IOP Publishing Physiol. Meas. 39 (2018) 115005 (13pp) https://doi.org/10.1088/1361-6579/aae943

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

Efficient sleep classification based on entropy features and a support vector machine classifier

Zhimin Zhang[1,2], Shoushui Wei[1,4], Guohun Zhu[2], Feifei Liu[1], Yuwen Li[2], Xiaotong Dong[1], Chengyu Liu[3,4] and Feng Liu[2,4]

[1] School of Control Science and Engineering, Shandong University, Jinan, People's Republic of China
[2] School of Information Technology and Electrical Engineering, University of Queensland, Queensland, Australia
[3] School of Instrument Science and Engineering, Southeast University, Nanjing, People's Republic of China
[4] Author to whom any correspondence should be addressed.

E-mail: sswei@sdu.edu.cn, chengyu@seu.edu.cn and feng@itee.uq.edu.au

Keywords: sleep stages classification, EEG, EOG, entropy, SVM

RECEIVED
18 May 2018

REVISED
10 October 2018

ACCEPTED FOR PUBLICATION
18 October 2018

PUBLISHED
26 November 2018

Abstract

Objective: Sleep quality helps to reflect on the physical and mental condition, and efficient sleep stage scoring promises considerable advantages to health care. The aim of this study is to propose a simple and efficient **s**leep **c**lassification method based on **en**tropy features and a support vector machine classifier, named SC-En&SVM. *Approach*: Entropy features, including fuzzy measure entropy (FuzzMEn), fuzzy entropy, and sample entropy are applied for the analysis and classification of sleep stages. FuzzyMEn has been used for heart rate variability analysis since it was proposed, while this is the first time it has been used for sleep scoring. The three features are extracted from 6 376 730 s epochs from Fpz-Cz electroencephalogram (EEG), Pz-Oz EEG and horizontal electrooculogram (EOG) signals in the sleep-EDF database. The independent samples *t*-test shows that the entropy values have significant differences among six sleep stages. The multi-class support vector machine (SVM) with a one-against-all class approach is utilized in this specific application for the first time. We perform 10-fold cross-validation as well as leave-one-subject-out cross-validation for 61 subjects to test the effectiveness and reliability of SC-En&SVM. *Main results*: The 10-fold cross-validation shows an effective performance with high stability of SC-En&SVM. The average accuracy and standard deviation for 2–6 states are 97.02 ± 0.58, 92.74 ± 1.32, 89.08 ± 0.90, 86.02 ± 1.06 and 83.94 ± 1.61, respectively. While for a more practical evaluation, the independent scheme is further performed, and the results show that our method achieved similar or slightly better average accuracies for 2–6 states of 94.15%, 85.06%, 80.96%, 78.68% and 75.98% compared with state-of-the-art methods. The corresponding kappa coefficients (0.81, 0.74, 0.72, 0.71, 0.67) guarantee substantial agreement of the classification. *Significance*: We propose a novel sleep stage scoring method, SC-En&SVM, with easily accessible features and a simple classification algorithm, without reducing the classification performance compared with other approaches.

1. Introduction

Sleep is a complex amalgam of physiologic processes and it can reflect the quality of the physical and mental condition. With the increasing pressures of modern life, sleep apnea, insomnia and narcolepsy are gradually raising people's concerns and efficiently identifying sleep stages greatly helps analyze and monitor sleep quality. Polysomnograms (PSGs) (Zhang and Wu 2017) are generally utilized to diagnose sleep disorders by experts, and sleep stage classification is traditionally performed based on visual interpretation of PSGs according to Rechtschaffen's and Kales's (R&K) recommendations (Rechtschaffen and Kales 1968) or a new guideline developed by the American Academy of Sleep Medicine (AASM) (Iber 2007). In this study, R&K is applied as the standard for the six sleep stages, including wake (W), non-rapid eye movement (NREM) 1 (S1), NREM 2 (S2), NREM 3 (S3), NREM 4 (S4) and rapid eye movement (REM).

The manual scoring by experts is always time consuming and error prone. Thus, an efficient automatic classification method is of great necessity since sleep wearable devices have developed greatly in recent years (Bianchi 2017). Different methods have been proposed for sleep stage classification (Oral *et al* 2017, Qureshi and Vanichayobon 2017, Karimzadeh *et al* 2018). Most of these methods are based on time or frequency domain features from EEG, EOG and electromyogram (EMG) signals. Spectral and non-linear candidate features were extracted from EEG and EMG signals, and Chapotot *et al* proposed a novel classification framework by selecting robust candidate features, emulating artificial neural network classifiers, and assigning sleep–wake stages based on flexible decision rules (Chapotot and Becq 2010). Krakovská and Mezeiová achieved a classification accuracy of 74% for 5-state classification using features extracted from different frequency bands of EEG, the power of EMG and variances of EOG (Krakovská and Mezeiová 2011). Goldberger *et al* combined entropy and spectral edge frequency features from a single EEG channel, and they were able to reach with an accuracy of 93.8% for five sleep stages by a multi-class SVM (Goldberger *et al* 2000), with a sensitivity of 49.1% for the S1 stage (Nakamura *et al* 2017). Charbonnier *et al* proposed an automatic sleep–wake stages classifier that dealt with the presence of artifacts and they claimed 85.5% of overall accuracy with an improved ability to distinguish the S1 stage from REM (Charbonnier *et al* 2011). Aboalayon *et al* proposed a new framework that can be implemented in an embedded hardware device for sleep scoring based on statistical features applied to single-channel EEG signals, and, using a decision tree as a classifier, they claimed an average classification sensitivity, specificity and accuracy of 89.06%, 98.61% and 93.13%, respectively (Aboalayon *et al* 2014). Zhu *et al* applied different visibility graphs (VGs) to study sleep EEG signals. They identified significant differences in mean degrees between different VGs and horizontal VGs associated with single-channel EEG signals and the accuracy and kappa coefficients of the 6-state classification were 87.5% and 0.81, respectively (Zhu *et al* 2014).

This paper proposes a novel sleep stage classification method based on entropy features and an SVM classifier, called SC-En&SVM, by employing fuzzy entropy (FuzzyEn), fuzzy measure entropy (FuzzyMEn) and sample entropy (SampEn) from EEG and EOG signals and performing classification by multi-class SVM using a one-against-all class approach. A statistical independent samples t-test shows that the applied entropy features of each sleep stage have significant differences. The 2–6 states sleep stage classifications are given based on the leave-one-subject-out cross-validation for 61 individual subjects and 10-fold cross-validation for non-independent training and testing. The overall average accuracy and the corresponding Cohen's kappa coefficient and per-class classification performance metrics are used to evaluate our method. The results show that SC-En&SVM has a similar or slightly better performance than the existing sleep scoring methods.

The remaining sections of this paper are organized as follows: the introduction of the experimental data is given in section 2 and the methodology is briefly presented in section 3. In section 4, the detailed experiments and results are presented for the leave-one-subject-out cross-validation and 10-fold cross-validation of the non-independent training and testing set. A discussion is given in section 5 by comparing the SC-En&SVM with state-of-the-art published methods and we also discuss the limitations of the SC-En&SVM. Finally, the conclusions are presented in section 6.

