Reliable Sleep Staging of Unseen Subjects with Fusion of Multiple EEG Features and RUSBoost

Ritika Jain, Ramakrishnan Angarai Ganesan

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

Abstract

Extensive experiments have been carried out in this study to classify sleep EEG from three different standard databases - Sleep EDF, DREAMS and Expanded sleep EDF databases. Both two-class (sleep-awake) and multiclass classifications have been performed using a fusion of various EEG features and an ensemble classifier called random undersampling with boosting technique (RUSBoost). The results achieved using a single channel EEG are comparable or better than the state-of-the-art methods in the literature for both types of classification, on all the databases. Two-class classification is useful to determine the preferred timings for sensory stimulation of patients with disorders of consciousness. 10-fold cross-validation accuracies of 92.6% and 97.9% have been obtained on Sleep EDF database for 6-class and 2-class problems, respectively. Using Expanded Sleep-EDF dataset, the accuracies improved to 96.3% for 6-state and 99.8% for 2-state classification. For DREAMS dataset, we achieved an accuracy of 96.6% for 2-state classification. Unlike most research in the literature where performance on unseen subjects is not considered, we report classification results on the data from unseen test subjects using both 50%-holdout and leave-one-out cross-validation approaches. Similar results were achieved using both validation techniques for different datasets emphasizing the reliability of our method. These results are very crucial for the method to be applicable for clinical use on new patients.

Email addresses: ritikajain@iisc.ac.in (Ritika Jain), agr@iisc.ac.in (Ramakrishnan Angarai Ganesan)

Keywords: AR model, band power ratios, disorders of consciousness, DWT, EEG, Higuchi fractal dimension, Hurst exponent, LZC, sample entropy, sleep staging

1. Introduction

Sleep is extremely crucial for the maintenance of physical and mental health of human beings. Sleep deprivation or poor quality of sleep can lead to numerous health problems, depression or even death [1]. In order to analyze the various sleep-related disorders, the quality of sleep is evaluated using polysomnography (PSG) which utilizes multiple signals such as electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG) and electromyogram (EMG). Sleep scoring is performed by the sleep experts based on the visual analysis of the PSG signals. However, this manual scoring is tedious, costly and time-consuming. An automated system may be more efficient for sleep scoring since it can reduce the time and cost involved in sleep stage classification. Based on Rechtschaffen & Kales's (R&K) scoring criteria, sleep is categorized into rapid eye movement (REM) stage, four non-rapid eye movement (NREM) stages namely S1, S2, S3 and S4, and awake stage [2]. Each of these stages is characterized by a specific signature in terms of the corresponding EEG, EOG and/or EMG patterns.

In 2007, American Academy of Sleep Medicine (AASM) proposed a new sleep scoring standard, which merged S3 and S4 stages into a single stage [3]. According to AASM standard, there are only three NREM sleep stages namely N1, N2 and N3, while REM and wake stages are the same as in R&K standard. N1 stage is essentially a changeover from wakefulness to sleep, with slowing down of heartbeat, breathing, eye movements and even brain waves [1]. As the sleep deepens in the N2 and N3 stages, all the physiological signals further slow down. N2 stage is characterised by the presence of micro-structures such as sleep spindles, K-complexes, or both [4]. N3 is the deepest stage of sleep prominently characterized by the low-frequency delta waves. REM stage is very similar to the awake state, also referred to as the dream stage [5]. Many researchers are working towards accurate and automatic sleep stage classification [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. Some of these works have utilized temporal features or spectral information, while others have considered time-frequency analysis using wavelet transforms or Wigner-Ville distribution (WVD) to extract both the time and frequency information embedded in the EEG signal [7, 9, 11, 20, 23, 30]. In [20], an analysis of covariance matrices of the wavelet decomposition of five EEG channels is utilized and an overall accuracy of about 83% is achieved for 5-stage sleep classification. Bajaj et. al. [7] employed time-frequency image of the EEG signals using pseudo WVD for the classification of sleep stages into 5 or 6 classes. They achieved overall accuracies of 88.5% and 92.9% for 6-stage and 5-stage classifications, respectively, using 10-fold cross-validation. However, this study considered only a small subset of the Sleep-EDF database (4700 out of 15188 epochs). Therefore, the results obtained are based on limited data, and need to be further validated by including the entire dataset (i.e. 8 subjects) as well as considering a larger database.

Various time-domain decomposition techniques such as empirical mode decomposition (EMD) and its variants like ensemble EMD (EEMD) have also been utilized in a few studies [9, 10, 11, 31]. Hassan et. al. [31] obtained an overall accuracy of 88.6% using EMD and statistical features for 6-state classification. In [11], the authors applied EEMD in conjuction with random undersampling boosting (RUSBoost) classification technique and obtained 88% accuracy for 6-state classification. Another study by the same authors utilized complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) which provided an accuracy of 86.89% [10]. Further improvement in classification accuracy (89.6%) was achieved by using tunable Q-factor wavelet transform (TQWT) with adaptive boosting (Adaboost) [32]. This method provided overall accuracies of 89.6%, 90.8%, 91.5%, 93.9% and 97.2% for 6-stage to 2-stage classifications, respectively. With a similar notion of EMD, Sharma et. al. [22] employed a method based on iterative filtering of EEG signals to obtain modes referred to as AM-FM (amplitude-modulated and frequency-modulated) components. They obtained overall classification accuracies ranging from 90% to 98% for various classifications involving multiple classes.

Non-linear features such as complexity, fractal dimension, sample entropy, higher order spectra, Lyapunov exponent, and Lempel-Ziv complexity have also been employed by several studies to distinguish between different sleep stages [33, 9, 17, 18, 16, 15, 25]. Liang et. al. [18] proposed a method based on multi-scale entropy and autoregressive modelling for sleep scoring. This study could achieve an overall accuracy of around 83% by using a single EEG channel, linear classifier (LDA) and post-processing based on smoothing. Fraiwan et. al. [9] employed time-frequency analysis and entropy measures to derive EEG features for 5-stage classification as per AASM standard. An accuracy of 83% could be achieved with the random forest classifier. Ganesan et. al. [16] achieved an accuracy of around 94% for 2-state classification by using Lempel-Ziv complexity and spectral band power ratios. Another study by the same authors obtained 98% accuracy for sleep-wake classification by using features from both EOG and EEG channels [15]. Lajnef et. al. [17] utilized multiple physiological signals, namely EEG, EOG and EMG to derive various linear and nonlinear features such as variance, skewness, kurtosis, linear prediction error energy, spectral power and power ratios, teager energy operator and permutation entropy. They obtained an overall accuracy of 88% for 5 classes using dendogram-based SVM (DSVM). However, all the above mentioned studies did not report the results of various multi-class (i.e. 2-, 3-, 4-, 5- and 6-class) classification problems.

Some researchers have particularly focused only on the detection of REM sleep stage [34, 35]. Imtiaz et. al. [35] proposed a new feature called spectral edge frequency (SEF) and used it along with the absolute and relative power in 8-16 Hz band for the detection of REM stage. They achieved a sensitivity of 81% and specificity of 75% for REM detection on the sleep data of 8 subjects. However, this work did not consider identification of multiple sleep stages. Similarly, Agrawal et. al. [34] proposed a REM sleep detection scheme with the help of two EOG channels with minimal parameter adjustments. They achieved an overall sensitivity of 67.2% and specificity of 77.5% for the test data on five subjects.

Phan et. al. [36] proposed a metric learning approach for the classification of various sleep stages. This approach could outperform many existing methods by utilizing Mahalanobis distance metric instead of the default Euclidean metric in kNN (K-nearest neighbors). They achieved an overall accuracy of 98.3% and 94.5% for 2-state and 4-state classifications, respectively. However, this study considered only four (all healthy) out of eight subjects' data (excluding the other four with sleeping difficulties) from sleep-EDF database. This requires further validation of the results on a larger database. Further, they merged S1 and REM stages together; resulting in a higher 4-state classification accuracy because S1 and REM are generally the most difficult stages to score accurately. A study by Liu et. al. [37] utilized multi-domain analysis of EEG signals to extract various features for sleep stage identification. The feature extraction process involved multifractal detrended fluctuation analysis (DFA), visibility graph algorithms, frequency analysis as well as non-linear analysis and the classifier used in this study was least-square SVM. However, they have considered only 6-class classification and data from limited subjects.

Some of the studies in the literature have employed neural network and deep learning techniques to classify multiple sleep stages. Hsu et. al. [12] utilized energy features derived from an EEG channel and a recurrent neural classifier. Accuracy of about 87% is obtained for 5-stage sleep classification. Dong et. al. [38] proposed a mixed neural network approach by combining the rectifier neural network and long short-term memory (LSTM) network to achieve optimal classification performance. Further, their aim was to minimize the number of channels without losing much information and they concluded that an EOG and a frontal EEG channel can provide sufficiently high accuracy for sleep stage classification. Tsinalis et. al. [27] incorporated convolutional neural networks (CNN) for sleep scoring without considering any prior domain knowledge. This work could achieve an overall 5-class accuracy of 74% across 20 healthy subjects. Ronzhina et. al. [39] used artificial neural networks for multi-class classification and obtained accuracy values of 76.4%, 81.6%, 88.9% and 96.9% for 6-class, 4-class, 3-class and 2-class classifications, respectively. In [40], Chambon et. al. employed deep learning approach that exploited multimodal and multivariate signals (EEG, EMG and EOG) without the need of any handcrafted features. They showed that the temporal context of these signals can be exploited to further improve the classification performance. This study considered six EEG (F3, F4, C3, C4, O1, O2), two EOG (left and right) and three EMG (chin) channels for 5-state sleep classification.

Andreotti et. al. [41] proposed a transfer learning approach in which the model can be trained on a large public database and then fine-tuned on each subject. This study compared various existing CNN approaches on four different databases. They demonstrated that the model's performance on a smaller and more challenging dataset can be improved by utilizing the technique of transfer learning. Phan et. al. [42] utilized deep bidirectional recurrent neural network (RNN) with attention mechanism for sleep-staging using a single EEG channel. They used this trained network only for extracting features while classification is performed using a linear SVM. A recent study by the same author [43] proposed a hierarchical RNN called SeqSleepNet utilizing multichannel time-frequency image to classify a sequence of multiple epochs at a go. This work could establish a standard baseline outperforming most state-of-the-art methods including DeepSleepNet [24] and [25] on a massive dataset comprising 200 subjects. The latest study [44] along the similar notion of transfer learning approach introduced a novel framework called MetaSleepLearner to assist the clinicians in sleep scoring. It utilized a massive dataset for pre-training the network and fine-tuned it to new subjects from other cohorts by using only a few samples from each subject. This approach resulted in a new benchmark for semi-automated sleep-stage classification. A recent study by Zhang et. al. [45] proposed orthogonal convolutional neural network (OCNN) in which they used Hilbert-Huang transform to convert the time series EEG signal into 2-D time frequency image followed by dimensionality reduction by the use of autoencoder. This study used two different datasets in which they achieved classification accuracy of 87-88%. But, since they did not evaluate their method on the commonly reported dataset (Sleep-EDF dataset), it is difficult to compare their performance with the rest of the studies.

Wavelet coefficients and artificial neural networks were utilized by Ebrahimi et. al. [30] to discriminate four sleep stages, namely awake, S1+REM, S2 and SWS (slow wave sleep i.e. S3+S4). They obtained 93% accuracy using a single EEG channel. However, they considered data of only seven subjects and also merged S1 and REM together (i.e. REM and an NREM stage) into a single stage, which are often the most difficult sleep stages to distinguish. Therefore, a high accuracy could be achieved, but it would be fair if the distinction between REM and S1 stage is considered in a 4-stage sleep classification.

Among the various PSG signals, EEG is considered to be the most informative for analyzing the different states of the brain. However, relying only on EEG signals to differentiate between the N1 and wake stage or REM stage is difficult, since N1 stage marks the transition from wakefulness to sleep and REM possesses patterns quite similar to the awake state. Therefore, a high percentage of the epochs belonging to REM or awake or N1 are generally misclassified. Utilizing EMG or EOG signal along with the EEG can further improve the classification accuracy. Hence, some studies have included EOG and EMG channels along with the EEG channels to achieve a better classification performance, especially for the N1 and REM stage [17, 40, 46].

