Complexity science for sleep stage classification from EEG

Takashi Nakamura*, Tricia Adjei*, Yousef Alqurashi†, David Looney*, Mary J. Morrell† and Danilo P. Mandic*
*Department of Electrical and Electronic Engineering, Imperial College London
London, SW7 2AZ, United Kingdom
†Sleep and Ventilation Unit, National Heart and Lung Institute, Imperial College London, and NIHR Respiratory Disease Biomedical Research Unit at the Royal Brompton and Harefield NHS Foundation Trust, and Imperial College London
London, SW3 6NP, United Kingdom
Email: {takashi.nakamura14, t.adjei15, y.alqurashi15, david.looney06, m.morrell, d.mandic}@imperial.ac.uk

***Abstract*—Automatic sleep stage classification is an important paradigm in computational intelligence and promises considerable advantages to the health care. Most current automated methods require the multiple electroencephalogram (EEG) channels and typically cannot distinguish the S1 sleep stage from EEG. The aim of this study is to revisit automatic sleep stage classification from EEGs using complexity science methods. The proposed method applies fuzzy entropy and permutation entropy as kernels of multi-scale entropy analysis. To account for sleep transition, the preceding and following 30 seconds of epoch data were used for analysis as well as the current epoch. Combining the entropy and spectral edge frequency features extracted from one EEG channel, a multi-class support vector machine (SVM) was able to classify 93.8% of 5 sleep stages for the SleepEDF database [expanded], with the sensitivity of S1 stage was 49.1%. Also, the Kappa's coefficient yielded 0.90, which indicates almost perfect agreement.**

I. Introduction

Sleep can be defined as a reversible behavioural state of perceptual disengagement and unresponsiveness to the environment. Sleep is also a complex amalgam of physiologic and behavioural processes [1]. The function of sleep is not fully understood yet, and among the many hypotheses proposed, the most widely accepted ones are brain thermoregulation, brain detoxification, tissue restoration and metabolic homeostasis [2, 3]. Quality of sleep also reflects the state of body and mind, and in addition, numerous sleep disorders are common [4]; these disorders include insomnia, breathing disturbances during sleep (i.e., sleep apnea), and narcolepsy, which are diagnosed using polysomnography (PSG).

The PSG recording requires multiple electrodes which includes at least two EEG channels, two electrooculography (EOG) channels to observe eye movements, at least one chin electromyography (EMG) channel. The PSG recording is usually conducted within a hospital or at a sleep centre. After the recording, the PSG is analysed by experts in order to identify individual sleep patterns. The classification of sleep stages has been classically performed based on the visual interpretation of each 30-second PSG epoch, according to the Rechtschaffen and Kales (R&K) sleep scoring manual [5], or the manual of the American Academy of Sleep Medicine (AASM) [6]. The sleep stages include: wake (W), stage1 (S1), stage2 (S2), stage3 (S3), stage4 (S4), and rapid eye movement (REM) [1]. The stages S3 and S4 are called slow wave sleep (SWS) and these two stages are merged into one condition.

The main limitation of epoch based visual sleep scoring is that it is extremely time-consuming. The scoring of 8 hours overnight PSG takes approximately $2-4$ hours for an expert [7]. In addition, the multiple electrode montage disturbs patients' sleep, also, with PSG recordings typically occurring in hospitals, or other unfamiliar environments, some patients find themselves unable to sleep as usual. In the last two decades, computer based sleep staging has been developed in order to minimise analysis time [8].

To alleviate the aforementioned problems, a large number of automatic sleep stage algorithms, based on a small number of electrodes have been reported. These methods are typically based on feature extraction techniques and pattern recognition algorithms. In terms of the analysis data, the Physionet SleepEDF database [9] has been available for more than 10 years. For datasets extracted from SleepEDF database, Imtiaz *et al.* [10] calculated spectral band power and spectral edge frequency (SEF) from two EEG channels. These methods correctly classified 82.2% of epochs for a 5-stage task. Zhu *et al.* [11] classified the sleep stages based on graph domain features, and achieved an accuracy of 89.0%. Hassan *et al.* [12] applied empirical mode decomposition (EMD) analysis for a single lead EEG montage, and calculated 1st to 4th statistical moments for each intrinsic mode function (IMF). Their proposed method achieved 90.7% accuracy by bootstrap aggregating decision tree. Recently, Silveria *et al.* [13] reported 91.5% accuracy for 5-stage classification problems when using discrete wavelet transform and random forests classifiers (RF).

Among research groups who recorded PSG by themselves, Liang *et al.* [14] recorded PSG from 20 healthy subjects, and calculated multi-scale sample entropy (MSSE) and autoregressive coefficients from a single EEG channel. The linear discriminant analysis (LDA) yielded 88.1% accuracy for testing data. Şen *et al.* [15] used PSG recordings from 25 subjects to calculate multiple features and applied a feature

[1]For the ASSM manual, the stages are noted as non-REM stage 1-3 (N1, N2, N3) and REM.

selection algorithm to reduce the dimension of the feature matrix. The RF correctly classified 97.3% of epochs for labeled 5-stage sleep. Overall, a large number of algorithms have been proposed, combining a wide range of features and classification algorithms, and the performance of each method is evaluated by datasets recorded by the researchers themselves or public datasets.