2. Experimental data

The experimental data used in this paper are obtained from the sleep-EDF database (Goldberger *et al* 2000), which contains 61 recordings from two studies: (1) the SC* files (SC = sleep cassette) were obtained in a 1987–1991 study of age effects on sleep for healthy Caucasians aged 25–101 without any sleep-related medication; (2) the ST* files (ST = sleep telemetry) were obtained in a 1994 study of temazepam effects on sleep for Caucasian males and females with mild difficulty falling asleep. For the SC* files, PSGs of about 20 h are recorded during two subsequent day–night periods at the subjects' homes, while the PSGs of about 9 h, for the ST* files, are recorded in hospital over two nights, one of which is after temazepam intake (Mourtazaev 1995, Kemp *et al* 2000). However, it is worth noting that the ST* recordings included in this paper were recorded during the night with no temazepam intake. Each SC* recording contains seven signals while each ST* recording contains five signals. The detailed information from the database is presented in table 1.

In this study, Fpz-Cz EEG, Pz-Oz EEG and horizontal EOG signals are selected to analyze and identify the sleep stages; these three channels are sampled at 100 Hz, providing detailed information for each sleep stage (Liang *et al* 2012, Ronzhina *et al* 2012). These stages are characterized by their spectral contents, patterns and duration. The wake stage is characterized by alpha or faster frequency rhythms, along with frequent eye movements and high EMG amplitudes. S1 corresponds to the intervals at which the alpha frequency (8–13 Hz) occupies less than 50% of the epochs and vertex waves are evident. S2 refers to an epoch in which sleep spindles and K-complexes exist. Stage 3 is scored when there are low-frequency waves with a frequency less than 2 Hz; sleep spindles and K-complexes may also occur. The deepest sleep stage, S4, is characterized by the existence of delta (<4 Hz) waves in more than 50% of the epochs. Finally, the REM stage is characterized by saw-tooth waves,

Table 1. Detailed information from the database used in this study.

Series	Recordings	No. of PSGs	Recording duration	Channels	Sampling frequency
SC* recordings	SC4ssnE0 $00 \leqslant ss^{\#1} \leqslant 19$ $n^{\#2} = 1$ or 2 (with some lost)	39	Day–night (containing a large number of pre-sleep and post-sleep wake epochs)	Fpz-Cz EEG Pz-Oz EEG Horizontal EOG	100 Hz
				EMG Body temperature Respiration Event maker	1 Hz
ST* recordings	ST7ssnJ0 $01 \leqslant ss^{\#1} \leqslant 24$ $n^{\#2} = 1$ or 2 (with some lost)	22	Overnight (most epochs are of sleep stages)	Fpz-Cz EEG Pz-Oz EEG Horizontal EOG Submental EMG	100 Hz
				Event maker	1 Hz

Note: $^{\#1}$ denotes the subject number. $^{\#2}$ denotes the night number. Both the SC* and ST* files are recorded for two separate nights. For example, SC4001E0 is the PSG of 00 subject recorded on the first night, while SC4012E0 is for 01 subject on the second night.

saccadic eye movements and very low EMG amplitudes, and it is similar to the wake stage, but REM has lower amplitude alpha activity (Karimzadeh *et al* 2018). Figure 1 illustrates an example of 30 s epoch Fpz-Cz EEG and EOG signals of six sleep stages from sc4002e0. Additionally, the 39 SC* recordings are recorded in high quality without any failed leads. In contrast, the small number of failed signals in the ST* series are firstly detected and discarded before further processing.

All recordings are stored in EDF format, each PSG is scored by well-trained technicians based on the R&K manual (Rechtschaffen and Kales 1968) and the hypnograms consist of the sleep stages W, REM, S1, S2, S3, S4, M (movement time) and '?' (not scored). As the M and '?' segments (mostly at the start and end of each recording) do not belong to the five sleep stages, we filter and remove these two stages (90 epochs overall for 61 subjects) before further processing (Iber 2007). Also, there are long periods of wake stages at the start and the end of each recording for the SC* recordings; thus, to balance the original data, we only include 30 min of such periods just before and after the sleep periods. Afterwards, a 4th-order Butterworth filter with a passband from 0.5 to 30 Hz is applied to remove the linear trends and baseline drift (Boostani *et al* 2017). Because the hypnogram is generated by the experts every 30 s, the data is divided into epochs every 30 s, and each epoch contains 3000 data points. We evaluate our proposed method using the Fpz-Cz, Pz-Oz and EOG channels without any further preprocessing, and a total of 6 376 730 s epochs are selected in this paper. Table 2 summarizes the number of epochs and its corresponding percentage for each stage for 61 subjects overall.

3. Methodology

Entropy is an information measure which determines the uniformity of proportion distribution (Inouye *et al* 1991). For a time series, the greater the probability of generating new patterns means the greater its irregularity, and entropy is widely used to evaluate the degree of randomness (or inversely, the degree of orderliness) of a time series (Costa *et al* 2005). In other words, entropy can measure the irregularity of a signal, and the higher the irregularity of a signal the greater its corresponding entropy will be. For example, the entropy of the EEG or EOG signal (Hassan and Subasi 2017) recorded when the subject is awake would be greater than that recorded when the subject is asleep because the brain and the eyeball are active and excited during the waking period. When the sleep turns into non-REM, the brain becomes inactive and its response to the external stimuli decreases as the sleep turns deeper, and accordingly, the entropy of the EEG and EOG signals would decrease correspondingly. Therefore, the idea of applying entropy as features for classification is theoretically possible, and several papers have reported their research on sleep based on entropy features (Popovic *et al* 2014, Rodríguezsotelo *et al* 2014, Chen *et al* 2015, Lajnef *et al* 2015).

3.1. Entropy features

Considering an EEG or EOG sequence $x = \{x(i) : 1 \leqslant i \leqslant N\}$ of length N (30 s, 3000 points), the fuzzy entropy, fuzzy measure entropy and sample entropy are calculated in this study.

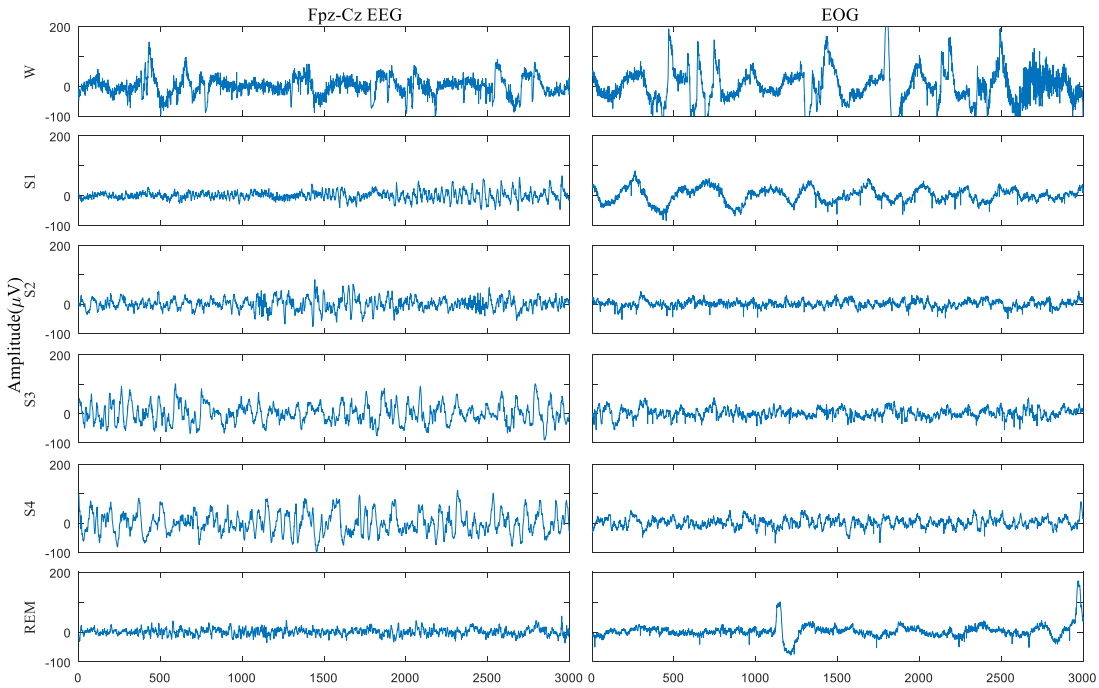

Figure 1. An example of 30 s epoch Fpz-Cz EEG and EOG signals of various sleep stages from sc4002e0.

