NOTE

Automatic sleep stage classification based on a two-channel electrooculogram and one-channel electromyogram

Yanjun Li, Zhi Xu, Yu Zhang, Zhongping Cao and Hua Chen
China Astronaut Research and Training Center, Haidian District, Beijing, People’s Republic of China
E-mail: yanjunli211@sina.com

Keywords: automatic sleep stage classification, sleep scoring, EEG-free sleep monitoring, EOG, EMG, sleep quality

Abstract

Objective. Sleep monitoring by polysomnography (PSG) severely degrades sleep quality. In order to reduce the load of sleep monitoring, an approach to automatic sleep stage classification without an electroencephalogram (EEG) was proposed. Approach. A total of 124 records from the public dataset ISRUC-Sleep incorporating American Academy of Sleep Medicine (AASM) standards were used: 10 records were from the healthy group while the others were from sleep disorder groups. The 124 records were collected from 116 subjects (eight subjects had two records each, the others had one record each) with ages ranging from 20 to 85 years. A total of 108 features were extracted from the two-channel electrooculograms (EOGs) and six features were extracted from the one-channel electromyogram (EMG). A novel ‘quasi-normalization’ method was proposed and used for feature normalization. Then the random forest algorithm was used to classify five stages, including wakefulness, rapid eye movement sleep, N1 sleep, N2 sleep and N3 sleep. Main results. Using 114 normalized features from the combination of EOG (108 features) and EMG (6 features) data, Cohen’s kappa coefficient was 0.749 and the accuracy was 80.8% by leave-one-out cross-validation. As a reference for AASM standards using a computer-assisted method, Cohen’s kappa coefficient was 0.801 and the accuracy was 84.7% for the same dataset based on 438 normalized features from a combination of EEG (324 features), EOG (108 features) and EMG (6 features) data. Significance. A combination of EOG and EMG can reduce the load of sleep monitoring, and achieves comparable performance to the ‘gold standard’ signals of EEG, EOG and EMG for sleep stage classification.

1. Introduction

Sleep is the body’s self-repair and self-recovery process (Matar et al 2018). Good-quality sleep will eliminate fatigue and restore physical strength. Generally the quality of sleep is assessed quantitatively from the macro-sleep structure. Following American Academy of Sleep Medicine (AASM) standards (Iber et al 2007), sleep stages are divided into wakefulness, rapid eye movement (REM) sleep and non-REM (NREM) sleep. Furthermore, NREM sleep is divided into stages N1, N2 and N3 (Iber et al 2007). Deep sleep (stage N3) is mainly conducive to physical recovery (Matar et al 2018). REM sleep is mainly conducive to the backtracking, encoding, consolidation, trimming and even strengthening of the memory (Matar et al 2018). Breathing and blood oxygen are other dimensions to sleep quality assessment (Mendona et al 2019). Frequent apnea can cause the reduction of both oxygenation and pH value in blood (Mendona et al 2019). Also, frequent movement during sleep is a characteristic of some types of sleep disorder (Mendona et al 2019).

Sleep stage classification (SSC) from PSG provides sleep stage information for the study of sleep patterns (Matar et al 2018). Correct identification of sleep stages is important in diagnosing and treating sleep disorders (Phan et al 2019). Hence, researchers have tried to obtain high accuracy with respect to manual SSC (Yan et al 2021). As the most accurate method, PSG is the decisive approach in many cases (Matar et al 2018).
A six-channel electroencephalogram (EEG), two-channel electrooculogram (EOG) and one-channel electromyogram (EMG) are recommended for sleep scoring according to AASM standards (Iber et al 2007). A framework based on a convolutional neural network (CNN) yielded an accuracy of 82.3% on Sleep-EDF Expanded (Sleep–EDF) dataset and an accuracy of 83.6% based on the Montreal Archive of Sleep Studies dataset (Phan et al 2019). Yan et al (2021) developed a versatile deep-learning architecture to automate sleep scoring with raw PSG, and obtained an accuracy of 86% and a kappa coefficient of 0.82 on ISRUC-Sleep data. A neural network yielded an accuracy of 83.1% and a kappa coefficient of 0.78 for 994 subjects when half the subjects were randomly assigned to the training set and the other half to the testing set (Kuo and Chen 2020). Rahman et al (2018) analyzed single-channel EOG data in the discrete wavelet transform (DWT) domain employing various statistical features, and achieved an accuracy of 86.0% with the random forest algorithm (RF) for ISRUC-Sleep data. The main shortcoming of their study is that only 10 records were used, with five records for training and the other five for testing.

Specifically, the inter-scorer agreement following the AASM standards (Iber et al 2007) was only approximately 82.6% (Rosenberg and Van-Hout 2013). Too high accuracy an from automatic sleep stage classification may be caused by over-fitting. It is reasonable to get a trade-off between boosting the classification performance by integrating as much features as possible and not over-fitting the sleep scoring model in certain sleep stage types (Mendonca et al 2019).

Although PSG is currently the ‘gold standard’ for sleep monitoring (Matar et al 2018), it requires the attachment of at least 10 electrodes to the head, face and body (Iber et al 2007), which seriously interferes with the subject’s natural sleep (Matar et al 2018). As a result, PSG is mainly used in hospitals to monitor patients with severe sleep disorders because of its complex operation and high level of discomfort (Mendonca et al 2019). Hence, some studies emphasized the importance of reducing the burden from data recording during the whole night sleep, such as using wearable, on-bed, and actigraphy devices instead of PSG (Mendonca et al 2019). Researchers have tried to use feasible devices for sleep monitoring in the community, such as ear-EEG (Nakamura et al 2017), dynamic ECG (García-Molina and Jiang 2022), breathing sounds (Myczak et al 2020), plethysmography (Beattie et al 2017), actigraphy-based devices (Mendonca et al 2019) and pressure sensor mattresses (Matar et al 2018).

An EEG-free sleep monitoring method would be of great significance for reducing the load of sleep monitoring. Can comparable performances on sleep classification be achieved by EOG and EMG compared with that of EEG, EOG and EMG? Focusing on the problem above, this paper studies a method for low-load sleep monitoring based on EOG and EMG, and evaluates the role of EOG and EMG signals in sleep staging.

2. Method

2.1. Data acquisition

A public dataset called ISRUC-Sleep (Khalighi et al 2016) with AASM standards was used, it includes sleep disorder groups (two subsets, namely subgroup I and subgroup II) and a healthy group (subgroup III). The dataset was provided by the Sleep Medicine Centre of Coimbra. It can be downloaded freely from the website http://sleeptight.isr.uc.pt/ISRUC_Sleep/. More diagnostic information is presented on the website.

The data comprise 126 PSG records and sleep stage labels from two experts. There are 19 channels of physiological data for most PSG records. However, record 8’ and record ‘40’ were excluded from subgroup I for the analysis in this paper. The former record does not provide EEG channels F3 and F4, while the latter one suffers some electrode problems. As a result, a total of 124 records were used in this research as shown in table 1, including 98 records from subgroup I, 16 records from subgroup II and 10 records from subgroup III. All records in table 1 were combined as the mixed group (10 healthy recordings and 114 sleep disorder recordings) for sleep scoring in this paper.