For the sleep stage classification, distinguishing between S1 sleep and REM is the most challenging with state-of-the-art papers. In the R&K sleep scoring guideline, S1 sleep is defined as 50% of the epoch consists of relatively low voltage mixed $(2-7Hz)$ activity, and $<50\%$ of the epoch contains alpha activity, whereas REM is represented by a relatively low amplitude and mixed frequency $(2-7Hz)$ EEG with episodic rapid eye movements and the absence of (or reduced) chin EMG activity [16]. Figure 1 shows the power spectral density for two EEG channels (Fpz-Cz and Pz-Oz) of the *SEDFx-S* dataset (see details in Section III-A1) with $\mathcal{C} = 5$ sleep stages. For the wake (W) condition, observe a peak in the alpha band $(8-13Hz)$. Since the alpha rhythm is predominantly observed from the occipital lobe, the Pz channel has stronger alpha than the Fpz. The alpha power reduced in the S1 stage, and the strong peak around $13-14Hz$ was present for the S2 stage. During the SWS stage, the delta activity become larger, however, the spectrum of REM mostly overlapped with that of S1 except for the beta band. In order to distinguish between REM and S1, we consider an automatic sleep stage classification problem based on structural complexity analyses by means of entropy.

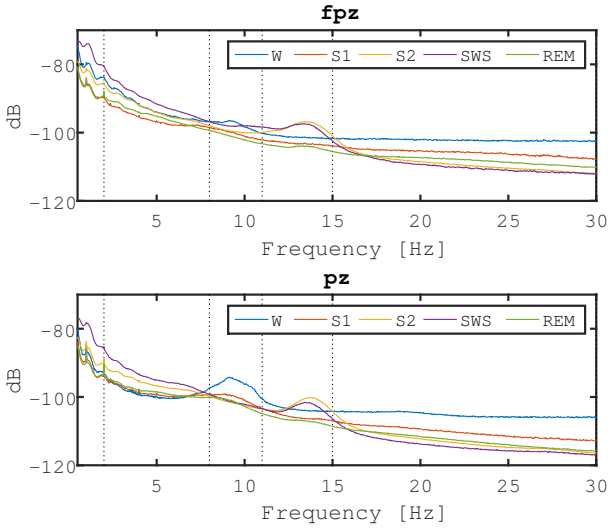

Fig. 1. Power spectral density for two EEG channels (Fpz-Cz, and Pz-Oz) from the *SEDFx-S* dataset with $\mathcal{C} = 5$ sleep stages

Multi-scale entropy (MSE) [17] was introduced as a non-parametric method for estimating dynamical complexity over multiple scales of a time series. The MSE has been widely utilised for physiological responses such as ECG [18] and EEG [19]. We calculate MSE for evaluating multi-scale complexity using fuzzy entropy (FE) and permutation entropy (PE). The FE [20] is feasible for relatively short physiological signals with a small embedding dimension, and is robust to noise. Furthermore, the FE is more independent of data length and has relative consistency. The PE detects dynamic changes in a time series based on neighbouring data points requires less computational time, and is robust for noisy real world time series. As the computation of PE uses self-generated partitions to create numeric symbols, the method is well suited to series with poor stationarity, such as physiological signals [21].

In this paper, we make use of the robustness of structural complexity measures to perform sleep stage classification which yields the following advantages:

- The required number of EEG channels is reduced to one or two channels of EEG, which is a prerequisite for both portable and wearable devices.
- Structural complexity analyses as well as spectral edge frequency (SEF) are able to distinguish between the S1 stage and REM sleep, which has been notoriously difficult with the state-of-the-art methods using only EEG.

II. Complexity Analysis

A. Multi-scale entropy

The multi-scale entropy (MSE) method [17] measures the amount of structural complexity in a time series. The MSE can be calculated from different types of entropy, such as approximate entropy and sample entropy, with multiple coarse-grained time series. Previously, multi-scale sample entropy (MSSE) has been utilised for sleep stage classification in [14]. In this study, we employ the fuzzy entropy [20] and the permutation entropy [21] as kernels for entropy calculation.

Given an EEG signal with N data points $\{x_i, i = 1:N\}$, a coarse-grained time series $\{y^{(\tau)}\}$ is first generated, where τ is the scale factor. This is achieved by dividing a given EEG signal into non-overlapping windows of length of τ, and by averaging over as

$$y^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \qquad 1 \leq j \leq \frac{N}{\tau}. \tag{1}$$

Therefore, $y^{(1)}$ corresponds to the original EEG signal, while the length of the coarse-grained time series, $y^{(\tau)}$, is simply the length of the original EEG signal divided by τ.

B. Multi-scale fuzzy entropy

Fuzzy entropy is calculated for each coarse-grained time series to yield the multi-scale fuzzy entropy (MSFE). For coarse-grained time series with M samples $\{y^{(\tau)}(k), k = 1:M\}$, a

vector sequence $\{\mathbf{Y}_k^{(\tau),m}, k = 1 : (M-m+1)\}$ is constructed with given m as follows:

$$\{\mathbf{Y}_k^{(\tau),m} = \{y^{(\tau)}(k), y^{(\tau)}(k+1), ..., y^{(\tau)}(k+m-1)\} - \overline{y^{(\tau)}(k)}, \quad (2)$$

$$where \quad \overline{y^{(\tau)}(k)} = \frac{1}{m}\sum_{l=0}^{m-1} y^{(\tau)}(k+l), \quad (3)$$

where $\mathbf{Y}_k^{(\tau),m}$ denotes m consecutive $y^{(\tau)}$ values starting from the kth point, with the average of this sequence $\overline{y^{(\tau)}(k)}$ removed. Here, the distance $d_{kl}^{(\tau),m}$ between $\mathbf{Y}_k^{(\tau),m}$ and $\mathbf{Y}_l^{(\tau),m}$ is defined as,

$$d_{kl}^{(\tau),m} = d\left[\mathbf{Y}_k^{(\tau),m}, \mathbf{Y}_l^{(\tau),m}\right] \quad (4)$$

$$= \max_{p\in(0,m-1)} |\{y^{(\tau)}(k+p) - \overline{y^{(\tau)}(k)}\} - \{y^{(\tau)}(l+p) - \overline{y^{(\tau)}(l)}\}|. \quad (5)$$

