frac{t^2}{2}cot\alpha - jutcsc\alpha} dt, \alpha \neq n\pi \\ x(u) \quad , \ \alpha = 2n\pi \\ x(-u) \quad , \ \alpha = (2n \pm 1)\pi \end{cases} \tag{1}$$

Eq. (1) can be rewritten as

$$X_p(u) = \sqrt{\frac{(1-jcot\alpha)}{2\pi}} e^{j\frac{u^2}{2}cot\alpha} \int_{-\infty}^{+\infty} x(t) e^{j\frac{t^2}{2}cot\alpha} e^{-jutcsc\alpha} dt \tag{2}$$

where angle $\alpha = p\pi/2$, order $p \in [-1, 1]$ and j stands for the imaginary part of the complex number. Fig. 2 shows an example of a pre-processed epoch and its amplitude with 0.5-order FRFT, respectively. The FRFT of the dimensionally normalised signal $x(t)$ can be divided into the following three steps:

(a) Modulate signal $x(t)$ with a chirp signal:

$$g(t) = exp\left[-j\pi t^2 tan(\alpha/2)\right] x(t) \tag{3}$$

(b) Convolute signal $g(t)$ with another chirp signal:

$$g'(t) = \sqrt{\frac{(1-jcot\alpha)}{2\pi}} \int_{-\infty}^{+\infty} exp\left[\frac{j\pi}{sin(a)}(t-\tau)^2\right] g(\tau) d\tau \tag{4}$$

(c) Modulate signal $g'(t)$ with the same chirp signal in (a):

$$f_a(t) = exp\left[-j\pi t^2 tan(\alpha/2)\right] g'(t) \tag{5}$$

2.2.3. Feature extraction

We extracted time, frequency and FRFT domain features from the pre-processed 30 s EEG epochs.

2.2.3.1. Time domain features. Features can be further divided into statistical and non-linear features. In this work, we selected {**mean**, **energy**, **variance**, **minimum**, **maximum**, **median**, **25 percentile**, **75 percentile**, **skewness**, **kurtosis**, **Hjorth parameters**, **AR coefficient**, **CID coefficient**, **AC coefficient**, **PAC coefficient**} as the statistical features and {**Hurst exponent**, **HFD**, **PFD**, **LZC**, **binned entropy**} as the nonlinear features.

A few relatively unusual algorithms of the time domain feature extraction are introduced below:

(a) Hjorth Parameters

Hjorth proposed three parameters in the time domain to determine the characteristics of EEG signals [37]. These parameters are activity (A), mobility (M) and complexity (C), respectively. A is equivalent to variance and denotes the energy of signals. M denotes the average frequency of signals. C denotes the frequency variation of signals. The definitions of mobility and complexity are

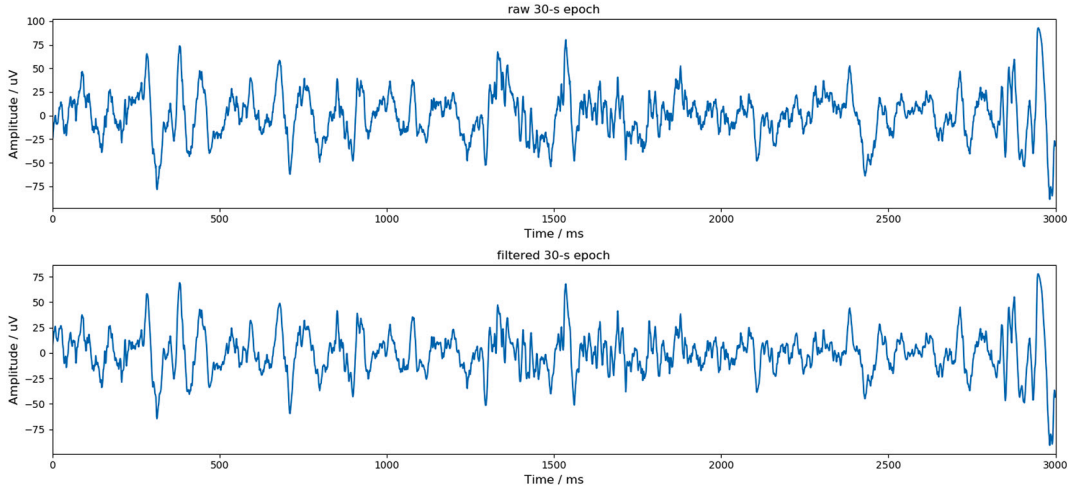


Fig. 1. A raw 30 s EEG epoch (top) and a 0.5–35 Hz bandpass filtered 30 s EEG epoch (bottom).

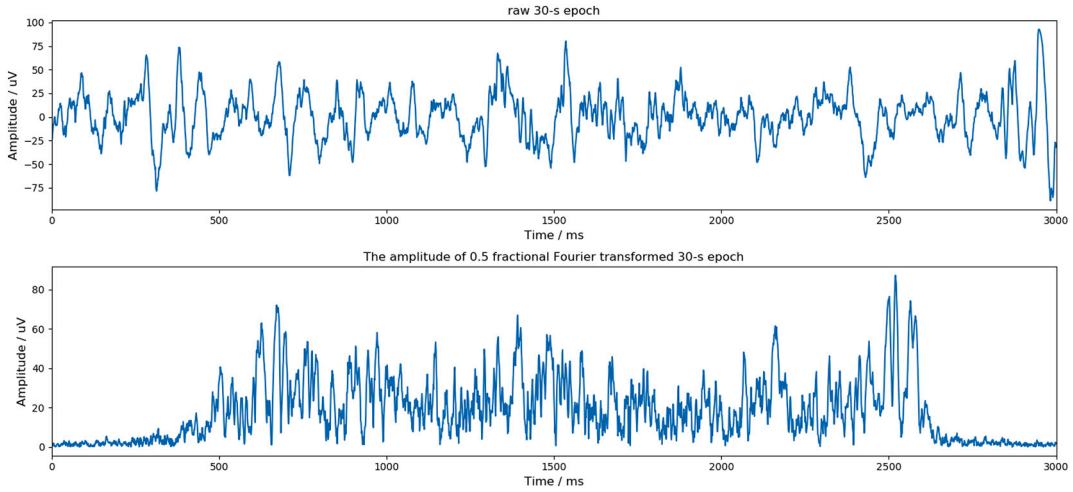


Fig. 2. A pre-processed 30 s epoch (top) and its amplitude after 0.5-order FRFT (bottom).

$$M = \frac{\sigma_1}{\sigma_0} \tag{6}$$

$$C = \sigma_2\sigma_0/{\sigma_1}^2 \tag{7}$$

where σ_0 is the standard deviation of signals and σ_1 and σ_2 are the standard deviation of the first order and second order differential of signals, respectively.

(b) AR Coefficient

This method fits the unconditional maximum likelihood of an autoregressive AR(k) process. This process can be seen as using the previous values of signals x to predict the current value [38]. The k parameter is the maximum lag of the process. Consider a time series $x(t)$; the definition of its AR process is

$$x(t) = c + \sum_{i=1}^{k} \varphi_i x(t-i) + \varepsilon_t \tag{8}$$

where c is a constant value; ε_t is a random disturbance value whose mean is 0 and its standard deviation is σ; and σ is a constant for any t. φ_i is the AR coefficient.

(c) CID Coefficient

The algorithm is an estimator for a time series complexity [39] (a more complex time series has more peaks, valleys and so on). Consider a time series $x(t)$ whose length is n. The definition of the CID coefficient is

$$CID = \sqrt{\sum_{t=0}^{n-2} (x(t) - x(t-1))^2} \tag{9}$$

(d) AC Coefficient

The autocorrelation (AC) coefficient measures the correlation degree of the same event in two different periods to measure the influence of the previous value on the current value [40]. The definition of the AC coefficient is

$$\frac{1}{(n-l)\sigma^2} \sum_{t=1}^{n-l} (x(t) - \mu)(x(t+l) - \mu) \tag{10}$$

where n is the length of time sequence x, σ^2 its variance, μ its mean. l denotes the lag.

(e) PAC Coefficient

This algorithm calculates the value of the partial autocorrelation (PAC) function at the given lag. The PAC of lag k of a time series $\{x_t, t = 1 \ldots T\}$ equals the partial correlation of x_t and x_{t-k} adjusted for the intermediate variables $\{x_{t-1}, \ldots x_{t-k+1}\}$ [41]. The PAC coefficient α_k can be defined as

$$\alpha_k = \frac{Cov(x_t, x_{t-k} | x_{t-1}, \ldots, x_{t-k-1})}{\sqrt{Var(x_t | x_{t-1}, \ldots, x_{t-k-1}) Var(x_{t-k} | x_{t-1}, \ldots, x_{t-k-1})}} \tag{11}$$

where $Cov(\bullet)$ denotes the covariance and $Var(\bullet)$ denotes the variance.

