Computers in Biology and Medicine 103 (2018) 71–81

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device

Xin Zhang[a], Weixuan Kou[a], Eric I-Chao Chang[c], He Gao[d], Yubo Fan[a,b], Yan Xu[a,b,c,*]

[a] School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China
[b] Research Institute of Beihang University in Shenzhen and the Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education and the State Key Laboratory of Software Development Environment and Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing, 100191, China
[c] Microsoft Research Asia, Beijing, 100080, China
[d] Clinical Sleep Medicine Center, The General Hospital of the Air Force, Beijing, 100142, China

ARTICLE INFO

Keywords:
Sleep stage
Classification
Feature learning
Recurrent neural networks
Wearable device

ABSTRACT

Background: Automatic sleep stage classification is essential for long-term sleep monitoring. Wearable devices show more advantages than polysomnography for home use. In this paper, we propose a novel method for sleep staging using heart rate and wrist actigraphy derived from a wearable device.

Methods: The proposed method consists of two phases: multi-level feature learning and recurrent neural networks-based (RNNs) classification. The feature learning phase is designed to extract low- and mid-level features. Low-level features are extracted from raw signals, capturing temporal and frequency domain properties. Mid-level features are explored based on low-level ones to learn compositions and structural information of signals. Sleep staging is a sequential problem with long-term dependencies. RNNs with bidirectional long short-term memory architectures are employed to learn temporally sequential patterns.

Results: To better simulate the use of wearable devices in the daily scene, experiments were conducted with a resting group in which sleep was recorded in the resting state, and a comprehensive group in which both resting sleep and non-resting sleep were included. The proposed algorithm classified five sleep stages (wake, non-rapid eye movement 1–3, and rapid eye movement) and achieved weighted precision, recall, and F_1 score of 66.6%, 67.7%, and 64.0% in the resting group and 64.5%, 65.0%, and 60.5% in the comprehensive group using leave-one-out cross-validation. Various comparison experiments demonstrated the effectiveness of the algorithm.

Conclusions: Our method is efficient and effective in scoring sleep stages. It is suitable to be applied to wearable devices for monitoring sleep at home.

1. Introduction

Sleep is a fundamental physiological activity of the human body, which contributes to self-recovery and memory consolidation [1,2]. Regular sleep facilitates the performance of daily work. However, many sleep disorders, such as insomnia, apnea, and narcolepsy, disturb sleep quality and thus threaten human health [3]. Effective diagnosis and treatment of these sleep disturbances rely on accurate detection of sleep stages and sleep cycles [4]. Therefore, sleep stage classification is a premise and significant step for sleep analysis.

The standard technique for scoring sleep is to use polysomnography (PSG) to synchronously record multichannel biomedical signals of the patient all through the night in a hospital which include electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), respiratory effort signals, blood oxygen saturation, and other measurements. These recordings are divided into nonoverlapping 30-s epochs. Domain experts evaluate sleep epoch by epoch, based on Rechtschaffen and Kales (R&K) rules [5] and the more recent American Academy of Sleep Medicine (AASM) guideline [6]. According to the AASM, sleep is categorized into five stages: wake (W), rapid eye movement (REM), and non-rapid eye movement (NREM, including N1, N2, and N3).

Monitoring sleep through the PSG system has many disadvantages when used at home. First, patients have to wear numerous sensors in different parts of the body. It may negatively impact patients' normal sleep and thus produce discrepant results. Such results are not able to reflect the real sleep state. Second, PSG is expensive. It is not available for most ordinary families. Third, PSG is not portable, making it

[*] Corresponding author. School of Biological Science and Medical Engineering, Beihang University, Beijing, 100191, China.

E-mail addresses: xinzhang0376@gmail.com (X. Zhang), weixuankou@outlook.com (W. Kou), echang@microsoft.com (E.I.-C. Chang), bjgaohe@sohu.com (H. Gao), yubofan@buaa.edu.cn (Y. Fan), xuyan04@gmail.com (Y. Xu).

https://doi.org/10.1016/j.compbiomed.2018.10.010
Received 13 June 2018; Received in revised form 10 October 2018; Accepted 10 October 2018

inappropriate for long-term home monitoring. To overcome the above shortcomings, it is a promising strategy to utilize a wearable device in place of the PSG system to classify sleep stages automatically. A wearable device can readily record the pulse rate from the photoplethysmography (PPG) signal and wrist actigraphy without causing many obstructions to natural sleep. During sleep, the former can be regarded as a surrogate measurement of heart rate [7], and the latter can evaluate body movement. It has been extensively investigated in previous studies that both heart rate and body movement are related to sleep stage transition [8,9]. Therefore, the wearable device can be an alternative choice to score sleep automatically at home.

Many physiological studies of sleep have indicated that sleep structure is associated with autonomic nervous system (ANS) regulation [8,10,11]. The contributions of parasympathetic and sympathetic activities vary among different sleep stages. Meanwhile, ANS activity during sleep can be measured using heart rate variability (HRV) as a quantitative index of parasympathetic or sympathetic output [12–14]. Hence, it is reasonable to use heart rate to determine sleep stages. More specifically, the sympathetic input is reduced, and parasympathetic activity predominates in NREM sleep. Thus with the transition of the sleep stage from N1 to N3, heart rate gradually decreases and reaches the minimum in N3 stage [15]. During REM sleep, in contrast, sympathetic activity shows more predominant influence. Accordingly, heart rate increases and becomes more unstable [16]. The spectral components of heart rate also exhibit distinct characteristics in sleep transition [10,17]. The ratio of the power in low frequency (LF, 0.04–0.15 Hz) to high frequency (HF, 0.15–0.40 Hz) tends to decrease in NREM sleep and significantly increase in REM sleep.

Quite a few sleep staging algorithms have been developed based on HRV, most of which emphasize the applicability in home-based scenarios [18–20]. Yoon et al. [18] designed thresholds and a heuristic rule based on automatic activations derived from HRV to determine the slow wave sleep (SWS). An overall accuracy of 89.97% was achieved. Ebrahimi et al. [19] extracted features from both HRV and respiratory signals. They used recursive feature elimination (RFE) enhanced support vector machine (SVM) to classify four sleep stages (W, N2, N3, and REM), achieving a classification accuracy of 89.23%. Xiao et al. [20] extracted 41 HRV features in a similar way and random forest (RF) [21] was used to classify three sleep stages (W, REM, and NREM). A mean accuracy of 72.58% was achieved in the subject independent scheme.

Actigraphy-based methods which capture body movement during sleep have long been investigated, especially to identify wake/sleep [9,22,23]. It is easy to understand since the body tends to remain stationary when falling asleep. The motion amplitude becomes distinctively smaller than that in the wake state. Various studies have been implemented to classify sleep stages with actigraphy. Herscovici et al. [24] presented a REM sleep detection algorithm based on the peripheral arterial tone (PAT) signal and actigraphy which were recorded with an ambulatory wrist-worn device. Kawamoto et al. [25] detected REM sleep based on the respiratory rate which was estimated from actigraphy. Long et al. [26] designed features based on dynamic warping (DW) methods to classify sleep and wake using actigraphy and respiratory efforts.

So far, few sleep staging methods have been developed that use both heart rate and wrist actigraphy. Furthermore, most protocols in previous studies focus on designing low-level features which are extracted in the time domain, frequency domain, and nonlinear analysis. This causes the effectiveness of feature extraction to be overly dependent on the expert analysis of signals, which makes these hand-engineered features not robust and flexible enough to adapt to different circumstances. In this paper, we propose a multi-level feature learning framework which extracts low- and mid-level features hierarchically. Low-level features capture temporal and frequency domain properties, and mid-level features learn compositions and structural information of signals. Specifically, to extract low-level features, the mean value and discrete cosine transform (DCT) [27] are adopted to heart rate and cepstral analysis [28] is adopted to wrist actigraphy. Mid-level features are learned based on low-level ones.

Recently, deep learning methods have been introduced into the sleep stage classification field, which produces encouraging results. Tsinalis et al. [29] utilized convolutional neural networks (CNNs) based on single-channel raw EEG to predict sleep stages without using prior domain knowledge. The sparse deep belief net (DBN) was applied in Ref. [30] as an unsupervised method to extract features from EEG, EOG and EMG. Sleep staging is a sequential problem [6] as sleep shows typically cyclic characteristics and NREM/REM sleep appears alternately. Moreover, manual sleep stage scoring depends on not only temporally local features, but also the epochs before and after the current epoch [31]. Recurrent neural networks (RNNs), particularly those using bidirectional long short-term memory (BLSTM) hidden units, are powerful models for learning from sequence data [32,33]. They are capable of capturing long-range dependencies, making RNNs quite suitable for modeling sleep data. Inspired by this, we apply a BLSTM-based RNN architecture for classification.

In this paper, we develop a sleep stage classification algorithm using heart rate and wrist actigraphy derived from the wearable device. The proposed method consists of two phases: multi-level feature learning and RNN-based classification. In the feature extraction phase, contrary to traditional methods that extract specific hand-engineered features using much prior domain knowledge, we aim to obtain main information of sleep data. Low-level features (mean value and DCT of heart rate, cepstral analysis of wrist actigraphy) are extracted from raw signals and mid-level representations are explored based on low-level ones. In the classification phase, the BLSTM-based RNN architecture is employed to learn temporally sequential patterns. The flowchart of the whole algorithm is shown in Fig. 1.