The distance $d_{kl}^{(\tau),m}$ represents the maximum absolute difference of the scalar components. The degree of similarity, $D_{kl}^{(\tau),m}$, between $\mathbf{Y}_k^{(\tau),m}$ and $\mathbf{Y}_l^{(\tau),m}$ is given by

$$D_{kl}^{(\tau),m}(n,r) = \mu(d_{kl}^{(\tau),m}, n, r) \quad (6)$$

$$= \exp\left(-\frac{\left(d_{kl}^{(\tau),m}\right)^n}{r}\right), \quad (7)$$

where n and r are given parameters, and the fuzzy function $\mu(d_{kl}^{(\tau),m}, n, r)$ was chosen to be the exponential function. Next, $\phi^{(\tau),m}(n,r)$ is set as follows,

$$\phi^{(\tau),m}(n,r) = \frac{1}{M-m}\sum_{k=1}^{M-m}\left(\frac{1}{M-m-1}\sum_{l=1,k\neq 1}^{M-m} D_{kl}^{(\tau),m}\right).$$

Finally, the MSFE is given by

$$MSFE(\tau, m, n, r, N) = \ln\phi^{(\tau),m}(n,r) - \ln\phi^{(\tau),m+1}(n,r).$$

Figure 2 illustrates MSFE analysis for two EEG channels (Fpz and Pz) of the *SEDFx-S* dataset (see details in Section III-A1) with $\mathcal{C} = 5$ sleep stages; the parameters used to calculate the MSFE were $\tau = 30$, $m = 2$, $n = 2$, $r = 0.15\times$(*standard deviation of each epoch*). The trends of MSFE are similar when two EEG channels except W condition; since the Pz has strong alpha rhythm (see Figure 1), the structural complexity of the signal becomes smaller with larger scale τ.

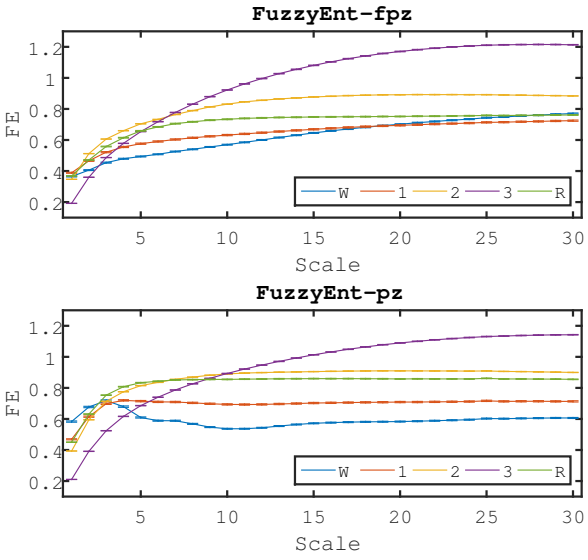

Fig. 2. MSFE analyses for two EEG channels from the *SEDFx-S* dataset with $\mathcal{C} = 5$ sleep stages, and scale $\tau = 30$. The error bars represent the standard error

C. Multi-scale permutation entropy

Permutation entropy (PE) is calculated for each coarse-grained time series as the multi-scale permutation entropy (MSPE) [22]. For a coarse-grained time series with M samples $\{y^{(\tau)}(k), k = 1 : M\}$, the series of vectors of length d, $v_d^{(\tau)}(k) = \{y^{(\tau)}(k), y^{(\tau)}(k+L), ..., y^{(\tau)}(k+(d-1)L)\}$ is first calculated, where d is the embedding dimension and L time delay. Then, the vector $v_d^{(\tau)}(k)$ is arranged in an increasing order: $\{y^{(\tau)}(k+j1-1), y^{(\tau)}(k+j2-1), ..., y^{(\tau)}(k+jk-1)\}$. For the sequence of length d, there is $d!$ possible patterns, π, called motifs. Let $f(\pi_j)$ define the frequency of occurrence in the time series for each motif π_j. The relative frequency is then given by,

$$p(\pi_j) = \frac{f(\pi_j)}{M-d+1}. \quad (8)$$

The PE for the time series is defined as the Shannon entropy for the $d!$ motifs,

$$H(d) = -\sum_{\pi_j}^{d!} p(\pi_j)\ln p(\pi_j), \quad (9)$$

while the normalised MSPE entropy is given by

$$0 \leq MSPE(\tau, d, L, N) = \frac{H(d)}{\ln d!} \leq 1. \quad (10)$$

Figure 3 illustrates MSPE analysis for two EEG channels (Fpz, and Pz) of *SEDFx-S* dataset with $\mathcal{C} = 5$ sleep stages; the parameters used to calculate the MSPE are $\tau = 20$, $d = 5$, and $L = 1$.

III. Methods

To evaluate the performance of the proposed method, the EEG recording was obtained from the both SleepEDF database [expanded] [9] and DREAMS Subjects database [23]. We extracted features from each EEG channel, and applied them to the classification algorithm. Figure 4 shows the flowchart of the proposed analysis.