Table 2. The number of epochs for each stage for 61 subjects.

Stages	W	S1	S2	S3	S4	REM	Total
Number of epochs	10 931 (17.14%)	4848 (7.60%)	27 292 (42.80%)	5075 (7.96%)	3773 (5.92%)	11 848 (18.58%)	63 767

3.1.1. *Fuzzy entropy*

Fuzzy entropy (FuzzyEn) is the entropy of a fuzzy set, representing the information of uncertainty for a series (Al-Sharhan *et al* 2001). Due to the fact that FuzzyEn contains vague and ambiguous un-certainties, it is quite different from the classical Shannon entropy because no probabilistic concept is needed to define it, while the Shannon entropy contains the randomness uncertainty (probabilistic) (Al-Sharhan *et al* 2001). The FuzzyEn is defined using the concept of membership degree and its calculation procedure is explained in Chen *et al* (2007). The FuzzyEn statistic, the given input value for embedding dimension m, fuzzy power n, and tolerance threshold r, is defined as

$$\mathrm{FuzzyEn}(m, n, r, N) = \ln \phi^m(n, r) - \ln \phi^{m+1}(n, r) \tag{1}$$

where

$$\phi^m(n, r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \left(\frac{1}{N-m-1} \sum_{j=1, j\neq i}^{N-m} (\exp(-(d_{ij}^m)^n / r)) \right) \tag{2}$$

in which $d_{ij}^m = d[X_i^m, X_j^m] = \max_{k\in(0,m-1)} \{|x(i+k) - x_0(i) - (x(j+k) - x_0(j))|\}$, indicating the maximum distance between two local sequence segments X_i^m and X_j^m, $X_i^m = \{x(i), x(i+1), \ldots, x(i+m-1)\} - x_0(i)$, $1 \leqslant i, j \leqslant N-m$. While $x_0(i)$ indicates the mean value of the sequence $\{x(i), x(i+1), \ldots, x(i+m-1)\}$.

3.1.2. *Fuzzy measure entropy*

Fuzzy measure entropy (FuzzyMEn) has been used for heart rate variability analysis in (Liu *et al* 2013), while this is the first time it has been used from EEG and EOG for sleep stage classification. This represents an evolution of FuzzyEn and is devised to integrate both local and global characteristics and could reflect the entire irregularity of a time series. FuzzyMEn is calculated based on the fuzzy set theory and constructed with the membership degree of a fuzzy function instead of using the '0–1' judgment of the Heaviside function that is typically used in the approximate entropy and sample entropy; it can improve the poor statistical stability in the approximate entropy and sample entropy. The algorithm of FuzzyMEn is inspired by the study of Chen *et al* (2007), in which the fuzzy sets are introduced to improve the statistical stability. The detailed description of the calculating procedure is given by Liu *et al* (2013). Similarly to FuzzyEn, FuzzyMEn is defined as

$$\text{FuzzyMEn}(m,n,r,N) = \text{FuzzyLMEn}(m,n,r,N) + \text{FuzzyFMEn}(m,n,r,N) \tag{3}$$

in which $\text{FuzzyLMEn}(m,n,r,N) = \text{FuzzyEn}(m,n,r,N)$ indicates the local fuzzy measure entropy while $\text{FuzzyFMEn}(m,n,r,N)$ represents the global fuzzy measure entropy. FuzzyFMEn is computed in a similar fashion to FuzzyLMEn:

$$\text{FuzzyFMEn}(m,n,r,N) = \ln \phi_F^m(n,r) - \ln \phi_F^{m+1}(n,r) \tag{4}$$

where

$$\phi_F^m(n,r) = \frac{1}{N-m}\sum_{i=1}^{N-m}\left(\frac{1}{N-m-1}\sum_{j=1,j\neq i}^{N-m}(\exp(-(dF_{ij}^m)^n/r))\right) \tag{5}$$

in which $dF_{ij}^m = d[XF_i^m, XF_j^m]$, indicating the maximum distance between two global sequence segments XF_i^m and XF_j^m, $XF_i^m = \{x(i), x(i+1), \ldots, x(i+m-1)\} - \bar{x}, 1 \leqslant i, j \leqslant N-m$, in which $\bar{x}$ is the mean value of the entire sequence $x = \{x(i) : 1 \leqslant i \leqslant N\}$.

3.1.3. Sample entropy

The reader is referred to Ge *et al* (2007) for the computing procedure. For the N length sequence x, given the parameters m and r, the sample entropy (SampEn) can be calculated as follows.

Considering the $N-m+1$ sequence segments $X^m(i) = \{x(i), x(i+1), \ldots, x(i+m-1)\}$ for $\{i | 1 \leqslant i \leqslant N-m+1\}$, B_i counts the number of pairs with distances smaller than r between $X^m(i)$ and $X^m(j)$, where $i \neq j$; and similarly A_i is the number of pairs with distances smaller than r between $X^{m+1}(i)$ and $X^{m+1}(j)$, where $i \neq j$.

Then, the probability that the distance between $X^m(i)$ and $X^m(j)$ is smaller than r is given by $B_i^m(r) = B_i/N-m-1$, and the density is $B^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m(r)$. Similarly, $A_i^m(r)$ and $A^m(r)$ can also be calculated in the same fashion. Then, the number of templates that matches in an m-dimensional (or $(m+1)$-dimensional) phase space within the tolerance r can be calculated as

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

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

Finally, SampEn is defined as

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

Therefore, SampEn calculates the negative natural logarithm of the conditional probability that an EEG or EOG sequence x of length N, having repeated itself for m samples within the tolerancer r, will also repeat itself for $m+1$ samples, with no self-matches (Richman and Moorman 2000).

3.2. One-against-all SVM

For classification, the multi-class SVM is applied in this paper for its simplicity and robustness, and it has been widely used in sleep stage classification by other researchers (Adnane *et al* 2012, Zhu *et al* 2012, 2014, Surantha *et al* 2017). We have tried several types of SVMs, and finally the multi-class SVM using the one-against-all class approach is employed for its simplicity and good performance. The main idea of the one-against-all class approach is simple. As it is designed for binary classification, we need to reconstruct the multi-class classifier, and the idea of the one-against-all class approach comes from the one-against-one class approach used by Nakamura *et al* (2017). For a multi-class classification task, we reconstruct the multi-class classifier by connecting several binary classifiers in series, and each binary classifier only solves the classification of one class against all the others, and the next binary classifier does the same thing, ignoring the classified class by the previous classifier.