The contributions of our algorithm include:

1 A complete sleep stage classification solution specially designed for wearable devices is proposed.
2 The mid-level feature learning is introduced into sleep domain, and its effectiveness is demonstrated.
3 The feasibility of using RNNs to model sleep signals is verified.

2. Materials and methods

We first describe how we collect experimental data (including heart rate and wrist actigraphy) and obtain ground truth of corresponding sleep stages in Section 2.1 and Section 2.2. In Section 2.3, we explain how multi-level features are extracted. Specifically, low-level features

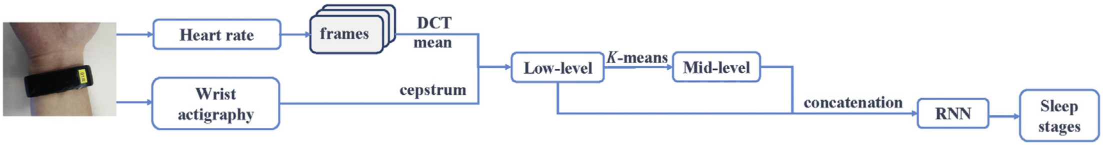


Fig. 1. The flowchart of the proposed method. The method consists of two phases: multi-level feature learning and RNN-based classification. In the first phase, low-level features are extracted from heart rate and wrist actigraphy signals. Then mid-level features are obtained based on low-level ones. Combining two levels of features, we arrive at the final representations. In the second phase, a BLSTM-based RNN architecture is applied for classification. The obtained features serve as inputs to the network and predictions of sleep stages are finally obtained by RNN.

are extracted directly from the raw heart rate and wrist actigraphy in Section 2.3.1. Based on low-level features, mid-level feature learning is explored in 2.3.2. Then we extract mid-level features in Section 2.3.3. In Section 2.4, the BLSTM-based RNN model is trained using multi-level features as input to classify sleep stages.

2.1. Sleep recordings

The study recruited 39 healthy subjects (30 males and 9 females) with an age range of 19–64 years old. None of them took drugs or medications that could affect sleep before the experiment. We collected sleep recordings at the Center for Sleep Medicine in the General Hospital of the Air Force, PLA, Beijing, China. During sleep, subjects were equipped with both the wearable device and PSG. The wearable device was used to collect heart rate and wrist actigraphy. PSG was used for labeling sleep stages to obtain “golden standard”. Actually, the wearable device used in this paper was Microsoft Band I. The Microsoft corporation provided the module for estimating pulse rate from PPG signals which we used as the alternative measurement of heart rate in our work. The choice of the wearable band and the estimation accuracy of pulse rate were not the focus of this article. PSG data were acquired using Compumedics $E-64$.

All subjects slept in the same sleep lab and used the same set of the band and PSG. Subjects were required not to take a nap during the daytime. Before sleep, the band was fully charged and synchronized with the PSG system. To better simulate the use of wearable bands in the daily scene, subjects were required to take bands according to their own habits. Ether left or right wrist was fine. The placement of the band and PSG electrodes was completed before 9:00 p.m. Subjects were awakened by the doctor at 6:00 a.m. the next day. After the subject awakened, heart rate (beats/min) and triaxial wrist actigraphy (g) were exported. The band was then recharged. Heart rate was calculated by converting the pulse-to-pulse intervals at each detected pulse peak. The actigraphy was sampled at 32 Hz. As exercise has a physiological influence on sleep [34], we considered two conditions. The condition that subjects stayed relaxed and took no exercise before sleep was defined as the resting state [35]. 28 recordings were collected in resting state. In the other condition, subjects were required to do moderate aerobic exercise one to two hours before sleep. 11 recordings were collected in this condition. We divided all these 39 recordings into two groups: the resting group with 28 resting state sleep recordings, and the comprehensive group with all 39 recordings to simulate the actual situation of the daily sleep.

2.2. Ground truth

The sleep physician assigned one of the five stages (W, N1, N2, N3, and REM) to the overnight sleep at each epoch according to the AASM guideline. To ensure the accuracy of ground truth, each subject's PSG recording was scored by 5 physicians from the General Hospital of the Air Force independently. Then we adopted the voting strategy to obtain final staging results. For the cases in which multiple stages tied for the most votes, we followed the stage of the previous epoch. An example of collected recordings is shown in Fig. 2. Details of subject demographics are listed in Table 1.

2.3. Feature extraction

2.3.1. Low-level feature extraction

We first divide each heart rate and actigraphy records into 30s epochs synchronizing in time with PSG classification results. For each sleep epoch, low-level features of heart rate and actigraphy are extracted, respectively. Then they are combined to learn mid-level features.

Both temporal and frequency properties of heart rate are considered. To make features more representative in context, the feature extraction procedure is carried out based on the sliding window technique. We extract low-level features of heart rate in a frame which includes 10 sleep epochs centered around the current one. Features are extracted within each epoch and then concatenated. First, we derive pulse intervals from heart rate,

$$PPI = 60/HR, \tag{1}$$

in which PPI refers to pulse intervals and HR refers to heart rate.

In the time domain, we compute the mean pulse intervals of each epoch in the frame to constitute a mean value vector. Discrete cosine transform (DCT) [36] is applied for frequency domain analysis. Compared with the discrete Fourier transform (DFT), DCT shows better performance with respect to energy concentration. We adopt DCT to pulse intervals through which dominant frequency components in each epoch are procured to form the frequency feature vector. To measure the frequency fluctuation, we calculate the first and second order differences of dominant frequency components. Then the zero order (dominant frequency components), first order, and second order differences of frequency components are joined together as frequency domain features of heart rate.

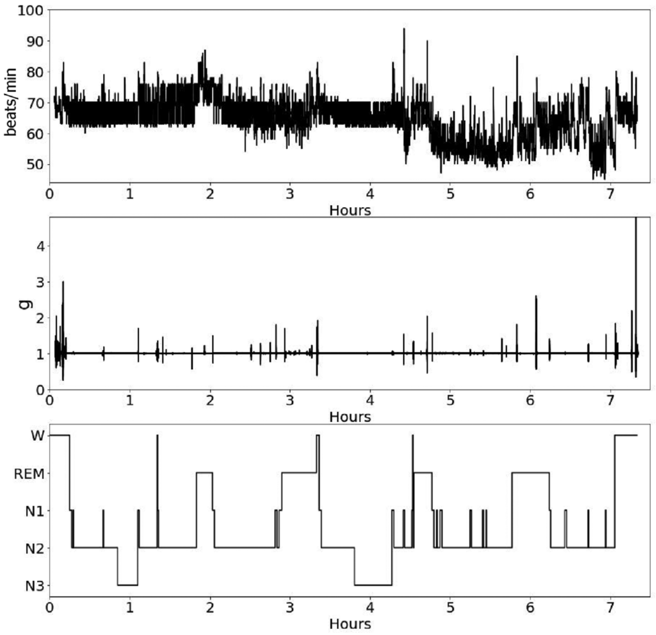


Fig. 2. Illustration of the recording. The top one represents an overnight heart rate signal. The middle one represents an actigraphy signal (Here, to save space, we integrate the actigraphy in X, Y, and Z axes). The bottom one represents the corresponding sleep stages.

Table 1
Subject demographics.

Parameter	Resting group		Comprehensive group	
	Mean±Std	Range	Mean±Std	Range
Num	28		39	
M/F	20/8		30/9	
Age (*y*)	26.25±7.77	19–48	27.72±10.12	19–64
BMI	21.81±2.66	17.07–29.05	21.81±2.68	17.07–29.05
TRT (*h*)	7.80±0.52	6.92–8.80	7.86±0.54	6.82–8.80
W (%)	19.14±8.80	4.81–40.24	19.37±9.13	4.81–40.24
N1 (%)	5.79±3.20	1.73–14.18	6.05±3.43	1.73–15.09
N2 (%)	45.00±9.48	26.04–62.31	46.14±8.84	26.04–62.31
N3 (%)	14.86±9.57	0.00–42.23	13.47±8.73	0.00–42.23
REM (%)	15.20±4.47	6.15–24.30	14.96±4.15	6.15–24.30

Num = number of recordings, M/F = male/female, BMI = body mass index, TRT = total recording time.

Actigraphy features are extracted only within the current epoch. As body movement during night tends to be transient and the sampling rate of actigraphy is high enough to capture movement details, there is no need for a wide range of context information. The cepstral analysis, which is widely used in action recognition area [37], has also been implemented in sleep studies to assess body movement [38]. In this study, we calculate the first order difference of the actigraphy along three axes, respectively. Then the dominant cepstrum components of the aforementioned difference in each axis are concatenated to form the actigraphy feature vector.

2.3.2. Mid-level feature learning

Mid-level feature learning methods are widely used in various kinds of pattern recognition tasks and give a nice performance boost [39–41]. Compared with low-level feature extraction, mid-level feature learning pays more attention to analyzing compositions and exploring the inherent structure of signals [42]. It can be assumed that sleep is comprised of different compositions. Weights of each composition vary among different stages. Thus bag-of-words (BOW), a kind of dictionary learning method is quite appropriate for obtaining mid-level sleep representations. In this work, we implement BOW based on low-level features of both heart rate and actigraphy signals to learn mid-level features.