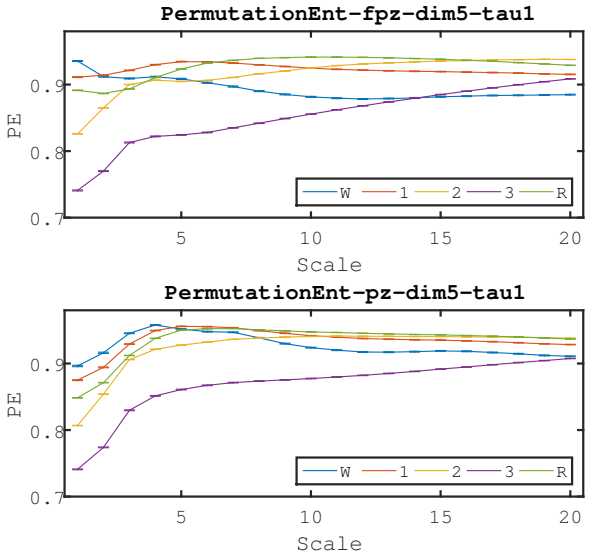

Fig. 3. MSPE analyses for two EEG channels from the *SEDFx-S* dataset with $C = 5$ sleep stages, and scale $\tau = 20$. The error bars represent the standard error

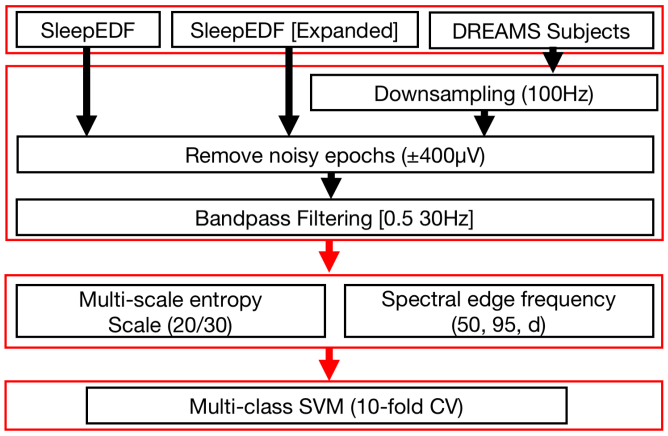

Fig. 4. Flowchart of an automatic sleep staging system

A. Data acquisition

1) SleepEDF database: The SleepEDF database [9] consists PSG data of 8 subjects, which include 2 EEG, Fpz-Cz and Pz-Oz (for simplicity, we denote Fpz and Pz later), 1 EOG (horizontal), and 1 submental chin EMG. Each PSG is sampled at 100 Hz and manually scored by well-trained technicians based on the R&K manual. Recently, the SleepEDF database [expanded] [24] has been made public, and contains in total PSG recordings of 61 subjects.

Although the SleepEDF database is extensively used for automatic sleep stage scorings [7, 10–14, 25], the usage of the database varies for research groups. One problem is that the database contains two types of PSG data: some of them are recorded over 24 hours (titled as *sc* $*$), and the others contain only overnight data (denoted by *st* $*$). In terms of *sc* $*$ data, the majority of epochs were labeled as the wake condition, because the data were obtained both over day and night periods; in other words, the data contains a large number of pre-sleep wake epochs (from afternoon to evening) and post-sleep wake epochs (from morning to afternoon). Therefore, most current analyses are biased towards the wake condition which dominates the 24-hour data. To this end, we focus on 8-hour segment of EEG.

The lack of consistency in the data structure makes the direct comparison between existing methods and the proposed idea difficult. Here, we partitioned the dataset from the SleepEDF database in three different ways in order to enable comparisons with existing methods.

1) *SleepEDF Whole (SEDF-W)* - The number of subjects is 8. We downloaded data from [9], and extracted whole scored data (from *hypnogram start time* to last scored data) and excluded epochs labeled "?". This dataset has a similar form to the dataset used in [11, 12, 25].
2) *SleepEDF [expanded] Whole (SEDFx-W)* - The number of subjects is 61. We downloaded data from [24], and extracted whole scored data (from *hypnogram start time* to last scored data) and excluded epochs labeled "?". This dataset has a similar form to the dataset used in [13]
3) *SleepEDF [expanded] Sleep (SEDFx-S)* - The number of subjects is 61. We downloaded data from [24]. Imtiaz *et al.* [26] proposed an open source toolbox for Matlab to extract only overnight data in order to create a standardised dataset. The toolbox truncates the data in the following way:
   - From *light off* time to *light off* time.
   - if *light off/on* time is not available, from 15 minutes before the first scored sleep epoch to 15 minutes after the last scored sleep epoch.

   This dataset has a similar form to the dataset used in [10].

TABLE I
COMPARISON OF DATASETS - AFTER REMOVING NOISY EPOCHS (NUMBER OF EPOCHS AND *Ratio*(*italic*)[%])

Dataset	Wake	Sleep Stage				
		S1	S2	S3	S4	REM
SEDF-W	7943	581	3596	667	619	1589
	53.0	*8.2*	*51.0*	*9.5*	*8.8*	*22.5*
SEDFx-W	74346	4715	27070	5056	3757	11755
	58.7	*9.0*	*51.7*	*9.7*	*7.2*	*22.5*
SEDFx-S	6448	4676	26849	4996	3705	11749
	11.0	*9.0*	*51.7*	*9.6*	*7.1*	*22.6*
DREAMS	3593	1181	8812	1381	1966	3017
	17.9	*7.2*	*53.6*	*8.4*	*12.0*	*18.4*

Table I summarises the number of epochs and the ratio of epochs with respect to each sleep stage for datasets extracted from the SleepEDF database after the pre-processing. The ratio for the wake condition, and the ratios for S1, S2, S3, S4, and REM are calculated as follows:

$$Wake\ Ratio = \frac{Number\ of\ wake\ epoch}{Number\ of\ whole\ epoch},$$

$$Sleep\ Ratio \quad = \quad \frac{Number\ of\ each\ stage\ epoch}{Number\ of\ whole\ sleep\ epoch}.$$

2) DREAMS database: The DREAMS Subjects database [23] consists of 20 overnight PSG recordings from healthy subjects, and the data are scored according to both the R&K and AASM criteria. The PSGs are at least from two EOG channels (P8-A1, P18-A1), three EEG channels (Cz-A1 or C3-A1, Fp1-A1 and O1-A1) and one sub-mental EMG channel. The sampling frequency is $200Hz$. For this analysis, we only used O1-A2 and Cz-A1 or C3-A1 channels (for simplicity, we denote O1 and Cz or C3 later). Since the PSG data were recorded overnight, not over 24-hours, we used all the data except epochs labeled "sleep stage movement" and "unknown sleep stage". We shall refer to this dataset as *DREAMS*, see Table I.