In this study, we have proposed a single-channel EEG-based automatic sleep-stage identification method (MEFF-R) by utilizing a wide variety of features and an ensemble classifier. The features are chosen such that they can characterize most of the aspects of the EEG signal such as irregularity, frequency and temporal information, entropy or periodicity. Further, in order to tackle the problem of class-imbalance, we have utilized a hybrid sampling-boosting technique i.e. random undersampling and adaptive boosting with an ensemble classifier (RUSBoost). The promising results of the proposed method show that MEFF-R can serve as a potential alternative to the tedious, time-consuming and expensive, manual sleep scoring procedure.

The major highlights of our work are: 1) the set of features and classifier utilized in this method are able to outperform most of the recent works in the literature; 2) better S1 detection accuracy than most of the existing studies; 3) high classification accuracies across three different datasets using various validation techniques i.e. 10-fold crossvalidation, leave-one-subject-out (LOSO) and 50%-holdout validation, and 4) promising results for the prediction of sleep stages on unseen subjects. In this work, our objective is to study various features that can capture the different aspects of the EEG signal and utilize a combination of such features to provide a high classification accuracy among the different sleep stages, especially on subjects unseen by the classifier during training. The rest of the paper is structured as follows: The proposed method is explained in detail in section 2, followed by the experimental results and their comparison with the previous studies in section 3. Discussion is provided in section 4 and section 5 concludes the paper.

2. Materials and Method

The flowchart of the proposed multiple EEG feature fusion with RUSBoost (MEFF-R) method is shown in figure 1. In this work, we have not considered EMG or EOG signals and utilized only the EEG signal. We have examined the performance of the proposed method in two scenarios: a) by considering only a single EEG channel and b) combination of two or three EEG channels. Further, unlike most research work reported in the literature, we have carried out subject-independent testing, in addition to subject-dependent testing. Reliable performance on test data from subjects not seen by the classifier is important to conclude that the technique is usable in real-life clinical setting on hitherto unseen subjects (patients).

2.1. Datasets used for the experiments

We have performed experiments on three different publicly available and widely used datasets, namely Sleep-EDF, Expanded Sleep-EDF, and DREAMS Subjects databases.

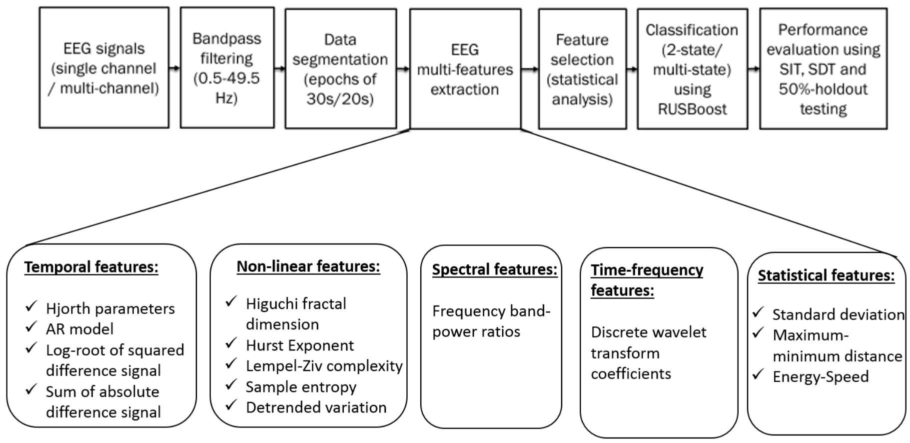

Figure 1: Flowchart of the proposed Multiple EEG feature fusion with RUSBoost (MEFF-R) technique for sleep staging

2.1.1. Sleep-EDF database

The first dataset used in this study is Sleep-EDF database from Physionet [47]. It has overnight sleep data from 8 subjects, four healthy and four with mild difficulty in falling asleep. These recordings consist of horizontal EOG, Fpz-Cz and Pz-Oz EEG channels sampled at 100 Hz. Each 30s epoch has been scored by experts as per R&K scale [2].

2.1.2. DREAMS Subjects database

This database comprises whole night polysomnographic recordings from 20 healthy subjects (16 females and 4 males). It consists of at least two EOG channels (P8-A1, P18-A1), three EEG channels (Cz-A1 or C3-A1, FP1-A1 and O1-A1) and one submental EMG channel, all sampled at 200 Hz [48]. The sleep stage annotations have been provided according to both R&K and AASM criteria on the basis of 20 and 30 second epochs, respectively.

2.1.3. Expanded Sleep-EDF database

The third dataset used in this study is the Expanded Sleep-EDF database which is also available on Physionet [47]. This is an expanded version of Sleep-

EDF database, containing 197 PSG sleep recordings, 153 of them from healthy subjects and 44 from people with mild difficulty falling asleep. These recordings have been scored on the basis of 30s epochs according to R&K manual. In this study, we have considered the EEG recordings of 30 subjects and the corresponding epoch distribution is shown in Table 1.

The sleep stages according to R&K criteria are: Wake, S1, S2, S3, S4 and REM (rapid eye movement) while as per AASM scoring rule, the stages are classified as: Wake, N1, N2, N3 and REM. Table 1 presents the total number of epochs analyzed for different sleep/wake stages for each of the three datasets.

Table 1: Sleep stage-wise distribution of the number of EEG epochs considered for experimentation in this study from each standard dataset.

Database	**Wake**	**REM**	**S1**	**S2**	**S3**	**S4**	**Total**
Sleep-EDF	8055	604	3621	672	627	1609	15188
DREAMS	5601	4555	1788	13274	2112	2929	30259
Expanded Sleep-EDF	55949	2174	13235	2430	1912	5858	81558

2.2. Extraction of different features

The raw EEG signal is preprocessed using an 8^{th} order butterworth IIR bandpass filter with the passband from 0.5 to 49.5 Hz. These cut-off frequencies are chosen in order to remove the DC, some slow drifts and the line noise (50 Hz). This signal is then segmented into 30 s or 20 s epochs (depending upon the ground truth format), from which the epochs with movement or/and no score are removed. From these segmented epochs, different features are extracted as explained below, and provided as inputs to the classifier for training. Various multi-domain features are chosen in this work so as to capture the different aspects of the sleep EEG signal.

In order to measure the temporal variations of the signal, we have utilized autoregressive modelling and Hjorth parameters, whereas frequency contents are studied using band power ratios. In general, sleep is dominated by low frequency signals (delta or theta), while wake or light sleep states are governed by high frequency bands such as alpha or beta. Hence, the ratios between powers in these frequency bands are robust markers of the various sleep stages and are utilized primarily in all the studies on sleep-stage classification. Further, to obtain time-frequency representation of the EEG signals, we have used discrete wavelet transform (DWT). EEG signal is highly non-linear in nature and its complexity varies across different sleep stages. Therefore, we calculated the Lempel-Ziv complexity (LZC) measure, which has been used in our earlier work to detect sleep and wake states of the brain [16]. Along with LZC, we have used other non-linear measures like sample entropy and Higuchi fractal dimension to capture the irregularity and structure of the signal at multiple scales. Both these non-linear features have proven their ability to classify sleep and wake stages in our previous work [15]. Some of the statistical features successfully employed in previous studies [19, 49] are also utilized in this work. A detailed description of all these features is given below.

2.2.1. Autoregressive (AR) model parameters

An autoregressive model predicts the current value of a signal based on its past values [50]. This is the most frequently used technique in linear predictive modelling of time series. The number of past samples used for prediction determines the order of the model. We have used AR model of order 8 to predict the EEG samples of each individual epoch.

$$x(n) = \sum_{j=1}^{p} a_j x(n-j) \tag{1}$$

where p is the order of the AR model and a_j are the AR coefficients, which are used as the features.

2.2.2. Higuchi fractal dimension (HFD)

Fractal dimension is an important characteristic of a system, because it contains information about the geometrical structures at multiple scales [51]. Higuchi's algorithm [52] is popular for calculating the fractal dimension. In order to obtain Higuchi's fractal dimension, a new set of series is first generated from the original time series, defined as follows:

$$X_k^m : x(m), x(m+k), x(m+2k), ...x(m + \lfloor \frac{N-m}{k} \rfloor k) \tag{2}$$

with $m = 1, 2, ...k$ and $\lfloor . \rfloor$ is the floor operation. Then the length of the curve associated with each of the newly generated time series X_k^m is calculated as,

$$L_k^m = \frac{1}{k} \Big(\sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} \mid x(m+ik) - x(m+(i-1)k) \mid \Big) \Big(\frac{N-1}{\lfloor \frac{N-m}{k} \rfloor k} \Big) \tag{3}$$

$$\langle L(k) \rangle \propto k^{-D} \tag{4}$$

The average value $\langle L(k) \rangle$ of the lengths associated with the set of time series follows a power law given by (4), where 'D' is the value of Higuchi fractal dimension. We have calculated HFD values for various frequency bands listed in Table 2.

2.2.3. Sample entropy (SE)

Entropy quantifies the irregularity in the signal based on the series of patterns embedded in it. Since entropy is an estimate of the degree of randomness, its value is higher when the sequences in a series are less ordered. For instance, the value of entropy would be lesser in sleep stages than in wake state. Sample entropy has also been utilized to monitor the depth of anaesthesia of patients during surgery [53]. Sample entropy (SE) has been shown to perform better than the approximate entropy (ApEn) [54]. Also, unlike ApEn, sample entropy is easier to implement and is independent of the length of data. Therefore, SE is favored in many studies involving stationary time series analysis [53, 55]. SE is defined as:

$$S_E(m, r, N) = -\ln \left[C_{m+1}(r)/C_m(r) \right] \tag{5}$$

where C_m is given by

$$C_m(r) = \frac{\{\text{Number of pairs } (i,j) \text{ with } \mid x_i^m - x_j^m \mid < r, i \neq j\ \}}{\{\text{Number of all probable pairs}\}} \tag{6}$$

The parameters m, r and N are the length of sub-series, tolerance for accepting matches and the total number of samples in the series, respectively. The values of the parameters used in this study, chosen based on the previous studies [14, 55], are $m = 2$ and $r = 0.15$ times the standard deviation of the original time series.

2.2.4. Lempel-Ziv complexity (LZC)

It is based on coarse-graining of the measurements, which means that the raw signal is transformed into a new time-series with very few symbols as the elements of the series [56]. LZC measures the number of distinct patterns embedded in the given sequence. In order to calculate LZC, we first convert the time-series into a binary sequence by thresholding it with respect to the mean value of the epoch. Then, we evaluate the number of distinct patterns contained in that sequence. It is also normalized to make it independent of the sequence length. This EEG complexity measure has already proved its ability to assess the depth of anaesthesia and it is computable in real time [57]. Hence, it can be a good feature to characterize the different sleep stages.

The following procedure has been adopted to calculate the values of LZC corresponding to the EEG signal.

Step 1: Convert EEG samples $\{X_i|i = 1, 2, 3....n\}$ into a binary sequence by setting the threshold as the mean of the samples, $X_m = (1/n)\sum_{i=1}^{n} X_i$, where n is the length of the epoch.

Step 2: The new sequence $Y = s_1, s_2, s_3....s_n$ obtained in the previous step is used to compute the distinct patterns embedded in it. This requires comparison of the present subsequence with the preceding one. If they are distinct, the complexity counter C is incremented.

Step 3: Normalisation of the counter C by $n/log_2(n)$. This gives us the value of LZC corresponding to the EEG epoch.

The values of LZC capture the information of different patterns across multiple sleep stages and therefore help in distinguishing amongst the sleep and wake stages.

2.2.5. Discrete wavelet transform (DWT) coefficients

Wavelet transform provides time-frequency information about the signal, unlike the Fourier transform which provides only the frequency representation of the signal. DWT represents the signal as the linear combination of dilated and translated versions of a chosen basis function, known as the mother wavelet. DWT applies a low-pass and a high-pass filter at each level of decomposition resulting in approximation and detail signals. We perform 4-level wavelet decomposition of each 30s epoch of the EEG signal using order-2 Daubechies wavelet. The detail coefficients (D_1 to D_4) contain high frequency content of the signal while the approximation coefficient (A_4) contains the low frequency information. Finally, we calculate the minimum, maximum, mean and standard deviation for each of the wavelet coefficients. Hence, for each coefficient (one approximation and four detail coefficients), we have four corresponding values which gives us a total of 20 features (referred to as $\text{DWT}_1-\text{DWT}_{20}$).