The dictionary is constructed upon low-level features of all sleep epochs from the training set using the K-means algorithm [43]. K clusters are thus generated. Each cluster center represents one composition. K cluster centers together constitute the whole sleep structure. For a set of sleep epochs, we define its corresponding set of low-level features as $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbf{R}^{d\times 1}$, $i \in \{1,2,\ldots,n\}$, in which $\boldsymbol{x}_i$ represents low-level features of i-th sleep epoch, and the dimension of low-level features is d. Each $\boldsymbol{x}_i$ is related to an index $z_i \in \{1,2,\ldots,K\}$. If $z_i = l \in \{1,2,\ldots,K\}$, $\boldsymbol{x}_i$ belongs to the l-th cluster. The center of the l-th cluster is denoted as

$$\boldsymbol{m}_l = \sum_{i=1}^{n} 1\{z_i = l\}\boldsymbol{x}_i / \sum_{i=1}^{n} 1\{z_i = l\},\ \boldsymbol{m}_l \in \mathbf{R}^{d\times 1}, \tag{2}$$

in which $\boldsymbol{m}_l$ refers to the l-th sleep composition.

2.3.3. Mid-level feature extraction

The dictionary is built as above. Given a sleep epoch, the corresponding mid-level features are extracted as follows: the Euclidean distances between its low-level features and each cluster center are first computed. Then we take the reciprocals of the Euclidean distances as mid-level features (K dimensions), which shows the weighted influence of each compositions to the current epoch.

We concatenate low-level and mid-level features as the final feature vector. The order of concatenation does not affect the performance. After all feature vectors of the training set are extracted, we normalize features along each dimension with Z-score strategy:

$$p_{ij} = (x_{ij} - \mu_j)/\delta_j, \tag{3}$$

in which $x_{ij} \in \boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$, μ_j and δ_j are the mean value and standard deviation of the j-th dimensional feature, respectively. D refers to the dimension of the final feature vector. $D = d + K$.

2.4. Recurrent neural networks

Recurrent neural networks are suitable models for sequential data and have gained great success in numerous sequence learning domains, including natural language processing, speech, and video [33]. Bidirectional long short-term memory (BLSTM) [44,45] which combines long-term dependency of LSTM [46] with bidirectional propagation, is a novel version of RNN to exploit long-range information in both input directions. In fact, the long-term information utilization is key to sleep staging. For example, there always exists a long-term memory of heart rate in the REM stage and wake stage, but weak long-term correlation in NREM sleep [47]. Given that sleep is an inherently dynamic and long-term dependency process, it seems natural to consider BLSTM as a potentially effective model.

Generally, we define input units $\boldsymbol{x} = (\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(T)})$, hidden units $\boldsymbol{h} = (\boldsymbol{h}^{(1)},\ldots,\boldsymbol{h}^{(T)})$ and output units $\hat{\boldsymbol{y}} = (\hat{\boldsymbol{y}}^{(1)},\ldots,\hat{\boldsymbol{y}}^{(T)})$, where each input $\boldsymbol{x}^{(t)} \in \mathbf{R}^{D\times 1}$, $\hat{\boldsymbol{y}}$ is the hypothesis of the true label $\boldsymbol{y}$ and each output $\hat{\boldsymbol{y}}^{(t)} \in [0,1]^M$. Here, D refers to the dimension of final features, t refers to the t-th sleep epoch, T refers to the total number of sleep epochs and M refers to the total classes of the sleep stage. The following three equations describe horizontal propagation of BLSTM:

$$\overrightarrow{\boldsymbol{h}}^{(t)} = \mathscr{H}\left(W_{x\overrightarrow{h}}\boldsymbol{x}^{(t)} + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{\boldsymbol{h}}^{(t-1)} + \boldsymbol{b}_{\overrightarrow{h}}\right), \tag{4}$$

$$\overleftarrow{\boldsymbol{h}}^{(t)} = \mathscr{H}\left(W_{x\overleftarrow{h}}\boldsymbol{x}^{(t)} + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{\boldsymbol{h}}^{(t+1)} + \boldsymbol{b}_{\overleftarrow{h}}\right), \tag{5}$$

$$\hat{\boldsymbol{y}}^{(t)} = \mathscr{S}\left(W_{\overrightarrow{h}y}\overrightarrow{\boldsymbol{h}}^{(t)} + W_{\overleftarrow{h}y}\overleftarrow{\boldsymbol{h}}^{(t)} + \boldsymbol{b}_y\right), \tag{6}$$

in which W refers to weight matrices and b refers to bias vectors with superscripts denoting time steps and subscripts denoting layer indices. $\mathscr{H}$ denotes the hidden layer function. $\mathscr{S}$ denotes the output layer function. $\overrightarrow{\boldsymbol{h}}$ and $\overleftarrow{\boldsymbol{h}}$ represent the hidden layers in the forwards and backwards directions, respectively. We apply a softmax function to output probabilities of M classes.

In this paper, we train a BLSTM-based RNN model with multiple hidden units. We also evaluate how RNN can benefit from the use of deep architectures. Specifically, by stacking multiple recurrent hidden layers on top of each other, the way that conventional deep networks do, we arrive at the deep BLSTM. The structure of the deep BLSTM is shown in Fig. 4. The version of LSTM memory cells is with forget gates [48] and peephole connections [32], whose structure is illustrated in the right part of Fig. 3. Assuming there are N hidden layers with hidden units of BLSTM, the hidden sequences $\boldsymbol{h}_n^{(t)}$ are computed from $n = 1$ to N and $t = 1$ to T as below:

$$\boldsymbol{h}_n^{(t)} = \mathscr{H}\left(W_{h_{n-1}h_n}\boldsymbol{h}_{n-1}^{(t)} + W_{h_n h_n}\boldsymbol{h}_n^{(t-1)} + \boldsymbol{b}_{h,n}\right). \tag{7}$$

Thus the output $\hat{\boldsymbol{y}}^{(t)}$ can be defined as:

$$\hat{\boldsymbol{y}}^{(t)} = \mathscr{S}\left(W_{h_N y}\boldsymbol{h}_N^{(t)} + \boldsymbol{b}_y\right). \tag{8}$$

According to the softmax output, cross entropy function is used as the loss function that we optimize:

$$loss(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\frac{1}{T}\sum_{t=1}^{t=T}\left(\boldsymbol{y}^{(t)}\cdot log(\hat{\boldsymbol{y}}^{(t)}) + (1 - \boldsymbol{y}^{(t)})\cdot log(1 - \hat{\boldsymbol{y}}^{(t)})\right), \tag{9}$$

in which $\boldsymbol{y}^{(t)}$ refers to the true label at time step t.

3. Results

3.1. Performance evaluation

8-fold cross-validation is conducted to present unbiased performance of the algorithm. On each iteration, we use 6 portions for training, 1 portion for validation, and 1 portion for testing. Finally, testing results of each iteration are averaged to form the overall performance of RNN classifiers. We randomly run cross-validation for three times and calculate average results.

Considering that the distribution of five sleep stages is severely unbalanced, to adapt to this characteristic, weighted precision (P), recall (R) and F_1 score (F_1) are selected to evaluate the performance. Evaluation measures are defined as:

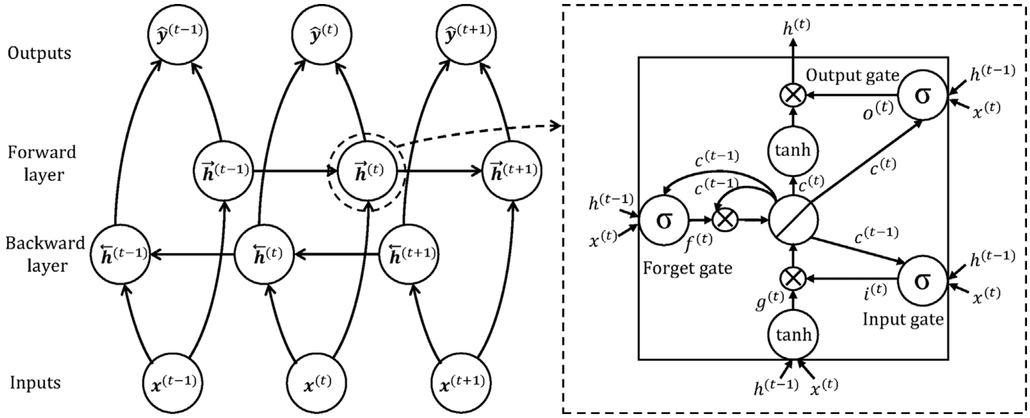


Fig. 3. Illustration of the BLSTM architecture. The left side is the overall view. There exist two hidden layers with opposite directions. Each hidden unit is a LSTM memory cell. The right side shows the detail of the LSTM memory cell.