(f) Hurst Exponent

The Hurst exponent is a measure for the 'long-term memory' of a time series, that is, the long statistical dependencies in the data that do not originate from cycles [42].

The rescaled range (R/S) approach is derived from Hurst's definition. The time series $\boldsymbol{X}$ (its length is N) is split into n non-overlapping subseries $\boldsymbol{x}_i$ ($i = 1, \ldots, n$), with a length of r. Then, R_i and S_i are calculated for each subseries and the mean is taken over all subseries yielding R and S. R_i and S_i are defined as

$$y_{ij} = x_{ij} - \overline{\boldsymbol{x}_i} \tag{12}$$

$$z_{ij} = \sum_{k=1}^{j} y_{ik} \tag{13}$$

$$R_i = max(z_{ij}) - min(z_{ij}) \tag{14}$$

$$S_i = \mathrm{std}(\boldsymbol{x}_i) \tag{15}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, r$.

In calculating the Hurst exponent, the number of subseries set with different lengths should be more than one. Take for example a subseries length set $\{r_1, r_2, \ldots, r_m\}$; then, R and S should be calculated for m times to yield $(R/S)_m$. Finally, the Hurst exponent is obtained by fitting a straight line to the plot of $\log((R/S)_m)$ vs. $\log(m)$. The Hurst exponent is the slope of the line.

(g) Higuchi Fractal Dimension (HFD)

This algorithm is regarded as the most stable estimator of the fractal dimension (FD) and is a computationally fast FD [43]. Previous authors found that the Higuchi fractal dimension (HFD) can successfully discriminate between individual stages of sleep and is particularly

suitable in distinguishing N3 from any other sleep stages [44].

For a given signal $x(n)$, its length is N. p new time series x_m^p are constructed as follows: $x_m^p = \{x(m), x(m+p), x(m+2p), \ldots, x(m+\lfloor (N-m)/p \rfloor p)\}$, $m = 1, 2, \ldots, p$. Here, m and p are integers that indicate the initial time and time interval, respectively. Then, for each time series x_m^p, the average length L(p) is obtained by

$$\mathrm{L}(p) = 1 \Bigg/ p \sum_{m=1}^{p} \left[\left(\sum_{i=1}^{\lfloor \frac{N-m}{p} \rfloor} |x(m+ip) - x(m+(i-1)p)| \right) \cdot \frac{N-1}{\lfloor \frac{N-m}{p} \rfloor p} \right] \Bigg/ p \tag{17}$$

where $\frac{N-1}{\lfloor \frac{N-m}{p} \rfloor p}$ is a factor for normalisation. From the curve of log[L(p)] vs log[1/p], the slope of the least-squares linear best fit is the estimate of HFD [45].

(h) Petrosian Fractal Dimension (PFD)

The Petrosian fractal dimension (PFD) provides a fast calculation of FD by turning a signal into a binary sequence [46]. PFD can be defined as

$$\mathrm{PFD} = log_{10}k \Bigg/ \left(log_{10}k + log_{10}\frac{k}{k + 0.4N_\delta} \right) \tag{18}$$

where k denotes the number of samples of the signal. N_δ denotes the number of signal changes in the signal derivatives.

(i) Lempel-Ziv Complexity (LZC)

Lempel-Ziv complexity (LZC) is a nonparametric measure of complexity whose large values correspond to high complexity data. It has been applied to the context of wake and sleeps stage diagnosis during anesthesia [47].

LZC is computed by transforming the EEG signal x(n) into a sequence of symbols (zero and one) on the basis of the comparison of samples given a predefined threshold T_d. The binary sequence is scanned from left to right, and the complexity counter $c(N)$ is increased by one unit every time a new subsequence of consecutive characters is encountered. Finally, the normalised complexity is defined as [45]

$$C(N) = (c(N)/(N/log_2 N)) \tag{19}$$

(j) Binned Entropy (BE)

Entropy represents the complexity of time series [40]. However, the computation of entropy, such as approximate entropy and sample entropy, is time-consuming. In this work, we applied the binned entropy (BE) due to its fast computation. The BE method first bins the values of time series x into n equidistant bins. Then, it calculates the value of

$$-\sum_{k=1}^{n} p_k log(p_k) \cdot 1_{(p_k > 0)} \tag{20}$$

where p_k is the percentage of samples in bin k.

2.2.3.2. Frequency domain features. EEG signals are mainly composed of alpha (α) (8–13 Hz), beta (β) (12–30 Hz), theta (θ) (4–8 Hz), delta (δ) (0.5–2 Hz) characteristic waves. Therefore, extracting the frequency domain features from EEG signals is meaningful. In this work, we selected {**Relative Spectral Power**, **Harmonic parameters**} as the frequency domain features.

(a) Relative spectral power

The spectral analysis provides several important features. For each signal X, an FFT squared modulus estimator is applied to estimate the power spectral density (PSD). The spectrum is divided into five frequency subbands (Table 3) [48]. For each frequency subband, the relative spectral power is computed. This parameter is given by the ratio between the subband spectral power (BSP) and the total spectral power, i.e. the sum of all five BSP subbands [49]. Moreover, the spectral band's delta, theta, and alpha can be highlighted over slow wavebands utilising slow-wave indexes defined by the following ratios:

$$\mathrm{DSI} = \frac{\mathrm{BSP}_{delta}}{\mathrm{BSP}_{theta} + \mathrm{BSP}_{alpha}} \tag{21}$$

$$\mathrm{TSI} = \frac{\mathrm{BSP}_{theta}}{\mathrm{BSP}_{delta} + \mathrm{BSP}_{alpha}} \tag{22}$$

$$\mathrm{ASI} = \frac{\mathrm{BSP}_{alpha}}{\mathrm{BSP}_{theta} + \mathrm{BSP}_{delta}} \tag{23}$$

where TSI, ASI [50] and DSI denote the theta, alpha and delta slow-wave index, respectively.

(b) Harmonic Parameters

The harmonic parameters of EEG signals include three parameters: the centre frequency (CF) f_c, bandwidth (BW) f_σ and the spectral value at the centre frequency S_{f_c}. These parameters are defined as follows [51]:

$$f_c = \sum_{f_L}^{f_H} f p_{xx}(f) \Bigg/ \sum_{f_L}^{f_H} p_{xx}(f) \tag{24}$$

$$f_\sigma = \sqrt{\sum_{f_L}^{f_H} (f - f_c)^2 f p_{xx}(f) \Bigg/ \sum_{f_L}^{f_H} p_{xx}(f)} \tag{25}$$

$$S_{f_c} = p_{xx}(f_c) \tag{26}$$

where $p_{xx}(f)$ denotes the PSD, which is calculated for the frequency bandf_L-f_H (Table 4). These parameters allow the analysis of a specific brand in the EEG spectrum [48].

2.2.3.3. FRFT domain features. To extract the FRFT domain features, we fractional Fourier-transformed the EEG signals at given orders (Section 2.2.6.1) and extracted specific features from each order of the fractional Fourier-transformed 30 s EEG epochs. We selected {**AR coefficient**, **CID coefficient**, **PAC coefficient**, **binned entropy with 4 bins**} as the FRFT domain features. These features, which were extracted from the fractional Fourier-transformed EEG signals, were the FRFT domain features.

2.2.4. Model

The architecture of the model shown in Fig. 3 consists of a batch normalisation (BN) layer [52], Bi-LSTM layer, dropout layer [53], and SoftMax layer (linear layer with SoftMax activation). We applied Adam as the optimiser and cross-entropy as the loss function. The BN layer was used to normalise the inputs and accelerate the convergence velocity. The Bi-LSTM layer was used to train to learn the temporal information, such as sleep stage transition rules, from the extracted features. The

Table 3
Frequencies corresponding to different decomposition levels.

Decomposition	Frequency range (Hz)
D1	25–50
D2	12.5–25
D3 (alpha)	6.25–12.5
D4 (theta)	3.125–6.25
D5 (delta)	0–3.125

Table 4
Spectral subbands used in PSD computation.

Bands	Subbands	Bandwidth f_L-f_H (Hz)
Delta	Delta 1	0.5–2.0
	Delta 2	2.0–4.0
Theta	Theta 1	4.0–6.0
	Theta 2	6.0–8.0
Alpha	Alpha 1	8.0–10.0
	Alpha 2	10.0–12.0
Sigma	Sigma 1	12.0–14.0
	Sigma 2	14.0–16.0
Beta	Beta 1	16.0–25.0
	Beta 2	25.0–35.0

dropout layer was used to alleviate overfitting. The SoftMax layer was used to classify the Bi-LSTM encoded features into five sleep stages. We should note that we applied the L2 weight decay technique to the Bi-LSTM and SoftMax layers to alleviate overfitting and the gradient clipping technique to prevent gradient explosion. We implemented our model by using the PyTorch machine learning framework.