B. Pre-processing

For the *DREAMS* dataset, the signal was downsampled to $100Hz$. Then, the epochs which contained amplitudes of more than $\pm 400\mu V$ were removed, since the amplitude of the K-complex is almost always less than $400\mu V$ [27]. Epochs containing NAN values were removed from the analyses. Afterwards, the 4th order Butterworth filter with passband from $0.5 - 30Hz$ was applied. In total, there were 58423 epochs for the *SEDFx-S* dataset; 11.0% of the data were scored as awake, and 9.0%, 51.7%, 9.6 %, 7.1%, and 22.6% of sleep data were scored as S1, S2, S3, S4, and REM, respectively.

C. Feature extraction

After the pre-processing, multi-scale fuzzy entropy (MSFE), multi-scale permutation entropy (MSPE), and spectral edge frequency (SEF) were calculated from each epoch of the EEG channels.

1) Epoch length: The length of epochs for manual sleep stage scoring was 30 seconds, and the epoch-based criteria are often dependent on the preceding and following epochs. Here, we extract epochs in two different ways: 1) use only the current 30 seconds epoch to calculate features, 2) enlarge the epoch size from 30 seconds to 90 seconds; in other words, utilise both the preceding and following 30 seconds of epoch data as well as the current epoch to take account of sleep transition. Since the sampling frequency was $100Hz$, the length of each epoch for the no overlap method is $N = 3000$, and $N = 9000$ data points were used for the proposed overlap methods. For the overlap method, the first and last epoch of EEG recordings were excluded from analyses.

2) Multi-scale entropy: For the MSPE, we chose maximum scale $\tau = 20$. When the window contains less than $d!$ samples, the PE can be calculated with the dimension equal to or less than the $d-1$. The length of coarse-grained time series for no overlapping epochs with scale $\tau = 20$ is $\frac{N=3000}{\tau=20} = 150$ which is larger than $5! = 120$, so the PE with dimension $d = 5$ can be calculated with $L = 1$. The parameters used to calculate the MSFE were $\tau = 30$, $m = 2$, $n = 2$, $r = 0.15\times$(*standard deviation of each epoch*).

3) Spectral edge frequency: The $r\%$ spectral edge frequency (SEF) is the frequency below which $r\%$ of the signal power is contained; in other words, SEFr corresponds to $r\%$ of the signal power and is given by

$$\sum_{f=f_{low}}^{f_{high}} \|magnitude(f)\|^2 \times \frac{r}{100} = \sum_{f=f_{low}}^{SEFr} \|magnitude(f)\|^2.$$

Imtiaz *et al.* [28] used the difference between SEF95 and SEF50, called SEFd, in order to detect REM stage; $SEFd = SEF95 - SEF50$. The SEF has been utilised for physiological response analysis, specifically for EOG [29] and for EEG [10] analysis. For the current analyses, we chose the frequency ranges for SEF50, SEF95, and SEFd in the following bands; $\delta - \beta = 0.5 - 30Hz$, $\delta - \alpha = 0.5 - 16Hz$, $\theta = 2 - 8Hz$, $\alpha = 8 - 15Hz$, $\alpha_l = 8 - 11Hz$, $\alpha_h = 11 - 15Hz$, and $\beta = 16 - 30Hz$.

D. Classification

The extracted features were normalised to between $[0, 1]$ before classification. As a classifier, the multi-class support vector machine (SVM) with one-against-one class approach, which uses voting strategies for each binary classification, was employed [30]. The radial basis function (RBF) kernel was used for the SVM given by

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\gamma|x - x|^2). \tag{11}$$

The parameters for classification were set to $\gamma = 0.6 - 4$.

IV. Results

Feature extraction parts was undertaken using Matlab 2015a, and the classification was conducted in Python 2.7.11 |Anaconda 2.3.0 (x86_64) operated on a MacBook Pro with 2.2GHz Intel Core i7, 16GB of RAM. First, we compared the classification results using different features with no overlap and overlap blocks. Then, we evaluated the classification performance for $C = 2 - 6$ class sleep stages (See, Table II). Finally, we compared our methods to the existing methods using the SleepEDF database.

TABLE II
Sleep stage classes ($\mathcal{C}$)

$\mathcal{C}$	Sleep stages
6	Wake, S1, S2, S3, S4, REM
5	Wake, S1, S2, SWS(S3-4), REM
4	Wake, S1-2, SWS, REM
3	Wake, NREM(S1-4), REM
2	Wake, Sleep(REM & NREM)

A. Evaluation

The multi-class SVM with 10-fold cross-validation (CV) was utilised, and the performance indicators to evaluate the proposed methods were accuracy, sensitivity, precision, and Cohen's Kappa coefficient κ. κ determines the agreement between scorers, with chance agreement removed. The values between $0.00-0.20$, $0.21-0.40$, $0.41-0.60$, $0.61-0.80$, and $0.81 - 1.00$ respectively correspond to slight, fair, moderate, substantial, and almost perfect agreement [31].

B. Overlapping performance

Table III shows the classification accuracy for the *SEDFx-S* dataset and *DREAMS* dataset for 5-stage classification, using different features extracted from EEG. The overlap epoch based method performed better than the conventional 30-second epoch method. The classification accuracy using MSPE and SEF features was the highest for both the *SEDFx-S* and *DREAMS* datasets; 88.6% and 86.8% of epochs were correctly predicted using the overlapping window, respectively. Figure 5 illustrates the labeled hypnogram and the predicted label for subject *sc4001e0* with no overlapping epochs (Upper) and with overlapping epochs (Lower). Since the overlapping epochs method extracted EEG not only from the current 30-second epoch but also from the preceding and following epochs, the predicted class was more smooth, e.g. 280-320 SWS epoch; on the other hand, the result without overlapping epochs predicted quick transitions, such epochs were not correctly predicted by the overlapping method, e.g. transition from SWS to S2 at 100. Overall, the overlap method performed slightly better than no overlap epoch based feature extraction.