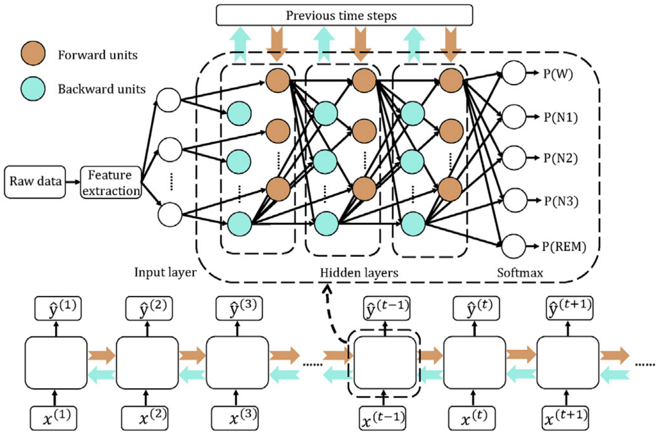


Fig. 4. Illustration of the proposed RNN classifier. The upper part is the detailed structure of the network. In order to express concisely, we omit the full connection between layers. The lower part depicts the workflow of the RNN. For a certain sleep epoch, data are processed in two opposite directions. The output layer predicts the sleep stage of the current epoch.

Table 2
Performance of 8-fold, 10-fold, and leave-one-out cross-validation for 5-class classification (%).

Method	RG			CG		
	P	R	F_1	P	R	F_1
8-fold	58.0	60.3	58.2	58.5	61.1	58.5
10-fold	65.3	65.9	62.1	63.1	64.0	59.9
Leave-one-out	**66.6**	**67.7**	**64.0**	**64.5**	**65.0**	**60.5**

RG = resting group, CG = comprehensive group.
The Bold numbers show the best performances in each tables.

Table 3
MCC and G-Mean of 8-fold, 10-fold, and leave-one-out cross-validation for 5-class classification (%).

Method	RG		CG	
	MCC	G-Mean	MCC	G-Mean
8-fold	52.9	65.1	49.1	62.6
10-fold	53.3	65.6	50.4	63.5
Leave-one-out	**55.8**	**67.1**	**51.9**	**64.7**

RG = resting group, CG = comprehensive group.
The Bold numbers show the best performances in each tables.

$$P = \sum_i \omega_i \cdot TP_i/(TP_i + FP_i), \tag{10}$$

$$R = \sum_i \omega_i \cdot TP_i/(TP_i + FN_i), \tag{11}$$

$$F_1 = \sum_i 2 \cdot \omega_i \cdot P_i \cdot R_i/(P_i + R_i), \tag{12}$$

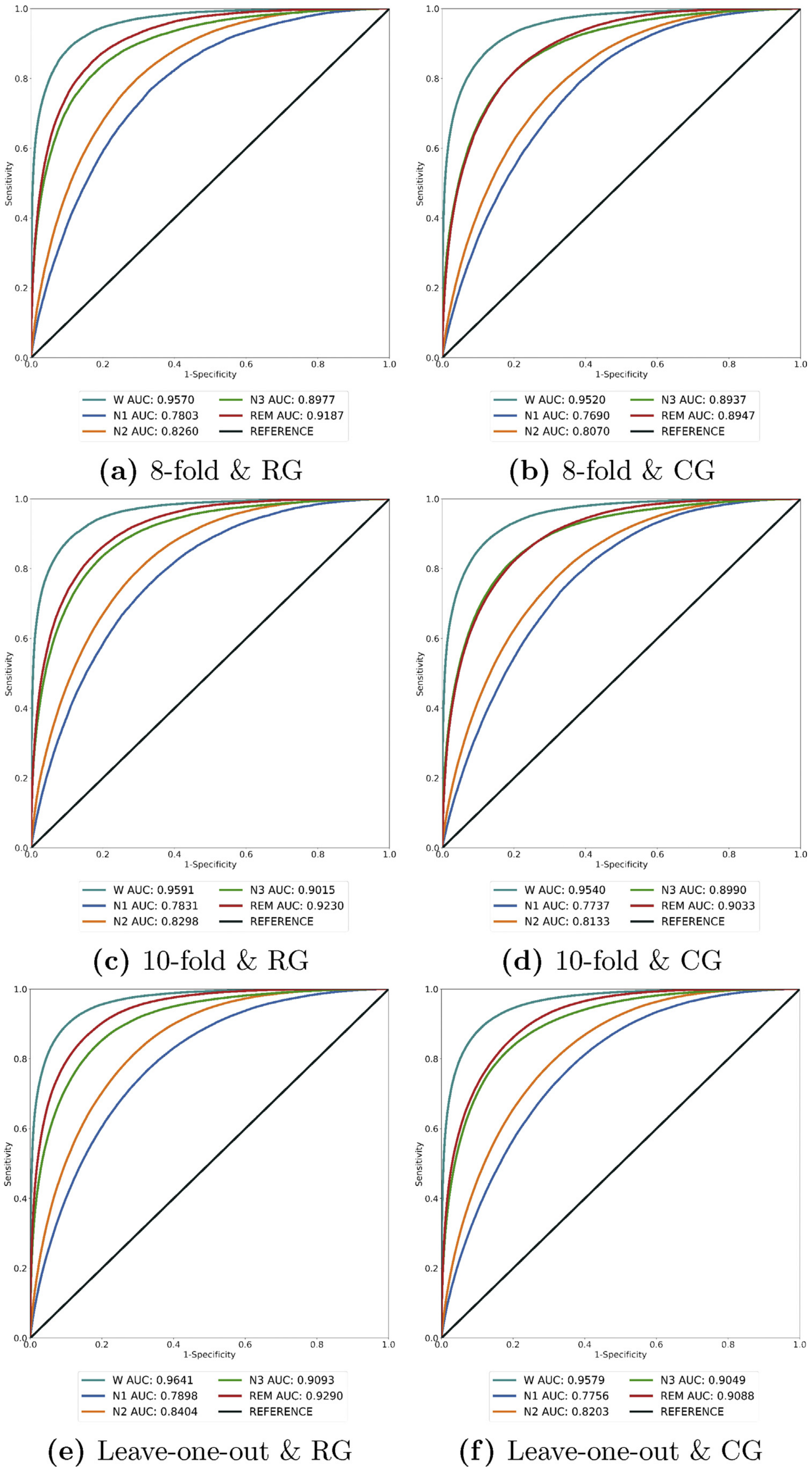
in which i refers to the stage category and ω_i is the proportion of the i-th stage class in all classes. TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives (Here, we omit the subscript i.).

3.2. Experiments

In order to fully explore the property of the proposed approach, we conduct the following experimental procedures: (1) we describe the detailed configuration and analyze the performance; (2) we evaluate the effectiveness of mid-level feature learning; (3) we make a comparison between the BLSTM-based RNN and two frequently-used classifiers for classification; (4) we explore the sensitivity of parameters in the feature extraction process; (5) we evaluate the performance of RNN models with different hidden layer width, depth, and unit types. As W and N1 stages are similar in HRV, we also conduct experiments that combine W and N1 into one stage, resulting in 4 classes of classification. Meanwhile, the resting group and comprehensive group are both experimented on, respectively.

3.2.1. Performance analysis

520 features are finally extracted, including 220 low-level features and 300 mid-level features. In low-level feature extraction, 10-dimension mean vector of heart rate is extracted in consecutive 10 epochs as temporal features. The first 5 frequency components of each epoch are collected by using DCT analysis. Then the first order and second order differences are calculated. Consequently, we obtain 120 frequency features of heart rate (40 dimensions for each order). For actigraphy, we join the first 30 cepstrum components in each axis to shape into 90 actigraphy features. Based on the above low-level features, mid-level features are learned. The size of the dictionary, K, is set to 300. By concatenating low-level and mid-level features, the final 520-dimension feature vector is formed. The RNN classifier is a three-layer structure: the input layer with the number of units equal to the dimension of final features, one hidden layer with 400 BLSTM cells, and the output layer with 5 units which mean 5 sleep stages. RNN is trained using stochastic gradient descent [49] with the learning rate of 10^{-6}. Network weights are randomly initialized with a Gaussian distribution of (0,0.1). To be more generalized, the Gaussian weighted noise ($\sigma = 0.005$) is added to the network. We implement the proposed RNN architecture under the



(a) 8-fold & RG

(b) 8-fold & CG

(c) 10-fold & RG

(d) 10-fold & CG

(e) Leave-one-out & RG

(f) Leave-one-out & CG

Fig. 5. ROC curves of 8-fold, 10-fold, and leave-one-out cross-validation for 5-class classification. (RG = resting group, CG = comprehensive group.)

Table 4
Comparison of performance with and without mid-level features on the resting group for 5-class classification (%).

HL	With mid-level features			Without mid-level features		
	P	R	F_1	P	R	F_1
1	**58.0**	**60.3**	**58.2**	52.8	54.7	52.7
2	57.7	59.6	57.1	53.5	55.5	53.3
3	56.3	59.6	57.1	53.1	55.7	53.1
4	55.4	58.7	56.4	52.2	54.7	51.5

HL = number of hidden layers.
The Bold numbers show the best performances in each table.

Table 5
Comparison of various classifiers on the comprehensive group for 5-class classification (%).

Classifier	P	R	F_1
SVM	**60.3**	60.6	55.6
RF	56.5	59.2	53.3
RNN	58.5	**61.1**	**58.5**

The Bold numbers show the best performances in each tables.