2.2.4.1. Batch normalisation. Whilst extracting the features from the 30 s EEG epochs, the features were not normalised. Therefore, the model performance would be influenced due to the ununified dimension of the features. Batch normalisation is an easy and ingenious method for normalisation. In training, suppose there are values of input x over a mini-batch $B = \{x_{1, \ldots, n}\}$; the values of outputs y can be defined as

$$\mu_B = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{27}$$

$$\sigma_B{}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_B)^2 \tag{28}$$

$$y_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B{}^2 + \varepsilon}}\gamma + \beta \tag{29}$$

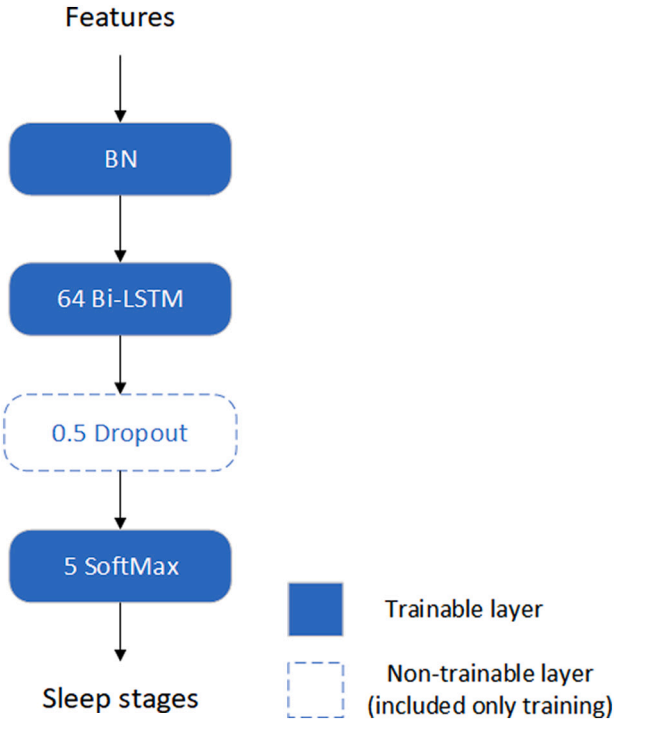


Fig. 3. Architecture of the model. 64 Bi-LSTM means that the hidden size of Bi-LSTM is 64 and its output size is 128 (forward + backward). A 0.5 dropout means that the possibility of dropout is 0.5. 5 SoftMax means that the number of units is 5 (5 sleep stages).

where μ_B is the mean of the mini-batch, $\sigma_B{}^2$ is its variance and ε is a small positive number used to avoid a divisor of 0 (1e-5 herein). γ and β are trainable parameter vectors. By default, the elements of γ and β are set to 1 and 0, respectively. The mean and variance of each mini-batch were recorded for inference.

During inference, the mean and variance for normalising the input are called running mean and running variance, respectively. The running mean μ and running variance σ^2 were fixed during inference, but they could be updated during training. The updating process is defined

$$\mu = (1 - m)\mu + m\mu_B \tag{30}$$

$$\sigma^2 = (1 - m)\sigma^2 + m\sigma_B{}^2 \tag{31}$$

where μ is the running mean, σ^2 is the running variance and m is the momentum set to 0.1 by default. μ_B is the mean of the mini-batch of training data and $\sigma_B{}^2$ is its variance.

2.2.4.2. Bidirectional LSTM (Bi-LSTM). Long short-term memory (LSTM) is capable of solving the long-term dependency problem. One of the advantages of LSTM which the classic recurrent neural network (RNN) does not have, is the capability to apply current information and long-term information in the past to perform the present task. Herein, a gate mechanism was designed for LSTM to memorise long-term memory. Gate is an approach which lets information pass selectively. It consists of a sigmoid (σ) neural network layer and a pointwise multiplication operation. An LSTM cell has three gates: forget gate, input gate and output gate.

Bi-LSTM consists of forward and backward LSTMs which process input sequences forward and backward, respectively. The forward and backward LSTMs are mutually independent. Hence, the model is capable of exploiting information from the past and future to determine the current state [2].

2.2.5. Training method

We first organised the manually extracted features into sequences by arranging the extracted features of each subject in chronological order. The sequence length was 25, and the batch size was 10.

Algorithm 1. Training Algorithm

```
Input: features
Output: model

model = init_model()
for i = 1 to epochs do
    for each subject in features do
        model = reset_lstm_states(model)
        serial_subject = serialize(subject)
        for each batch in serial_subject do
            model = adam(model, batch)
        end for
    end for
end for
return model
```

Then, we performed the training algorithm shown in Algorithm 1.

The elements of the cell and hidden states should be re-initialised to zero at the beginning of each subject's features during the training and inference.

2.2.6. Parameters selection

2.2.6.1. Feature extraction parameters.

1) The order set of FRFT = {0.1, 0.3, 0.5, 0.6} for Sleep-EDF/{0.1, 0.3, 0.5, 0.6, 0.7} for MASS;

2) AR coefficient: k = 10, coefficient index = 4;

3) AC coefficient: lag = 3;

4) PAC coefficient: lag = 2

5) Hurst exponent: subseries length set =exp({start + 0/15 × span, …, start + 14/15 × span }). Where start = l × (1 – ratio) × 0.5, span = l × ratio. l = log(N). N is the length of series;

6) HFD: the maximum of p = 10;

7) LZC: threshold = median of series;

8) Binned entropy: number of bins = {3, 4, 5}.

2.2.6.2. Model parameters.

1) Hidden size of Bi-LSTM = 64;

2) Probability of dropout = 0.5 (during the inference, the probability is 0);

3) Number of units of SoftMax layer = 5.

2.2.6.3. Training parameters.

1) epochs = 100;

2) Batch size = 10, sequence length = 25;

3) Adam: learning rate = 5e-5 (1st-50th epoch), 5e-6 (51st-100th epoch), beta1 = 0.9, beta2 = 0.999;

4) L2 weight decay rate = 1e-4;

5) Threshold of gradient clipping = 10.

2.2.7. Performance metrics

We evaluated the performance of the proposed model using the per-class F1 score (F1), macro-averaging F1 score (MF1), overall accuracy (ACC), and Cohen's kappa coefficient (κ) [54,55]. The ACC and MF1 are calculated as follows:

$$\mathrm{ACC} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c}{N} \tag{32}$$

$$\mathrm{MF1} = \frac{\sum_{c=1}^{C} \mathrm{F1}_c}{C} \tag{33}$$

where TP_c denotes the true positives of class c, $\mathrm{F1}_c$ is the per-class F1-score of class c, C is the number of sleep stages, and N is the total number of test samples [2].

2.2.8. Statistical analysis and feature selection

We performed statistical hypothesis testing to validate the extracted features. This step was taken for two reasons. Firstly, hypothesis is the key to determining the discriminatory capability of the selected features and discovering whether this discriminatory capability is statistically significant or not [56]. Secondly, features that are not statistically significant, if selected mistakenly, can infact end up reducing the performance. These features have to be identified and eliminated from the feature (train and test) matrices before feeding them into the classifier [57].

To assess whether the values of the features in the five sleep stages differ significantly, we performed a one-way analysis of variance [58]. The test was carried out at a 95% confidence level. Hence, a difference was statistically significant if $p < (\alpha = 0.05)$. Any feature having a p-value greater than α was discarded and eliminated from the feature matrices. Tables 1, 2 and 3 in the Supplementary section present the results of the hypothesis testing of the features extracted from Fpz-Cz, Pz-Oz EEG from Sleep-EDF and F4-EOG (left) EEG from MASS, respectively. Some of the features did not pass the test (highlighted in bold). These features were eliminated before feeding the training and test matrices into the model.

In addition to the statistical analysis, we utilised the feature selector from the project of Will Korhrsen (GitHub: https://github.com/WillKoehrsen/feature-selector) to further evaluate the quality of the selected features. The test results showed that the overall accuracy only increased by 0.1 percentage point. Hence, the quality of the selected features was acceptable, but we still discarded the features with low importance and high collinearity before training and testing.

3. Results

3.1. Initial experiments

We initially constructed classifiers on the basis of MLP, Bi-RNN, and Bi-LSTM and then compared their classification performances. MLP performed the worst, and Bi-LSTM performed slightly better than Bi-RNN. The result of the comparison demonstrated that architectures based on RNNs can improve classification performance significantly. We finally chose Bi-LSTM due to its capability of solving the gradient vanishing problem [29]. The model and training parameters mentioned in Sections 2.2.6.2 and 2.2.6.3 were determined through several experiments.