Table IV, V, and VI depict the confusion matrix obtained from the classification results with multi-scale entropy and SEF features by overlapped epoch obtained from the *SEDF-W* dataset, the *SEDFx-S* dataset and the *SEDFx-W* dataset, respectively. Since the majority of epochs were labeled as the wake (W) condition in the *SEDFx-W* dataset, the sensitivity of the W stages was approximately 10% higher than that of the *SEDFx-S* dataset. Therefore, the accuracy for the *SEDFx-W* dataset became higher.

TABLE III
CLASSIFICATION ACCURACY WITH 10-FOLD CV FOR $C = 5$ CLASS WITH 2 EEG CHANNELS (NO OVERLAPPING[%]/OVERLAPPING[%])

Dataset	No of features	*SEDFx-S* (Fpz and Pz)	*DREAM* (O1 and Cz(or C3))
MSFE	60	81.2/84.7	80.4/83.2
MSPE	40	81.2/83.5	80.0/83.8
MSFE + SEF	102	85.5/87.9	84.1/86.5
MSPE + SEF	82	86.2/**88.6**	84.3/**86.8**

C. Different class accuracy

Table VII shows the sensitivity and precision of the classification for different sleep stages $\mathcal{C} = 2 - 6$ with MSPE and SEF features by overlapped epoch obtained from the *SEDFx-S* dataset. For the all 2- to 6-stages classification, the Kappa values achieved almost perfect agreement, as 0.86, 0.88, 0.86, 0.84, and 0.81, respectively.

D. Comparison with previous approaches

Table VIII shows the S1 sleep stage sensitivity of the proposed methods. In order to evaluate the performance, the *SEDF-W*, the *SEDFx-S*, and the *SEDFx-W* were compared to the results in [11, 12, 25], [10], and [13], respectively. The sensitivities of the S1 stage are 51.7%, 57.8%, 49.1% for the *SEDF-W*, *SEDFx-S*, and *SEDF-W* datasets, respectively.

TABLE IV
SEDF-W - CONFUSION MATRIX FOR $C = 5$-STAGE SLEEP CLASSIFICATION USING MSFE AND SEF FEATURES WITH OVERLAPPING WINDOWS EXTRACTED FROM THE FPZ CHANNEL

		Algorithm W	S1	S2	SWS	REM	S(%)/*P(%)*
Reference	W	7841	60	18	1	10	98.9/*98.5*
	S1	65	300	127	4	84	**51.7**/*61.1*
	S2	27	75	3330	115	47	92.7/*90.3*
	SWS	7	1	150	1128	0	87.7/*90.4*
	REM	18	55	62	0	1454	91.5/*91.2*
Kappa κ = 0.90, Accuracy = 93.8% (S:Sensitivity, *P:Precision)*							

TABLE V
SEDFx-S - CONFUSION MATRIX FOR $C = 5$-STAGE SLEEP CLASSIFICATION USING MSPE AND SEF FEATURES WITH OVERLAPPING WINDOWS EXTRACTED FROM THE FPZ AND PZ CHANNELS

		Algorithm W	S1	S2	SWS	REM	S(%)/*P(%)*
Reference	W	5559	489	150	21	124	87.6/*87.0*
	S1	545	2699	926	13	487	**57.8**/*68.6*
	S2	149	490	24951	810	442	93.0/*90.3*
	SWS	35	7	1161	7495	2	86.2/*89.9*
	REM	104	248	452	3	10939	93.1/*91.2*
Kappa κ = 0.84, Accuracy = 88.6% (S:Sensitivity, *P:Precision)*							

TABLE VI
SEDFx-W - CONFUSION MATRIX FOR $C = 5$-STAGE SLEEP CLASSIFICATION USING MSFE AND SEF FEATURES WITH OVERLAPPING WINDOWS EXTRACTED FROM THE FPZ CHANNEL

		Algorithm W	S1	S2	SWS	REM	S(%)/*P(%)*
Reference	W	73419	448	199	24	150	98.9/*98.3*
	S1	758	2312	1009	10	620	**49.1**/*65.4*
	S2	275	452	24929	845	563	92.1/*89.0*
	SWS	55	7	1213	7530	7	85.5/*89.5*
	REM	187	319	652	2	10592	90.1/*88.8*
Kappa κ = 0.90, Accuracy = 93.8% (S:Sensitivity, *P:Precision)*							

TABLE VII
SEDFx-S - SENSITIVITY AND *Precision* (*in italic*) FOR SLEEP CLASSIFICATION USING MSPE AND SEF WITH OVERLAPS

(%)	$\mathcal{C} = 2$	$\mathcal{C} = 3$	$\mathcal{C} = 4$	$\mathcal{C} = 5$	$\mathcal{C} = 6$
W	85.8 *89.4*	86.9 *88.4*	86.7 *87.8*	87.6 *87.0*	87.6 *86.8*
S1				57.8 *68.6*	57.9 *68.1*
S2		96.3 *96.3*	92.8 *91.6*	93.0 *90.3*	92.9 *90.2*
S3	98.8 *98.3*				64.9 *69.6*
S4			86.1 *90.1*	86.2 *89.9*	83.7 *85.2*
REM		92.4 *91.7*	92.2 *91.8*	93.1 *91.2*	93.1 *91.2*
Acc	97.4	94.5	91.0	88.6	86.6
κ	0.86	0.88	0.86	0.84	0.81