CURRENNT framework [50].

To fully report robust performance results of the proposed method, we add 10-fold and leave-one-out cross-validation besides 8-fold cross-validation. Experiments are conducted on both the resting group and the comprehensive group for 5-class classification. Table 2 shows performance results of weighted precision, recall, and F_1 score. For 8-fold cross-validation, F_1 score of 58.2% and 58.5% are achieved in the resting group and comprehensive group, respectively. 62.1% and 59.9% are yielded using 10-fold cross-validation. 64.0% and 60.5% are achieved using leave-one-out cross-validation. Two groups show similar results in three types of cross-validation. This proves that the algorithm is robust enough to adapt to different sleep conditions. It can be noted that the performance results of 10-fold cross-validation are higher than those of 8-fold cross-validation. Leave-one-out cross-validation achieves the highest performance. This may be due to that more data are used to train the model in every iteration. As five sleep stages are heavily unbalanced, to further evaluate the classification capability of each sleep stage, Matthews correlation coefficient (MCC) and Geometric mean (G-Mean) are presented in Table 3. Receiver operation characteristic (ROC) curves and areas under curves (AUCs) of different experiments are generated in Fig. 5. It can be noticed that except N1 stage, AUCs of other four sleep stages all exceed 0.80 in six experiments. The consistent results prove that the proposed method is discriminative in the case of data unbalance.

3.2.2. Effectiveness of mid-level learning

To demonstrate the effectiveness of mid-level feature learning, we design an experiment in which low-level features directly serve as the input to RNN without mid-level feature learning. Parameter settings remain the same. We conduct experiments with the resting group for 5-class classification. 1 to 4 hidden layers of RNNs are implemented.

As shown in Table 4, the performance is improved significantly when mid-level features are involved. The dictionary helps us obtain the spatial distribution of sleep compositions and explore inherent structures, which can describe sleep in a more representative way.

3.2.3. Comparison with various classifiers

We make a comparison between RNN and two classic classifiers, including support vector machine (SVM) [51] and random forest (RF). All three classifiers use the same features extracted in the first experiment. The comparison is performed with the comprehensive group for 5-class classification.

The SVM uses the radial basis function (RBF) kernel with the kernel coefficient *gamma* of 0.0021. Penalty parameter C is set to 1 to regularize the estimation. The shrinking heuristic is also utilized. 600 estimators are set in the RF, and the function to measure the quality of a split is the Gini impurity. The number of features to consider when looking for the best split is denoted as p, which is equal to the square root of the number of total features. The minimum number of samples required to split an internal node and be at a leaf node are $p^{1/2}$ and 1.

Results shown in Table 5 indicate that RNN is superior to RF in all three metrics and superior to SVM in weighted recall and F_1 score. Despite both SVM and RF being able to deal with high dimension situations, RNN works better at learning long-term dependencies.

3.2.4. Sensitivity of parameters

This section elucidates the sensitivity of variables in feature extraction, including the dominant frequency component size of DCT and the dictionary size in mid-level feature learning. We evaluate the sensitivity with the comprehensive group for 4-class classification. Hidden layers from 1 to 4 are employed.

The dominant frequency component size ranges from 5 to 25. The results are shown in Table 6 and Fig. 6(a). It can be seen that performance decreases as frequency component size increases. This may be because the high-frequency components are mostly noise and thus impact classification.

We change the dictionary size in mid-level feature learning from 100 to 500. The results are shown in Table 7 and Fig. 6(b). The variation trend of performance is small. It can be attributed to the lack of data. Although a larger dictionary generates a more detailed description of sleep compositions, there are more parameters to optimize in RNN. Sleep data of totally around 37,000 epochs are not sufficient for training the RNN model.

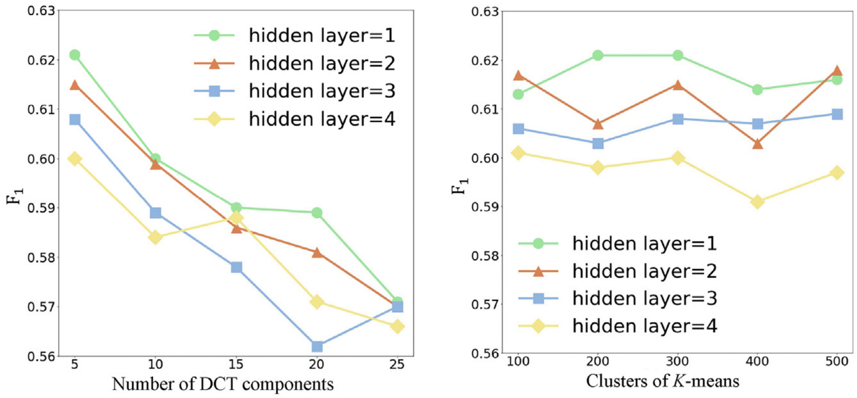
3.2.5. Neural networks configurations

We carry out extensive experiments to assess the performance of neural networks with different hidden layer width, depth, and unit types. We consider 6 scales of hidden units numbers: 100 to 600, 4

Table 6
The effect of the dominant frequency component size on the comprehensive group for 4-class classification (%).

Components	HL											
	1			2			3			4		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
5	**63.0**	**63.1**	**62.1**	62.1	61.9	61.5	61.7	61.9	60.8	60.6	60.8	60.0
10	60.3	61.2	60.0	60.6	60.9	59.9	60.0	60.3	58.9	59.5	60.1	58.4
15	60.1	61.2	59.0	59.6	60.3	58.6	58.7	59.4	57.8	59.9	60.3	58.8
20	59.4	60.6	58.9	58.5	59.2	58.1	57.9	58.5	56.2	57.6	59.0	57.1
25	58.0	59.5	57.1	57.4	58.9	57.0	58.0	59.0	57.0	57.3	58.9	56.6

HL = number of hidden layers.
The Bold numbers show the best performances in each table.



(a) Frequency component size

(b) Dictionary size

Fig. 6. The sensitivity of DCT components and dictionary size.

Table 7
The effect of dictionary size in the mid-level feature learning on the comprehensive group for 4-class classification (%).

K	HL											
	1			2			3			4		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
100	61.6	62.2	61.3	62.4	62.7	61.7	60.9	61.3	60.6	60.9	61.1	60.1
200	62.4	62.7	62.1	61.2	61.4	60.7	60.5	61.0	60.3	61.5	61.4	59.8
300	**63.0**	**63.1**	**62.1**	62.1	61.9	61.5	61.7	61.9	60.8	60.6	60.8	60.0
400	62.6	62.9	61.4	60.7	61.2	60.3	61.1	61.3	60.7	60.1	60.3	59.1
500	62.0	62.7	61.6	62.1	62.5	61.8	62.2	62.0	60.9	60.6	60.9	59.7

HL = number of hidden layers.
The Bold numbers show the best performances in each table.

Table 8
Comparison of different network architectures for 4-class classification (%).

Sub.	HT	HL	HU																	
			100			200			300			400			500			600		
			P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
MLP	RG	1	48.7	54.3	46.0	51.5	55.8	49.0	52.0	56.3	50.0	52.2	56.5	50.5	53.0	56.5	50.4	53.5	56.8	51.3
		2	45.6	54.9	45.3	50.8	55.9	47.9	52.8	56.4	49.4	54.8	57.6	52.1	55.4	57.7	52.2	56.3	57.9	52.2
		3	43.4	54.4	43.2	53.3	56.7	49.2	53.1	56.6	49.8	54.8	57.3	50.7	55.7	58.2	52.6	56.2	58.9	54.4
		4	41.4	54.1	42.4	47.2	55.7	46.2	51.4	56.9	49.5	52.2	57.5	50.9	54.7	57.4	51.0	58.2	59.2	55.0
	CG	1	50.6	56.3	48.2	52.7	56.8	49.8	53.8	57.7	51.5	53.1	56.8	49.5	53.2	57.2	50.6	54.5	57.5	51.1
		2	47.3	55.5	45.5	52.0	57.0	49.0	54.2	57.4	50.7	55.3	58.4	52.7	55.1	57.9	51.7	56.4	59.0	53.1
		3	44.8	55.5	45.2	50.5	56.2	47.6	56.2	58.1	51.3	55.8	58.4	52.3	56.7	58.7	53.2	55.5	58.9	52.2
		4	44.8	55.7	45.0	49.2	55.8	46.2	49.1	56.9	48.9	56.2	58.3	52.7	56.1	58.9	54.1	55.7	58.6	52.6
LSTM	RG	1	58.7	61.0	58.5	58.1	60.2	57.6	58.2	60.8	58.6	58.4	60.5	58.3	57.7	60.2	57.9	58.0	60.2	58.3
		2	56.4	60.1	57.6	56.6	60.3	57.7	57.3	60.4	57.7	55.8	58.6	56.4	56.1	59.1	56.9	56.0	59.0	56.6
		3	56.9	60.6	57.9	56.4	59.8	57.3	55.7	59.2	56.6	55.8	58.8	56.1	55.5	58.4	55.7	55.0	58.2	55.9
		4	55.7	59.3	56.4	55.9	59.4	56.7	54.4	57.5	55.2	54.2	57.3	54.7	54.8	57.5	55.0	54.1	57.2	54.8
	CG	1	58.9	61.3	58.3	59.0	61.1	58.2	58.4	61.0	58.3	57.3	59.8	57.3	58.0	60.6	58.2	57.8	60.3	58.0
		2	57.0	60.8	57.9	56.7	60.6	57.6	56.0	59.9	56.9	55.8	58.9	56.0	55.4	58.8	56.1	56.2	59.3	56.5
		3	56.4	60.4	57.5	56.1	60.1	57.2	56.1	59.4	56.5	55.3	59.1	55.9	55.8	59.2	56.2	55.0	58.4	55.5
		4	55.3	59.3	56.6	55.7	59.7	56.6	54.7	58.5	55.7	54.9	58.6	55.4	54.9	58.1	55.0	55.3	58.5	55.5
BLSTM	RG	1	58.6	61.0	58.5	58.6	60.8	58.5	58.3	60.6	58.2	58.0	60.3	58.2	58.7	61.0	58.7	58.0	60.3	58.2
		2	57.5	60.2	57.7	57.2	60.5	58.1	58.5	61.0	58.6	57.7	59.6	57.1	56.6	59.3	57.0	57.3	59.8	57.7
		3	55.6	59.3	56.7	55.7	59.2	56.9	56.0	59.5	56.9	56.3	59.6	57.1	55.7	58.6	56.4	56.3	59.1	56.7
		4	54.6	58.3	55.7	56.1	59.7	57.2	57.6	59.2	56.9	55.4	58.7	56.4	55.7	58.8	56.1	55.7	58.4	56.1
	CG	1	59.0	61.5	58.9	58.4	60.9	58.2	57.6	60.5	57.9	58.5	61.1	58.5	58.2	60.7	58.3	58.1	60.7	58.2
		2	56.8	60.7	58.0	57.1	60.9	58.0	56.3	60.2	57.5	56.7	59.9	57.0	56.9	59.9	56.9	56.7	59.7	57.2
		3	58.6	60.4	57.5	57.0	60.6	57.8	56.8	60.7	57.7	55.9	59.4	56.6	56.4	59.5	56.4	55.5	58.9	56.0
		4	55.1	59.2	56.1	55.5	59.4	56.8	58.0	59.8	56.9	55.0	58.7	56.3	55.3	58.9	56.0	54.7	58.1	55.6