As for the feature selection, we initially extracted the time and frequency domain features from the pre-processed Fpz-Cz 30 s EEG epochs of Sleep-EDF, and extracted the same features from the fractional Fourier-transformed 30 s EEG epochs (the orders are {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}). The number of features reached 737. The hidden size of Bi-LSTM was set to 256. The performance metrics obtained from the 20-fold cross-validation on the 737 features are shown in Table 5. The performance of the 737 features was much worse than that of the time and frequency domain features extracted from pre-processed 30 s EEG epochs (Table 7). The results demonstrated that numerous redundant features caused overfitting, thereby causing the generalized performance to drop steeply. The phenomenon could be explained by the FRFT being an averaging process which focuses the energy of signals on a small region and makes the difference between signals of different types small. The higher the order is, the more obvious the effect of averaging is. Therefore, the overall classification performance declined. However, the results shown in Table 6 and Table 7 suggested that certain features in the FRFT domain could still facilitate the classification of several specific sleep stages. The topic is discussed in Section 4.

The result shown in Fig. 4 suggested that if we only utilise the features extracted from the fractional Fourier-transformed 30 s EEG epochs of the Fpz-Cz channel with a certain order (the same features are extracted from pre-processed 30 s EEG epochs), then the higher-order we set, the worse metrics we will obtain. To ensure the difference

between the selected orders whilst removing the highest-orders, we finally selected {0.1, 0.3, 0.5, 0.6} as the order set of FRFT for Sleep-EDF and {0.1, 0.3, 0.5, 0.6, 0.7} for MASS through many experiments.

The time and frequency domain features mentioned in Section 2.2.3 have been proved effective for sleep stage classification in the existing literature. Hence, we mainly aimed to remove the useless FRFT domain features. For this purpose, we visualised the distributions of the features corresponding to different sleep stages. Fig. 5 shows two examples of the distributions of certain features extracted from pre-processed and 0.6-order fractional Fourier-transformed Fpz-Cz EEG at different sleep stages. On the basis of these figures, we attempted to select the features with the greatest difference in distribution amongst the different sleep stages. Then, through several experiments, we finally selected four features as the FRFT domain features, i.e. {AR coefficient, CID coefficient, PAC coefficient, binned entropy with 4 bins}.

3.2. Classification performance

Table 6 shows the confusion matrix, ACC, MF1, κ and F1 obtained from the 20-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF. The metrics in Table 7 were obtained from the 20-fold cross-validation on the time and frequency domain features from the 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF.

Most metrics on the features, including the FRFT domain features, were better than the metrics on the features without FRFT domain features, except W—F1. For the N2 stage, the F1 increased from 84.61 to 86.24 and ACC, MF1 and κ rose by 1.01%, 0.6% and 0.0116, respectively. However, given the class imbalance problem in the training set, the F1 of the N1 stage was much lower than that of the other stages. In conclusion, the FRFT domain features made contribution to sleep stage classification.

In addition to using Sleep-EDF's Fpz-Cz EEG, we also used Sleep-EDF's Pz-Oz EEG and MASS's F4-EOG (left) EEG to further evaluate our model. Table 8 shows the metrics obtained from the 20-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the Pz-Oz channel of Sleep-EDF, and Table 9 shows the metrics obtained from the 31-fold cross-validation on the features extracted from the 30 s EEG epochs of F4-EOG (left) channel of MASS. Different k values were chosen for the k-fold cross-validation for different datasets because the training and test data were supposed to be composed of different subjects, to consequently make the testing results recognised and objective. The Sleep-EDF dataset consists of 20 subjects (resulting in data on 39 nights; data for one night were collected from the 14th subject, and data for two nights were collected from each of the other subjects). The MASS dataset consists of 62 subjects (resulting in data on 62 nights; data for one night were collected from each subject). Therefore, similar to the authors of Ref. [2], we also used 20- and 31-fold cross-validation for Sleep-EDF and MASS, respectively. We should note that we did not list the metrics obtained from the cross-validation on the time and frequency domain features extracted from the 30 s EEG epochs of the Pz-Oz and F4-EOG (left) channels. Nonetheless, the 30 s EEG epochs of the three channels yielded similar results.

Table 5
Confusion matrix, ACC, MF1, κ and F1 obtained from 20-fold cross-validation on 737 features extracted from Sleep-EDF's Fpz-Cz 30 s EEG epochs.

	ACC/%		MF1/%		κ	
	66.41		57.87		0.5289	
	Predicted					F1/%
	W	N1	N2	N3	REM	
W	**5039**	261	875	968	277	77.62
N1	206	**388**	1480	259	196	22.78
N2	214	125	**12,260**	3011	415	70.85
N3	18	0	245	**4891**	4	67.06
REM	86	103	3723	299	**2657**	51.01

Table 6
Confusion matrix, ACC, MF1, κ and F1 obtained from 20-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF.

	ACC/%		MF1/%		κ	
	81.61		**74.66**		**0.7468**	
	Predicted					F1/%
	W	N1	N2	N3	REM	
W	**6139**	329	211	100	641	84.57
N1	336	**818**	695	25	655	**39.71**
N2	415	181	**13,826**	969	634	**86.24**
N3	52	0	429	**4669**	8	**85.42**
REM	157	263	879	11	**5558**	**77.39**

Table 7
Confusion matrix, ACC, MF1, κ and F1 obtained from 20-fold cross-validation on the time and frequency domain features extracted from Sleep-EDF Fpz-Cz 30 s EEG epochs.

	ACC/%		MF1/%		κ	
	80.60		74.06		0.7352	
	Predicted					F1/%
	W	N1	N2	N3	REM	
W	**6214**	379	206	94	527	**85.82**
N1	348	**850**	638	43	650	39.66
N2	347	184	**13,219**	1312	963	84.61
N3	36	0	304	**4798**	20	84.08
REM	116	344	854	7	**5547**	76.12

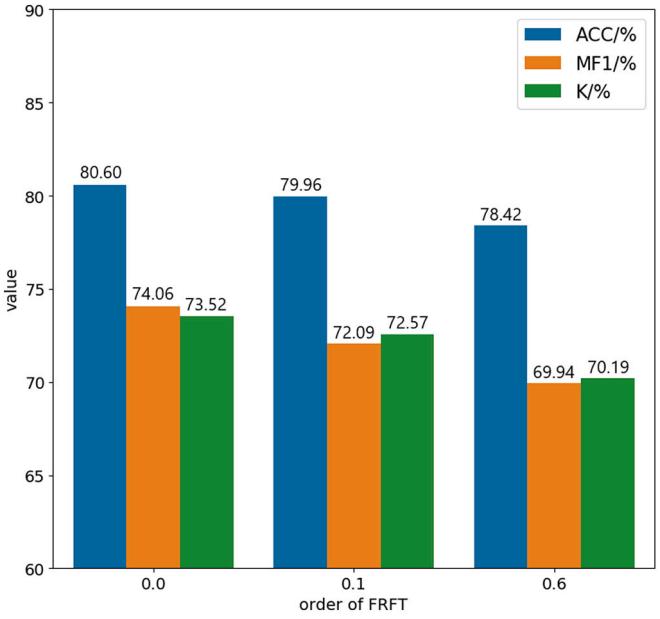
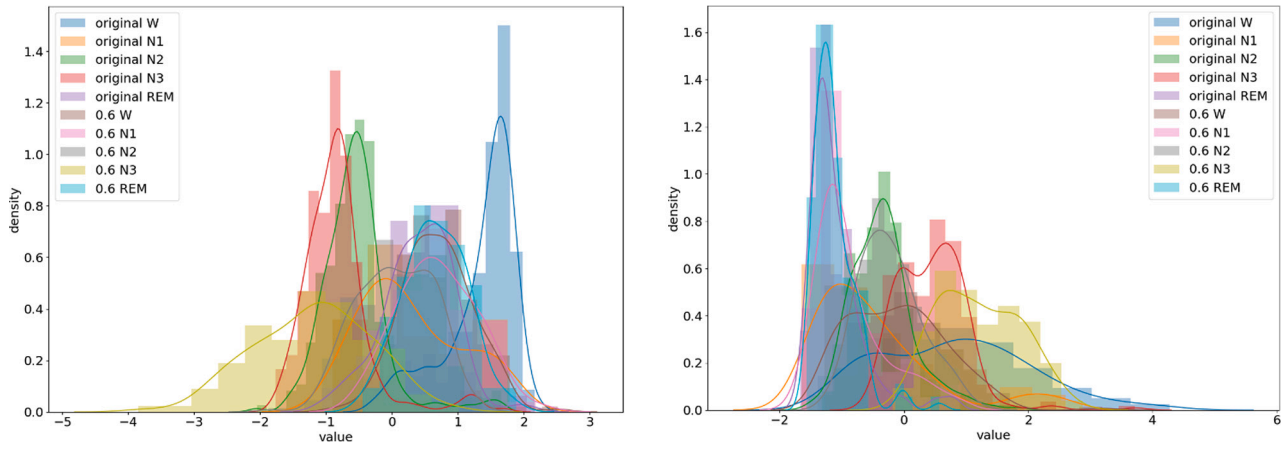


Fig. 4. ACC, MF1, and κ obtained from 20-fold cross-validation on time and frequency domain feature extracted from the fractional Fourier-transformed 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF.