The proposed method focused on two-channel EOG and one-channel EMG, while a combination of six-channel EEG, two-channel EOG and one-channel EMG was used for comparison. The sampling frequency was 200 Hz for each channel. All these channels in ISRUC-Sleep had been filtered to eliminate noise and undesired background by the dataset itself, aiming to enhance the PSG signal quality and increase the signal to noise ratio (SNR). The filtering stage comprised: (1) a notch filter to eliminate the 50 Hz electrical noise; (2) a band-pass Butterworth filter with a low cutoff frequency at 0.3 Hz and a high cutoff frequency at 35 Hz for both EEG and EOG channels, and a low cutoff frequency at 10 Hz and a high cutoff frequency at 70 Hz for EMG channels.

The sleep stages of each subject in the dataset were labeled by two experts individually. Therefore, small differences existed in annotations between two experts. If sleep scores from only one expert were used, a bias would produce from a rater’s style. As a result, only those stages with consistent labels from two experts were extracted for analysis in this paper to avoid personal inclination.
The fractal dimension of EMG is defined as 

\[ E_{\text{Ent}} = \frac{\sum_{i=1}^{N} E_{\text{sum}}}{N} \]

where 

\[ E_{\text{sum}} = E_{\text{delta}} + E_{\text{theta}} + E_{\text{alpha}} + E_{\text{beta}} \]

The Shannon entropy derived from each sub-band is defined as

\[ S_{\text{delta}} = -\sum_{i=1}^{N} \frac{E_{\text{delta}}}{E_{\text{sum}}} \log \left( \frac{E_{\text{delta}}}{E_{\text{sum}}} \right) \]

\[ S_{\text{theta}} = -\sum_{i=1}^{N} \frac{E_{\text{theta}}}{E_{\text{sum}}} \log \left( \frac{E_{\text{theta}}}{E_{\text{sum}}} \right) \]

\[ S_{\text{alpha}} = -\sum_{i=1}^{N} \frac{E_{\text{alpha}}}{E_{\text{sum}}} \log \left( \frac{E_{\text{alpha}}}{E_{\text{sum}}} \right) \]

\[ S_{\text{beta}} = -\sum_{i=1}^{N} \frac{E_{\text{beta}}}{E_{\text{sum}}} \log \left( \frac{E_{\text{beta}}}{E_{\text{sum}}} \right) \]

Similarly, the sum of the absolute value in these sub-bands in each 30 s period is defined as

\[ S_{\text{sum}} = S_{\text{delta}} + S_{\text{theta}} + S_{\text{alpha}} + S_{\text{beta}} \]

The Shannon entropy derived from each sub-band is defined as

\[ S_{\text{entropy}} = \sum_{i=1}^{N} \frac{E_{\text{sum}}}{E_{\text{delta}} + E_{\text{theta}} + E_{\text{alpha}} + E_{\text{beta}}} \log \left( \frac{E_{\text{sum}}}{E_{\text{delta}} + E_{\text{theta}} + E_{\text{alpha}} + E_{\text{beta}}} \right) \]

and

\[ S_{\text{sum}} - S_{\text{entropy}} \]

The feature vector derived from the four sub-bands is defined as

\[ \text{EogBand4}_{\text{Ft11}} = [E_{\text{Ent}}, E_{\text{delta}}, E_{\text{theta}}, E_{\text{alpha}}, E_{\text{beta}}, S_{\text{delta}}, S_{\text{theta}}, S_{\text{alpha}}, S_{\text{beta}}] \]

In the same way, for the 11 sub-bands, the number of features for single-channel EOG is 49. The feature vector is as follows:

\[ \text{OneLeadEog} = [\text{EogBand4}_{\text{Ft11}}, \text{EogBand1}_{\text{Ft13}}] \]

where OneLeadEog represents LeadEog_{LOC} for lead ‘LOC-A2’ and LeadEog_{ROC} for lead ‘ROC-A1’.

### 2.2.2. Correlation features between two-channel EOG

Temporal signals within sub-bands delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz) and beta (13–30 Hz) are derived from individual Finite Impulse Response band-pass filters from original EOG data within the frequency band 1–4 Hz, 4–8 Hz, 8–13 Hz and 13–30 Hz, respectively. The Pearson correlation coefficients between two-channel EOG in four sub-bands during each 30 s period are defined as

\[ r_{\text{delta}}, r_{\text{theta}}, r_{\text{alpha}}, r_{\text{beta}} \]

The number of features between two-channel EOG is 10, and the feature vector is as follows:

\[ \text{EogBtw} = [r_{\text{delta}}, r_{\text{theta}}, r_{\text{alpha}}, r_{\text{beta}}, PLV_{\text{delta}}, PLV_{\text{theta}}, PLV_{\text{alpha}}, PLV_{\text{beta}}, PLV_{\text{org}}] \]

The number of features for one-channel EOG is 49, and the number of features between two-channel EOG is 10. Therefore, the total number of features for two-channel EOG ‘LOC-A2’ and ‘ROC-A1’ is

\[ 49 \times 2 + 10 = 108 \]

and the whole vector for EOG is defined as

\[ \text{EogFeat} = [\text{LeadEog}_{\text{LOC}}, \text{LeadEog}_{\text{ROC}}, \text{EogBtw}] \]

### 2.2.3. Features from single-channel EMG

The fractal dimension of EMG is defined as EmgFD, and the root mean square is defined as EmgStd during each 30 s.

### Table 1. Records used in this paper from the ISRUC-Sleep database (ave., average; Sd, standard deviation).

| Group | Type of participants | No. of records | No. of subjects (gender) | Age (years) |
|-------|----------------------|----------------|--------------------------|-------------|
| Subgroup I | Participants with sleep disorder | 98* | 98 subjects (54 male, 44 female) | 20–85, ave. = 50.7, Sd = 15.9 |
| Subgroup II | Participants with sleep disorder | 16 | 8 subjects (6 male, 2 female) | 26–79, ave. = 46.9, Sd = 18.7 |
| Subgroup III | Healthy participants | 10 | 10 subjects (9 male, 1 female) | 30–58, ave. = 39.6, Sd = 10.1 |

* Record 8’ and record 40’ were excluded from the analysis. The former record does not provide EEG channels F3 and F4, while the latter has some electrode problems.
The EMG signal of every 30 s period is transformed by the Hilbert transform to obtain the enveloping signal. After that, the enveloping mean is defined as EnvlpMean. The enveloping maximum is defined as EnvlpMax. The enveloping root mean square is defined as EnvlpStd, and the ratio of EnvlpMax to EnvlpMean is defined as RtMaxdMean. The total number of features for single-channel EMG is 6, and the whole vector for EMG is defined as:

\[
\text{EmgFeat} = [\text{EmgFD EmgStd RtMaxdMean EnvlpMean EnvlpMax EnvlpStd}]
\]  

(5)

2.2.4. Features from six-channel EEG

This method is compared with AASM standards in this paper using a computer-assisted method. Therefore, EEG features are calculated. Six-channel EEGs can be divided into three groups, including \{F3, F4\}, \{C3, C4\} and \{O1, O2\}. For each group, a total of 108 features can be obtained in the same way as in formula (4), defined as F34Feat, C34Feat and O12Feat, respectively. Consequently, there are \(108 \times 3 = 324\) features for all six-channel EEGs.