Sub. = subjects, RG = resting group, CG = comprehensive group, HT = hidden unit type, HL = number of hidden layers, HU = number of hidden units.

scales of hidden layers: 1 to 4, and 3 types of neural networks: multilayer perceptron (MLP), LSTM, and BLSTM.

Tables 8 and 9 summarize results of multiple neural network architectures with the resting group and comprehensive group for 4- and 5-class classification. Fig. 7 illustrates the general trend of the weighted F_1 score along with the number of hidden units. It is obvious

Table 9
Comparison of different network architectures for 4-class classification (%).

Sub.	HT	HL	HU																	
			100			200			300			400			500			600		
			P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
MLP	RG	1	53.1	55.8	49.9	55.5	57.1	51.9	53.7	56.8	50.9	57.7	58.5	54.0	58.0	59.3	55.4	58.1	58.8	55.1
		2	44.9	51.0	41.9	56.8	58.2	52.6	57.3	59.4	54.6	59.4	59.8	55.8	57.3	58.9	53.5	58.1	59.6	55.7
		3	52.9	56.6	48.8	51.9	56.6	49.1	56.3	57.4	50.8	57.2	59.4	54.6	59.4	59.8	55.6	61.2	61.0	57.7
		4	49.9	57.1	48.4	56.7	58.2	51.8	56.9	59.2	54.2	60.7	60.6	57.1	59.9	60.5	56.8	60.9	61.2	58.5
	CG	1	56.5	58.2	52.8	56.5	58.7	53.8	58.6	59.8	55.6	59.5	60.4	56.2	59.1	60.1	55.8	59.3	60.4	56.6
		2	54.0	57.9	51.3	57.9	59.6	54.8	59.2	60.4	56.3	59.5	60.5	56.2	59.9	60.7	56.8	60.1	60.9	57.0
		3	51.8	57.8	49.7	57.9	59.3	53.4	60.2	60.8	56.5	59.6	60.9	57.0	60.1	61.2	57.6	60.2	60.9	57.3
		4	53.4	57.7	49.4	58.0	59.7	54.4	59.9	60.7	57.0	59.6	60.7	57.2	59.7	60.9	57.3	60.5	61.4	58.4
LSTM	RG	1	62.3	62.6	62.0	62.8	63.0	62.3	63.0	63.1	62.4	62.0	62.1	61.6	62.6	62.8	62.2	62.2	62.4	61.7
		2	62.5	62.5	61.8	61.8	61.7	61.2	61.4	61.4	60.6	60.5	60.7	59.8	60.3	60.4	59.7	60.5	60.6	59.9
		3	60.7	60.8	60.1	61.3	61.2	60.7	60.7	60.6	59.7	60.0	60.0	59.3	60.2	60.1	59.5	60.5	60.3	59.7
		4	60.6	60.5	59.9	60.7	60.7	59.9	59.4	59.2	58.4	58.7	58.6	57.6	59.3	59.2	58.6	59.1	59.1	58.4
	CG	1	62.9	63.3	62.4	62.8	63.2	62.3	62.3	62.6	61.8	62.9	63.2	62.5	62.0	62.4	61.6	61.6	62.0	61.1
		2	62.6	62.8	61.9	61.8	61.9	61.1	61.3	61.6	60.5	61.2	61.4	60.6	61.7	61.9	60.9	60.3	60.7	59.5
		3	61.8	61.9	61.2	61.1	61.2	60.3	60.9	61.0	59.9	60.6	60.8	59.7	59.9	60.3	59.0	59.7	59.9	59.1
		4	61.6	61.7	60.8	61.5	61.3	60.5	60.1	60.3	59.2	59.5	59.7	58.5	59.5	59.7	58.5	59.4	59.7	58.6
BLSTM	RG	1	62.8	62.9	62.4	62.2	62.5	61.9	62.3	62.4	61.9	61.9	62.0	61.4	62.3	62.5	61.6	62.8	63.0	62.5
		2	62.6	62.5	62.0	62.0	62.1	61.6	62.1	62.1	61.6	61.1	61.1	60.6	61.3	61.4	60.7	60.8	61.0	60.3
		3	61.4	61.4	60.9	61.4	61.2	60.6	61.1	60.8	60.3	60.2	60.4	59.7	59.9	60.0	59.2	60.3	60.3	59.5
		4	60.0	60.0	59.4	61.1	61.0	60.5	60.5	60.5	60.1	60.2	60.3	59.6	60.2	60.3	59.6	60.0	60.1	59.2
	CG	1	62.5	62.9	61.9	62.4	62.9	62.0	63.1	63.6	62.7	63.0	63.1	62.1	62.1	62.7	61.8	62.0	62.4	61.7
		2	62.5	62.7	61.9	62.2	62.1	61.3	61.8	62.0	60.9	62.1	61.9	61.5	61.2	61.5	60.5	60.6	61.0	60.3
		3	60.8	61.0	60.1	61.5	61.3	60.6	60.9	61.2	60.5	61.7	61.9	60.8	60.3	60.7	59.8	60.3	60.7	59.8
		4	61.4	61.2	60.3	61.2	61.2	60.4	60.8	61.2	60.3	60.6	60.8	60.0	60.8	61.1	60.2	60.0	60.5	59.6

Sub. = subjects, RG = resting group, CG = comprehensive group, HT = hidden unit type, HL = number of hidden layers, HU = number of hidden units.

that RNNs with LSTM and BLSTM units achieve much better results than MLP in all experiments. This demonstrates the efficacy of RNNs to learn temporal relationships between sleep sequences. BLSTM outperforms LSTM slightly. As the number of hidden units increases, the improvement of performance in MLP is most significant. It is reasonable because MLP models are relatively simple. The collected sleep data are sufficient to optimize MLP models with deeper and wider structures. Performance results of LSTM and BLSTM show a slight decrease as the number of hidden units and layers increase. This may be caused by the mismatch between excessive parameters and relatively sparse data.

4. Discussion

It is observed that high performances were achieved using conventional approaches [18–20]. However, the remarkable performances rely on elaborate feature engineering which is data dependent. Furthermore, HRV features they used are extracted based on ECG which is not convenient to collect for wearable devices. In this work, we combine heart rate and actigraphy to predict sleep stages. They are much easier to collect via wearable devices.

Different experiment settings, data, and evaluation methods are applied to classify different numbers of sleep stages in previous literature. Hence, it is quite difficult to make a direct comparison between different algorithms. We compare our algorithm with the method described in Ref. [20]. The baseline method extracts 41-dimension hand-engineered features based on HRV in the time domain (8 dimensions), frequency domain (20 dimensions) and nonlinear analysis (13 dimensions). These features are then trained and tested through RF. Here we implement it using pulse intervals in our dataset. To make a fair comparison, we also implement our proposed algorithm without actigraphy. Moreover, we add our extracted actigraphy features to 41-dimension HRV features to perform the baseline method. Both the resting group and the comprehensive group are used to evaluate experiments for 5-class classification. Table 10 shows the results.