In this paper, the Gaussian radial basis function (RBF) (Chriskos *et al* 2017) is selected as the kernel function, as it is shown to perform better than linear or polynomial kernels (Bsoul *et al* 2011) in the context of sleep classification. In the SVM training phase, tuning of the parameters is necessary for a better classification performance, including choosing the box constraint (C) and γ. Specifically, C is a tradeoff parameter between regularization and accuracy, which influences the behavior of the support vector selection, and γ is an important factor to control the RBF kernel in transmitting data to a new hyperspace. Optimization of the hyperparameters is investigated by following a 10-fold cross-validation that divides the training set into ten subgroups and iteratively minimizes the error using nine training groups and testing against the remaining subgroup. We implement a grid-search by scanning over the range [0.10:0.01:10] for both parameters, and the best C and γ are found at 2.97 and 0.74, respectively. These two selected hyperparameters are then used for the leave-one-subject-out cross-validation.

3.3. Performance evaluation

We evaluated the performance of the proposed method using overall accuracy (ACC), Cohen's kappa coefficient (κ), per-class precision (PR), and per-class recall (RE). With chance agreement removed, κ determines the agreement between scorers. The κ values between 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00 correspond to slight, fair, moderate, substantial and almost perfect agreement, respectively. The per-class metrics are computed by considering a single class as a positive class and all other classes combined as a negative one. The ACC, PR, RE and κ are calculated as follows:

$$\mathrm{ACC} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}, \mathrm{PR} = \frac{T_P}{T_P + F_P}, \mathrm{RE} = \frac{T_P}{T_P + F_N}, \kappa = \frac{P_o - P_e}{1 - P_e},$$

where T_P is the number of positive class epochs classified correctly, T_N is the number of negative class epochs identified correctly, F_P is the number of negative class epochs classified incorrectly as a positive class, F_N is the number of positive class epochs identified incorrectly as a negative class, P_o is the observed agreement ratio and P_e is the chance agreement probability (Sors *et al* 2018). Since we are dealing with a multi-class classification problem in this study, we calculate the class-specific PR and class-specific RE, as shown in tables 4 and 6.

4. Experiments and results

To evaluate the classification performance of the proposed method SC-En&SVM, a set of experiments are conducted using MATLAB 2015a on a Dell computer with a 3.40 GHz Intel Core i7-2600 CPU and 16.0 GB RAM. This section consists of three main parts: entropy features and an independent samples *t*-test, independent training and testing, and non-independent training and testing. The experiments and results of each part will be given as follows.

4.1. Entropy features extraction

As discussed in section 2, the data is divided into epochs every 30 s, and there are 63 767 epochs in total for all 61 subjects. Following the data preprocessing, FuzzyEn, FuzzyMEn and SampEn are calculated for each epoch of the three channels. For parameter selection, the embedding dimension m is chosen as the default value 2, and different choices of tolerance threshold $r = r' * \mathrm{SD}$ (SD is the standard deviation of each 30 s epoch) and fuzzy power n influence the standard deviation of the calculated entropy. It is worth noting that n and r are the gradient and width of the boundary of the fuzzy function $\mu(d_{ij}^m, n, r)$, respectively. We test different r' values from 0.05 to 0.2 and different n values from 1 to 5, and for a comprehensive consideration of a steady standard deviation of the calculated entropy and a small computation, we set the tolerance threshold $r = 0.15 * \mathrm{SD}$ and fuzzy power $n = 2$. In addition, the time lag tau is set to the default value 1. The parameter selection in this study refers to the suggestions in Azami *et al* (2017) and Richman and Moorman (2000).

The box plots of the three entropies is given in figure 2, from which it is clearly shown that the entropy values change as the sleep grows deeper and the entropy of the stage W is generally higher than all other sleep stages. The FuzzyEn and SampEn of the three channels show a similar trend in that the entropy values decrease slightly as the sleep grows deeper from wake to REM. However, the values of FuzzyMEn are relatively smaller and the REM stage is higher than the others for both EEG channels. In contrast, the FuzzyMEn of the EOG channel shows a different trend in that the value of the S2 stage is higher than other sleep stages while that of REM is the lowest.

4.2. Independent samples *t*-test

To test the differences in the entropy features, the independent samples *t*-test is performed between every possible stage-pair for three entropy values of the three channels. However, since we are performing 15 stage-pair tests in total, some tests may result in statistically significant differences just by chance. Thus, we compensate for this random effect by performing Bonferroni's corrections, in which a new threshold of 0.0033 is given by dividing 0.05 by 15. Then, it would be more dependable to be considered as significant if the *p*-value is still lower than this new threshold.

The results are given in table 3, in which the *p*-value of most stage-pairs is 0.000 ($p < 0.0033$), meaning significant differences between the corresponding pairs. Meanwhile, it is worth noting that the *p*-value in bold with * annotated indicates the stage-pair that have no significant differences, including the FuzzyMEn of the EOG between W–S4; the FuzzyEn of Pz-Oz between S2–S3, S2–S4 and S3–S4; the FuzzyEn of Fpz-Cz between S1–S3; the SampEn of Fpz-Cz between S1–S2, S1–S3, S1–S4 and S3–S4; and the SampEn of Pz-Oz between S1–S3, S1–S4 and S3–S4. The independent samples *t*-test has shown significant differences between most possible stage-pairs, which establishes the foundation for sleep stage scoring based on entropy features.

4.3. Non-independent training and testing

Non-independent training and testing sets are commonly used in sleep stage scoring, and like Zhu *et al* (2014) and Hassan (2015), we thus first give the performance of the SC-En&SVM for the non-independent scheme.