2.2.6. Band power ratios

Features such as ratios of powers in different frequency bands can provide useful information regarding the sleep stages. For instance, low frequency delta waves are dominant in the deep sleep stages like S3 and S4, while in waking state, higher frequency bands such as alpha or beta waves are dominant. We have employed Welch's method to estimate the power spectral density for each epoch and computed power in the different frequency bands listed in Table 2.

The following eight ratios of powers in the different frequency bands have been computed and added to the list of features used.

Table 2: EEG bands considered in this study and their frequency ranges in Hz. Sp Band refers to the spindle band.

Band	δ	θ	α	β	γ	Sp Band	Band-I	Band-II
Freq	1-4	4-8	9-12	13-30	30-49.5	11.5-16	0.5-3.5	3.5-8

1. Alpha to delta power ratio :

$$R_{\alpha,\delta} = P_\alpha / P_\delta$$

2. Alpha to theta power ratio:

$$R_{\alpha,\theta} = P_\alpha / P_\theta$$

3. Beta to delta power ratio:

$$R_{\beta,\delta} = P_\beta / P_\delta$$

4. Beta to theta power ratio:

$$R_{\beta,\theta} = P_\beta / P_\theta$$

5. Theta to delta power ratio:

$$R_{\theta,\delta} = P_\theta / P_\delta$$

6. Alpha to slow wave (SW) power ratio:

$$R_{\alpha,\mathrm{SW}} = P_\alpha / P_{(\text{Band-I + Band-II})}$$

7. Theta to Alpha-Delta ratio:

$$R_{\theta,\alpha-\delta} = P_{\text{Band-II}} / P_{(\alpha+\text{Band-I})}$$

8. Spindle power ratio (SPR): The spindles are oscillatory-structured sleep elements found in N2 stage. Since they occupy the frequency range of 11.5-16 Hz, this frequency band is referred to as the spindle band. Spindle

power ratio is calculated as the ratio of power in the spindle band to that of the total power P_T in the complete signal (0.5 to 49.5 Hz). This ratio can be highly informative for sleep staging, especially for the detection of NREM stage 2.

$$R_{SP} = P_{\text{Sp Band}}/P_T$$

2.2.7. Hurst exponent

Hurst analysis is generally used to estimate the long-term memory of a time series [58]. Hurst exponent has been successfully used for the prediction of epileptic seizures and sleep stage analysis [33, 59]. It is a nonlinear measure, defined as

$$H = \frac{\log(R/S)}{\log(T)} \tag{7}$$

where T is the duration of the signal, R is the difference between the maximum and minimum deviations from the mean and S is the standard deviation.

We compute Hurst exponent corresponding to the filtered signals obtained by passing the raw signal through five different frequency bands i.e. alpha, beta, delta, theta and gamma bands as defined in Table 2. Thus, Hurst exponents add five to the overall feature dimension.

2.2.8. Sum of absolute differences (SAD)

It is defined as the sum of the absolute first difference signal samples in the series, computed for each epoch. We have evaluated this feature for all the frequency bands mentioned in Table 2, which adds eight to the feature dimension.

$$SAD = \sum_{n=1}^{N} \mid x(n) - x(n-1) \mid \tag{8}$$

where N is number of samples in an epoch.

2.2.9. Log root of squared difference signal (LRSD)

This feature was originally proposed by Memar et. al. [19] to measure the sequential variations among the samples of a signal. It is calculated as the logarithm of square root of the sum squared differences between the consecutive samples. The values are obtained for the eight frequency bands considered in this study.

$$LRSD = \log_{10}\left(\sqrt{\sum_{n=1}^{N}(x(n) - x(n-1))^2}\right) \tag{9}$$

where N is the number of samples in an epoch.

2.2.10. Detrended variation

This is also a nonlinear feature, which has been explored by several studies to determine the self-similarity in a time series [60, 61, 62]. To compute this feature, each epoch is divided into segments of length p, whose local linear fit is subtracted and then root-mean square variation F is computed. Here, a window length of 10 samples ($p = 10$) is considered and the detrended variation for each epoch is computed as,

$$\begin{aligned} y(n) &= x(n) - \mu \\ F &= \sqrt{(1/N)\sum_{n=1}^{N}[y(n) - y_p(n)]^2} \end{aligned} \tag{10}$$

where N is the epoch length, μ is the mean of the epoch and y_p is the piecewise straight-line fit of y with length p.

2.2.11. Hjorth parameters

These are related to the variances of the signal and its first and second derivatives [63]. These comprise Hjorth activity (A_H), mobility (M_H), and complexity (C_H), which are defined as follows:

$$\begin{aligned} A_H &= \sigma_x^2 \\ M_H &= \sigma_{x'}/\sigma_x \\ C_H &= \frac{\sigma_{x''}/\sigma_{x'}}{\sigma_{x'}/\sigma_x} \end{aligned} \tag{11}$$

where $\sigma_x, \sigma_{x'}$ and $\sigma_{x''}$ are the standard deviations of the signal $x(n)$, its first derivative $x'(n)$ and second derivative $x''(n)$, respectively. Again, each of the three values are calculated for the filtered signals corresponding to the different frequency bands listed in Table 2.

2.2.12. Statistical features

We have calculated two statistical features epoch-wise for alpha, beta, delta, theta and gamma frequency bands. The two novel features namely maximum-minimum distance and energy_speed, proposed by Aboalayon et. al. [49] are utilized to capture the underlying statistics and provide useful information to distinguish between the different sleep stages.

Maximum-minimum distance (MMD): To begin with, the distance (d) between the maximum and minimum points is obtained for each sliding window of 100 samples within the epoch. For each such mini-epoch (window of 100 samples), d is calculated as described in (12). Finally, the sum of the (d) values of all the mini-epochs is the MMD value for an epoch and this is obtained for different frequency bands.

$$\begin{aligned} d &= \sqrt{\Delta x^2 + \Delta y^2} \\ MMD &= \sum_{i=1}^{i=n} d_i \end{aligned} \tag{12}$$

where Δx is the difference between the indices of the maximum and minimum samples and Δy is the corresponding amplitude difference; n is the total number of mini-epochs in an epoch.

Energy_speed (EnSp): This is the product of the energy and speed of the signal. We have computed the EnSp value of each epoch for alpha, beta, delta, theta and gamma frequency bands. Energy of an epoch is obtained as:

$$E = \sum_{k=1}^{k=N} x_k^2 \tag{13}$$

where x_k are the EEG samples of the epoch. The speed of the signal is defined as

$$v = f * \lambda \tag{14}$$

Table 3: The p-values obtained for the various EEG-derived features using Kruskal-Wallis one-way ANOVA test. Five DWT features, whose p-values are high, are not considered in the feature-set.

Features	$p-$value	**Features**	$p-$value	**Features**	$p-$value	**Features**	$p-$value	**Features**	$p-$value	**Features**	$p-$value	**Features**	$p-$value
AR_1	0	AR_2	0	AR_3	0	AR_4	$8.4e-70$	AR_5	0	AR_6	0	AR_7	$8.1e-266$
AR_8	0	DWT_1	0	DWT_2	0	DWT_3	0	DWT_4	0	DWT_5	0	DWT_6	0
DWT_7	0	DWT_8	0	DWT_9	0	DWT_{10}	0	DWT_{11}	0.9671	DWT_{12}	0.099	DWT_{13}	0.7335
DWT_{14}	0.4781	DWT_{15}	0.1582	DWT_{16}	0	DWT_{17}	0	DWT_{18}	0	DWT_{19}	0	DWT_{20}	0
HFD_δ	0	HFD_θ	0	HFD_α	0	HFD_β	0	HFD_γ	0	HFD_{SB}	0	HFD_{Band-I}	0
$\mathrm{HFD}_{Band-II}$	0	LZC	0	$R_{\alpha,\delta}$	0	$R_{\alpha,\theta}$	0	$R_{\beta,\delta}$	0	$R_{\beta,\theta}$	0	$R_{\theta,\delta}$	0
$R_{\alpha,SW}$	0	$R_{\theta,\alpha-\delta}$	0	R_{SP}	0	H_γ	0	H_δ	0	SAD_δ	0	SAD_θ	0
SAD_α	0	SAD_β	0	SAD_γ	0	SAD_{SB}	0	SAD_{Band-I}	0	$SAD_{Band-II}$	0	$LRSD_\delta$	0
$LRSD_\theta$	0	$LRSD_\alpha$	0	$LRSD_\beta$	0	$LRSD_\gamma$	0	$LRSD_{SB}$	0	$LRSD_{Band-I}$	0	$LRSD_{Band-II}$	0
H_α	0	$A_{H,\delta}$	0	$A_{H,\theta}$	0	$A_{H,\alpha}$	0	$A_{H,\beta}$	0	$A_{H,\gamma}$	0	$A_{H,SB}$	0
$A_{H,Band-1}$	0	$A_{H,Band-II}$	0	$M_{H,\delta}$	0	$M_{H,\theta}$	0	$M_{H,\alpha}$	0	$M_{H,\beta}$	0	$M_{H,\gamma}$	0
$M_{H,SB}$	0	$M_{H,Band-1}$	0	$M_{H,Band-II}$	0	SE	0	$C_{H,\delta}$	0	$C_{H,\theta}$	0	$C_{H,\alpha}$	0
$C_{H,\beta}$	0	$C_{H,\gamma}$	0	$C_{H,SB}$	0	$C_{H,Band-1}$	0	$C_{H,Band-II}$	0	σ	0	$EnSp_\delta$	0
$EnSp_\theta$	0	$EnSp_\alpha$	0	$EnSp_\beta$	0	$EnSp_\gamma$	0	MMD_δ	0	MMD_θ	0	MMD_α	0
MMD_β	0	MMD_γ	0	F	0	H_θ	0	H_β	0				

where λ is the length of the mini-epoch i.e. 100 samples and f is the mid-frequency of the pass band.

Standard deviation: If $\bar{x}$ is the mean of the samples, the standard deviation $\sigma(x)$, also used as a feature, is given by,

$$\sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{15}$$

2.3. Feature selection by statistical significance

In order to examine the statistical significance of the features considered in the study, we have used the non-parametric Kruskal-Wallis one-way analysis of variance test at a confidence level of 95% [64]. 23 sets of features, namely AR model coefficients, HFD, LZC, power ratios such as alpha to delta ratio, beta to theta, beta to delta, alpha to theta, theta to delta, alpha to slow wave ratio, theta to alpha-delta ratio, and spindle power ratio, DWT coefficients, Hurst exponent, SAD, LRSD, detrended variation, standard deviation, EnSp, MMD, Hjorth's mobility, complexity and activity and sample entropy are considered in this study. Further, some of these features are calculated for different frequency bands listed in Table 2. The features are normalized using z-score normalisation method. The corresponding $p-$values obtained by using Kruskal-Wallis test are shown in Table 3.

All the features, except for the mean values of wavelet coefficients ($DWT_{11} - DWT_{15}$) have $p-$values less than 0.005, and hence considered relevant for sleep-stage classification. The features $DWT_{11} - DWT_{15}$ are not statistically significant and therefore excluded from the feature-set. The final set of all the features included in the study are shown in Table 4. The total dimension of the feature vector is 98.

Table 4: List of all the features used for classification of sleep EEG.