2.2.5. Whole feature vector

For classification based on EOG and EMG, the whole feature vector with 114 features from two-channel EOG (108 features) and one single-channel EMG (6 features) is defined as follows:

\[
\text{Feat1} = [\text{EogFeat EmgFeat}].
\]

(6)

For comparison with AASM standards using a computer-assisted method, the whole feature vector with 438 features from two-channel EOG (108 features), one single-channel EMG (6 features) and six-channel EEG (324 features) is defined as follows:

\[
\text{Feat2} = [\text{EogFeat EmgFeat F34Feat C34Feat O12Feat}].
\]

(7)

2.3. Characteristic normalization

One normal sleep in adults may last for 8 h. During such a long period, the recording conditions, such as skin humidity, body temperature, body position, body movements or even loss of electrode contact, may change from time to time. The discriminant information for sleep stage classification lies in relative amplitudes rather than absolute amplitudes. Therefore, amplitudes of electric physiological signals will change from time to time. Hence normalization of features is important.

If the maximum and the minimum values in the feature sequence are taken as the reference for feature normalization it may cause an error, because both the maximum and the minimum values may be noise points. For example, suppose most values in a feature sequence are near 1, but one noise point is 100 and another noise point is \(-10\). If the scale is normalized according to the maximum and the minimum values, i.e. 100–\((-10) = 110\), then most values in the normalized feature series are clustered around 0.01. Besides, only the noise point 100 is scaled to near 1 and the noise point \(-10\) is scaled to near 0, which is obviously not the expected result of normalization.

A new ‘quasi-normalization’ method is designed in this paper. First, the original feature sequences \{\(a(n)\)\} are arranged in order from small to large, which is defined as \{\(f(n)\)\}. We set the series number of \{\(f(n)\)\} at the position of 10% of the distance from the beginning as \(n1\), the position of 50% distance from the beginning as \(n2\) and the position of 90% distance from the beginning as \(n3\).

The standard deviation (Sd) of the sequences \{\(f(n1:n3)\)\} is defined by formula (8)

\[
\text{Sd} = \sqrt{\frac{1}{n3 - n1} \sum_{i=n1}^{n3} f(i) - \frac{1}{n3 - n1} \sum_{j=n1}^{n3} f(j)}
\]

(8)

\[
Ku = f(n3) - f(n2)
\]

(9)

\[
Kd = f(n2) - f(n1)
\]

(10)

\[
S = 2 \times \min([\text{Sd} Ku Kd])
\]

(11)

\[
b(n) = (a(n) - f(n2)) / S.
\]

(12)

Then, using formula (12) for ‘quasi-normalization’, most elements in \{\(b(n)\)\} are transformed into the interval \([-2, 2]\), but a few elements are outside that range. In order to locate all elements in the interval \([-2, 2]\), the following transform is applied:
Finally, the feature sequences \( \{c(n)\} \) are used for classification.

Figure 1 shows an example of quasi-normalization for EnvlpStd of EMG. For the original index EnvlpStd, most elements are lower than 2 in figure 1(a), but none is lower than 0. After arranging in order from small to large, sequences \( f(n) \) are obtained. As it is shown in figure 1(b), \( n1 \) is at the position of 10% of the distance of \( f(n) \) from the beginning, \( n2 \) is at the position of 50% distance from the beginning, and \( n3 \) is at the position of 90% distance from the beginning. Values of \( \{f(n1), f(n2), f(n3)\} \) are used for quasi-normalization. After using formula (12) for ‘quasi-normalization’, most elements in \( \{b(n)\} \) are transformed into the interval \([-2, 2] \), as in figure 1(c), but some elements are still higher than 2. After using formula (13) for data truncation, elements that are higher than 2 are reset as 2.

Figure 2 is an example of quasi-normalization for EmgStd of EMG. For the original index EmgStd, most elements are lower than 2 in figure 2(b), but none is lower than 0. After using formula (12) for ‘quasi-normalization’, most elements in \( \{b(n)\} \) are transformed into the interval \([-2, 2] \), as in figure 2(c), but some elements are still higher than 2. After using formula (13) for data truncation, elements that higher than 2 are reset as 2.

Figure 3 is an example of quasi-normalization for PLVorg of EOG. For the original index \( PLV_{org} \) of EOG, as in figure 3(b), all elements are in the range \([-1, 0.5] \). After using formula (12) for ‘quasi-normalization’, most elements in \( \{b(n)\} \) are transformed into the interval \([-2, 2] \), as in figure 3(c), but some elements are lower than \(-2 \). After using formula (13) for data truncation, elements that lower than \(-2 \) are reset as \(-2 \).

2.4. Classification model selection

The RF algorithm (Pan et al 2018) has some wonderful advantages, including strong generalization ability, strong anti-over-fitting ability, rapid model training, simple structure and easy construction, making it suitable for processing high-dimensional data sets without the need for feature selection.

In this paper, the number of trees is also set as 500 (Pan et al 2018). When the number of trees is too huge, the calculation speed will slow down, but the accuracy is almost unaffected.

2.5. Comparison of classification results

The leave-one-record-out cross-validation (LOOCV) strategy is applied to the mixed group (10 healthy recordings and 114 sleep disorder recordings). The training dataset contains 123 records while the other dataset is used as the validation set. This step repeats 124 times until each record has been tested. The combination of the 124 times of testing forms the final results.

Furthermore, results are obtained that are derived from each signal type among EEG, EOG and EMG. Evaluation indices are employed, including accuracy, the multi-class weighted F1-score (Khalighi et al 2013) and
Cohen's kappa coefficient. Cohen's kappa coefficient is a more effective evaluator because it takes prior probability into account (Pan et al. 2018).

3. Results

3.1. Classification results for individual records

The results of sleep stage classification of subgroup III from the combination of EOG (108 features) and EMG (6 features) are shown in Table 2. According to Cohen's kappa coefficients, results from 8 out of 10 records are in substantial agreement, with kappa coefficients in the range [0.6, 0.8]. Record no. 1 shows almost perfect agreement (0.801), but record no. 10 is only in moderate agreement (0.542).

Classification results for record no. 6 from subgroup III are shown in Figure 4. Classification results are similar with manual scoring, supported by F-scores (N3, 0.930; N2, 0.862; awake, 0.811; REM, 0.772). The percentage of each stage is also similar to the manual scoring, as shown in the row for subject 6 in Table 3. However, the REM stage percentage (17.7%) is significantly higher than with manual scoring (10.8% from the first expert and 10.0% from the second expert).

The worst-case result from Table 2 is for subject 10, as shown in Figure 6. Classification results are to some extent similar to manual scoring, supported by the F-scores (REM 0.822; N3 0.778). The REM stage percentage is similar to manual scoring, as shown in the row for subject 3 in Table 3. However, the N3 stage percentage (22.0%) derived from the classification is significantly lower than with manual scoring (14.1% from the second expert). The visual evidence is very obvious in Figure 5, as the second N3 stage in the second expert's manual scoring is mistaken for N2 stage by the proposed method, and almost half of the last two N3 stages in the second expert's manual scoring is also mistaken for N2 stages. The wakefulness stage percentage (17.7%) is significantly higher than with manual scoring (10.8% from the first expert and 10.0% from the second expert).