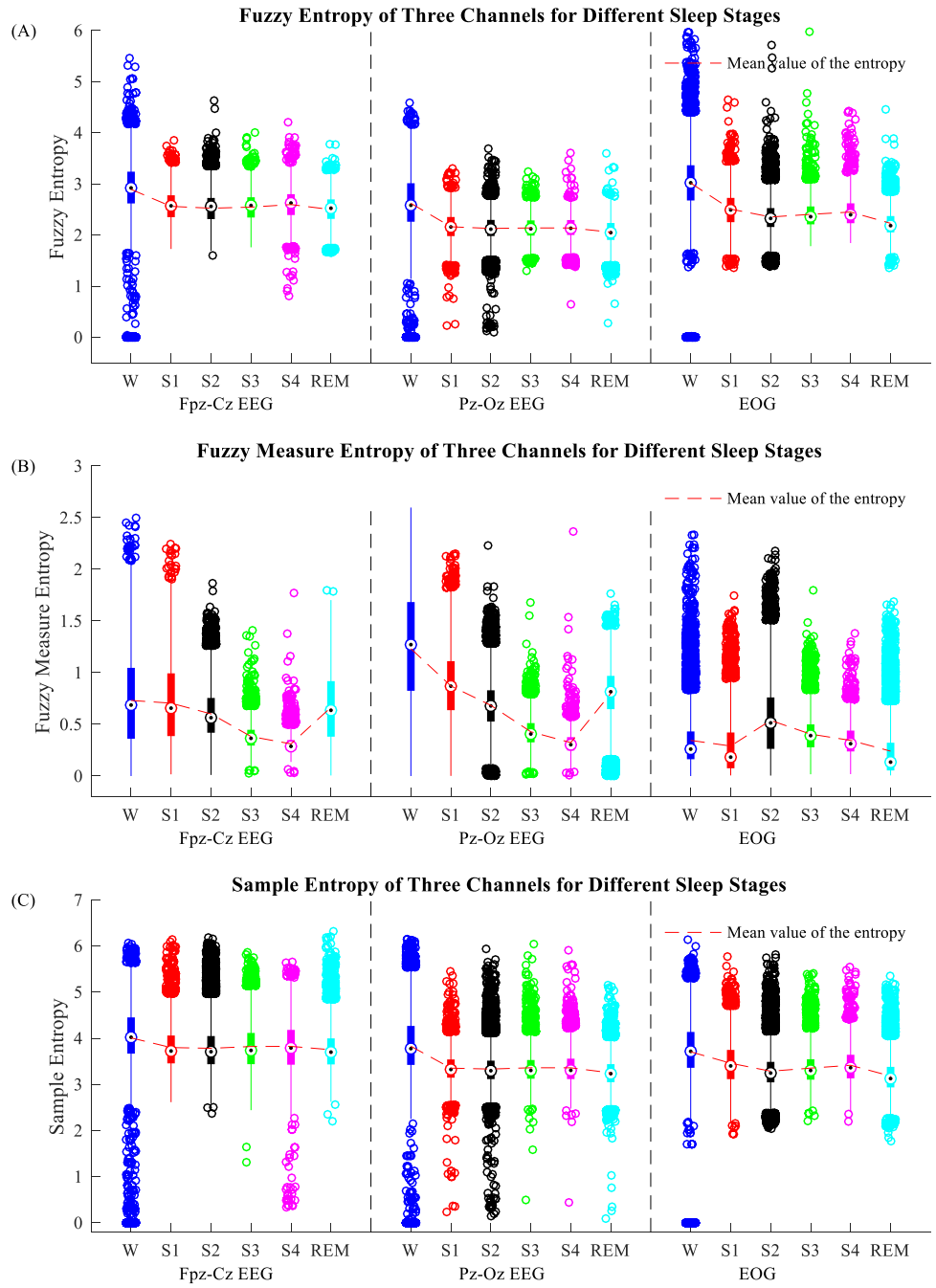

Figure 2. Box plots of the entropy of three channels for different sleep stages. (A) Fuzzy entropy. (B) Fuzzy measure entropy. (C) Sample entropy.

The 10-fold cross-validation is performed for $C = 2 - 6$ with the non-independent training and testing set. In this study, ten experiments are performed for each C, and in each experiment a random 90% of the 63 767 epochs are chosen as the training data and the remaining 10% epochs as the testing data. Table 4 presents the average RE, PR, ACC and κ of ten experiments for each C. It is noted that we only present the standard deviation of ACC and κ ($\pm$standard deviation) in the last two rows, ignoring those of RE and PR for a simple and clear table. It shows that the non-independent scheme results in a good performance; nevertheless, we believe that the independent scheme is more suitable for practical evaluation in clinical diagnosis. Thus, we test our method using these two schemes.

4.4. Independent training and testing

We further evaluate our method using the leave-one-subject-out cross-validation scheme for 61 subjects and the average evaluation criteria, i.e. ACC, PR, RE and κ, are presented to verify the reliability of the proposed SC-En&SVM. Specifically, we perform 61 experiments in total for each classification (C, $C = 2, 3, 4, 5, 6$) so that each recording can be tested, and in each experiment, we select one individual subject as the testing subject and the remaining subjects constitute the training set.

Table 3. The p-values of the independent samples t-test of the entropy features between all possible pairs.

	Fuzzy entropy			Fuzzy measure entropy			Sample entropy		
Pairs	Fpz-Cz	Pz-Oz	EOG	Fpz-Cz	Pz-Oz	EOG	Fpz-Cz	Pz-Oz	EOG
W–S1	.000	.000	.000	.000	.000	.000	.000	.000	.000
W–S2	.000	.000	.000	.000	.000	.000	.000	.000	.000
W–S3	.000	.000	.000	.000	.000	.000	.000	.000	.000
W–S4	.000	.000	.000	.000	.000	**.948***	.000	.000	.000
W–REM	.000	.000	.000	.000	.000	.000	.000	.000	.000
S1–S2	.000	.000	.000	.000	.000	.000	**.030***	.000	.000
S1–S3	**.017***	.000	.000	.000	.000	.000	**.027***	**.231***	.000
S1–S4	.000	.000	.000	.000	.000	.000	**.035***	**.360***	.000
S1–REM	.000	.000	.000	.000	.000	.000	.000	.000	.000
S2–S3	.000	**.006***	.000	.000	.000	.000	.000	.000	.000
S2–S4	.000	**.028***	.000	.000	.000	.000	.000	.000	.000
S2–REM	.000	.000	.000	.000	.000	.000	.000	.000	.000
S3–S4	.000	**.907***	.000	.000	.000	.000	**.811***	**.881***	.000
S3–REM	.000	.000	.000	.000	.000	.000	.000	.000	.000
S4–REM	.000	.000	.000	.000	.000	.000	.000	.000	.000

Note: It is noted that there are large numbers of .000 ($p < 0.0033$) in this table, which is actually expected, indicating the significant differences between the corresponding stage-pairs. Meanwhile, the cells in bold ($p > 0.0033$) indicate the stage-pairs that have no significant differences.

Table 4. The average RE, PR, accuracy and κ of the 10-fold cross-validation.

	$C = 2$ RE(%)/PR(%)	$C = 3$ RE(%)/PR(%)	$C = 4$ RE(%)/PR(%)	$C = 5$ RE(%)/PR(%)	$C = 6$ RE(%)/PR(%)
W	95.31/86.88	96.75/89.84	95.71/89.63	96.63/87.52	95.74/87.52
S1	97.35/99.11	91.61/99.05	85.58/94.75	38.91/47.97	34.61/46.63
S2				89.67/88.84	90.14/88.91
S3			92.67/82.79	87.32/86.43	75.74/63.34
S4					78.45/92.05
REM		93.53/74.25	90.27/79.97	88.57/76.98	87.22/76.64
ACC	97.02 ± 0.58	92.74 ± 1.32	89.08 ± 0.90	86.02 ± 1.06	83.94 ± 1.61
κ	0.89 ± 0.02	0.85 ± 0.01	0.84 ± 0.01	0.80 ± 0.01	0.78 ± 0.02

We combine the predicted sleep stages from all 61 subjects and compute the performance metrics compared with the annotated labels from experts for all recordings. Table 5 shows the confusion matrix obtained from the leave-one-subject-out cross-validation for the 5-state classification ($C = 5$). For simplicity, we only include the confusion matrix for $C = 5$, in which S3 and S4 are merged, because the 5-state classification is an issue of common concern with the comprehensive comparison. The numbers in bold indicate the number of epochs that are correctly classified, while the others represent the incorrectly classified epochs. The last two columns in each row indicate per-class performance metrics computed from the confusion matrix.

Table 6 gives the average accuracy and κ of 61 subjects for each C. Examples of the predicted sleep stages compared with the expert annotation are given in figure 3 (SC*) and figure 4 (ST*), where the blue line corresponds to the expert annotation, the red line corresponds to the predicted sleep stages and the black line on the bottom indicates the misclassified epochs.