Feature	**Dimension**	**Feature**	**Dimension**
AR model parameters	8	Max Min Distance	$5 \times 1 = 5$
3 values X 5 DWT coefficients	15	Energy-Speed	$5 \times 1 = 5$
Alpha to delta band power ratio	1	Hurst Exponent	$5 \times 1 = 5$
Alpha to theta band power ratio	1	Log Root Sq Diff	$8 \times 1 = 8$
Beta to delta band power ratio	1	Hjorth's complexity	$8 \times 1 = 8$
Beta to theta band power ratio	1	Hjorth's mobility	$8 \times 1 = 8$
Theta to delta band power ratio	1	Hjorth's activity	$8 \times 1 = 8$
Alpha to slow wave band power ratio	1	Higuchi FD	$8 \times 1 = 8$
Theta to alpha-delta band power ratio	1	Detrended variation	1
Relative spindle band power ratio	1	Standard deviation	1
Lempel-Ziv Complexity	1	Sample entropy	1
SAD values for 8 frequency bands	8	**All the features**	**98**

2.4. RUSBoost Classifier

The major issue in sleep stage classification is that of class-imbalance. The distribution across different sleep stages is highly skewed: stages such as wake and S4 have more epochs than that of REM. Hence, there is a need to overcome this problem; otherwise, the classifier may end up performing poorly on the minority class. An efficient way to tackle it is by using sampling methods or modelling algorithms like boosting. We can either go for undersampling the majority class or oversampling the minority class, such that both these classes have similar number of samples and hence give a balanced dataset to work with.

In this work, we utilize random undersampling technique (RUS) since it helps in reducing the size of data and computational time. Other method is SMOTE i.e. synthetic minority oversampling technique, which oversamples the minority class by generating synthetic data. Oversampling results in a larger training dataset, thus increasing the model training time. Also, it has been reported in the literature that RUS outperforms SMOTE in most of the scenarios [65]. Hence, we adopt the RUS technique along with adaptive boosting algorithm (Adaboost) by utilizing an ensemble of decision trees as weak learners. The choice of using ensemble classifier with RUS approach provides us with an improved performance than that of the traditional SVM or kernel-SVM, especially for the minority classes.

We have found that as the datasize increases, the performance of SVM deteriorates and it also requires higher computational time. The SVM-RBF (i.e. SVM with a radial basis function as kernel) classifier provides an accuracy of around 68% for 6-state classification using a subset of 8 subjects from the Expanded Sleep-EDF dataset, while the RUSBoost classifier is able to reach up to 94% accuracy on the same data. The ensemble classifier therefore turns out to be a much better choice than SVM. This is because of the fact that the ensemble learning algorithms generally tend to outperform other machine learning algorithms when dealing with unbalanced data [65].

So, we have used two methods to tackle this problem: a) data-level method by re-sampling approach (we use RUS technique) and b) algorithm-level method by utilizing ensemble learning and boosting. RUSBoost (random undersampling with boosting) merges the two key techniques i.e. undersampling and boosting to improve the classifier's performance [65]. It is useful in dealing with the problem of class imbalance, such as in sleep stage classification, where there is a lot of variability in the number of epochs for different sleep stages. For instance, Table 1 shows that the number of REM epochs in Sleep-EDF database is 604 which is far less than the 8055 wake epochs. Also in the other two datasets, the epoch distribution across different sleep stages is very uneven, which results in the class-imbalance problem.

Further, while evaluating the performance of unbalanced datasets, accuracy is not generally considered as the best metric. Hence, we have used various other metrics that can provide better insight such as confusion matrix, precision, recall, F1-score and MCC (Matthews correlation coefficient) [66].

The EEG derived features and the corresponding hypnogram are provided as input to the classifier. We have compared the performance using single EEG channel with that of using a combination of multiple EEG channels on three different datasets. Further, 10-fold cross-validation is performed to evaluate the generalization capability of the model.

2.5. Performance Evaluation of the proposed MEFF-R technique

In order to evaluate the performance of the proposed technique on the test subjects, we adopt three different approaches: (i) subject-independent testing (SIT) (ii) subject-dependent testing (SDT), and (iii) 50% hold-out testing. In the first case (SIT), we consider leave-one-out strategy to keep one subject's data as the test set, while training and validating via 2-fold cross-validation on the rest of the dataset. This process is repeated k times, where k is the number of subjects considered in the respective dataset. This approach ensures that none of the instances of the test data is seen by the model, and hence the prediction is independent of the test subject. This provides a more reliable measure of the classifier's ability to predict the sleep stages of a new test subject which is desired practically in a clinical setting.

The second approach, which has generally been followed in most of the studies in the literature, involves training and test sets that share epochs from the same subjects. In this method, the partition occurs with respect to the epochs, and not subject-wise. So, some of the epochs from the test subject will be part of the training data too. Hence, it is biased and overestimates the generalizability of the classifier.

The third approach involves a random 50%-50% splitting of the entire dataset such that the data from half of the total subjects are utilized for training the model, and the remaining half, for testing. In this way, half of the subjects from the dataset are unseen by the model. This allows more subjects (half the size of the database) to be available for testing as compared to the leave-one-out strategy, where only a single subject is available for testing at a time. Further, we carry out five different runs, each time choosing a random set of subjects for training and testing. This provides a robust estimation of the generalizability of the model.

3. Experimental Results

We have experimented with a variety of features that can be used to classify different sleep stages accurately. We have also compared the performance with single EEG channel against that with the use of multiple EEG channels. Three different datasets have been used to test the accuracy and reliability of the proposed method. Different $n-$class classification problems (common structure for all the three datasets) considered in this study are shown in Table 5.

Table 5: Various number of classes (sleep/wake states) considered for classification in this study under the two different scoring standards.

Classes	AASM score	R&K score
6	-	W Vs REM vs S1 vs S2 vs S3 vs S4
5	W vs REM vs N1 vs N2 vs N3	W Vs REM vs S1 vs S2 vs (S3+S4)
4	W vs REM vs (N1+N2) vs N3	W vs REM vs (S1+S2) vs (S3+S4)
3	W vs REM vs NREM (N1+N2+N3)	W vs REM vs NREM (S1+S2+S3+S4)
2	W vs Sleep (N1+N2+N3+REM)	W vs Sleep (S1+S2+S3+S4+REM)

All the experiments are performed using MATLAB 2017a environment on Intel i7-3770@3.40GHz with 8 GB RAM. It requires around 25 sec on an average to extract the features from an EEG epoch and about 6 min to train the model. When the data of a new subject is used for testing the model, it takes around 1.3 sec to classify a 30s epoch. Thus, once the model is trained, testing can be performed for each epoch, along with the acquisition of the signal for the next epoch. However, the computational complexity has not been stated in most of the papers in the literature. A study by Jiang et. al. [67] has mentioned the computational time to be $\sim 57s$ for the training of classifier using Intel i7-6700 processor with 24 GB RAM. Though the training time in our method is higher than the above study, the training is a one-time process and not to be performed every time. So once the model is trained, the sleep stage classification for an epoch of a test subject requires much less time ($\sim 25sec$).

The classification performance is reported in terms of confusion matrices, kappa coefficients (κ) and overall accuracies. Confusion matrix presents a summary of the number of misclassifications corresponding to the different classes. For instance, the confusion matrices in Figs. 4, 6 and 8 show that the REM sleep stage is generally misclassified as S1 stage and vice-versa in all the three datasets. It also provides us with the values of the sensitivity and specificity of each class, thereby indicating the performance of the classifier corresponding to each sleep stage. Further, it can be seen from the confusion matrices shown in figures 4, 6, 8 that sleep stage S1 provides the least sensitivity values among all the stages. The reason for this is that S1 sleep stage is very similar to REM as well as wake stages. The confusion matrix also tells us where we need to work further if we are interested to further improve the classification accuracy. For example, we can now set up a next level classifier, which disambiguates REM from S1, leading to better classification of REM stage.

The sensitivity, precision and specificity values of individual classes are reported for each dataset. For the expanded sleep-EDF and DREAMS datasets, we have evaluated the classification accuracy on the test data using three different approaches as mentioned before.

3.1. Results of MEFF-R technique on Sleep-EDF database

Table 6 compares the subject-dependent classification results on the Sleep-EDF database using single as well as both EEG channels (Fpz-Cz & Pz-Oz) with the performance of the techniques in the literature. It is evident from the table that by using a single EEG channel, the proposed method outperforms the rest of the studies for 6-state, 4-state, and 3-state classification problems. Using both EEG channels provides better results for all the classes than using a single-channel. Classification accuracy for 2-class problem is higher in [10, 11, 22, 68] than the proposed method. However, the studies by Hassan et. al.

[11] and Yildirim et. al. [68] did not consider k-fold cross-validation technique which provides a more robust measure of the classification performance of the model. Also, they have distributed equal number of epochs for each of the sleep stages in the training and test sets. As a consequence, the model provides increased accuracy. Further in [10], the authors have not reported the sensitivity or specificity values for 2-state classification and hence those results cannot be compared directly. Also, the sensitivity values of S4 and REM sleep stages reported in [10] are significantly lower ($< 40\%$ and $< 80\%$, respectively) than that of our work ($\approx 82\%$ and $\approx 85\%$, respectively).

Most of the studies [10, 11, 19, 22, 29, 31, 39] have followed the existing literature, which suggests that Pz-Oz channel provides better performance than Fpz-Cz channel. However, we have found that for the Sleep-EDF dataset, Fpz-Cz channel provides better results for all the number of classes except the 2-class problem. This particular observation coincides with those of a few recent studies [16, 25, 24, 26, 67, 68]. Figure 2 shows the performance curve for 6-stage classification with the sequential inclusion of each feature into the feature-set for Sleep-EDF and DREAMS databases. It can be seen that the classification accuracy increases with the increase in number of features for both the datasets. Also, the improvement is much higher for the set of features included initially than the features added later. Inclusion of more than 15 set of features presents a slow and steady improvement with the addition of each new feature. It is also evident from the figure that each feature-set is able to provide some information regarding the different stages of sleep.

In order to examine the agreement between the scores of the expert and our method, the confusion matrix corresponding to 6-state classification using single EEG channel (Fpz-Cz) is shown in Figure 3. The confusion matrix for 6-class using both channels (FPz-Cz and Pz-Oz) is presented in Figure 4. It can be inferred that the classification accuracy improves by utilizing the information from both EEG channels. The confusion plots for Pz-Oz channel and combination of both channels for the other number of classes are not shown here, but the accuracy for each class is listed in Table 6. Also, the comparison

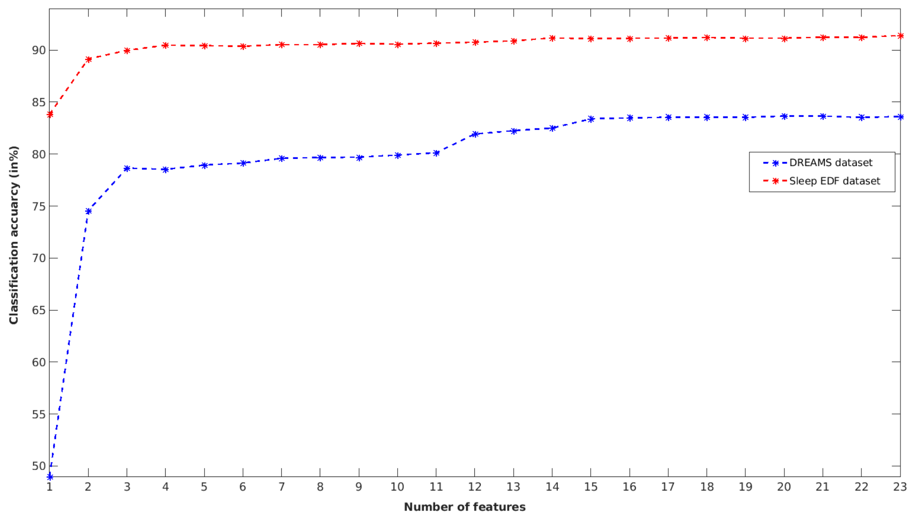

Figure 2: 5-fold cross-validation accuracies for 6-class classification as a function of the number of sets of features, using sequential selection of feature-set for sleep EDF and DREAMS databases with single EEG channel (FPz-Cz and Cz-A1, respectively)

of S1 sensitivity values across the existing literature for 6-class and 5-class on the Sleep-EDF dataset is presented in Table 7. It can be seen that the MEFF-R method performs comparable to or better than others for both the number of classes. The performance of the model for three broad sleep stages i.e. wake, NREM and REM is also evaluated using the metrics of precision, recall, F1-score [69] and Matthews correlation coefficient (MCC) [66]. The values (mean and standard error of mean i.e. SEM) obtained for all these metrics for the three sleep states are shown in Figure 5. The wake state has the maximum recall, precision, F1-score and MCC, followed by NREM and REM. Further, all three stages show high MCC values implying all the classes are predicted quite well. It is because MCC takes into account all the four values in the confusion matrix and hence the value is high only if the classifier is able to perform well on both positive and negative classes. The sensitivity, precision and specificity values corresponding to each of the sleep stages in different n-class classification problems (n varying from 6 to 2) are shown in Table 13.