The worst-case result from Table 2 is for subject 10, as shown in Figure 6. Classification results are to some extent similar to manual scoring, supported by the F-scores (REM 0.822; N3 0.778). The REM stage percentage (17.7%) is similar to the result with manual scoring (14.1%), as shown in the row for subject 10 in Table 3. However, the N1 stage percentage (6.9%) derived from the classification is significantly lower than the...
and the N2 stage percentage (42.7%) is significantly higher than the manual scoring result (22.7%).

### 3.2. Classification results for all records

The results of sleep stage classification from the combination of EOG (108 features) and EMG (6 features) are shown in table 4, and results from the combination of EEG (324 features), EOG (108 features) and EMG (6 features) are shown in table 5. More details are shown in table 6.

When classification results were derived from only one type of the signal from EEG, EOG and EMG, Cohen’s kappa coefficients from high to low were in the order C-EEG > F-EEG > O-EEG > EOG > EMG, as shown in table 6.

In multi-class sleep staging, the best discrimination was achieved by a combination of EEG, EMG and EOG, in which the highest F1-score was 0.894 for both awake and N3 stages in table 6, followed by REM (0.884) and N2 (0.836) stages. However, the lowest F1-score was for the detection of the N1 stage (0.528). The order of F1-scores for detection of a single stage is consistent with the order of accuracy of the research of Khalighi et al (2013), with awake (88.59%) > N3 (87.13%) > REM (86.99%) > N2 (79.06%) > N1 (66.91%) (Khalighi et al 2013).

Most studies in table 6 did not use the whole ISRUC-Sleep dataset for cross-validation. It is not a fair comparison when one study uses more than 100 records for validation and another one only uses five.

Furthermore, for training and testing sets as in table 7, our method has comparable performance with the method of Rahman et al (2018).

Using 99 records for validation from the combination of six-channel EEG, two-channel EOG, one-channel EMG and one-channel ECG, deep learning (Yan et al 2021) in table 7 achieved an accuracy of 86% and Cohen’s kappa coefficient of 0.82. In comparison with that, our proposed method used 124 records for validation, and obtained an accuracy of 84.7% and Cohen’s kappa coefficient of 0.807 from the combination of six-channel EEG, two-channel EOG and one-channel EMG. The main difference is that the F-score with our method is only 0.528 but the deep-learning method (Yan et al 2021) had a better F-score of 0.67 for the N1 stage.
| Record no. | Accuracy (%) | Kappa  | Balanced F-score | Awake F-score | REM F-score | N1 F-score | N2 F-score | N3 F-score |
|-----------|--------------|--------|------------------|--------------|------------|------------|------------|------------|
| 1         | 85.2         | 0.801  | 0.826            | 0.875        | 0.901      | 0.625      | 0.871      | 0.858      |
| 2         | 80.8         | 0.745  | 0.748            | 0.771        | 0.884      | 0.404      | 0.822      | 0.861      |
| 3         | 70.9         | 0.613  | 0.686            | 0.695        | 0.822      | 0.482      | 0.654      | 0.778      |
| 4         | 78.8         | 0.724  | 0.734            | 0.898        | 0.743      | 0.368      | 0.767      | 0.895      |
| 5         | 81.2         | 0.752  | 0.751            | 0.797        | 0.792      | 0.327      | 0.901      | 0.940      |
| 6         | 83.1         | 0.769  | 0.769            | 0.811        | 0.772      | 0.469      | 0.862      | 0.930      |
| 7         | 78.1         | 0.708  | 0.664            | 0.946        | 0.810      | 0.109      | 0.649      | 0.806      |
| 8         | 78.4         | 0.715  | 0.729            | 0.912        | 0.896      | 0.358      | 0.617      | 0.862      |
| 9         | 77.4         | 0.697  | 0.735            | 0.824        | 0.642      | 0.484      | 0.813      | 0.871      |
| 10        | 63.4         | 0.542  | 0.643            | 0.808        | 0.627      | 0.339      | 0.575      | 0.867      |
| Mean ± Sd | 77.7 ± 6.3   | 0.707 ± 0.076 | 0.729 ± 0.053 | 0.834 ± 0.075 | 0.793 ± 0.091 | 0.397 ± 0.136 | 0.753 ± 0.119 | 0.867 ± 0.049 |
3.3. Best features for two-stage classification

The best features from the combination of EEG (324 features), EOG (108 features) and EMG (6 features) are different for the two-stage classification by RF, including \{W, R\}, \{W, N1\}, \{W, N2\}, \{W, N3\}, \{R, N1\}, \{R, N2\}, \{R, N3\}, \{N1, N2\}, \{N1, N3\} and \{N2, N3\}. Overall, the N1 stage is the most difficult to distinguish. In figure 7, the kappa coefficient with two-classification of LOOCV is arranged in order from large to small for each single feature. The highest kappa coefficient by single feature is 0.721 for \{W, R\}, 0.431 for \{W, N1\}, 0.638 for \{W, N2\}, 0.900 for \{W, N3\}, 0.454 for \{R, N1\}, 0.590 for \{R, N2\}, 0.927 for \{R, N3\}, 0.392 for \{N1, N2\}, 0.831 for \{N1, N3\} and 0.632 for \{N2, N3\}, respectively. This reveals that the two-classification from easy to hard is in the following order: \{R, N3\}, \{W, N3\}, \{N1, N3\}, \{W, R\}, \{W, N2\}, \{N2, N3\}, \{R, N2\}, \{R, N1\}, \{W, N1\}, \{N1, N2\}.

The adjacent relations among five stages in this paper are defined as figure 8.

(1) Three types are easy to classify, including \{N1, N3\}, \{R, N3\} and \{W, N3\}, which all contain N3 but with the other stage nonadjacent to N3. They are labeled as ‘o’ in the legend of figure 7. As shown in table 8, for the \{N1, N3\} classification, the best five features are all related to EEG in the 18–25 Hz band. For \{R, N3\} classification, the best ten features are all related to EEG with low frequency in the delta (1–4 Hz) band or 0.4–4 Hz band. For \{W, N3\} classification, the best six features are all related to EEG index ratios of beta to delta.

(2) Three types are very hard to classify, including \{N1, N2\}, \{W, N1\} and \{R, N1\}, which all contain N1 and another stage adjacent to N1. They are labeled as ‘<’ in the legend of figure 7. As shown in table 8, for \{N1, N2\} classification, the best 15 features are all related to EEG with frequency higher than 18 Hz. For \{W, N1\} classification, the best four features are all related to the EEG alpha band (8–13 Hz), as alpha bursts in the N1 stage. For \{R, N1\} classification, the best feature is the enveloping mean of EMG, while the second to the sixth best features are all related to EOG with a frequency higher than 41 Hz.