5. Discussion

5.1. Comparison with state-of-the-art sleep classification methods

Table 7 shows a comprehensive comparison of the classification accuracy with eight state-of-the-art methods. We classify these methods into two groups: non-independent and independent training and testing sets. As presented in section 4, the non-independent training and testing scheme is the method that randomly selects epochs from all subjects to construct the testing set, while the independent one only selects epochs from one individual subject to construct the testing set.

Table 5. The confusion matrix for $C = 5$.

		Expert annotation					
		W	S1	S2	SWS	REM	RE(%)/PR(%)
Predicted classification	W	**10 244**	1410	821	178	634	93.72/77.10
	S1	52	**1124**	358	60	124	23.18/65.42
	S2	279	997	**22 532**	1655	1533	82.56/83.40
	SWS	69	85	1634	**6904**	172	78.03/77.89
	REM	287	1232	1947	51	**9365**	79.04/72.40

Table 6. The average accuracy and κ for each C.

	$C = 2$	$C = 3$	$C = 4$	$C = 5$	$C = 6$
Average accuracy for 61 individual subjects (%)	94.15 ± 0.95	85.06 ± 0.8	80.96 ± 1.12	78.68 ± 0.93	75.98 ± 1.05
Average kappa for 61 individual subjects	0.81 ± 0.01	0.74 ± 0.01	0.72 ± 0.02	0.71 ± 0.01	0.67 ± 0.02

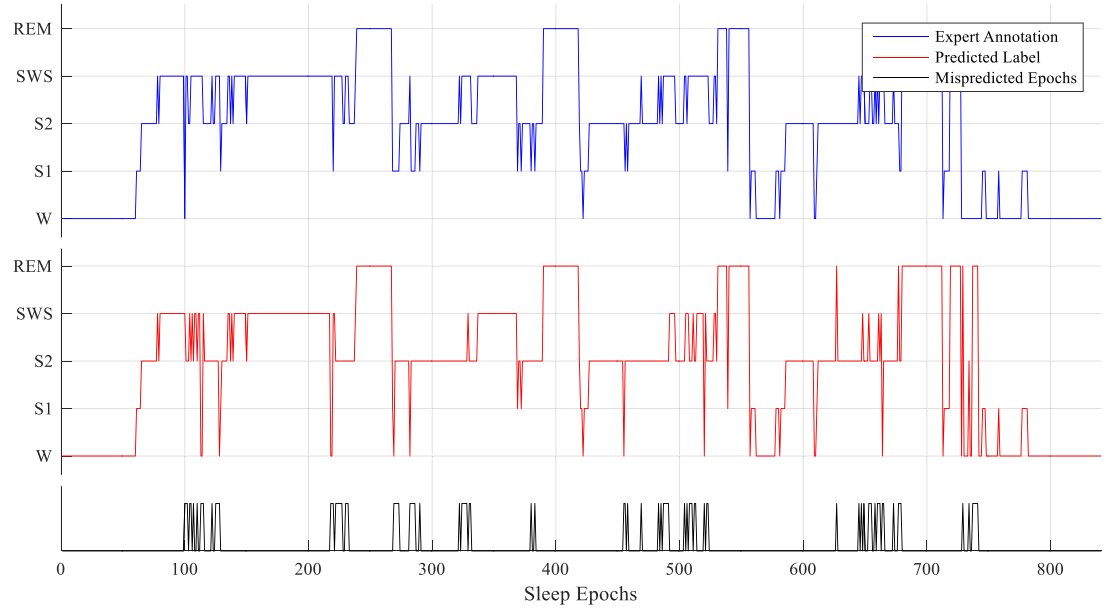

Figure 3. The sleep stages classification for subject SC4001E0. From top to bottom: expert annotation, predicted sleep stages and epochs for which there is a mismatch between the expert annotation and predicted sleep stages.

We should note that the comparison with other methods is very difficult for different EEG databases, different numbers of EEG channels and different numbers of sleep epochs. Based on this consideration, we include in this review all studies that report using the same EEG database as in this study, and they all used 30 s as a sleep epoch, and different kinds of features and classifiers are applied for sleep stage scoring. It is noted that most of the studies only use eight or 20 recordings from the sleep-EDF database, while to the best of our knowledge, this study is the first attempt to classify 2–6 sleep stages of 30 s epochs using both EEG and EOG channels with Fuzzy-MEn features, especially using all 61 recordings. Furthermore, we only give the comparison for $C = 5$, because most of the reviewed methods only report their results for the 5-state classification. We have 63 767 epochs in this study, much more than the referred studies, as shown in table 7. Thus, the results of this study are statistically reliable and clinically credible.

Compared with the methods in both groups, it can be seen that our method achieves a similar or slightly better classification accuracy compared with the state-of-the-art methods. Furthermore, the κ coefficient shows substantial agreements (0.71) between the sleep experts and our method. Considering that we use three common and feasible statistical features and a simple but efficient classifier that can be easily implemented in hardware, the classification performance is quite satisfying and competitive.

5.2. Why the S1 stage is not well detected

S1 stage discrimination is always the most challenging for automatic sleep stage classification tasks, because S1 is a transition phase between the change from wakefulness to other sleep stages. In this study, the most misclassified pair of sleep stages is S1–W, more than 29% of the S1 stage epochs are misclassified as W, as table 5 shows. The

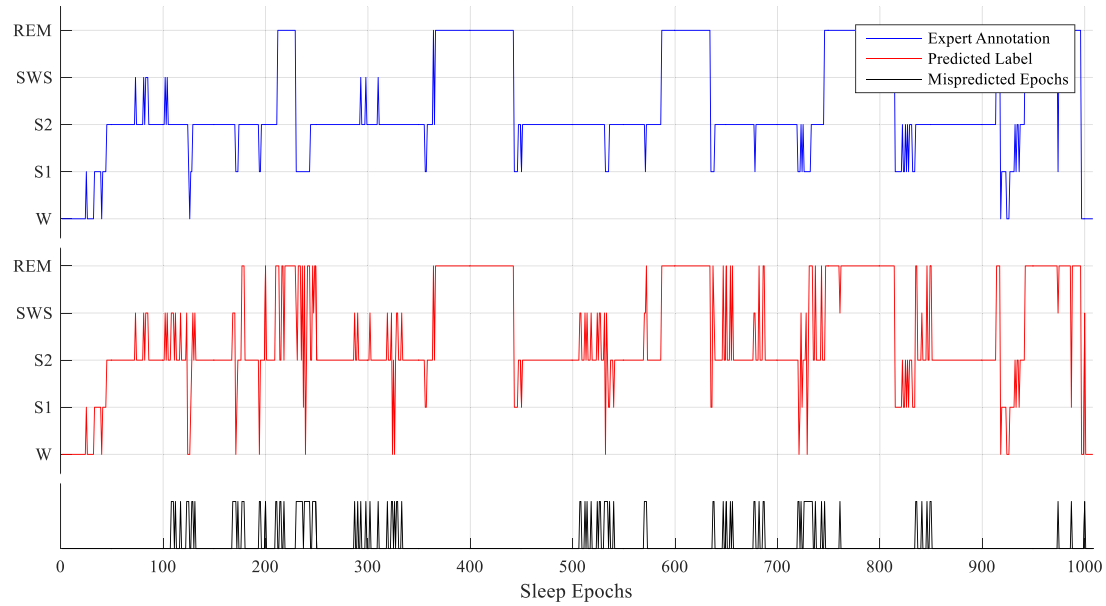

Figure 4. The sleep stages classification for subject ST7061J0. From top to bottom: expert annotation, predicted sleep stages and epochs for which there is a mismatch between the expert annotation and predicted sleep stages.