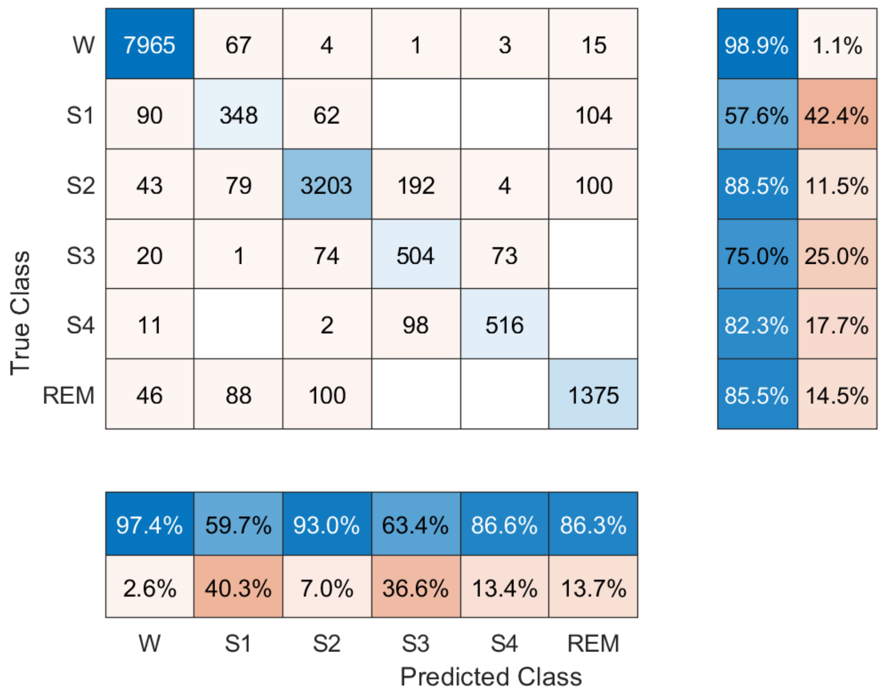

Figure 3: Confusion matrix for 6-stage sleep classification by MEFF-R method using single EEG channel (FPz-Cz) on Sleep-EDF database using 10-fold cross-validation. Significant number of S1 epochs are misclassified as REM (17.2%) or waking (14.9%) state.

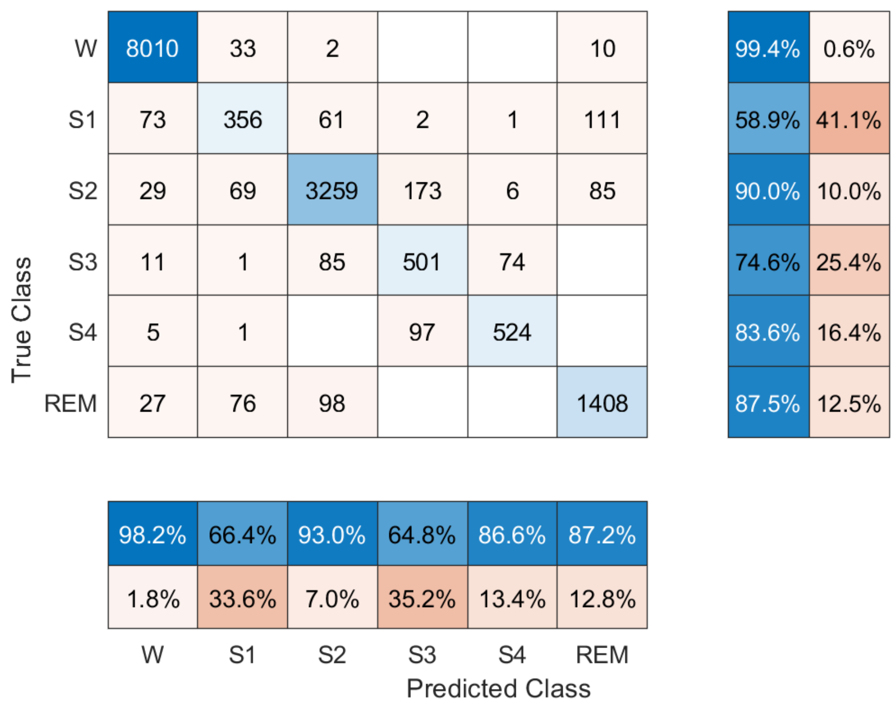

Figure 4: Confusion matrix for 6-stage sleep classification using both EEG channels (FPz-Cz and Pz-Oz) on Sleep-EDF database using 10-fold cross-validation. A sizable fraction of S1 epochs (111 out of 604) are misclassified as REM state.

Table 6: Comparison of performance (subject-dependent test accuracy in %) of MEFF-R technique on Sleep-EDF database with those of techniques in the literature. For each classification problem, the best result is shown in bold.

Method	Channel	Validation	2-class	3-class	4-class	5-class	6-class
Bajaj et al. [7]	Pz-Oz	10-fold CV (4700 epochs)	-	-	-	92.9	88.5
Berthomier et al. [6]	Pz-Oz	Not mentioned	95.4	88.3	74.5	71.2	-
Doroshenkov et al. [8]	Pz-Oz	Not mentioned	95.4	88.3	74.5	71.2	61.1
Ghimatgar et. al. [70]	Fpz-Cz	50% /50%	97.9	94.9	92.7	91.7	90.8
Hassan et. al. [11]	Pz-Oz	50% /50%	98.2	94.2	92.7	83.5	88.1
Hassan et. al. [10]	Pz-Oz	7592/7596 epochs	**99.5**	94.1	92.1	90.7	86.9
Hsu et. al. [12]	Fpz-Cz	1920/960 epochs	-	-	-	87.2	-
Jiang et al. [67]	Fpz-Cz	10-fold CV	-	-	-	92.7	-
Liang et al. [18]	Pz-Oz	50% /50%	-	-	-	83.6	-
Phan et al. [36]	Pz-Oz	70% /30% (11314 epochs)	98.3	-	94.5	-	-
Rahman et. al. [21]	EOG	50% /50%	98.2	94.1	92.9	91.0	90.3
Ronzhina et al. [39]	Pz-Oz	10-fold CV	96.9	89.0	81.6	–	76.4
Sharma et. al. [22]	Pz-Oz	10-fold CV	98.0	94.7	92.3	91.1	90.0
Vural et al. [71]	Fpz-Cz & Pz-Oz	-	-	-	-	-	70.0
Yildirim et. al. [68]	Fpz-Cz	70% /15% /15%	98.3	94.2	91.4	90.8	89.5
Zhu et. al. [29]	Pz-Oz	10-fold CV	97.9	92.6	89.3	88.9	87.5
Zhu et. al. [72]	FPz-Cz	70% /30%				**93.7**	
MEFF-R	Pz-Oz	10-fold CV	97.6	95.0	92.7	91.9	90.6
MEFF-R	Fpz-Cz	10-fold CV	97.0	**95.1**	**93.6**	92.8	**91.6**
MEFF-R	Pz-Oz & Fpz-Cz	10-fold CV	98.0	**95.8**	**94.4**	**93.7**	**92.6**

CV: cross-validation

3.2. Results of MEFF-R technique on DREAMS Subjects database

In DREAMS Subjects database, both R&K and AASM scored labels are available. So, we have performed sleep-stage classification according to both the scoring criteria. The classification performance for the different testing approaches, namely SIT and SDT using a single EEG channel (FP1-A2 or CZ-A1) and the combination of a few channels are presented in Table 8 for R&K scoring. The classification accuracy achieved by MEFF-R for SDT is higher than those of all other studies in the literature for all classification problems. The best performance using a single channel is obtained with the central electrode (Cz-A1) followed by the fronto-parietal channel (FP1-A2). The classification accuracy is further improved by utilizing the combination of the two EEG channels FP1-A2 and Cz-A1. The occipital channels (O1-A2 and O2-A1) resulted in low performance and hence are not reported here. Further, utilizing three EEG channels

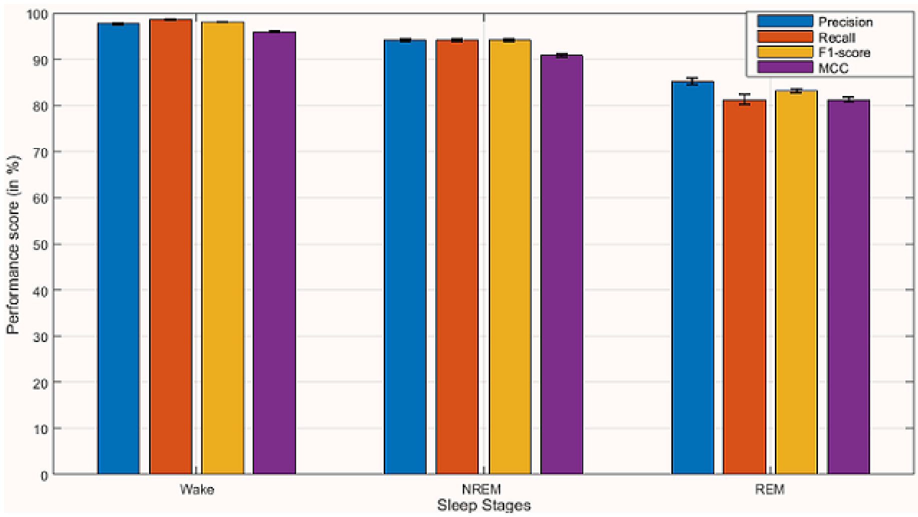

Figure 5: Performance metrics of MEFF-R for 3-stage (Wake, REM and NREM) classification on Sleep-EDF database using single EEG channel (FPz-Cz).

(FP1-A2, Cz-A1 and O1-A2) resulted in an improved performance for 4-class, 5-class and 6-class classification problems. Only a few studies have reported results based on subject independent testing for this dataset and among those, MEFF-R provides higher accuracies for all classification problems except for 5-class. Yang et. al. [74] obtained better accuracy than the proposed method for 5-state classification using 1D-CNN combined with hidden Markov model (HMM). However, they achieved a lower sensitivity of $\sim 34\%$ for S1 sleep stage than obtained by our method ($\sim 54\%$).

The confusion matrix as per R&K scoring criteria for 6-stage classification using Cz-A1 is shown in Figure 6. The values of sensitivity, specificity and precision corresponding to each sleep stage for different n-class classification problems ($n \in \{6, 5, 4, 3, 2\}$) using Cz-A1 channel are presented in Table 13. We have achieved sensitivities of 90.1% and 97.9% for the detection of the wake and sleep states, respectively, which are higher than those of all the previously reported studies. Figure 7 shows the precision, recall, F1-score and MCC (mean and SEM values) corresponding to the three broad sleep stages namely NREM, REM and wake for the DREAMS dataset. The values of all the metrics, except

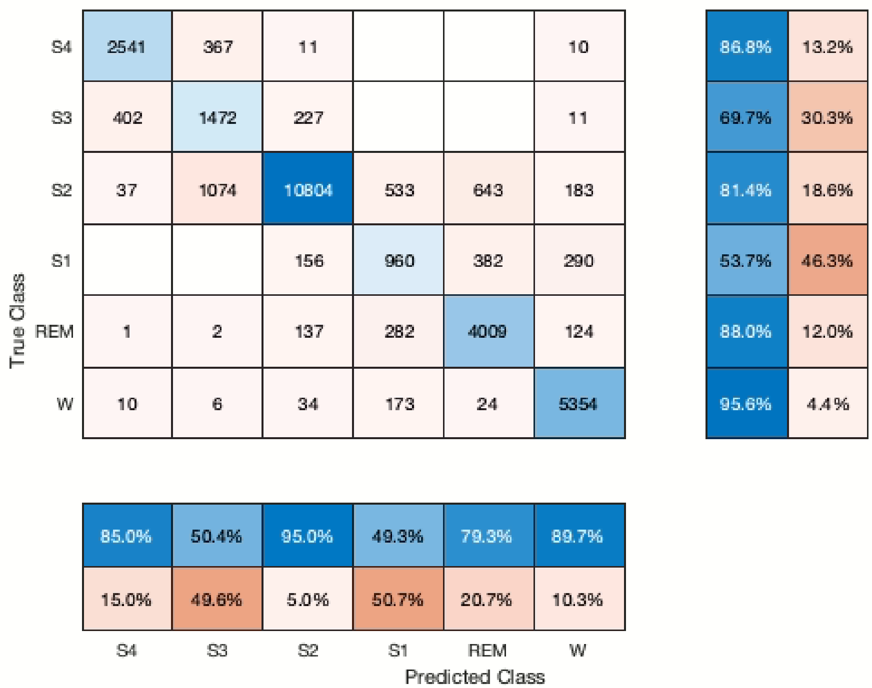

Figure 6: Confusion matrix for 6-stage sleep classification using single EEG channel (Cz-A1) on DREAMS database using 10-fold cross-validation. 382 (21.4%) and 290 (16.2%) out of the total 1788 epochs belonging S1 stage are misclassified as REM and waking state, respectively.