TABLE VIII
PERFORMANCE COMPARISON OF S1 STAGE SENSITIVITY(%) AMONG EXISTING METHODS FOR $C = 5$-STAGE SLEEP CLASSIFICATION

Zhu *et al.* [11]	15.8	Imtiaz *et al.* [10]	29.8	Silveria *et al.* [13]	6.1
Hassan *et al.* [12]	47.0				
Hassan *et al.* [25]	37.4				
Proposed *SEDF-W*	**51.7**	*SEDFx-S*	**57.8**	*SEDFx-W*	**49.1**

The proposed method was able to distinguish the S1 stage better than other methods; however, discriminating the S1 stage still remains challenging. The sensitivity of the S1 is smaller than the sensitivities for the other sleep conditions. Table IX depicts the accuracies of various methods utilised for

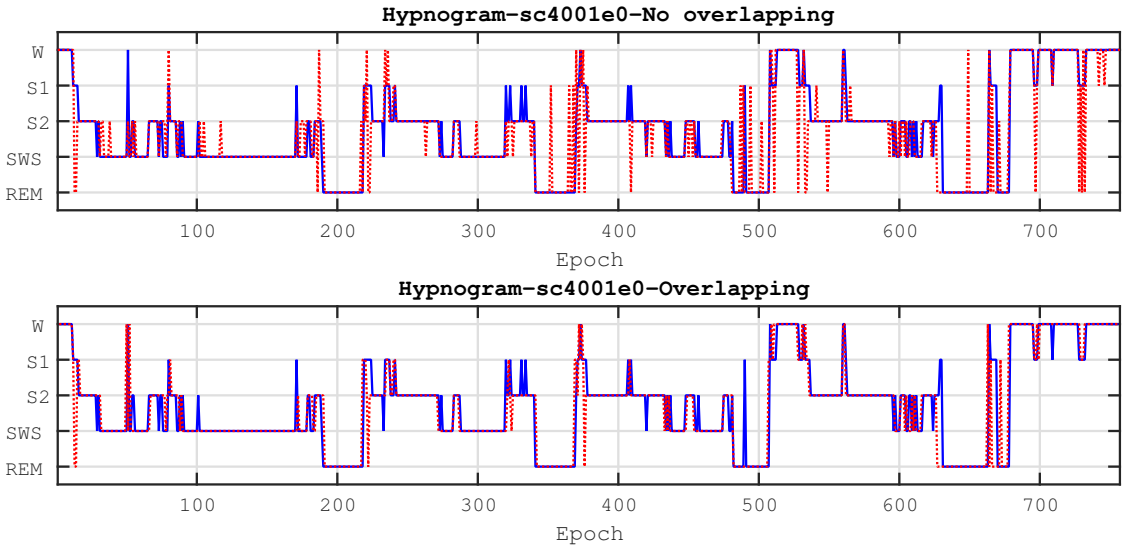

Fig. 5. Hypnogram (blue) and the predicted label (red) of subject *sc4001e0* in the *SEDFx-S* dataset for the $C = 5$ class classification problem

a different SleepEDF database. The accuracies were highest for the proposed method in $C = 2 - 6$ sleep state classification, except $C = 2$ stage classification in the *SEDF-W* dataset. Overall, the proposed method outperformed almost all other methods, regardless of the dataset extracted from the SleepEDF database.

V. Conclusion

We have investigated the structural complexity analyses of sleep EEG signals in sleep stage classification. This has been achieved by using the MSFE and MSPE, instead of using only spectral indices for sleep stage analysis. From the analyses using both the SleepEDF and DREAMS Subjects databases, we have found that the proposed features could be effectively utilised for classifying sleep stages from a limited number of EEG channels. For the *SEDFx-W* dataset, the proposed method was able to classify 93.8% of 5 sleep stages with the Kappa coefficient of 0.90 from a single channel EEG. Additionally, it has been found that multi-scale entropy has been able to improve the performance of discriminating the S1 sleep stages, which has been the most challenging for automatic sleep stage classification tasks.

References

[1] M. A. Carskadon and W. C. Dement, "Chapter 2 – Normal Human Sleep: An Overview," *Principles and Practice of Sleep Medicine, 4th edition*, pp. 13–23, 2005.

[2] P. Maquet, "The Role of Sleep in Learning and Memory," *Science*, vol. 294, no. 5544, pp. 1048–1052, 2001.

[3] L. Xie, H. Kang, Q. Xu, M. J. Chen, Y. Liao, M. Thiyagarajan, J. O'Donnell, D. J. Christensen, C. Nicholson, J. J. Iliff, T. Takano, R. Deane, and M. Nedergaard, "Sleep drives metabolite clearance from the adult brain," *Science*, vol. 342, no. 6156, pp. 373–377, 2013.

[4] A. Sehgal and E. Mignot, "Genetics of sleep and sleep disorders," *Cell*, vol. 146, no. 2, pp. 194–207, 2011.

[5] A. Rechtschaffen and A. Kales, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. 1968.

[6] C. Iber, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2007.

[7] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, and I. Provazník, "Sleep scoring using artificial neural networks," *Sleep Medicine Reviews*, vol. 16, no. 3, pp. 251–263, 2012.

[8] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131–148, 2000.

[9] PhysioNet, "The Sleep-EDF Database." https://www.physionet.org/physiobank/database/sleep-edf/. [Online; accessed 14-Nov-2016].

[10] S. A. Imtiaz and E. Rodriguez-Villegas, "Automatic sleep staging using state machine-controlled decision trees," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pp. 378–381, 2015.

[11] G. Zhu, Y. Li, and P. P. Wen, "Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1813–1821, 2014.

[12] A. R. Hassan and M. I. H. Bhuiyan, "Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating," *Biomedical Signal Processing and Control*, vol. 24, pp. 1–10, 2016.