Table 7. Comparisons with other sleep stage scoring methods using the same dataset, sleep-EDF, across the overall accuracy and κ coefficient for $C = 5$.

Methods	Extracted features	Classifier	Subjects & channels	Accuracy	κ
Non-independent training and testing					
Zhu *et al* (2014)	14 963 epochs; difference visibility graph (DVG) & horizontal visibility graph (HVG)	SVM	8 recordings, & Pz-Oz	88.9%	—
Hassan (2015)	15 188 epochs; four standard statistics features & five spectral features	Bootstrap aggregating (bagging)	8 recordings, & Pz-Oz	86.53%	—
SC-En&SVM	63 767 epochs; FuzzyEn, FuzzyMEn, SampEn	One-against-all SVM	61 recordings, & Fpz-Cz, Pz-Oz, EOG	86.02%	0.80
Independent training and testing					
Rodríguez-Sotelo *et al* (2014)	40 826 epochs; multiscale entropy	Artificial neural network (ANN)	20 recordings, & Fpz-Cz, Pz-Oz	69%	0.42
Rodríguez-Sotelo *et al* (2014)	40 826 epochs; approximation entropy	ANN	20 recordings, & Fpz-Cz, Pz-Oz	74%	0.54
Sanders (2014)	9830 epochs; average spectral power, preferential frequency band & cross-frequency-coupling (CFC) method	Linear discriminant analysis (LDA)	10 recordings, & Fpz-Cz	75%	—
Tsinalis *et al* (2016a)	37 022 epochs; wavelet transform (WT) coefficients	Stacked sparse auto-encoders neural network (NN)	20 recordings, & Fpz-Cz	78.9%	—
Tsinalis *et al* (2016b)	37 022 epochs; raw EEG input	Convolutional neural network (CNN)	20 recordings, & Fpz-Cz	74.8%	0.65
Supratak *et al* (2017)	41 950 epochs; time invariant features	Deep learning	20 recordings, & Pz-Oz	79.8%	0.72
SC-En&SVM	63 767 epochs; FuzzyEn, FuzzyMEn, SampEn	One-against-all SVM	61 recordings, & Fpz-Cz, Pz-Oz, EOG	78.68%	0.71

main reason for this problem, we think, is the similarity in the characteristic EEG frequency patterns between W and S1, as described in Iber (2007). Specifically, both stages are characterized by the low voltage mixed 2–7 Hz and alpha activity. Furthermore, S1–REM is another easily misclassified stage-pair, with 26% of the S1 stage epochs being misclassified as REM. The characteristic EEG frequency patterns of both stages are also similar. In addition, S2 is the next stage that S1 is easily confused with (about 21%). The classification between S1 and S2

depends on the transition patterns that partly rely on the detection of arousals, body movement and slow eye movements, which are very difficult to capture (Tsinalis *et al* 2016a). Moreover, another reason that S1 could not be well detected is that it has the lowest number of samples fed into the classifier resulting in a failure of the learning process to this stage. Also, the possibilities that experts assign wrong labels to S1 in clinical diagnosis may also contribute to this problem.

5.3. Main characteristics of entropy features

The FuzzyEn and SampEn, as figure 2 illustrates, show a similar variation trend for the three channels and six sleep stages, while the FuzzyMEn is smaller on the whole and holds relatively more significant differences among different stages. The three features all show the highest value for stage W and basically have the least value for S4. This is because W is the most active stage and it is characterized by the presence of continuous alpha waves, faster frequency rhythms, along with frequent eye movements, while S4 is the deepest sleep stage, characterized by the existence of delta waves in more than 50% of the epochs.

5.4. Performance analysis for leave-one-subject-out cross-validation

It can be seen from table 5 that the poorest performance is noted for stage S1, with RE less than 30%, while the RE for other stages is significantly better, with the range between 78.03% to 93.72%. Most of the S1 epochs are misclassified as W, REM and S2. In addition, the S2 stage is easily misclassified as SWS (slow wave sleep, including S3 and S4) and REM. At the same time, SWS and REM are also easily misclassified as S2. It can also be seen that the confusion matrix is basically symmetric via the diagonal line (except in the pairs of S1–W and S1–REM). This indicates that the misclassifications are less likely to be due to the problem of data imbalance.

5.5. Limitations and future work

Finally, we should also point out the limitations of the proposed method, which we will focus on improving in our future work. On one hand, as a supervised learning process, the robustness of our method still needs to be tested for other EEG databases, and especially applied to different electrode positions to understand the best channels for sleep classification. Moreover, we only tested healthy people in the current study; the classification performance for people with sleep disorders needs to be tested in the future. In addition, we will also further investigate the correlations between features and the performance may be improved by feature selection. Lastly, even though our classification performance is encouraging, the overall classification accuracy and per-class performance metrics still need improvement, especially when applied to clinical diagnosis.

6. Conclusions

A novel sleep stage classification method, SC-En&SVM, based on entropy features and the SVM classifier was proposed in this paper. We applied a novel entropy feature, i.e. FuzzyMEn, to study sleep EEG and EOG signals. Together with FuzzyEn and SampEn, we found that the entropy values showed significant differences between different sleep stages, and specifically, the entropy value for the W stage was obviously higher than the sleeping stages. What is more, the entropy features of each sleeping stage also showed evident differences between each other. As a consequence, the entropy features were extracted from two EEG channels and one EOG channel to perform 2–6 state classifications. The 61 subjects from the sleep-EDF database were applied and the signals were divided into epochs every 30 s. Afterwards, we designed the multi-class SVM using the one-against-all class approach to perform the classification. The leave-one-subject-out cross-validation for 61 individual subjects and 10-fold cross-validation for the non-independent training and testing set were performed, and the results showed a similar or slightly better classification performance compared with the eight state-of-the-art methods. The sleep scoring performance of the proposed SC-En&SVM proved encouraging and competitive, considering that we applied three common and feasible statistical features and a simple but efficient classifier that can be easily implemented in hardware.

In the future, we will aim to improve the classification performance of the S1 stage, as S1 stage discrimination has always been the most challenging for automatic sleep stage classification tasks. In the meantime, we also plan to improve the SC-En&SVM so that it can be applied to the single-channel EEG or EOG collected from wearable devices.

Acknowledgment

This work is supported by the Shandong Province Key Research and Development Plan (Grant No: 2018GSF118133) and the UQ 2016 Philanthropic Grant for Early Career Engineering Researchers (Biomedical Engineering) under Grant PG005-2016 and in part by the Guangxi Cloud Computing and Big Data Collaborative Innovation Centre (No: YD16E18). This work is also supported by China Postdoctoral Science Foundation (No.

2017M612280). The authors gratefully acknowledge the financial support from China Scholarship Council (No. 201706220221).