MCC, are the highest for NREM stage followed by Wake and REM stages. The n-state (n varying from 6 to 2) classification results on unseen subjects using 50% hold-out technique are shown in Figure 9. In this approach, 10 (out of 20) subjects are chosen randomly to train the classifier and the remaining 10 subjects are used to test the classification performance for different states. It is evident from the figure that the classification accuracy increases from around 75% for 6-state classification to 95% for 2-state classification. This is because it becomes more challenging to accurately classify each of the states as we increase the number of classes in the classification problem.

Table 9 presents the classification accuracy for all the number of classes using single EEG channel as per both R&K and AASM scoring criteria. Evidently the MEFF-R approach performs equally well for AASM standard. Our method is consistent with both the scoring standards and provides high accuracies for all the states irrespective of the criteria. The results for multiple classification problems (2-state to 6-state) are shown for the top two single EEG channels. For this dataset, Cz-A1 yields the best classification accuracies for all the states.

3.3. Results of MEFF-R technique on Expanded Sleep EDF database

For this dataset, we have trained the RUSBoost classifier on the EEG data from 30 subjects. Table 10 compares the classification performance on Expanded Sleep-EDF database in terms of accuracy (in %) for the different number of classes with those of the techniques reported in the literature. For subject-dependent testing, the best overall accuracy has been achieved by MEFF-R for all the classification problems from 2 to 6-classes. Further, we achieved a very high accuracy of above 99% in every case (except 6-class) by combining features from both the EEG channels. This is the highest sleep staging performance reported so far in the existing literature. We have also reported the subject-independent test results which provide a far better measure of the generalizability of the model. The accuracy for this testing would be less than that of subject-dependent testing, since SIT tests the model on an unseen subject's data which is not the case in SDT. However, most of the studies in the literature

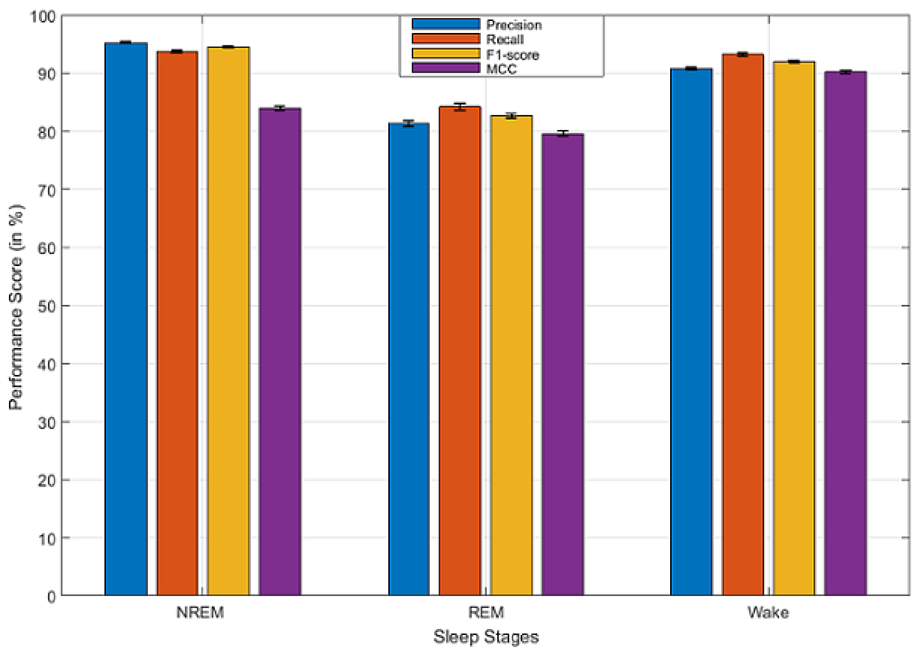

Figure 7: Performance metrics for 3-stage (Wake, REM and NREM) classification on DREAMS database using single EEG channel (Cz-A1).

have reported only 5-state classification. Among all the reported studies, the best overall classification accuracy for subject-independent testing is achieved by our method for all the multi-class problems except for 5-class. Jiang et. al. achieved higher 5-state classification accuracy for SIT than the proposed method; however they have considered lesser number of epochs for evaluating their model. Also, our method has achieved a higher kappa value of 0.90 than that of 0.86 reported in their study.

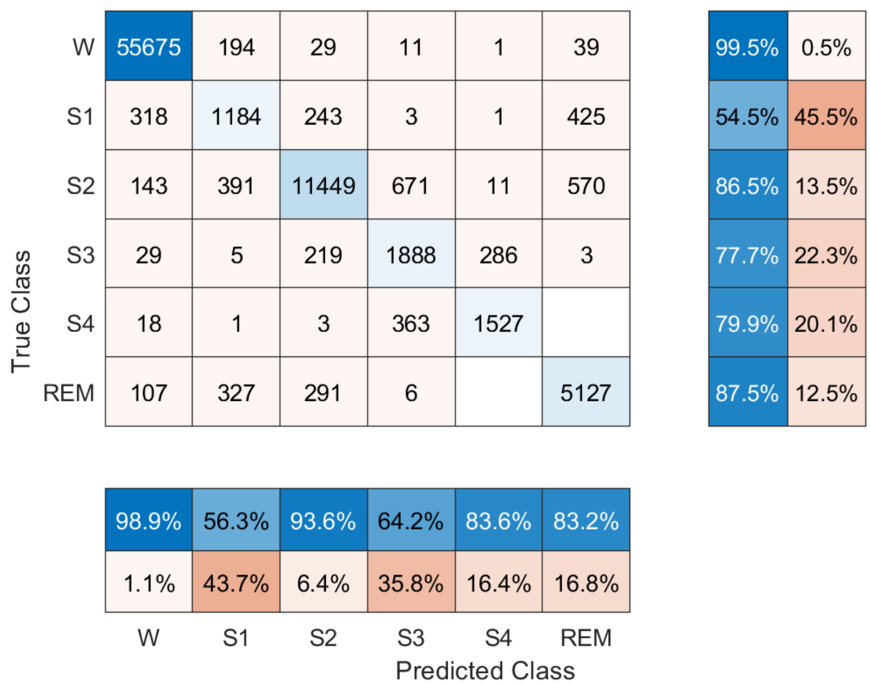

Figure 8: Confusion matrix for 6-stage sleep classification using single EEG channel (Pz-Oz) on Expanded Sleep-EDF database. 425 (19.5%) S1 epochs out of the total 2174 epochs are classified wrongly as belonging to REM stage.

We have also presented the prediction accuracy of the model for eight new subjects whose data is not included for training (30 subjects) and hence, not seen by the model previously. Table 12 shows the prediction accuracy of the model (in %) on these eight unseen subjects for multiple sleep states. On the average, it is able to predict the different sleep stages with an accuracy of about

89%, 90% and 95% for 6-, 5- and 2-classes, respectively.

The confusion matrix between the evaluation of the experts and our results for 6-state classification using the Pz-Oz channel is shown in Figure 8. The sensitivity for S1 stage is less than that of all other sleep stages. It can be seen that most of the misclassified S1 epochs are categorized as wake or REM, the reason being its similarity to the waking state. Since S1 stage is the transition from waking to sleep, it shows characteristics similar to that of the wake state. Further, S1 and REM stages exhibit similar EEG patterns [77]. Hence the accurate classification of this particular sleep stage is challenging. In the case of 2-state classification, our method achieves high sensitivities of 98.7% and 95.7% for wake and sleep states, respectively, using a single EEG channel. These sensitivity values could be further increased to 99.9% and 99.6% for wake and sleep stages, respectively, by using both EEG channels. The performance measures such as sensitivity, specificity and precision corresponding to each stage for different classification problems using single EEG channel (Pz-Oz) are presented in Table 13.

For this dataset, Pz-Oz provides the best classification performance among the single channels for all the classes. Further, the fusion of two EEG channels results in the best overall classification accuracies for all the classification problems (2-state to 6-state), as seen in Table 10. The precision, recall, F1-score and MCC for the three main sleep stages (NREM, REM and wake) are presented in Figure 10. The values of all the four metrics are maximum for wake stage, followed by NREM and REM. In fact, the values of precision, recall and F1-score are around 99% in the wake stage, revealing that the wake stage has been accurately identified by the model.

Figure 9 presents the mean classification accuracy and standard error across the different n-states ($n \in \{6, 5, 4, 3, 2\}$) for DREAMS and Expanded sleep EDF databases, for 20 and 30 subjects, respectively. These results are obtained by utilizing randomly chosen 50% of the subjects for training and remaining 50% for testing across multiple runs, separately with each database. Here, we have not considered Sleep-EDF database since it has only 8 subjects' data, which is not sufficient for the 50% - 50% splitting of train-test set.

The performance improves with the reduction in the number of sleep states, reaching the maximum for the 2-state classification problem; which is as expected. Further, the DREAMS database is a more challenging dataset than the Sleep-EDF, which is apparent from Figure 9. Only a few studies have reported the performance on DREAMS database and also the accuracies reported by those studies are lower (see Table 9). However, the improvement in performance from $(n+1)-$state to $n-$state classification problem ($n = 2$ to 5) is high in DREAMS dataset. Further, it is interesting to note that the classification accuracies on unseen test subjects are comparable between the 50%-holdout technique and the cross-validated subject independent testing (SIT) method across multiple classes for both DREAMS and Expanded sleep EDF datasets.

We have also computed the kappa (κ) coefficient for all n-class classification problems ($n \in \{6, 5, 4, 3, 2\}$) across all the three datasets, which are listed in Table 11. The obtained values of κ ranges from 0.87 to 0.96 for both sleep EDF and its expanded version, which shows an excellent agreement between the scoring by experts and MEFF-R method. For DREAMS dataset, κ shows a good agreement with values ranging from 0.77 to 0.88 across the different classes. However, it is slightly less than the other two datasets and the possible reason is as mentioned in the previous paragraph.

4. Discussion

Tables 6, 8 and 10 compare the performance of our method with those of the previous studies for the different datasets considered in this work. For all the three databases, the MEFF-R method provides promising results with classification accuracies higher than or comparable to those in the literature.