As indicated by the metrics listed in Table 6, Table 8 and Table 9, the Fpz-Cz channel gave a slightly better performance than the Pz-Oz channel in the Sleep-EDF dataset, but the model using the F4-EOG (left) channel performed the best. The results is explained as follows. Firstly, the larger the amount of training data is, the better the performance is. The amount of data from MASS is larger than that of Sleep-EDF. Thus, the two datasets gave significantly different performances.



(a) AR coefficient (b) CID coefficient

Fig. 5. Examples of the distributions of features extracted from pre-processed and 0.6 order fractional Fourier transformed 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF at different sleep stages.

Table 8
Confusion matrix, ACC, MF1, κ and F1 obtained from 20-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the Pz-Oz channel of Sleep-EDF.

	ACC/%		MF1/%		κ	
	80.64		73.17		0.7318	
	Predicted					F1/%
	W	N1	N2	N3	REM	
W	**6628**	394	180	32	186	85.88
N1	636	**769**	632	41	451	37.43
N2	233	211	**13,889**	1075	617	85.05
N3	16	0	970	**4171**	1	79.55
REM	501	206	963	10	**5188**	77.95

Secondly, in the same dataset, the data quality determines the performance. Thus, we inferred that the quality of the Fpz-Cz channel may be relatively good in Sleep-EDF. Similar results were also obtained in Ref. [2] and Ref. [15].

To further highlight the performance of the proposed model, we also compared the performance of the proposed model with that of some benchmark classifiers, i.e. MLP [59], k-NN [60], SVM [61] and random forest [62]. Note that we only compared the metrics obtained from the first-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the F4-EOG (left) channel of MASS. The results shown in Table 10 indicated that our model performed much better than these benchmark classifiers and that our model is thus reliable and competent for sleep stage classification.

Table 9
Confusion matrix, ACC, MF1, κ and F1 obtained from 20-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the F4-EOG (left) channel of MASS.

	ACC/%		MF1/%		κ	
	84.28		77.78		0.7666	
	Predicted					F1/%
	W	N1	N2	N3	REM	
W	**4573**	445	206	16	195	83.78
N1	554	**1673**	1139	5	671	47.93
N2	197	471	**22,886**	1167	566	89.30
N3	20	1	1211	**5338**	1	81.45
REM	147	359	527	11	**7881**	86.42

The MLP comprises three layers, and the unit numbers of each layer are 64, 64 and 5. For k-NN, the number of neighbours is 30. For SVM, the kernel is rbf, the degree of the polynomial kernel function is 3. For random forest, the number of estimators is 50, the maximum of features for one estimator is $\log_2 n$, and n is the total number of features.

3.3. Comparison of proposed method with other methods

In the related literature, most methods can be classified in to two groups: non-independent and independent training and test datasets. The non-independent ones were the methods that included parts of the test subjects' epochs in the training data whilst the independent ones were the methods that excluded all epochs of the test subjects from the training data. We believe that the practical evaluation scheme should not include any epochs from the test subjects [2]. Furthermore, the non-independent scheme has to been shown to result in an improvement in performance [63]. Thus, we only compared the performance of our method with that of the independent group. Table 11 summarises the comparison between our method and other sleep stage classification methods in the independent group in terms of ACC, MF1, κ and F1. In the models using hand-engineered features, the proposed method was better than most methods. Even though the overall accuracy of the model reported in Ref. [6] reached 0.8973 and its W—F1 reached 0.97, its MF1 was much lower than ours. This difference was due to the 30 s epochs at the W stage accounted for 68% of the training set as reported in Ref. [12]; in this case, a severe class imbalance problem existed in the dataset, as shown in the F1 scores of the other stages. Our method was also better than the method based on raw 30 s epochs that applied convolution neural network (CNN) [20]. DeepSleepNet performed better than our method using the Fpz-Oz and F4-EOG (left) EEG channels, but it performed worse than our model using the Pz-Oz EEG channel. DeepSleepNet consists of two main parts, namely, CNN and Bi-LSTM.

Table 10
Comparison between our model and some benchmark classifiers (MLP, k-NN, SVM and random forest). The metrics were obtained from the first-fold cross-validation on the time, frequency and FRFT domain features extracted from the 30 s EEG epochs of the F4-EOG (left) channel of MASS.

Model	ACC/%	MF1/%	κ
MLP	75.7	71.9	0.67
k-NN	79.5	67.6	0.68
SVM	83.6	73.8	0.75
Random Forest	82.9	73.1	0.74
Ours	**86.8**	**80.1**	**0.80**

Without Bi-LSTM, the performance of DeepSleepNet would decrease sharply [2]. Hence, Bi-LSTM can be trained to learn temporal information, such as sleep stage transition rules, to help improve the classification performance in a given task. As for the reason why DeepSleepNet performed so well, we think that specific architectures based on CNNs can be trained to learn the representations of samples efficiently on the basis of large datasets. In other words, if the size of the training set is sufficient enough, the generalized performance of the CNN would be relatively better than that of other statistical methods. Similar to CNN, Bi-LSTM, which is a type of DNN (in the sense of time depth), can also achieve high generalized performance on the basis of large datasets.

Although classification performance is an important metric to evaluate whether a model is good or bad, we should also consider the running efficiency of models. Most models based on DNNs can only be run on high-performance devices due to their large parameters; hence, they are hardly trained and run on mini AI computers. Fig. 6 shows the relationship between ACC–MF1 and the size of the parameters (params) of our model and other models based on neural networks. We should note that the Fpz-Cz channel of the training and test data was used herein to obtain the metrics. To show the differences in the params of the models properly, we transformed the params into log(1 + params) at the x-axis, where the parameters were stored as Float32 and shown in MB. The ACC and MF1 of our model were worse than those of Ref. [2], but the params of Ref. [2] (59 MB) were much larger than ours (0.31 MB). Hence, our model can be run and trained relatively rapidly after feature extraction, but its performance is similar to that in Ref. [2].

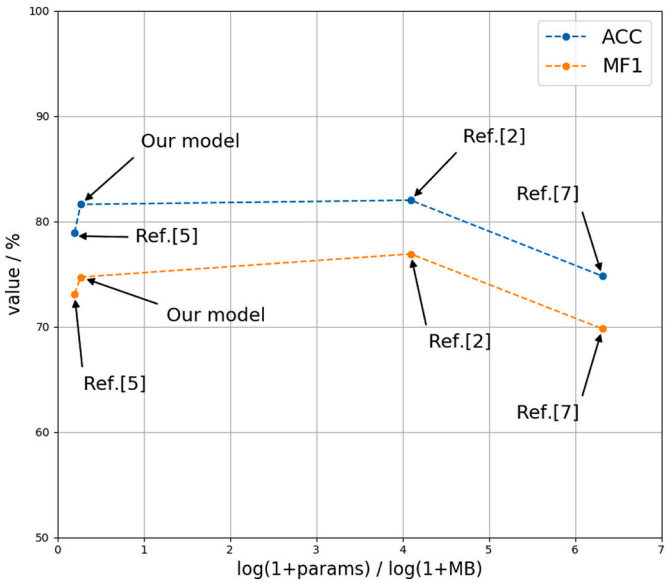


Fig. 6. ACC–MF1 vs. size of model parameters (params). The channel of the training and test data is Fpz-Cz.

3.4. Activation states of Bi-LSTM cells

To demonstrate the explicability of our model, we attempted to visualise the activation states (i.e. output) of the Bi-LSTM cells on different sleep stages. We initially recorded the activation states of the Bi-LSTM cells obtained from the validation set. Then we calculated the means of the activation states at the same sleep stage. Fig. 7 illustrates the means of the activation states of the Bi-LSTM cells at different sleep stages. The red colour denotes the positive activation state, and the blue colour denotes the negative activation state. The darker the colour is, the more active the cell is. Columns 1–4 are the activation states of the forward Bi-LSTM cells, and columns 5–8 are the activation states of the backward Bi-LSTM cells. As indicated in our observations, some cells, such as cell B3-x15 and B4-x13, became increasingly active as sleep deepened. Meanwhile, other cells were only active at specific sleep stage. For instance, the cell F1-x0 was only positively active at the W stage. Therefore, some Bi-LSTM cells can be trained to be sensitive to the temporal evolution of sleep stages whilst other cells can be trained to classify specific sleep stages. Our model utilised the combination of these cells to determine the current sleep stage and to formulate the transition rules.

3.5. Influence of FRFT on classification

The results shown in Table 6 and Table 7 demonstrated that FRFT enabled features to be increasingly contributive. To further validate the availability of the features in the FRFT domain for the improvement of classification performance, we visualised the distributions of the features extracted from the pre-processed 30 s EEG epochs and fractional Fourier-transformed 30 s EEG epochs with the selected order set. For instance, Fig. 8 illustrates the distributions of the AR coefficient extracted from pre-processed and {0.1, 0.3, 0.5, 0.6}-order Fractional Fourier-transformed 30 s EEG epochs of the Fpz-Cz channel. As shown in Fig. 8, the W stage was easily distinguishable from the AR coefficient extracted from the pre-processed 30 s EEG epochs. However, as the order increased, the W and REM stages became increasingly difficult to classify whilst the N3 stage was becoming easily distinguishable. The contributions of some features to the classification of specific sleep stages obviously differ in different FRFT domains. The mechanism of FRFT which influenced the contributions of the features herein is discussed in Section 4.