[13] T. L. T. da Silveira, A. J. Kozakevicius, and C. R. Rodrigues, "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in

TABLE IX
PERFORMANCE COMPARISON AMONG EXISTING METHODS FOR THE SLEEPEDF DATABASE [ACCURACY(%)/KAPPA]

Dataset	Methods	No of epochs	EEG ch.	$C = 2$	$C = 3$	$C = 4$	$C = 5$	$C = 6$	γ
SEDF-W	Zhu *et al.* [11]	14963	Pz	97.9/0.96	92.6/0.87	89.3/0.83	88.9/0.83	87.5/0.81	
	Hassan *et al.* [12]	15188	Pz	**99.5**/-	94.1/-	92.1/-	90.7/-	86.9/-	
	Hassan *et al.* [25]	15188	Pz	97.5/0.95	94.8/0.91	92.1/0.87	91.5/0.86	90.4/0.84	
	MSFE + SEF	**14979**	**Fpz**	98.8/**0.98**	**97.1/0.95**	**95.3/0.93**	**93.8/0.90**	**93.0/0.89**	1
			Pz	98.7/0.97	**97.1/0.95**	94.8/0.92	93.6/**0.90**	92.8/**0.89**	1
	MSPE + SEF		**Fpz**	98.6/0.97	97.0/**0.95**	94.8/0.92	93.4/**0.90**	92.6/0.88	4
			Pz	98.5/0.97	96.9/**0.95**	94.4/0.91	93.3/0.89	92.1/0.88	4
SEDFx-S	Imtiaz *et al.* [10]	59316	Fpz, Pz	-/-	-/-	-/-	82.2/-	-/-	
	MSFE + SEF	**58301**	**Fpz, Pz**	97.2/0.85	**94.5/0.88**	90.6/0.85	87.9/0.83	85.8/0.80	1
	MSPE + SEF			**97.4/0.86**	**94.5/0.88**	**91.0/0.86**	**88.6/0.84**	**86.6/0.81**	0.6
SEDFx-W	Silveria *et al.* [13]	106376	Pz	97.3/0.94	93.9/0.87	92.3/0.84	91.5/0.83	90.5/0.80	
	MSFE + SEF	**126577**	**Fpz**	**98.5/0.97**	**96.7/0.94**	**95.1/0.91**	**93.8/0.90**	**92.9/0.88**	1
			Pz	**98.5/0.97**	96.6/**0.94**	94.7/**0.91**	93.6/0.89	92.6/**0.88**	1
	MSPE + SEF		**Fpz**	98.3/**0.97**	**96.7/0.94**	94.7/**0.91**	93.5/0.89	92.4/0.87	4
			Pz	**98.5/0.97**	96.6/**0.94**	94.5/0.90	93.3/0.89	92.2/0.87	4

wavelet domain," *Medical & Biological Engineering & Computing*, pp. 1–10, 2016.

[14] S. F. Liang, C. E. Kuo, Y. H. Hu, Y. H. Pan, and Y. H. Wang, "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1649–1657, 2012.

[15] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms," *Journal of Medical Systems*, vol. 38, no. 3, pp. 1–21, 2014.

[16] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel, M. R. Pressman, and C. Iber, "The visual scoring of sleep in adults," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 121–131, 2007.

[17] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, p. 068102, 2002.

[18] M. U. Ahmed and D. P. Mandic, "Multivariate multiscale entropy: A tool for complexity analysis of multichannel data," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 84, no. 6, pp. 1–10, 2011.

[19] Y. Tonoyan, D. Looney, D. P. Mandic, and M. M. Van Hulle, "Discriminating multiple emotional states from EEG using a data-adaptive, multiscale information-theoretic approach," *International Journal of Neural Systems*, vol. 26, no. 02, p. 1650005, 2016.

[20] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface EMG signal based on fuzzy entropy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 2, pp. 266–272, 2007.

[21] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series.," *Physical Review Letters*, vol. 88, no. 17, p. 174102, 2002.

[22] F. C. Morabito, D. Labate, F. La Foresta, A. Bramanti, G. Morabito, and I. Palamara, "Multivariate multiscale permutation entropy for complexity analysis of Alzheimer's disease EEG," *Entropy*, vol. 14, no. 7, pp. 1186–1202, 2012.

[23] S. Devuyst, "The DREAMS Subjects Database." http://www.tcts.fpms.ac.be/~devuyst/Databases/DatabaseSubjects/. [Online; accessed 14-Nov-2016].

[24] PhysioNet, "The Sleep-EDF Database [Expanded]." https://physionet.org/pn4/sleep-edfx/. [Online; accessed 14-Nov-2016].

[25] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *Journal of Neuroscience Methods*, vol. 271, pp. 107–118, 2016.

[26] S. A. Imtiaz and E. Rodriguez-villegas, "An open-source toolbox for standardized use of PhysioNet Sleep EDF Expanded Database," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pp. 6014–6017, 2015.

[27] G. Bremer, J. R. Smith, and I. Karacan, "Automatic detection of the K-complex in sleep electroencephalograms," *IEEE Transactions on Biomedical Engineering*, vol. BME-17, no. 4, pp. 314–323, 1970.

[28] S. A. Imtiaz and E. Rodriguez-Villegas, "A low computational cost algorithm for REM sleep detection using single channel EEG," *Annals of Biomedical Engineering*, vol. 42, no. 11, pp. 2344–2359, 2014.

[29] J. W. Sleigh and J. Donovan, "Comparison of bispectral index, 95% spectral edge frequency and approximate entropy of the EEG, with changes in heart rate variability during induction of general anaesthesia," *British Journal of Anaesthesia*, vol. 82, no. 5, pp. 666–671, 1999.

[30] C.-C. Chang and C.-J. Lin, "LIBSVM : A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[31] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.