It can be observed from Table 10 that our method surpasses the baseline both with and without actigraphy features. The best performances, weighted precision, recall, and F_1 score of 58.0%, 60.3%, and 58.2% in the resting group and 58.5%, 61.1%, and 58.5% in the comprehensive group are achieved through our method using heart rate combined with actigraphy. Compared with the hand-engineered feature extraction method in the baseline which is overly dependent on expert knowledge, our feature learning method aims to obtain main information of signals and RNN is able to further refine features. With little prior domain knowledge used, the method has the potential to generalize sleep disorder detection. Approximately, the differences in results between the resting group and comprehensive group using our method are smaller than that using the baseline. Since sleep data in the comprehensive group are more diverse, it shows the robustness of the whole algorithm. Furthermore, the performances of both methods are improved when actigraphy features are considered, which suggests that body movement during sleep contains useful information for sleep stage classification.

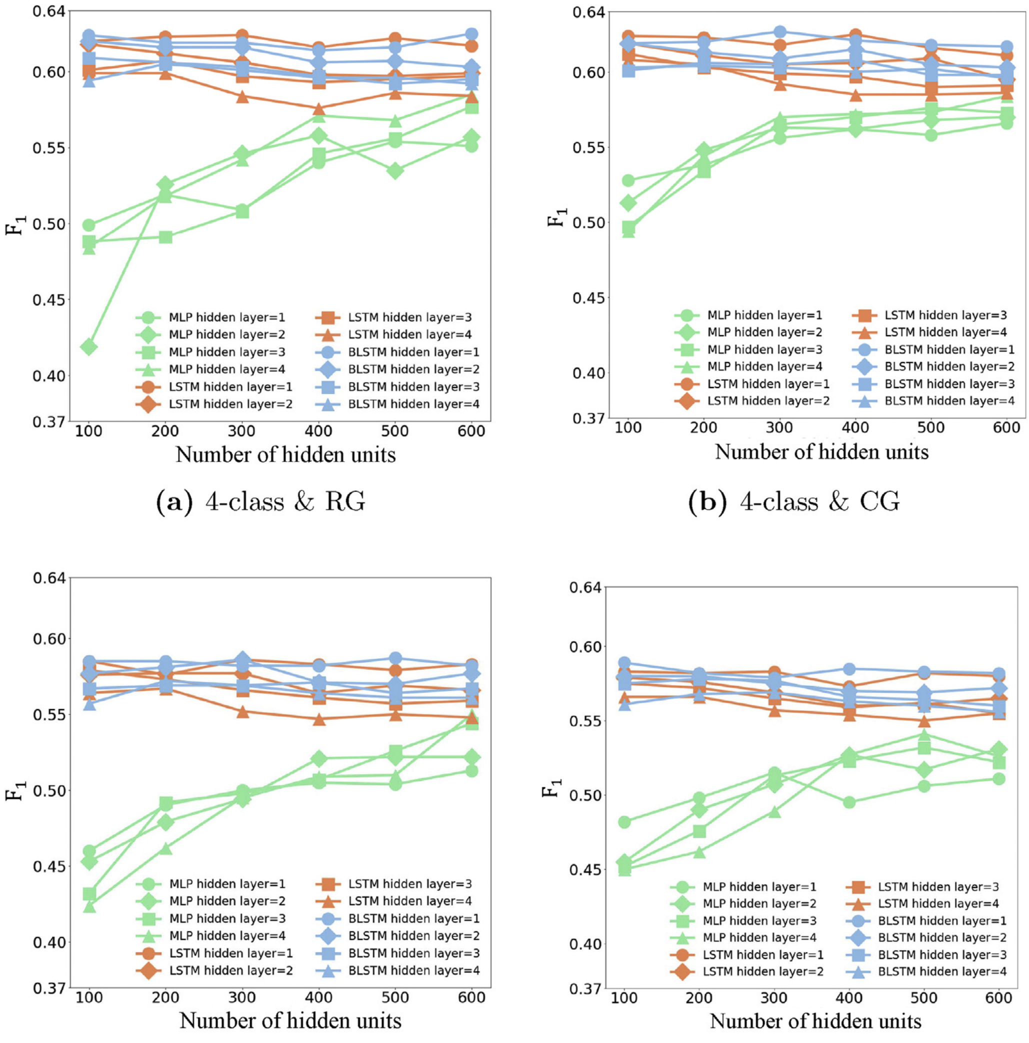
As heart rate is estimated from the pulse wave, accumulative error is inevitably introduced. Estimation accuracy of heart rate would affect classification performance. In future, we intend to develop algorithms that predict sleep stages from pulse wave directly.

5. Conclusion

We present a novel method for automatic sleep stage classification using heart rate and wrist actigraphy, which is quite suitable for wearable devices. The method consists of two phases: the multi-level feature learning framework and the BLSTM-based RNN classifier. Unlike traditional approaches with hand-engineered features, feature extraction is designed to capture properties of raw sleep data and composition-based structural representation. RNN learns temporally sequential patterns of sleep. Experiments have demonstrated the effectiveness of the proposed method.

Conflicts of interest

There are no conflicts of interest that could inappropriately



(a) 4-class & RG

(b) 4-class & CG

(c) 5-class & RG

(d) 5-class & CG

Fig. 7. Performance of RNN networks with different hidden layers and units. (RG = resting group, CG = comprehensive group.)

Table 10
Comparison with the existing method for 5-class classification (%).

Method	RG			CG		
	P	R	F_1	P	R	F_1
Baseline (HR)	50.2	47.5	43.2	48.8	47,2	42.1
Baseline (HR + Act)	53.5	51.7	46.0	52.2	51.9	45.5
Proposed (HR)	53.9	56.0	53.2	53.2	55.5	52.8
Proposed (HR + Act)	**58.0**	**60.3**	**58.2**	**58.5**	**61.1**	**58.5**

RG = resting group, CG = comprehensive group, HR = heart rate, Act = actigraphy.
The Bold numbers show the best performances in each table.

influence this research work.

Acknowledgment

This work was supported by Microsoft Research under the eHealth program, the National Natural Science Foundation of China under Grant 81771910, the National Science and Technology Major Project of the Ministry of Science and Technology in China under Grant 2017YFC0110903, the Beijing Natural Science Foundation in China under Grant 4152033, the Technology and Innovation Commission of Shenzhen in China under Grant shenfagai 2016–627, the Beijing Young Talent Project in China, the Fundamental Research Funds for the Central Universities of China under Grant SKLSDE-2017ZX-08 from the State Key Laboratory of Software Development Environment in Beihang University in China, the 111 Project in China under Grant B13003.

References

[1] R. Stickgold, Sleep-dependent memory consolidation, Nature 437 (7063) (2005) 1272, https://doi.org/10.1038/nature04286.

[2] M.A. Carskadon, W.C. Dement, et al., Normal human sleep: an overview, Principles Pract. Sleep Med. 4 (2005) 13–23, https://doi.org/10.1016/j.mcna.2004.01.001.

[3] A.A. of Sleep Medicine, et al., International classification of sleep disorders, Diagnos. Coding Manual (2005) 148–152, https://doi.org/10.1378/chest.14-0970.

[4] R. Boostani, F. Karimzadeh, M. Nami, A comparative review on sleep stage classification methods in patients and healthy individuals, Comput. Methods Progr. Biomed. 140 (2017) 77–91, https://doi.org/10.1016/j.cmpb.2016.12.004.

[5] A. Rechtschaffen, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, Public health service.

[6] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. Vaughn, The Aasm Manual for the Scoring of Sleep and Associated Events, Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine.

[7] A. Schäfer, J. Vagedes, How accurate is pulse rate variability as an estimate of heart rate variability? Int. J. Cardiol. 166 (1) (2013) 15–29, https://doi.org/10.1016/j.ijcard.2012.03.119.

[8] T. F. of the European Society of Cardiology, et al., Heart rate variability: standards of measurement, physiological interpretation, and clinical use, Circulation 93 (1996) 1043–1065.

[9] R.J. Cole, D.F. Kripke, W. Gruen, D.J. Mullaney, J.C. Gillin, Automatic sleep/wake

identification from wrist activity, Sleep 15 (5) (1992) 461–469, https://doi.org/10.1093/sleep/15.5.461.

[10] A. Baharav, S. Kotagal, V. Gibbons, B. Rubin, G. Pratt, J. Karin, S. Akselrod, Fluctuations in autonomic nervous activity during sleep displayed by power spectrum analysis of heart rate variability, Neurology 45 (6) (1995) 1183–1187, https://doi.org/10.1212/WNL.45.6.1183.

[11] J. Trinder, J. Kleiman, M. Carrington, S. Smith, S. Breen, N. Tan, Y. Kim, Autonomic activity during human sleep as a function of time and sleep stage, J. Sleep Res. 10 (4) (2001) 253–264, https://doi.org/10.1046/j.1365-2869.2001.00263.x.

[12] H. Otzenberger, C. Gronfier, C. Simon, A. Charloux, J. Ehrhart, F. Piquard, G. Brandenberger, Dynamic heart rate variability: a tool for exploring sympathovagal balance continuously during sleep in men, Am. J. Physiol. Heart Circ. Physiol. 275 (3) (1998) H946–H950, https://doi.org/10.1152/ajpheart.1998.275.3.H946.