4. Discussion

We proposed a novel method for automatic sleep stage classification on the basis of the time, frequency and FRFT domain features which were extracted from a single-channel EEG. The results shown in Table 6 and Table 7 suggested that FRFT domain features may improve classification performance. We also proposed a classifier based on Bi-LSTM which can be trained to learn temporal information, such as sleep stage transition rules, from sequential features vectors. Our method combines a statistical learning method and DNN and thus takes

Table 11
Comparison between our method and other sleep stage classification methods in terms of ACC, MF1, κ and F1.

Methods	Dataset	EEG channel	Test Epochs	Overall Metrics			Per-class F1-Score (F1)/%				
				ACC	MF1	κ	W	N1	N2	N3	REM
Ref. [19]	Sleep-EDF	Fpz-Cz	37,022	78.9	73.7	–	71.6	47.0	84.6	84.0	81.4
Ref. [12]	Sleep-EDF	Fpz-Cz	106,376	89.7	62.4	–	97.9	10.7	80.1	52.5	71.1
Ref. [20]	Sleep-EDF	Fpz-Cz	37,022	74.8	69.8	–	65.4	43.7	80.6	84.9	74.5
Ref. [2]	Sleep-EDF	Fpz-Cz	41,950	82.0	76.9	0.76	84.7	46.6	85.9	84.8	82.4
This work	Sleep-EDF	Fpz-Cz	38,000	81.6	74.7	0.75	84.6	39.7	86.2	85.4	77.4
Ref. [2]	Sleep-EDF	Pz-Oz	41,950	79.8	73.1	0.72	88.1	37.0	82.7	77.3	80.3
This work	Sleep-EDF	Pz-Oz	38,000	80.6	73.2	0.73	85.9	37.4	85.1	79.6	78.0
Ref. [2]	MASS	F4-EOG (left)	58,600	86.2	81.7	0.80	87.3	59.8	90.3	81.5	89.3
This work	MASS	F4-EOG (left)	50,250	84.3	77.8	0.77	83.8	47.9	89.3	81.5	86.4

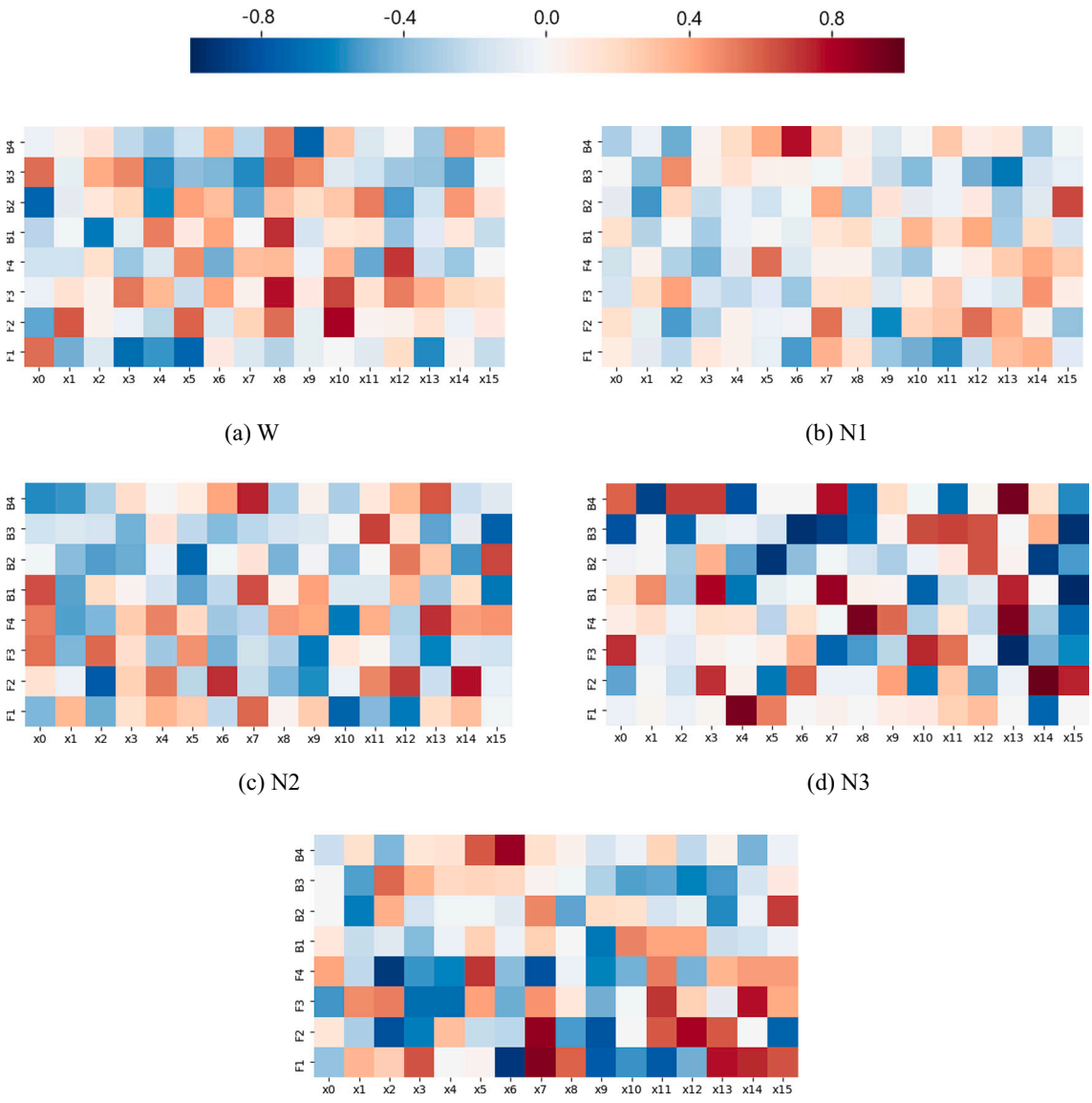
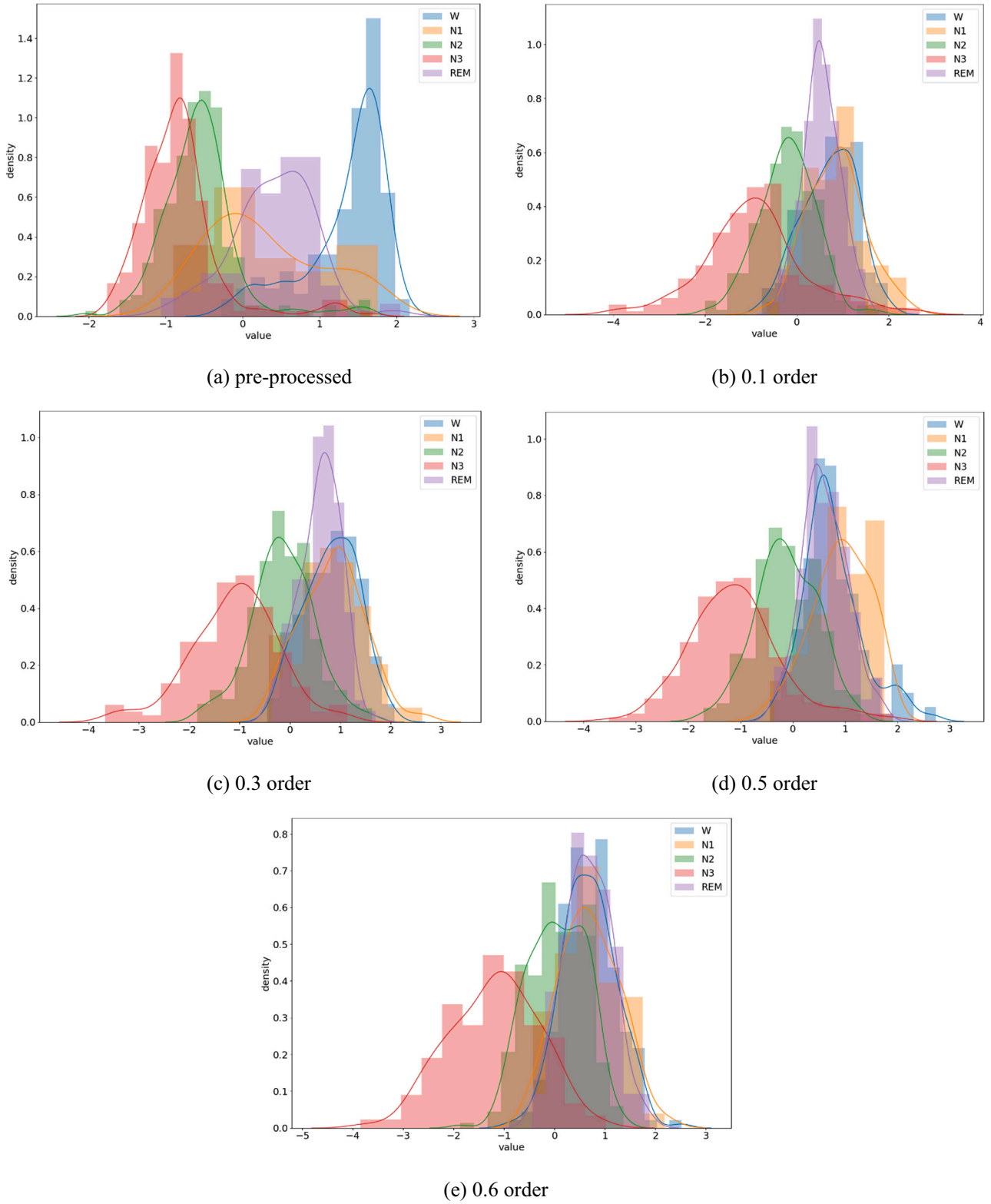