(3) Four types are somewhat hard to classify, including \{N2, N3\}, \{R, N2\}, \{W, R\} and \{W, N2\}. They are labeled as ‘[]’ in the legend of figure 7. As shown in table 8, for \{N2, N3\} classification, the best two features are all related to EOG within the frequency band 13–18 Hz due to the presence of spindles in the N2 stage. For \{R, N2\} classification, the best feature is the correlation coefficient between two-channel EOG with the original waveform during each 30 s period. For \{W, R\} classification, the best four features are all related to EEG energy with the frequency band 41–46 Hz. For \{W, N2\} classification, the best six features include three
Table 3. Percentage of each stage from the combination of EMG and EOG for subgroup III.

| Subject | W     | N1    | N2    | N3     | REM    | W     | N1    | N2    | N3     | REM    | W     | N1    | N2    | N3     | REM    |
|---------|-------|-------|-------|--------|--------|-------|-------|-------|--------|--------|-------|-------|-------|--------|--------|
| 1       | 17.3  | 12.5  | 39.1  | 18.7   | 12.5   | 15.4  | 11.5  | 41.1  | 19.9   | 12.1   | 19.1  | 8.5   | 40.0  | 18.7   | 13.7   |
| 2       | 12.2  | 15.3  | 34.3  | 20.9   | 17.2   | 9.4   | 11.3  | 34.2  | 25.8   | 19.3   | 17.7  | 6.3   | 42.8  | 16.7   | 16.5   |
| 3       | 10.8  | 7.9   | 30.3  | 38.5   | 12.5   | 10.0  | 8.9   | 31.9  | 35.0   | 14.3   | 17.7  | 11.4  | 35.9  | 22.0   | 15.0   |
| 4       | 21.9  | 17.3  | 29.5  | 20.0   | 11.3   | 19.8  | 16.9  | 27.5  | 19.9   | 16.0   | 21.4  | 6.4   | 41.3  | 21.4   | 9.4    |
| 5       | 32.0  | 7.6   | 31.1  | 20.7   | 8.6    | 30.1  | 11.3  | 32.9  | 15.4   | 10.3   | 22.1  | 14.7  | 34.3  | 21.1   | 7.2    |
| 6       | 9.1   | 15.9  | 34.7  | 29.0   | 11.3   | 8.4   | 8.1   | 39.7  | 27.7   | 16.1   | 11.4  | 13.6  | 35.5  | 29.7   | 9.8    |
| 7       | 27.3  | 8.4   | 19.8  | 32.2   | 12.4   | 24.6  | 2.1   | 31.7  | 29.0   | 12.7   | 28.6  | 11.7  | 30.8  | 19.7   | 9.2    |
| 8       | 37.6  | 11.9  | 20.1  | 14.3   | 16.1   | 37.6  | 11.9  | 20.1  | 14.3   | 16.1   | 33.5  | 2.8   | 30.1  | 19.6   | 14.0   |
| 9       | 15.4  | 17.2  | 37.7  | 23.2   | 6.5    | 14.9  | 17.3  | 37.9  | 23.4   | 6.5    | 18.2  | 8.5   | 39.0  | 23.3   | 11.0   |
| 10      | 18.5  | 27.8  | 22.7  | 14.1   | 17.0   | 18.5  | 27.8  | 22.7  | 14.1   | 17.0   | 25.8  | 6.9   | 42.7  | 13.9   | 10.7   |
| Mean ± Sd | 20.2 ± 9.5 | 14.2 ± 6.1 | 29.9 ± 7.0 | 23.2 ± 7.8 | 12.5 ± 3.5 | 18.9 ± 9.5 | 12.7 ± 6.8 | 32.0 ± 6.9 | 22.5 ± 7.0 | 14.0 ± 3.7 | 21.6 ± 6.3 | 9.1 ± 3.7 | 37.2 ± 4.6 | 20.6 ± 4.2 | 11.5 ± 2.7 |
EEG indices of ratios of beta to delta, and three EEG indices of percentage of with frequency higher than 36 Hz.

According to Table 8, most good features are associated with EEG. However, there are a few features related to EOG or EMG. The enveloping mean from EMG is the seventh best feature for discriminating stages \{W, R\}. Energy powers of the alpha band in EOG are the third and the fourth best features for discriminating stages \{W, N1\}. One remarkable finding is that the best 13 features for discriminating stages \{R, N1\} are all non-EEG features.
4. Discussion

4.1. The need for feature normalization

As shown in table 9, our proposed method for feature normalization improves performance in comparison with original features. For traditional feature normalization, i.e. $y(n) = (x(n) - \text{mean}(X))/\text{max}(X) - \text{min}(X)$, the performance is even worse than using original features. The reason is that physiological signals often have significant individual characteristics. For example, although the lowest EMG amplitudes in most subjects occurred during deep or REM sleep, a few subjects tended to be different when the SNR was too low, so had the highest EMG amplitudes during wakefulness. This is the reason why traditional feature normalization should be reconsidered for noisy physiological signals.

Positions of indices $n_1$ and $n_3$ were chosen according to the noise level of the signal. Let $r\%$ be the key index, then $n_1$ is at the position of $r\%$ of the distance of $f(n)$ from the beginning and $n_3$ is at the position of $1 - r\%$ of the distance from the beginning. If the noise level is very low, then $r\%$ should be close to 0. If the noise level is very high, then $r\%$ should be close to the noise level but less than 50%. Therefore, the noise level could be estimated as $r\%$, i.e. about $r\%$ of the features are invalid values. The influence of positions of $n_1$ and $n_3$ in equations (8)–(10) was also tested in this paper. As shown in table 9, the changes of their positions induce fluctuations in accuracy of less than 1%. Obviously, the proposed classification method benefited from the feature normalization proposed in this paper.

4.2. Selection of signal type for sleep monitoring

Currently, the gold standard for sleep monitoring is PSG recorded in the hospital or in a sleep laboratory. PSG remains a complex, highly demanding and obtrusive procedure (Matar et al 2018). Several shortcomings limit the utility of PSG, such as discomfort during sleep assessment, high cost and labor-intensive procedure (Nakamura et al 2017). On the contrary, unattended and portable non-medical devices, also referred to as electronic gadgets, deliver highly unobtrusive measurements at the expense of accuracy and reliability (Matar et al 2018). The problem, therefore, is ‘how to assess sleep stages and sleep quality less intrusively but more reliably’ (Matar et al 2018). It is contradictory to improve the classification accuracy and reduce the degree of intrusion during sleep.

When signals from sleep recordings are selected from EEG, EOG and EMG signals, the more the signal types, the higher the classification accuracy. However, to reduce the burden on subjects and reduce sleep disturbance, the fewer signal types the better. According to the AASM, six-channel unipolar EEG acquisition needs eight electrodes, two-channel EOG acquisition needs three electrodes and one-channel EMG acquisition needs two electrodes. These electrodes and their cables will greatly influence sleep quality.

Table 4. Classification results from the combination of EMG and EOG (number of 30 s segments).

| Classification results | Awake | REM | N1 | N2 | N3 | All |
|------------------------|-------|-----|----|----|----|-----|
| Reference              | Awake | 20 376 | 464 | 671 | 902 | 338 | 22 751 |
| REM                    | 705   | 11 917 | 418 | 538 | 164 | 13 742 |
| N1                     | 1668  | 593  | 3191| 1933| 141 | 7526 |
| N2                     | 1778  | 997  | 1397| 22 463| 2611| 29 246 |
| N3                     | 293   | 163  | 7   | 1787| 15 755| 18 005 |
| All                    | 24 820| 14 134| 5684| 27 623| 19 009 |

Table 5. Classification results from the combination of EEG, EMG and EOG (number of 30 s segments).

| Classification results | Awake | REM | N1 | N2 | N3 | All |
|------------------------|-------|-----|----|----|----|-----|
| Reference              | Awake | 20 991| 468 | 614 | 624 | 54 | 22 751 |
| REM                    | 625   | 12 292| 362 | 422 | 41  | 13 742 |
| N1                     | 1403  | 645  | 3520| 1882| 76  | 7526 |
| N2                     | 1029  | 652  | 1300| 24 157| 2108| 29 246 |
| N3                     | 170   | 0    | 2  | 1447| 16 386| 18 005 |
| All                    | 24 218| 14 057| 5798| 28 532| 18 665| /   |
Table 6. Comparison of the performance by LOOCV on the ISRUC-Sleep dataset.