[13] R. Jerath, K. Harden, M. Crawford, V.A. Barnes, M. Jensen, Role of cardiorespiratory synchronization and sleep physiology: effects on membrane potential in the restorative functions of sleep, Sleep Med. 15 (3) (2014) 279–288, https://doi.org/10.1016/j.sleep.2013.10.017.

[14] M. Bonnet, D. Arand, Heart rate variability: sleep stage, time of night, and arousal influences, Electroencephalogr. Clin. Neurophysiol. 102 (5) (1997) 390–396, https://doi.org/10.1016/S0921-884X(96)96070-1.

[15] E. Tobaldini, L. Nobili, S. Strada, K.R. Casali, A. Braghiroli, N. Montano, Heart rate variability in normal and pathological sleep, Front. Physiol. 4 (2013) 294, https://doi.org/10.3389/fphys.2013.00294.

[16] H. Nazeran, Y. Pamula, K. Behbehani, Heart Rate Variability (Hrv): Sleep Disordered Breathing, Wiley Encyclopedia of Biomedical Engineering, https://doi.org/10.1002/9780471740360.ebs1387.

[17] M. Malik, Task force of the european society of cardiology and the north american society of pacing and electrophysiology. heart rate variability. standards of measurement, physiological interpretation, and clinical use, Eur. Heart J. 17 (1996) 354–381, https://doi.org/10.1111/j.1542-474X.1996.tb00275.x.

[18] H. Yoon, S.H. Hwang, J.-W. Choi, Y.J. Lee, D.-U. Jeong, K.S. Park, Slow-wave sleep estimation for healthy subjects and osa patients using r–r intervals, IEEE J. Biomed. Health Inf. 22 (1) (2018) 119–128, https://doi.org/10.1109/JBHI.2017.2712861.

[19] F. Ebrahimi, S.-K. Setarehdan, H. Nazeran, Automatic sleep staging by simultaneous analysis of ecg and respiratory signals in long epochs, Biomed. Signal Process. Control 18 (2015) 69–79, https://doi.org/10.1016/j.bspc.2014.12.003.

[20] M. Xiao, H. Yan, J. Song, Y. Yang, X. Yang, Sleep stages classification based on heart rate variability and random forest, Biomed. Signal Process. Control 8 (6) (2013) 624–633, https://doi.org/10.1016/j.bspc.2013.06.001.

[21] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[22] A. Sadeh, K.M. Sharkey, M.A. Carskadon, Activity-based sleep-wake identification: an empirical test of methodological issues, Sleep 17 (3) (1994) 201–207, https://doi.org/10.1093/sleep/17.3.201.

[23] J. Paquet, A. Kawinska, J. Carrier, Wake detection capacity of actigraphy during sleep, Sleep 30 (10) (2007) 1362–1369, https://doi.org/10.1093/sleep/30.10.1362.

[24] S. Herscovici, A. Peer, S. Papyan, P. Lavie, Detecting rem sleep from the finger: an automatic rem sleep algorithm based on peripheral arterial tone (pat) and actigraphy, Physiol. Meas. 28 (2) (2007) 129–140, https://doi.org/10.1088/0967-3334/28/2/002.

[25] K. Kawamoto, H. Kuriyama, S. Tajima, Actigraphic detection of rem sleep based on respiratory rate estimation, J. Med. Bioeng. 2 (1) (2013) 20–25, https://doi.org/10.12720/jomb.2.1.20-25.

[26] X. Long, P. Fonseca, J. Foussier, R. Haakma, R.M. Aarts, Sleep and wake classification with actigraphy and respiratory effort using dynamic warping, IEEE J. Biomed. Health Inf. 18 (4) (2014) 1272–1284, https://doi.org/10.1109/JBHI.2013.2284610.

[27] A. Quiceno-Manrique, J. Alonso-Hernandez, C. Travieso-Gonzalez, M. Ferrer-Ballester, G. Castellanos-Dominguez, Detection of obstructive sleep apnea in ecg recordings using time-frequency distributions and dynamic features, EMBC, IEEE, Minneapolis, MN, USA, 2009, pp. 5559–5562, , https://doi.org/10.1109/IEMBS.2009.5333736.

[28] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Trans. Acoust. Speech Signal Process. 29 (2) (1981) 254–272, https://doi.org/10.1109/TASSP.1981.1163530.

[29] O. Tsinalis, P. M. Matthews, Y. Guo, S. Zafeiriou, Automatic Sleep Stage Scoring with Single-channel Eeg Using Convolutional Neural Networks, preprint arXiv:1610.01683.

[30] J. Zhang, Y. Wu, J. Bai, F. Chen, Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers, Trans. Inst. Meas. Contr. 38 (4) (2016) 435–451, https://doi.org/10.1177/0142331215587568.

[31] H. Dong, A. Supratak, W. Pan, C. Wu, P.M. Matthews, Y. Guo, Mixed neural network approach for temporal sleep stage classification, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (2) (2018) 324–333, https://doi.org/10.1109/TNSRE.2017.2733220.

[32] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with lstm recurrent networks, J. Mach. Learn. Res. 3 (Aug) (2002) 115–143, https://doi.org/10.1162/153244303768966139.

[33] Z. C. Lipton, J. Berkowitz, C. Elkan, A Critical Review of Recurrent Neural Networks for Sequence Learning, preprint arXiv:1506.00019.

[34] M. Chennaoui, P.J. Arnal, F. Sauvet, D. Léger, Sleep and exercise: a reciprocal issue? Sleep Med. Rev. 20 (2015) 59–72, https://doi.org/10.1016/j.smrv.2014.06.008.

[35] M. H. Lee, C. D. Smyser, J. S. Shimony, Resting-state fmri: a review of methods and clinical applications, AJNR (Am. J. Neuroradiol.). https://doi.org/10.3174/ajnr.A3263.

[36] N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. 100 (1) (1974) 90–93, https://doi.org/10.1109/T-C.1974.223784.

[37] W.-J. Kang, J.-R. Shiu, C.-K. Cheng, J.-S. Lai, H.-W. Tsao, T.-S. Kuo, The application of cepstral coefficients and maximum likelihood method in emg pattern recognition [movements classification], IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 42 (8) (1995) 777–785, https://doi.org/10.1109/10.398638.

[38] J.M. Kortelainen, M.O. Mendez, A.M. Bianchi, M. Matteucci, S. Cerutti, Sleep staging based on signals acquired through bed sensor, IEEE Trans. Inf. Technol. Biomed. 14 (3) (2010) 776–785, https://doi.org/10.1109/TITB.2010.2044797.

[39] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, J.I. Mars, Multivariate temporal dictionary learning for eeg, J. Neurosci. Methods 215 (1) (2013) 19–28, https://doi.org/10.1016/j.jneumeth.2013.02.001.

[40] T. Liu, Y. Si, D. Wen, M. Zang, L. Lang, Dictionary learning for vq feature extraction in ecg beats classification, Expert Syst. Appl. 53 (2016) 129–137, https://doi.org/10.1016/j.eswa.2016.01.031.

[41] J. Wang, P. Liu, M. She, S. Nahavandi, A.Z. Kouzani, Biomedical time series clustering based on non-negative sparse coding and probabilistic topic model, Comput. Methods Progr. Biomed. 111 (3) (2013) 629–641, https://doi.org/10.1016/j.cmpb.2013.05.022.

[42] Y. Xu, Z. Shen, X. Zhang, Y. Gao, S. Deng, Y. Wang, Y. Fan, I. Eric, C. Chang, Learning multi-level features for sensor-based human action recognition, Pervasive Mob. Comput. 40 (2017) 324–338, https://doi.org/10.1016/j.pmcj.2017.07.001.

[43] A. Coates, A.Y. Ng, Learning feature representations with k-means, Neural Networks: Tricks of the Trade, Springer, 2012, pp. 561–580, , https://doi.org/10.1007/978-3-642-35289-8_30.

[44] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures, Neural Network. 18 (5) (2005) 602–610, https://doi.org/10.1016/j.neunet.2005.06.042.

[45] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: B.C. Vancouver (Ed.), ICASSP, IEEE, Canada, 2013, pp. 6645–6649, , https://doi.org/10.1109/ICASSP.2013.6638947.

[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[47] A. Bunde, S. Havlin, J.W. Kantelhardt, T. Penzel, J.-H. Peter, K. Voigt, Correlated and uncorrelated regions in heart-rate fluctuations during sleep, Phys. Rev. Lett. 85 (17) (2000) 3736, https://doi.org/10.1103/PhysRevLett.85.3736.

[48] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with lstm, Neural Comput. 12 (10) (2000) 2451–2471, https://doi.org/10.1162/089976600300015015.

[49] L. Bottou, O. Bousquet, The tradeoffs of large scale learning, Advances in Neural Information Processing Systems, 2008, pp. 161–168.

[50] F. Weninger, J. Bergmann, B. Schuller, Introducing currennt: the munich open-source cuda recurrent neural network toolkit, J. Mach. Learn. Res. 16 (1) (2015) 547–551.

[51] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300, https://doi.org/10.1023/A:1018628609742.