Fig. 7. The means of the activation states obtained from the extracted features at the same sleep stage. Red denotes the positive activation state, and blue denotes the negative activation state. The darker the colour is, the more active the cell is. Column 1–4 present the activation states of the forward Bi-LSTM cells, and column 5–8 show the activation states of the backward Bi-LSTM cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

advantages of both methods. As the results indicated, the performance of the proposed model is similar to that of the SOTA model (i.e. DeepSleepNet), but the parameters of our model are only 5% of those of DeepSleepNet. Hence, our model is a light and efficient model based on deep neural networks.

For automatic sleep stage classification, real-time inference is especially important. The acquisition of EEG signals relies on the electrodes attached to the scalp of the subjects. The acquisition device should be designed to be as small as possible in consideration of wearing comfort because subjects are required to wear such devices all night to collect sufficient data. The operating speed of this type of device is not always fast. Although most models based on DNNs perform well, they can only be run on high-performance devices, such as GPUs and TPUs, because of the large number of parameters they require. In this case, models with few parameters have an advantage in deployment on edge devices. Although our model is also based on DNNs (in the sense of time depth), the number of parameters it requires is only around 5% of that required by DeepSleepNet. The results in Table 11 demonstrated that the performance of the proposed model is similar to that of DeepSleepNet. The efficiency of our model relies on the prior feature extraction. The feature extraction algorithms based on heuristic knowledge are an essential part of statistical learning. They make the cost of obtaining effective representations smaller than that required by the deep learning methods.

As described in Section 3.4, we utilised the activation states of Bi-LSTM cells to demonstrate that the proposed model based on Bi-LSTM is capable of learning the sleep stage transition rules from the sequences of features extracted from EEG. For instance, in our work, some cells became increasingly active as sleep deepened whilst other cells were only active at specific sleep stages. Our model utilised a combination of these cells to determine the current sleep stage and to formulate the transition rules. However, we could not ignore the fact that the cells were less active at the N1 stage. In other words, classifying the samples of the N1 stage was difficult. The difficulty probably stemmed from the class imbalance problem; as the number of samples at



(a) pre-processed

(b) 0.1 order

(c) 0.3 order

(d) 0.5 order

(e) 0.6 order

Fig. 8. Distributions of the AR coefficient extracted from pre-processed and {0.1, 0.3, 0.5, 0.6}-order Fractional Fourier-transformed 30 s EEG epoch of the Fpz-Cz channel.

the N1 stage was the least in the training set, the model was likely trained to prefer to classify samples as the stages with more samples, such as N3. Hence, the cells were less sensitive to the stages with fewer samples, such as N1. This inference is consistent with the performance metrics revealed in Section 3.2.

The results in Section 3.5 show that FRFT is capable of changing the contributions of features to different sleep stages by changing the distribution of features, thereby improving the classification performance. As FRFT rotates a signal with a specific angle in the time–frequency plane, we can find new characteristics of signals which cannot be

observed in the time or frequency domain. Through FRFT, the distribution of signals is changed, and the distribution of the features extracted from them changes naturally. To some degree, the change of distribution may be useful for the classification of some specific sleep stages. Moreover, the rotating angle (i.e. the order of FRFT) is a parameter, which determines the degree of the frequency domain transform of the signals (if the order is 1, then FRFT is equivalent to Fourier transform). As shown in Fig. 8, the variation of the order of FRFT can also cause a change in the contributions of features. In conclusion, FRFT is more flexible than traditional frequency analysis techniques, such as Fourier transform, and the features in the FRFT domain can improve the classification performance. To our knowledge, this work is the first to apply FRFT to sleep stage classification. Hence, this work is original and prospective.

DNNs generally require numerous training samples to obtain a high performance in complex tasks. Therefore, DNNs have high computational complexity that enables them to have a strong learning capability and high generalization performance. According to the performance comparison, our model shares this property with other models based on DNNs. Nevertheless, our model is much lighter.

Our model has two feature extraction processes. The first is the hand-engineered feature extraction. The features were selected according to related statistical learning works [64], optimised by statistical analysis and an automatic feature selector to ensure minimal information loss. Under this premise, the number of model parameters is greatly reduced, thereby accelerating the training and inference processes. The second is the feature extraction based on deep learning. We utilised Bi-LSTM to capture the deep temporal information between epochs. Such an approach is difficult for other models to do. LSTM has also been found efficient in capturing long-term sleep stage transitions [65]. Using a model based on deep learning ensures the generalized performance of the model and makes it generalisable to a large population [2].

However, two limitations are identified in this work. Firstly, the feature extraction process is relatively time-consuming. Secondly, model training still relies on a large amount of data.

In the future, we will focus on tackling the practical deployment issues of the proposed model. We also aim to simplify the model and explore the possibility of few-shot learning, which can facilitate the deployment.

5. Conclusions

We proposed a novel method for automatic sleep stage classification based on the time, frequency, and FRFT domain features extracted from single-channel 30 s EEG epochs. Furthermore, we constructed a classifier based on Bi-LSTM.

The features extracted from the fractional Fourier-transformed single-channel 30 s EEG epochs may improve the performance of sleep stage classification. Using the 30 s EEG epochs of the Fpz-Cz channel of Sleep-EDF, the overall accuracy increased by circa 1% with the help of the FRFT domain features. It demonstrated that FRFT is positively contributive to sleep staging. Our model's performance is similar to that of the SOTA model (i.e. DeepSleepNet), but the parameters of our model are only around 5% of those of DeepSleepNet. Hence, the running efficiency of our model is relatively high, and it can be said that the model has a bright prospect for on-device machine learning.

Through analysis and discussion, we demonstrated that our model is capable of learning the sleep stage transition rules effectively. The features extracted from the fractional Fourier-transformed 30 s EEG epochs can help improve classification performance due to the effects of FRFT on the change in data distribution.

In this work, we made the application of FRFT to automatic sleep stage classification possible. The proposed model based on Bi-LSTM is small-scaled and efficient. At present, our model is a sleep staging model based on DNNs that has great potential in balancing performance and complexity well.

CRediT authorship contribution statement

Xuyang Zhong and Yuyang You contributed to the paper equally and share the first author. Xuyang Zhong, Yuyang You, Guozheng Liu, and Zhihong Yang conducted the experiments and analyses. Xuyang Zhong and Yuyang You wrote the manuscript. Zhihong Yang and Yuyang You provided computational resources and experimental guidance.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

This work is supported by grants from the National Natural Science Foundation of China (81973744, 81473579 and 81273654) and the Beijing Natural Science Foundation (7173267). The funding sources are not involved in the experiments, the results analysis, the writing of the manuscript and in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2022.102279.