| Method                                | No. of records | Signal type                  | Model              | No. of features | Accuracy (%) | Kappa | Balanced F-score | Awake F-score | REM F-score | N1 F-score | N2 F-score | N3 F-score |
|---------------------------------------|----------------|------------------------------|--------------------|-----------------|--------------|-------|------------------|----------------|-------------|------------|------------|------------|------------|
| Proposed method (this paper)          | 124            | 6EEG + 2EOG + 1EMG          | RF                 | 438             | 84.7         | 0.801 | 0.807            | 0.894          | 0.884       | 0.528      | 0.836      | 0.894      |
| 6EEG                                  | 324            | RF                           | 324                | 83.5            | 0.784        | 0.790 | 0.886            | 0.848          | 0.496       | 0.824      | 0.894      |            |
| C-EEG                                 | 108            | RF                           | 108                | 82.1            | 0.766        | 0.774 | 0.871            | 0.837          | 0.469       | 0.808      | 0.884      |            |
| F-EEG                                 | 108            | RF                           | 108                | 82.0            | 0.764        | 0.771 | 0.867            | 0.843          | 0.458       | 0.807      | 0.881      |            |
| O-EEG                                 | 108            | RF                           | 108                | 81.2            | 0.735        | 0.764 | 0.879            | 0.804          | 0.457       | 0.797      | 0.882      |            |
| 2EOG and EMG                          | 114            | RF                           | 114                | 80.8            | 0.749        | 0.767 | 0.857            | 0.855          | 0.483       | 0.790      | 0.851      |            |
| 2EOG                                  | 108            | RF                           | 108                | 80.3            | 0.742        | 0.759 | 0.854            | 0.835          | 0.465       | 0.791      | 0.848      |            |
| EMG                                   | 6              | RF                           | 6                  | 50.1            | 0.359        | 0.469 | 0.622            | 0.627          | 0.272       | 0.364      | 0.462      |            |
| Shen et al (2020)                     | 10             | C3–A2 EEG                    | State space model  | NP              | 81.7         | 0.763 | NP               | 0.903          | NP          | NP         | NP         | NP         |
| Khalighi et al (2013)                 | 40             | 6EEG + 2EOG + 1EMG          | Support vector     | 326             | 84.7         | NP    | 0.747            | NP             | NP          | NP         | NP         | NP         |
| Khalighi et al (2013)                 | 40             | 6EEG                         | SVM                | 200             | 80.2         | NP    | 0.671            | NP             | NP          | NP         | NP         | NP         |
| Ghasemzadeh et al (2019)              | 10             | C3–A2 EEG                    | Stockwell transform, SVM | 44             | 82.3         | 0.771 | NP               | NP             | NP          | NP         | NP         | NP         |

Note: leave-one-out cross-validation, NP = Not Provided.
Table 7. Comparison of the performance with the training and testing sets from the dataset ISRUC-Sleep.

| Method                        | No. of records | Validation method                     | Signal type               | Model            | No. features | Accuracy (%) | Kappa | Balanced F-score | Awake F-score | REM F-score | N1 F-score | N2 F-score | N3 F-score |
|-------------------------------|----------------|---------------------------------------|---------------------------|-------------------|--------------|---------------|--------|------------------|---------------|-------------|------------|------------|------------|
| Yan et al (2021)              | 99             | 5-fold cross-validation               | 6EEG + 2EOG + 1EMG + 1ECG| Deep learning     | –            | 86            | 0.82  | NP               | 0.94          | 0.84        | 0.67       | 0.86       | 0.89       |
| Rahman et al (2018)           | 10             | 5 records for training, 5 records for testing | Left EOG channel           | RUSBoost          | 30           | 84.7          | NP    | NP               | NP            | NP          | NP         | NP         | NP         |
| Rahman et al (2018)           | 10             | 5 records for training, 5 records for testing | Left EOG channel           | RF                | 30           | 86.0          | NP    | NP               | NP            | NP          | NP         | NP         | NP         |
| Rahman et al (2018)           | 10             | 5 records for training, 5 records for testing | Left EOG channel           | SVM               | 30           | 85.4          | NP    | NP               | NP            | NP          | NP         | NP         | NP         |
| Proposed method (this paper)  | 10             | 5 records for training, 5 records for testing | 2EOG and EMG              | RF                | 114          | 83.5          | 0.779 | 0.805            | 0.861         | 0.892       | 0.573      | 0.856      | 0.844      |
| Shi et al (2019)              | 126            | 106 records for training, 20 records for testing | 6EEG + 2EOG + 1EMG        | Long short-term memory (LSTM) | NP           | 81            | NP    | 0.80             | 0.82          | 0.89        | 0.78       | 0.70       | 0.82       |
| Shi et al (2019)              | 126            | 106 records for training, 20 records for testing | 6EEG + 2EOG + 1EMG        | LSTM with fuzzy entropy | NP           | 86            | NP    | 0.84             | 0.88          | 0.86        | 0.90       | 0.70       | 0.84       |
Using only one type of physiological signal can further reduce the number of electrodes attached to the body. The brain provides the most useful information about sleep regulation. The EEG is the most important signal that directly reflects the state of the brain during sleep (Sun et al 2019). When only one physiological signal is selected from EOG and EMG, as shown in table 6, the precision for sleep staging derived from EOG alone is much higher than that from EMG alone, but lower than EEG alone. EOG-based sleep staging is relatively poorer than the method using EEG, because the criterion of scoring sleep stages mainly depends on the characteristics of EEG signals (Sun et al 2019).

EOG is very useful for identifying wakefulness and REM sleep, since there are eye movements during these stages. Eye movements tend to slow down with increase in the depth of sleep. EOG records the movement of the
Table 8. Best 15 features with the highest kappa coefficient for two-classification.