In this study, we have experimented with all the available EEG channels so as to report the best channel to use for sleep staging. We find that the central electrode provides the best classification accuracy. In the case of DREAMS database, we have obtained the maximum classification accuracy by utilizing Cz-

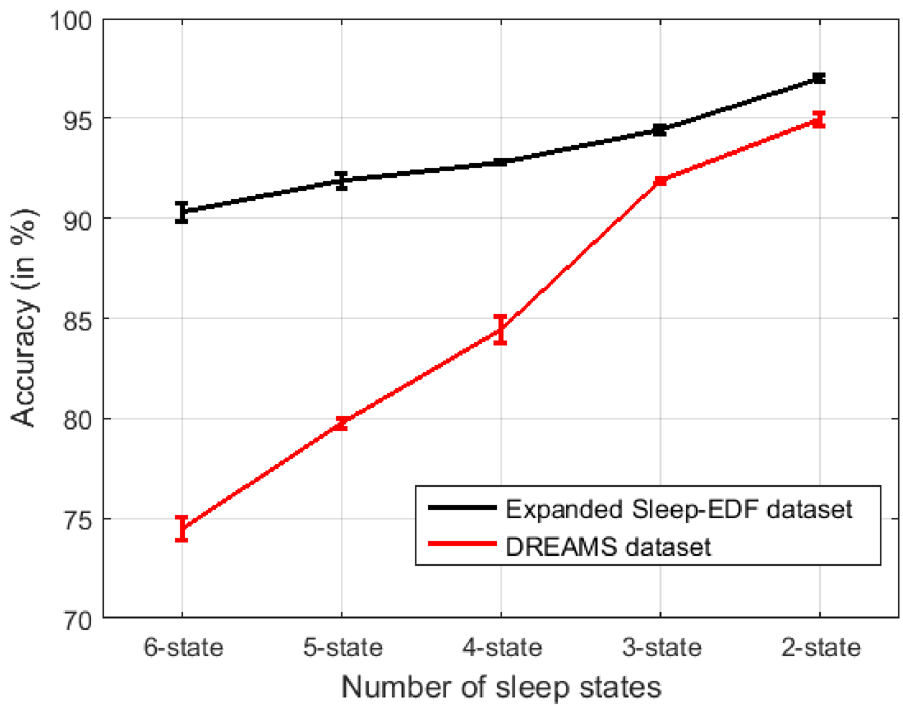

Figure 9: Sleep staging performance of MEFF-R technique on unseen subjects for different n-state classifications ($n \in \{6, 5, 4, 3, 2\}$) on Expanded Sleep EDF (15 subjects' data each for training and testing) and DREAMS (10 subjects' data each for training and testing) databases using a single EEG channel. Values plotted are the mean accuracies over five separate trials with random choice of training and testing subjects.

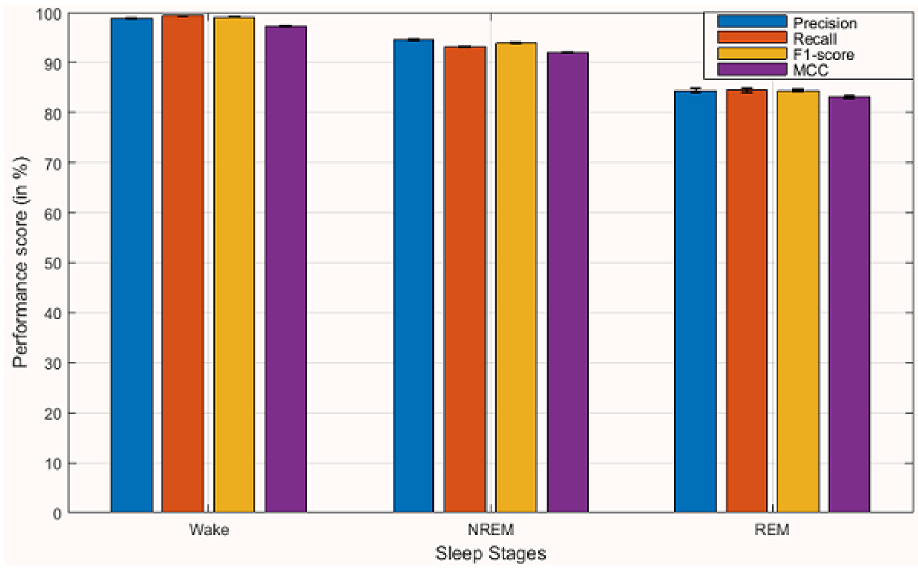

Figure 10: Performance metrics for 3-stage (Wake, REM and NREM) classification on the Expanded Sleep-EDF database using single EEG channel (Pz-Oz).

A1 channel (Cz channel referenced to mastoid), and for sleep-EDF dataset, we have got the best accuracy for FPz-Cz channel (bipolar frontal-central channel). On the other hand, Pz-Oz (bipolar parietal-occipital channel) channel provides the best classification performance for the Expanded Sleep EDF database. Further, the combination of FPz-Cz and Pz-Oz channels provide better performance than the single EEG channels for both Sleep-EDF and Expanded Sleep-EDF datasets. Similarly, for the DREAMS database, the combination of FP1-A2 and Cz-A1 provides better accuracy than the individual EEG channels.

It has been suggested in the literature that most of the observations such as delta activity, k-complexes and lower sleep spindles are predominantly frontal phenomena, and are thus best captured by the FPz-Cz channel. On the other hand, theta activity and higher frequency sleep spindles, being parietal phenomena, are best captured using the Pz-Oz channel [53]. Hence, any of these locations would be sufficient to provide information regarding the different sleep stages. However, the frontal location should be preferred since it is more convenient and also minimizes sleep disturbances.

As compared to the previously reported studies, our work has the following key advantages. Firstly, we have experimented with a variety of features as well as different classifiers and finally reported the feature-set and classifier that provides the best classification performance for all the classes. We have compared the performances among different EEG channels and their combinations. The fusion of two EEG channels results in a higher classification accuracy than using a single channel. Further, in the case of larger datasets, the RUSBoost classifier outperforms the traditional SVM classifier in terms of both computational time and classification accuracy. We have utilized the RUSBoost classifier to overcome the class imbalance problem. Secondly, many studies [7, 6, 8] have considered only the Sleep-EDF dataset to validate their method, which is a small database with only 8 subjects. Hence, these methods need to be validated on larger datasets, which we have taken care of in this work. Unlike some of the studies [7, 6, 8], we have utilized three different datasets and the results are promising on all the three, indicating the efficacy and generalizability of our method.

Thirdly, the proposed method has achieved better S1 sensitivity values than the rest for both 6- and 5-classes, except for the study by Tsinalis et. al. [26]. However, the latter has performed only 5-state classification. A high sensitivity for S1 stage is crucial since it is used to mark the onset of sleep and transition from awake to sleep state. This is particularly important for neurorehabilitation in coma, where providing sensory stimulation in awake state is more effective than in sleep state [82]. Fourthly, many studies have used a single splitting of train-test data to evaluate the performance of the model [6, 11, 12, 18, 68], which is not as robust as cross-validation based evaluation. In our work, we have reported the k-fold cross-validation performances for all the datasets. Further, testing of unseen subjects has not been considered in most of the studies [6, 10, 11, 75, 80] and hence they do not provide any information regarding the performance on new unseen subjects whose data has not been seen at all by the model. This is an important requirement from a practical point of view and has been addressed in our work. Finally, we have considered various n-state classification problems ($n \in \{6, 5, 4, 3, 2\}$), unlike many studies which have reported only 5-state classification [12, 55, 18, 24, 25, 27].

The major application of our study is that it can be readily deployed to perform automatic sleep staging. Thus, it can assist the sleep experts to analyse any anomaly in the sleep parameters such as sleep efficiency, percentage of REM and slow wave sleep (SWS), percentage of NREM stage, and spindle density. Further, it has been reported in the literature that alterations in the spindle density can be a symptom of various neurological disorders such as dementia, schizophrenia, depression or Parkinson's disease [83, 84, 85, 86]. Therefore, early detection of any anomaly is extremely crucial for the diagnosis of these disorders; and an automatic sleep staging method such as MEFF-R can assist the experts. Since spindles are mostly localized in N2-stage of sleep, the experts can analyse only the N2-stage provided by MEFF-R to detect any abnormal pattern. Similarly, to analyse any irregularity in percentage of REM sleep or SWS, MEFF-R can provide the input to experts in the form of annotated sleep stages. This may reduce the burden on the sleep experts in annotating the whole overnight PSG and then looking for a particular trait of interest.

We have achieved comparable or better S1 sensitivity values than the previous studies; however, still the values are lower than those for all the other sleep stages. In future work, we intend to improvise the detection accuracy of S1 sleep stage. Another limitation of this study is that all the datasets studied include only healthy subjects or some with mild sleeping difficulty. However, patients with various sleeping disorders such as insomnia, narcolepsy or REM-sleep behavior disorder (RBD) [87] are not encompassed here. Thus it is desirable to validate the MEFF-R method on patients' data in our next study.

Table 7: Comparison of S1 sensitivity values (in %) of the proposed method with the literature on Sleep-EDF dataset. The values are shown for both 6-class and 5-class classifications.

Authors	6-class	5-class
Doroshenkov et. al. [8]	4.8	-
Zhu et. al. [29]	-	15.8
Liang et. al. [18]	-	18.7
Sharma et. al. [22]	18.9	18.7
Ghimatgar et. al. [70]	-	23.7
Vural et. al. [71]	33.7	-
Liang et. al. [73]	-	35.1
Ronzhina et. al. [39]	35.8	-
Hsu et. al. [12]	-	36.4
Kim et. al. [46]	-	38.6
Hassan et. al. [31]	39.1	39.7
Sharma et al. [23]	41.4	17.4
Hassan et al. [10]	47.3	47.0
Fraiwan et. al. [9]	-	43.2
Sun et. al. [25]	-	52.5
Tsinalis et. al. [26]	-	**59.8**
Zhu et. al. [72]	-	52.2
MEFF-R	**58.9**	59.4

Table 8: Comparison of sleep-staging performance of MEFF-R (in %) on DREAMS database with those of existing techniques in the literature, using R&K standard. Cross-validation results obtained using one, two and three channels are tabulated. For each type of classification, the best result is shown in bold.

Method	EEG channel	Validation	2-class	3-class	4-class	5-class	6-class
Subject-Dependent Testing							
Hassan et. al. [11]	Not mentioned	10-fold CV	93.3	84.4	79.1	73.5	70.7
Shen et. al. [75]	Cz-A1	10-fold CV	94.9	87.7	82.7	80.9	78.2
Shen et. al. [76]	Cz-A1	10-fold CV	96.2	88.7	84.0	82.3	78.9
MEFF-R	FP1-A2	10-fold CV	96.0	92.0	88.0	84.5	81.8
MEFF-R	Cz-A1	10-fold CV	96.5	92.3	88.7	85.7	83.1
MEFF-R	FP1-A2 & Cz-A1	10-fold CV	**96.6**	**93.0**	89.2	86.0	84.4
MEFF-R	FP1-A2 & Cz-A1 & O1-A2	10-fold CV	96.6	92.4	**90.5**	**87.4**	**85.1**
Subject-Independent Testing (i.e. average prediction on unseen subject)							
Yang et. al. [74]	Cz-A1	20-fold CV	-	-	-	**81.7**	-
MEFF-R	Cz-A1	LOSO	**93.8**	87.3	83.4	79.8	76.3
MEFF-R	FP1-A2 & Cz-A1	LOSO	93.5	**87.6**	**84.4**	80.5	**76.9**

CV: cross-validation; LOSO: leave one subject out

Table 9: Performance of MEFF-R technique on DREAMS database for two to six classes based on R&K and AASM criteria using the top two single EEG channels.

	FP1-A2 Channel		**Cz-A1 Channel**	
Class	R&K	AASM	R&K	AASM
2	96.0 %	95.1 %	96.5 %	96.0 %
3	92.0 %	91.3 %	92.3 %	91.8 %
4	88.0 %	87.3 %	88.9 %	87.3 %
5	84.5 %	83.8 %	85.7 %	83.3 %
6	81.8 %	-	83.1 %	-

Table 10: Performance of MEFF-R on Expanded Sleep EDF database compared with those in the literature (both subject-dependent and independent testing). For each type of classification, the best result is shown in bold.