References

[1] Zoubek L, Charbonnier S, Lecocq S, Buguet A, Chapotot F. Feature selection for sleep/wake stages classification using data-driven methods. BiomedSignal ProcessControl 2007;2:171–9.
[2] Supratak A, Dong H, Wu C, et al. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans Neural Syst Rehabil Eng 2017;25(11):1998–2008.
[3] Hobson JA. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Electroencephalogr Clin Neurophysiol 1969;26(6):644.
[4] Iber C, Ancoli-Israel S, Chesson Jr AL, Quan Jr SF. The AASM manual for the scoring of sleep and associated events. Westchester, IL, USA: American Academy of Sleep Medicine; 2007.
[5] Hassan AR, Bhuiyan MIH. Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. BiomedSignal ProcessControl 2016;24(Feb.):1–10.
[6] Siddharth T, et al. EEG-based detection of focal seizure area using FBSE-EWT rhythm and SAE-SVM network. IEEE Sensors J 2020;20(19):11421–8.
[7] Sharma Rajeev, Pachori Ram Bilas, Upadhyay Abhay. Automatic sleep stages classification based on iterative filtering of electroencephalogram signals. Neural ComputApplic 2017;28(10):2959–78.
[8] Gupta Vipin, Pachori Ram Bilas. FBDM based time-frequency representation for sleep stages classification using EEG signals. BiomedSignal ProcessControl 2021;64:102265.
[9] Nishad Anurag, Pachori Ram Bilas, Rajendra Acharya U. Application of TQWT based filter-bank for sleep apnea screening using ECG signals. J Ambient Intell HumComput 2018:1–12.
[10] Singh Himali, Tripathy Rajesh Kumar, Pachori Ram Bilas. Detection of sleep apnea from heart beat interval and ECG derived respiration signals using sliding mode singular spectrum analysis. Digital Signal Process 2020;104:102796.
[11] Bajaj Varun, Pachori Ram Bilas. Automatic classification of sleep stages based on the time-frequency image of EEG signals. Comput Methods Programs Biomed 2013;112.3:320–8.
[12] Da Silveira TLT, Kozakevicius AJ, Rodrigues CR. Single-channel EEG sleep stage classification based on a streamlined set of statistical features in the wavelet domain. Med Biol Eng Comput 2017;55(2):343–52.
[13] Alickovic E, Subasi A. Ensemble SVM method for automatic sleep stage classification. IEEE TransInstrumMeas June 2018;67(6).
[14] Diykh M, et al. EEG sleep stages identification based on weighted undirected complex networks. Comput Methods Programs Biomed February 2020;184:105116.
[15] Seifpour S, et al. A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal. Expert Syst Appl 2018;104:277–93.
[16] Krakovská A, et al. Automatic sleep scoring: a search for an optimal combination of measures. Artificial Intelligence in Medicine September 2011;53(1):25–33.
[17] Abdulla S, et al. Sleep EEG signal analysis based on correlation graph similarity coupled with an ensemble extreme machine learning algorithm. Expert Systems with Applications 2019;138:112790.
[18] Yu Yunkai, et al. FASSNet: fast apnea syndrome screening neural network based on single-lead electrocardiogram for wearable devices. Physiological Measurement 2021;42:085005.

[19] Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann Biomed Eng 2016;44(5):1587–97.
[20] Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. Joint classification and prediction CNN framework for automatic sleep stage classification. IEEE TransBiomedEng May 2019;66(5):1285–96.
[21] Kanwal S, Uzair M, Ullah H, Khan SD, Ullah M, Cheikh FA. An image based prediction model for sleep stage identification. In: 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan; 2019. p. 1366–70.
[22] Phan H, Andreotti F, Cooray N, Chén OY, Vos MD. Automatic sleep stage classification using single-channel EEG: learning sequential features with attention-based recurrent neural networks. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA; 2018. p. 1452–5.
[23] Zhu T, Luo W, Yu F. Convolution- and attention-based neural network for automated sleep stage classification. Int J Environ Res Public Health 2020;17:4152.
[24] Cai Q, Gao Z, An J, Gao S, Grebogi C. A graph-temporal fused dual-input convolutional neural network for detecting sleep stages from EEG signals. IEEE TransCircSystIIExpress Briefs Feb. 2021;68(2):777–81.
[25] Pei SC, Ding JJ. Relations between Gabor transforms and fractional Fourier transforms and their applications for signal processing. IEEE TransSignal Process 2007;55(10):4839–50.
[26] XiuJie Zhang, Yi Shen, ShiYong Li, et al. Medical image registration in fractional Fourier transform domain. Optik 2013;124(12):1239–42.
[27] Sun HB, Liu GS, Gu H, et al. Application of the fractional Fourier transform to moving target detection in airborne SAR. IEEE TransAerospElectronSyst 2002;38(4):1416–24.
[28] Zhenli W, Xiongwei Z. Biomedical signal processing and control on the application of fractional Fourier transform for enhancing noisy speech. In: Linchang Z, Yinghong W, editors. Microwave, antenna, propagation and EMC technologies for wireless communications, pp. 1. Piscataway: IEEE; 2005. p. 289–92.
[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[30] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81.
[31] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation Jun. 2000;101:e215–20.
[32] Kemp B, Zwinderman AH, Tuk B, Kamphuisen HAC, Oberyé JJL. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Trans Biomed Eng Sep. 2000;47(9):1185–94.
[33] O'Reilly C, Gosselin N, Carrier J, Nielsen T. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. J Sleep Res 2014;23(6):628–35.
[34] Lagerlund TD. Manipulating the magic of digital EEG: montage reformatting and filtering. AmJElectroneurodiagnTechnol 2000;40(2):121–36.
[35] Hsu YL, Yang YT, Wang J, et al. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. Neurocomputing 2013;104:105–14.
[36] Ozarks M, Arikan O. Digital computation of the fractional Fourier transform. IEEE Trans Signal Process 1996;44(9):2141–50.
[37] Hjorth B. Frequency domain descriptors and their relation to a particular model for the generation of EEG activity. In: Dolce G, Künkel H, editors. CEAN – computerized EEG analysis. Stuttgart: Gustav Fischer Verlag; 1975. p. 3–8.
[38] Akaike H. Fitting autoregressive models for prediction. AnnInstStatMath 1969;21(1):243–7.
[39] Batista, Gustavo EAPA. CID: an efficient complexity-invariant distance for time series. Data Min Knowl Discov 2014;28(3):634–69.
[40] Yan Wang. Applied time series analysis. China Renmin University Press; 2008.
[41] Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons; 2015.
[42] Hurst HE. A suggested statistical model of some time series which occur in nature. Nature 1957;180:494.
[43] Higuchi T. Approach to an irregular time series on the basis of the fractal theory. PhysD 1988;31(2):277–83.
[44] Susmakova K, Krakovska A. Discrimination ability of individual measures used in sleep stages classification. Artif Intell Med 2008;44(3):261–77.
[45] Koley B, et al. An ensemble system for automatic sleep stage classification using single-channel EEG signal. Comput Biol Med 2012;42(12):1186–95.
[46] Genes S, Polat K, Yosunkaya S. Efficient sleep stage recognition system based on EEC signal using k-means clustering based feature weighting. Expert SystApplic 2010;37(12):7922–8.
[47] Zhang XS, Roy RJ, Jensen EW. EEG complexity as a measure of depth of anesthesia for patients. IEEE TransBiomedEng 2001;48(12):1424–33.
[48] Khalighi Sirvan, Sousa Teresa, Pires Gabiel, et al. Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels. Expert SystApplic 2013;40(17):7046–59.
[49] Mormann Florian, Andrzejak Ralph G, Elger Christian E, Lehnertz Klaus. Seizure prediction: the long and winding road. Brain 2007;130(2):314–33.
[50] Agarwal R, Gotman J. Computer-assisted sleep staging. IEEE TransBiomedEng 2001;48(12):1412–23.
[51] Tang WC, Lu SW, Tsai CM, Kao CY, Lee HH. Harmonic parameters with HHT and wavelet transform for automatic sleep stages scoring. ProcWorld AcadSciEngTechnol 2007;22:414–7.
[52] Ioffe S, Szegedy C. Batch normalisation: accelerating deep network training by reducing internal covariate shift. 2015. arXiv preprint arXiv:1502.03167.
[53] Hinton GE, Srivastava N, Krizhevsky A. Improving neural networks by preventing co-adaptation of feature detectors. 2012.
[54] Cohen JA. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
[55] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009;45(4):427–37.
[56] Hassan AR, Haque MA. Computer-aided obstructive sleep apneascreening from single-lead electrocardiogram using statistical and spectralfeatures and bootstrap aggregating. Biocybern. Biomed. Eng. 2016;36(1):256–66.
[57] Hassan Anhaf Rashik, Bhuiyan Mohammed Imamul Hassan. J Neurosci Methods 2016;271:107–18.
[58] Hassan AR, Haque MA. Computer-aided gastrointestinal hemorrhagedetection in wireless capsule endoscopy videos. Comput Methods Programs Biomed 2015;122(3):341–53.
[59] Longstaff Ian D, Cross JF. A pattern recognition approach to understanding the multi-layer perception. Pattern RecogLett 1987;5(5):315–9.
[60] Peterson Leif E. K-nearest neighbor. Scholarpedia 2009;4.2:1883.
[61] Saunders C, et al. Support vector machine. Comput Sci 2002;1(4):1–28.
[62] Breiman L. Random forest. MachLearn 2001;45:5–32.
[63] Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann Biomed Eng 2016;44(5):1587–97.
[64] Boonyakitanont Poomipat. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. Preprint at arXiv. Available, https://arxiv.org/abs/1908.00492v1.
[65] Huy Phan, Fernando, et al. Joint classification and prediction CNN framework for automatic sleep stage classification. In: IEEE TransBio-medEng; 2018.