| Order/Stage | W,R | W,N1 | W,N2 | W,N3 | R,N1 | R,N2 | R,N3 | N1,N2 | N1,N3 | N2,N3 |
|-------------|-----|------|------|------|------|------|------|-------|-------|-------|
| 1           | 1   | C3_Energy (41-46) | O1_Palpha | C4_Ebeta/Edelta | C4_Ebeta/Edelta | EnvlpMean | Eog_rorg | O2_Pdelta | F3_Prb_Norm2 (30-36) | C4_Prb_Norm2 (18-25) | Eog1_Prb_Norm2 (13-18) |
| 2           | 2   | F3_Energy (41-46) | O2_Palpha | F4_Prb_Norm2 (41-46) | C4_Ebeta/Edelta | Eog2_Prb_Norm2 (41-46) | O1_Pdelta | C3_Pdelta | C3_Prb_Norm2 (18-25) | C4_Prb_Norm2 (18-25) | Eog1_Prb_Norm1 (13-18) |
| 3           | 3   | C4_Energy (41-46) | C3_Ebeta/Edelta | C4_Sbeta/Sdelta | C4_Sbeta/Sdelta | Eog1_Prb_Norm2 (41-46) | O1_Pdelta | C3_Pdelta | C3_Prb_Norm2 (18-25) | C4_Prb_Norm2 (18-25) | C3_Prb_Norm2 (13-18) |
| 4           | 4   | O1_Energy (41-46) | Eog2_Palha | F4_Prb_Norm2 (36-41) | C4_Sbeta/Sdelta | Eog2_Prb_Norm2 (46-50) | O1_Ptheta | C4_Pdelta | C4_Prb_Norm2 (30-36) | F3_Prb_Norm2 (18-25) | Eog1_Prb_Norm2 (10-13) |
| 5           | 5   | C4_Energy (46-50) | O2_Energy (10-13) | F3_Prb_Norm2 (36-41) | F3_Sbeta/Sdelta | Eog2_Prb_Norm1 (41-46) | O1_Energy (4-8) | F3_Prb_Norm2 (30-36) | F4_Prb_Norm2 (18-25) | C4_Prb_Norm2 (13-18) | Eog1_Prb_Norm2 (10-13) |
| 6           | 6   | C3_Energy (36-41) | O1_Energy (10-13) | F4_Ebeta/Edelta | F3_Ebeta/Edelta | Eog1_Prb_Norm2 (46-50) | O2_Pdelta | O1_Energy (0.4-4) | F4_Prb_Norm2 (25-30) | C4_Prb_Norm2 (18-25) | C4_Prb_Norm2 (13-18) |
| 7           | 7   | EnvlpMean | Eog1_Energy (8-10) | F3_Prb_Norm2 (41-46) | F3_Prb_Norm2 (18-25) | Eog1_Prb_Norm2 (36-41) | O1_Pdelta | C3_Pdelta | C3_Prb_Norm2 (18-25) | F3_Prb_Norm2 (18-25) | Eog1_Ebeta/Edelta |
| 8           | 8   | F4_Energy (41-46) | O2_Energy (10-10) | F4_Prb_Norm2 (36-41) | C4_Prb_Norm2 (18-25) | Eog1_Prb_Norm2 (30-36) | F3_Pdelta | F3_Prb_Norm2 (18-25) | F3_Ebeta/Edelta | O2_Energy (0.4-4) | Eog1_Ebeta/Epsilon |
| 9           | 9   | C4_Pbeta | Eog2_Energy (8-10) | F4_Prb_Norm2 (41-46) | F4_Prb_Norm1 (18-25) | Eog1_Prb_Norm2 (36-41) | F4_Edelta | C3_Pdelta | C3_Prb_Norm2 (18-25) | C4_Prb_Norm2 (18-25) | Eog1_Ebeta/Epsilon |
| 10          | 10  | O1_Energy (30-36) | C3_Palpha | F4_Prb_Norm2 (46-50) | F3_Prb_Norm1 (18-25) | Eog2_Prb_Norm1 (36-41) | O2_Energy (0.4-4) | F4_Pdelta | C4_Prb_Norm2 (18-25) | C3_Prb_Norm1 (18-25) | Eog1_Ebeta/Epsilon |
| 11          | 11  | C4_Energy (36-41) | O2_Pbeta | F3_Prb_Norm2 (30-36) | C4_Prb_Norm2 (18-25) | Eog2_Prb_Norm2 (36-41) | O2_Pbeta | C3_Prb_Norm2 (18-25) | C4_Prb_Norm2 (36-41) | C4_Ebeta/Esum | Eog1_Ebeta/Epsilon |
| 12          | 12  | F4_Energy (46-50) | Eog2_Energy (30-36) | C4_Sbeta/Sum | C4_Sbeta/Sum | Eog2_Prb_Norm1 (30-36) | F4_Pbeta | C3_Palpha | F3_Prb_Norm2 (18-25) | F4_Ebeta/Sum | F3_Prb_Norm2 (13-18) |
| 13          | 13  | F4_Energy (36-41) | C4_Palpha | F4_Prb_Norm2 (30-36) | F4_Prb_Norm2 (18-25) | Eog1_Prb_Norm2 (25-30) | C4_Energy (10-13) | O1_Pbeta | C3_Palpha | C3_Pbeta | Eog1_Ebeta/Esum |
| 14          | 14  | C3_Pbeta | F3_Energy (8-10) | C3_Pbeta | F3_Ebeta/Edelta | C4_Energy(46-50) | C3_Energy (10-13) | C4_Prb_Norm2 (18-25) | O1_Pbeta | C3_Pbeta | Eog1_Ebeta/Esum |
| 15          | 15  | O1_Pbeta | Eog1_Energy (10-13) | C3_Prb_Norm2 (41-46) | C4_Sbeta/Sum | Eog2_Prb_Norm1 (46-50) | O1_Pbeta | C3_Pbeta | F3_Prb_Norm2 (18-25) | C3_Ebeta/Esum | Eog1_Energy(0.4-4) |
Table 9. Comparison of performance with normalized or original features by LOOCV on subgroup III of the dataset ISRUC-Sleep using two-channel EOG and one-channel EMG.

| Feature                      | n1 (%) | n3 (%) | Accuracy (%) | Kappa | Balanced F-score | Awake F-score | REM F-score | N1 F-score | N2 F-score | N3 F-score |
|------------------------------|--------|--------|--------------|-------|------------------|---------------|-------------|------------|------------|------------|
| Normalized by proposed method (this paper) | 10     | 90     | 77.7         | 0.709 | 0.739            | 0.851         | 0.809       | 0.412      | 0.762      | 0.862      |
| 5                            | 95     | 77.5   | 0.706        | 0.734 | 0.845            | 0.805         | 0.413       | 0.761      | 0.863      |
| 15                           | 85     | 77.3   | 0.703        | 0.733 | 0.850            | 0.810         | 0.392       | 0.758      | 0.855      |
| 25                           | 75     | 77.1   | 0.701        | 0.732 | 0.849            | 0.806         | 0.392       | 0.757      | 0.855      |
| 20                           | 80     | 77.1   | 0.700        | 0.734 | 0.848            | 0.799         | 0.413       | 0.754      | 0.857      |
| Original feature             | —      | —      | 77.0         | 0.702 | 0.741            | 0.834         | 0.820       | 0.462      | 0.754      | 0.836      |
| normalized by the traditional method | —      | —      | 74.5         | 0.665 | 0.693            | 0.822         | 0.782       | 0.281      | 0.735      | 0.843      |
eyes, and is a fundamental indicator for distinguishing between REM and NREM stages (Sun et al 2019). As shown in figure 3(b), $PL_{org}$ of EOG is usually greater than 0 during N2 and N3 sleep and it is usually less than 0 during wakefulness and REM sleep. Visual examples are shown in figures 9 and 10. During REM sleep in figure 9, the amplitude polarity between left EOG and right EOG is opposite, i.e. when a peak shows in the left EOG, a valley will show in the right EOG, such as the position at 12 s.

During N3 sleep in figure 10, the amplitude polarity between the left EOG and the right EOG is consistent, i.e. when a peak shows in the left EOG, a similar peak will also show in the right EOG, such as the position at 5 s. Furthermore, the waveforms in the EOG are to some extent similar to those of the EEG, which is obvious in figure 10.

Compared with the microvolt values of EEG’s small-amplitude variations, EOG and EMG values are in millivolts and require less stable contact with the body (Sun et al 2019), which make them less sensitive to noise and more suitable for unobtrusive measurement apparatus. Hence, the combination of EOG and EMG is a good choice for sleep monitoring.