ORCID iDs

Zhimin Zhang https://orcid.org/0000-0002-0531-9112
Guohun Zhu https://orcid.org/0000-0003-3356-8236
Chengyu Liu https://orcid.org/0000-0003-1965-3020
Feng Liu https://orcid.org/0000-0002-1074-2601

References

Aboalayon K A I, Ocbagabir H T and Faezipour M 2014 Efficient sleep stage classification based on EEG signals *Systems, Applications and Technology Conf.* pp 1–6
Adnane M, Jiang Z and Yan Z 2012 Sleep–wake stages classification and sleep efficiency estimation using single-lead electrocardiogram *Exp. Syst. Appl.* **39** 1401–13
Al-Sharhan S, Karray F, Gueaieb W and Basir O 2001 Fuzzy entropy: a brief survey *The IEEE Int. Conf. on Fuzzy Systems* pp 1135–9
Azami H, Fernández A and Escudero J 2017 Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis *Med. Biol. Eng. Comput.* **55** 2037–52
Bianchi M T 2017 Sleep devices: wearables and nearables, informational and interventional, consumer and clinical *Metab. Clin. Exp.* **84** 99–108
Boostani R, Karimzadeh F and Nami M 2017 A comparative review on sleep stage classification methods in patients and healthy individuals *Comput. Methods Prog. Biomed.* **140** 77–91
Bsoul M, Minn H and Tamil L 2011 Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG *IEEE Trans. Inf. Technol. Biomed.* **15** 416–27
Chapotot F and Becq G 2010 Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules *Int. J. Adapt. Control Signal Process.* **24** 409–23
Charbonnier S, Zoubek L, Lesecq S and Chapotot F 2011 Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging *Comput. Biol. Med.* **41** 380
Chen L L, Zhao Y, Zhang J and Zou J Z 2015 Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning *Exp. Syst. Appl.* **42** 7344–55
Chen W, Wang Z, Xie H and Yu W 2007 Characterization of surface EMG signal based on fuzzy entropy *IEEE Trans. Neural Syst. Rehabil. Eng.* **15** 266–72
Chriskos P, Kaitalidou D S, Karakasis G, Frantzidis C, Gkivogkli P T, Bamidis P and Kourtidou-Papadeli C 2017 Automatic sleep stage classification applying machine learning algorithms on EEG recordings *IEEE Int. Symp. on Computer-Based Medical Systems* pp 435–9
Costa M, Goldberger A L and Peng C K 2005 Multiscale entropy analysis of biological signals *Phys. Rev.* E **71** 021906
Ge J, Zhou P, Zhao X and Wang M 2007 Sample entropy analysis of sleep EEG under different stages *IEEE/ICME Int. Conf. on Complex Medical Engineering* pp 1499–502
Goldberger A L *et al* 2000 PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals *Circulation* **101** E215
Hassan A R and Subasi A 2017 A decision support system for automated identification of sleep stages from single-channel EEG signals *Knowl.-Based Syst.* **128** 115–24
Hassan A R, Bashar S K and Hassan Bhuiyan M I 2015 On the classification of sleep states by means of statistical and spectral features from single channel electroencephalogram *2015 Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI)* pp 2238–43
Iber A-I S, Chesson A L and Quan S F 2007 The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications *American Academy of Sleep Medicine, 2007* (https://doi.org/10.5664/jcsm.6576)
Inouye T, Shinosaki K, Sakamoto H, Toi S, Ukai S, Iyama A, Katsuda Y and Hirano M 1991 Quantification of EEG irregularity by use of the entropy of the power spectrum *Electroencephalogr. Clin. Neurophysiol.* **79** 204–10
Karimzadeh F, Boostani R, Seraj E and Sameni R 2018 A distributed classification procedure for automatic sleep stage scoring based on instantaneous electroencephalogram phase and envelope features *IEEE Trans. Neural Syst. Rehabil. Eng.* p 1
Kemp B, Zwinderman A H, Tuk B, Kamphuisen H A C and Oberye J J L 2000 Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG *IEEE Trans. Biomed. Eng.* **47** 1185–94
Krakovská A and Mezeiová K 2011 Automatic sleep scoring: a search for an optimal combination of measures *Artif. Intell. Med.* **53** 25–33
Lajnef T, Chaibi S, Ruby P, Aguera P E, Eichenlaub J B, Samet M, Kachouri A and Jerbi K 2015 Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines *J. Neurosci. Methods* **250** 94–105
Liang S F, Kuo C E, Hu Y H, Pan Y H and Wang Y H 2012 Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models *IEEE Trans. Instrum. Meas.* **61** 1649–57
Liu C, Ke L, Zhao L, Feng L, Zheng D, Liu C and Liu S 2013 Analysis of heart rate variability using fuzzy measure entropy *Comput. Biol. Med.* **43** 100–8
Mourtazaev M S, Kemp B, Zwinderman A H and Kamphuisen H A 1995 Age and gender affect different characteristics of slow waves in the sleep EEG *Sleep* **18** 557–64
Nakamura T, Adjei T, Alqurashi Y, Looney D, Morrell M J and Mandic D P 2017 Complexity science for sleep stage classification from EEG *Int. Joint Conf. on Neural Networks* pp 4387–94
Oral E A, Çodur M M and Ozbek I Y 2017 Sleep stage classification based on filter bank optimization *Signal Processing and Communications Applications Conf.* pp 1–4
Popovic D, Khoo M and Westbrook P 2014 Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults *J. Sleep Res.* **23** 211–21
Qureshi S and Vanichayobon S 2017 Evaluate different machine learning techniques for classifying sleep stages on single-channel EEG *Int. Joint Conf. on Computer Science and Software Engineering* pp 1–6

Rechtschaffen A and Kales A 1968 A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects *Natl Inst. Health* **1** 204–10

Richman J S and Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *Am. J. Physiol. Heart Circ. Physiol.* **278** H2039

Rodríguez-Sotelo J, Osorio-Forero A, Jiménez-Rodríguez A, Cuesta-Frau D, Cirugeda-Roldán E and Peluffo D 2014 Automatic sleep stages classification using EEG entropy features and unsupervised pattern analysis techniques *Entropy* **16** 6573–89

Ronzhina M, Janoušek O, Kolářová J, Nováková M, Honzík P and Provazník I 2012 Sleep scoring using artificial neural networks *Sleep Med. Rev.* **16** 251–63

Sanders T H, McCurry M and Clements M A 2014 Sleep stage classification with cross frequency coupling *36th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 4579–82

Sors A, Bonnet S, Mirek S, Vercueil L and Payen J F 2018 A convolutional neural network for sleep stage scoring from raw single-channel EEG *Biomed. Signal Process. Control* **42** 107–14

Supratak A, Dong H, Wu C and Guo Y 2017 DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG *IEEE Trans. Neural Syst. Rehabil. Eng.* p 1

Surantha N, Isa S M, Lesmana T F and Setiawan I M A 2017 Sleep stage classification using the combination of SVM and PSO *Int. Conf. on Informatics and Computational Sciences* pp 177–82

Tsinalis O, Matthews P M and Guo Y 2016a Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders *Ann. Biomed. Eng.* **44** 1587–97

Tsinalis O, Matthews P M, Guo Y and Zafeiriou S 2016b Automatic sleep stage scoring with single-channel EEG using convolutional neural networks (https://arxiv.org/abs/1610.01683)

Zhang J and Wu Y 2017 A new method for automatic sleep stage classification *IEEE Trans. Biomed. Circuits Syst.* **11** 1097–110

Zhu G, Li Y and Wen P P 2012 An efficient visibility graph similarity algorithm and its application on sleep stages classification *Int. Conf. Brain Inform.* **7670** 185–95

Zhu G, Li Y and Wen P P 2014 Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal *IEEE J. Biomed. Health Inform.* **18** 1813