4.3. Sleep monitoring by the combination of EOG and EMG

EOG makes the main contribution to sleep scoring derived from the combination of EOG and EMG. The placements of EOG electrodes are close to the placements of Fp1 and Fp2 EEG electrodes (Sun et al 2019). Hence, the EOG recordings are highly influenced by a portion of EEG activity (Sun et al 2019). During NREM sleep, EOG is similar in waveform to the EEG signals that are recorded at the frontal poles (Sun et al 2019).

Using 114 normalized features from a combination of EOG (108 features) and EMG (6 features), Cohen’s kappa coefficient by RF from LOOCV ($N = 124$) was 0.749 and the accuracy was 80.8%. The F1-scores were 0.857, 0.855, 0.483, 0.790 and 0.851 for wakefulness, REM sleep, N1 sleep, N2 sleep and N3 sleep, respectively.

In addition, using 438 normalized features from the combination of EEG (324 features), EOG (108 features) and EMG (6 features), Cohen’s kappa coefficient by RF from LOOCV ($N = 124$) was 0.801 and the accuracy was 84.7%. The F1-scores were 0.894, 0.884, 0.528, 0.836 and 0.894 for wakefulness, REM sleep, N1 sleep, N2 sleep and N3 sleep, respectively. Consequently, the sleep classification performance achieved by EOG and EMG is comparable to that of EEG, EOG and EMG.
4.4. Limitation

ISRUC-Sleep is an extremely unbalanced dataset, including 116 records from subjects with sleep disorders and only 10 records from healthy subjects. This research used 114 records from sleep disorder subjects and all 10 records from healthy subjects. Therefore sleep disorder records make the main contribution to the classification of sleep stages. Whether sleep scoring models derived from a healthy group can be applied to a sleep disorder group or not is still unknown, and vice versa. Balanced numbers in sleep disorder subjects and healthy subjects would be very important to answer that question. In the future, it will be necessary to study the classification effects on a balanced dataset from sleep disorder and healthy subjects, respectively.

According to the percentage of wakefulness and REM stage based on visual scoring from two experts in table 3, the healthy subjects did not have normal sleep but disturbed sleep. It holds true that PSG monitoring requires the attachment of many electrodes to the head, face and body, therefore it seriously interferes with a subject’s natural sleep. Therefore bad-quality sleep records make the main contribution to the classification of sleep stages, not only from sleep disorder subjects but also from healthy subjects.

The third limitation is that there is no process for feature selection in this paper. Also, the best combination of features with as few features as possible was not found either.

5. Conclusion

On a public dataset called ISRUC-Sleep, comparative analysis suggests that the sleep classification performance achieved by EOG and EMG is comparable to that of EEG, EOG and EMG. EOG makes the main contribution to sleep scoring derived from the combination of EOG and EMG. The proposed method from the combination of EOG and EMG can achieve comparable performance to the use of EEG signals for sleep staging.

Acknowledgments

This study was funded by State Key Laboratory of Space Medicine Fundamentals and Application, China Astronaut Research and Training Center (SMFA15B06, SMFA15A01). We acknowledge the support from the public dataset ISRUC-Sleep, provided by the Sleep Medicine Centre of Coimbra.
Conflict of interest

The authors have no financial interest in any related entities.

Ethical approval

All data used in this paper were from the public dataset ISRUC-Sleep website http://sleeptight.isr.uc.pt/ISRUC_Sleep/. According to the ISRUC-Sleep statement all patients referred were submitted to an initial briefing with the support of an informed consent document. The ethics committee of the Central Hospital of University of Coimbra (CHUC) approved the use of the data of the referred patients as anonymous for the research purposes.

ORCID iDs

Yanjun Li @ https://orcid.org/0000-0001-9732-5175

References

Beattie Z et al 2017 Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals Physiol. Meas. 38 1968–79

Garcia-Molina G and Jiang J W 2022 Interbeat interval-based sleep staging: work in progress toward real-time implementation Physiol. Meas. 43 025004

Ghasemzadeh P, Kalbkhani H and Shayesteh M G 2019 Sleep stages classification from EEG signal based on Stockwell transform IET Signal Proc. 13 242–52

Iber C et al 2007 The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (Westchester, IL: Westbrook Corporate Center AASM) pp 19–37

Khalighi S et al 2016 ISRUC-Sleep: a comprehensive public dataset for sleep researchers Comput. Methods Programs Biomed. 124 180–92

Khalighi S, Sousa T, Fires G and Nunes U 2015 Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels Expert Syst. Appl. 40 7046–59

Kuo C E and Chen G T 2020 Automatic sleep staging based on a hybrid stacked LSTM neural network: verification using large-scale dataset IEEE Access 11

Li Y J, Tang X Y, Xu Z, Liu W F and Li J 2016 Temporal correlation between two channels EEG of bipolar lead in the head midline is associated with sleep-wake states Australas. Phys. Eng. Sci. Med. 39 147–55

Matar G, Lina J M, Carrier J and Kaddoum G 2018 Unobtrusive sleep monitoring using cardiac, breathing and movements activities: an exhaustive review IEEE Access (https://doi.org/10.1109/ACCESS.2018.2865487)

Mendonça F, Mostafa S S and Morgado-Dias F 2019 A review of approaches for sleep quality analysis IEEE Access (https://doi.org/10.1109/ACCESS.2019.2900345)

Myczak M, Valdez T A and Kukwa W 2020 Joint apnea and body position analysis for home sleep studies using a wireless audio and motion sensor IEEE Access (https://doi.org/10.1109/ACCESS.2020.3024122)

Nakamura T, Goverdovsky V, Morrell M J and Mandic D P 2017 Automatic sleep monitoring using ear-EEG IEEE Access (https://doi.org/10.1109/ITEHM.2017.2702558)

Pan H, Xu Z and Yan H 2018 Lying position classification based on ECG waveform and random forest during sleep in healthy people Biomed. Eng. Online 17 116

Phan H, Andreotti F and Cooray N 2019 Joint classification and prediction CNN framework for automatic sleep stage classification IEEE Trans. Biomed. Eng. 66 1285–96

Rahman M M, Imazum M, Bhuiyan H and Hassan A R 2018 Sleep stage classification using single-channel EOG Comput. Biol. Med. 102 211–20

Rosenberg R S and Van-Hout S 2013 The American Academy of Sleep Medicine inter-rater reliability program: sleep stage scoring J. Clin. Sleep Med. 9 811–7

Shen H, Ran F, Xu M, Guez A, Li A and Guo A 2020 An automatic sleep stage classification algorithm using improved model based essence features Sensors 20 4677

Shi P, Zheng X, Du P and Yuan F 2019 Automatic sleep stage classification based on LSTM ed Y Sun et al Computer Supported Cooperative Work and Social Computing. Chinese CSCW 2018. Communications in Computer and Information Science 917 (Singapore: Springer) pp 478–86

Sun C, Chen C, Fan J, Li W, Zhang Y and Chen W 2019 A hierarchical sequential neural network with feature fusion for sleep staging based on EOG and RR signals J. Neural Eng. 16 1–17

Yan R, Li F, Zhou D D, Ristaniemi T and Cong F 2021 Automatic sleep scoring: a deep learning architecture for multi-modality time series J. Neurosci. Methods 348 108971