Method	EEG channel	Validation	Classifier	Epochs	2-class	3-class	4-class	5-class	6-class
Subject-Dependent Testing									
Jadhav et. al. [78]	Fpz-Cz	70%/10%/20%	CNN (Squeezenet)	62177	-	-	-	83.3	-
Jiang et. al. [67]	Fpz-Cz	10-fold CV	HMM-RF	36972	-	-	-	92.6	-
khalili et. al. [79]	Fpz-Cz	20-fold CV	RL+TCNN+CRF	41950	-	-	-	85.4	-
Silveria et. al. [80]	Pz-Oz	10-fold CV	RF	106376	-	93.9	92.3	91.5	90.5
Shen et. al. [76]	Pz-Oz	10-fold CV	Bagged Trees	104368	98.7	94.9	93.9	92.5	92.0
Yildirim et. al. [68]	Fpz-Cz	70% /15% /15%	1D-CNN	127512	97.9	94.2	92.2	90.5	89.4
MEFF-R	Pz-Oz	10-fold CV	RUSBoost	81558	98.4	96.9	95.8	95.1	94.2
MEFF-R	Fpz-Cz	10-fold CV	RUSBoost	81558	97.8	96.2	95.3	94.5	93.7
MEFF-R	Pz-Oz & Fpz-Cz	10-fold CV	RUSBoost	81558	**99.8**	**99.6**	**99.6**	**99.5**	**96.3**
Subject-Independent Testing (i.e. average prediction on unseen subject)									
Ghimatgar et. al. [70]	Fpz-Cz	LOSO	RF+HMM	40100	94.9	88.3	83.5	81.2	78.2
Ghimatgar et. al. [70]	Fpz-Cz	50%-holdout	RF+HMM	40100	95.5	88.2	83.5	80.5	76.9
Jiang et. al. [67]	Fpz-Cz	LOSO	HMM-RF	54728	-	-	-	**93.0**	-
Phan et. al. [42]	Fpz-Cz	LOSO	ARNN-SVM	46236	-	-	-	82.5	-
Supratak et. al. [24]	Pz-Oz	20-fold CV	CNN+BLSTM	41950	-	-	-	79.8	-
Supratak et. al. [24]	Fpz-Cz	20-fold CV	CNN+BLSTM	41950	-	-	-	82.0	-
Mousavi et. al. [81]	Pz-Oz	20-fold CV	BiRNN+EDNA	42308	-	-	-	82.8	-
Mousavi et. al. [81]	Fpz-Cz	20-fold CV	BiRNN+EDNA	42308	-	-	-	84.3	-
Sun et. al. [25]	Fpz-Cz	LOSO	WDBN+BLSTM	41950	-	-	-	85.5	-
Tsinalis et. al. [27]	Fpz-Cz	20-fold CV	CNN	37022	-	-	-	74.8	-
Tsinalis et. al. [26]	Fpz-Cz	20-fold CV	Autoencoders	37022	-	-	-	78.9	-
Yang et. al. [74]	Fpz-Cz	20-fold CV	CNN+HMM	42308	-	-	-	84.0	-
Zhu et. al. [72]	FPz-Cz	LOSO	CNN+ABNN	42269	-	-	-	82.8	-
MEFF-R	Pz-Oz	LOSO	RUSBoost	81558	**97.1**	**95.2**	**93.3**	92.4	90.9
MEFF-R	Fpz-Cz	LOSO	RUSBoost	81558	96.0	94.3	93.1	92.2	**91.0**
MEFF-R	Pz-Oz & Fpz-Cz	LOSO	RUSBoost	81558	**97.3**	95.1	**93.9**	92.9	**91.6**

ABNN: Attention-based neural network; ARNN: Attention-based recurrent neural network; BLSTM: Bidirectional long short-term memory network; BiRNN: Bidirectional recurrent neural network; CV: cross-validation; CNN: Convolutional Neural Network; CRF: Conditional Random Field; EDNA: Encoder-Decoder network with attention mechanism; HMM: Hidden Markov model; LOSO: Leave one subject out; RF: Random Forest; RL: Representation Learning; TCNN: Temporal CNN; WDBN: Window deep belief network.

Table 11: Kappa coefficients of MEFF-R on Sleep-EDF, DREAMS and Expanded Sleep EDF databases using the single EEG channels of FPz-Cz, Cz-A1 and Pz-Oz, respectively.

Dataset	6-class	5-class	4-class	3-class	2-class
Sleep EDF	0.87	0.88	0.89	0.91	0.94
DREAMS	0.77	0.81	0.84	0.85	0.88
Expanded Sleep EDF	0.88	0.90	0.91	0.93	0.96

Table 12: Prediction accuracy (in %) of MEFF-R on 8 unseen subjects from Expanded Sleep EDF dataset for 6-class, 5-class and 2-class classifications.

Unseen test subject	**6-class**	**5-class**	**2-class**
S1	90.7	92.0	98.1
S2	88.6	92.3	97.8
S3	89.6	90.5	94.1
S4	83.5	84.5	86.9
S5	87.8	88.0	93.5
S6	91.3	91.6	97.8
S7	90.1	91.2	97.0
S8	89.8	89.5	99.1

Table 13: Comparison of performance (sensitivity, specificity and precision) of MEFF-R on Sleep-EDF, DREAMS and Expanded sleep EDF databases using single EEG channel.

Dataset	Performance measure	6-class						5-class					4-class				3-class			2-class	
		S1	S2	S3	S4	REM	W	S1	S2	S3+S4	REM	W	S1+S2	S3+S4	REM	W	NREM	REM	W	S	W
Sleep EDF	Sensitivity	57.6	88.5	75.0	82.3	85.5	98.9	56.0	88.6	92.8	85.3	99.0	88.0	89.0	85.4	99.0	94.1	81.5	98.6	96.8	97.1
	Specificity	98.4	97.9	98.0	99.4	98.4	97.1	98.4	97.9	98.6	98.4	97.1	96.8	98.9	98.2	96.9	96.7	98.3	97.3	97.1	96.8
	Precision	59.7	92.9	63.4	86.6	86.3	97.4	59.8	92.2	86.1	85.9	97.5	91.4	88.5	84.6	97.3	94.2	85.0	97.7	96.7	97.2
DREAMS	Sensitivity	53.7	81.4	69.7	86.7	88.0	95.6	53.9	81.5	94.9	87.8	95.7	86.4	89.8	88.8	94.8	93.8	84.4	93.3	97.9	90.1
	Specificity	96.5	96.7	94.8	98.3	95.9	97.5	96.4	96.7	95.6	96.0	97.6	92.7	97.6	95.9	97.5	90.8	96.6	98.0	90.1	97.9
	Precision	49.3	95.0	50.4	85.0	79.3	89.6	48.6	95.0	81.3	79.4	90.0	92.1	88.3	79.4	89.7	95.3	81.3	90.9	97.7	90.9
Expanded Sleep EDF	Sensitivity	54.5	86.5	77.7	79.9	87.5	99.5	54.5	86.7	94.2	87.3	99.5	86.2	91.3	88.7	99.6	93.2	84.5	99.4	97.2	98.9
	Specificity	98.8	98.8	98.7	99.6	98.6	97.6	98.8	98.8	99.1	98.7	97.7	98.4	99.3	98.5	97.3	98.3	98.8	97.7	98.9	97.2
	Precision	56.3	93.6	64.2	83.6	83.2	98.9	81.5	53.9	87.8	95.7	94.9	86.4	88.8	94.8	89.8	84.4	93.3	93.8	90.1	98.7

5. Conclusion

In this work, we have considered a variety of features such as fractal dimension, sample entropy, wavelet coefficients, complexity, power-ratios, statistical features and autoregressive model parameters, that can extract maximum information from the EEG signal. These features characterize most of the aspects of an EEG signal like irregularity, frequency content, time-frequency information, and periodicity. In order to eliminate the class-imbalance issue in all the sleep datasets, we have utilized the random under-sampling with boosting (RUSBoost) classifier. The classification performance has been evaluated for various number of classes from two to six. We have employed three different datasets to evaluate the model's ability to classify distinct sleep stages and the performance on all the datasets are promising. This shows that our method is able to generalize well by providing good classification accuracies across all the three datasets. Further, three different approaches to assess the performance of the trained model are considered in our study. One approach is subject-dependent testing which is mostly reported in the literature, but does not provide an unbiased estimate of the model's prediction ability. On the other hand, the subject-independent testing approach provides a more reliable measure of the generalizability of the model, since it evaluates the performance on unseen test subjects. Also, the third approach utilizes 50% of the total subjects available in the dataset for training and tests the model's performance on the remaining unseen (50%) subjects. We have obtained similar results for both SIT and 50%-holdout approaches, which again confirms the reliability of the MEFF-R technique.

We have achieved overall accuracies of 98.0%, 95.8%, 94.4%, 93.7% and 92.6% for 2-class to 6-class classifications, respectively, for the Sleep-EDF dataset. The accuracies obtained by our method are the highest among the reported studies for all the number of classes, except 2-class. In the case of DREAMS Subjects database, we have attained maximum overall accuracies across all the number of classes employing single channel as well as multiple channels of data using SDT.

Further, the classification performance in the case of SIT is also promising. To the best of our knowledge, results based on prediction accuracy of unseen subjects have not been reported much in the literature. For the Expanded Sleep EDF database, we have attained maximum overall accuracies of 99.8%, 99.6%, 99.6%, 99.5% and 96.3% for 2-class to 6-class problems, respectively, which are again higher for all n-class classifications ($n \in \{2, 3, 4, 5, 6\}$) than the previous studies. Also in the case of SIT, MEFF-R technique provides better or comparable results for all the classes. The results show that the MEFF-R method has the potential to provide accurate classification for multiple sleep stages. It has been validated by utilizing three different datasets and three different validation techniques and it has provided results consistent with the experts' score using either scoring standard. Hence, it can be used to automate sleep staging and mitigate the problem of expensive, subjective and tedious manual sleep scoring procedure.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] B. Basics, Understanding sleep, National Institute of Neurol. Disorders and Stroke, Bethesda (2006).

[2] A. Rechtschaffen, A manual of standardized terminology, technique and scoring system for sleep stages of human subjects, Public Health Service (1968).

[3] C. Iber, S. Ancoli-Israel, A. L. Chesson, S. F. Quan, et al., The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, volume 1, American academy of sleep medicine Westchester, IL, 2007.

[4] L. De Gennaro, M. Ferrara, Sleep spindles: an overview, Sleep medicine reviews 7 (2003) 423–440.

[5] S. Keenan, M. Hirshkowitz, Monitoring and staging human sleep, Principles and practice of sleep medicine 5 (2011) 1602–1609.

[6] C. Berthomier, X. Drouot, M. Herman-Stoïca, P. Berthomier, J. Prado, D. Bokar-Thire, O. Benoit, J. Mattout, M.-P. d'Ortho, Automatic analysis of single-channel sleep eeg: validation in healthy individuals, Sleep 30 (2007) 1587–1595.

[7] V. Bajaj, R. B. Pachori, Automatic classification of sleep stages based on the time-frequency image of eeg signals, Computer methods and programs in biomedicine 112 (2013) 320–328.

[8] L. Doroshenkov, V. Konyshev, S. Selishchev, Classification of human sleep stages based on eeg processing using hidden markov models, Biomedical Engineering 41 (2007) 25–28.

[9] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier, Computer Methods and Programs in Biomedicine 108 (2012) 10–19.

[10] A. R. Hassan, 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 24 (2016) 1–10.

[11] A. R. Hassan, M. I. H. Bhuiyan, Automated identification of sleep states from eeg signals by means of ensemble empirical mode decomposition and random under sampling boosting, Computer methods and programs in biomedicine 140 (2017) 201–210.

[12] Y.-L. Hsu, Y.-T. Yang, J.-S. Wang, C.-Y. Hsu, Automatic sleep stage recurrent neural classifier using energy features of eeg signals, Neurocomputing 104 (2013) 105–114.

[13] A. R. Hassan, M. I. H. Bhuiyan, Dual tree complex wavelet transform for sleep state identification from single channel electroencephalogram, in: 2015 IEEE International Conference on Telecommunications and Photonics (ICTP), IEEE, 2015, pp. 1–5.

[14] J.-R. Huang, S.-Z. Fan, M. F. Abbod, K.-K. Jen, J.-F. Wu, J.-S. Shieh, Application of multivariate empirical mode decomposition and sample entropy in eeg signals via artificial neural networks for interpreting depth of anesthesia, Entropy 15 (2013) 3325–3339.

[15] R. A. Ganesan, R. Jain, Binary state prediction of sleep or wakefulness using eeg and eog features, in: 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–7.

[16] R. A. Ganesan, R. Jain, Sleep-awake classification using eeg band-power-ratios and complexity measures, in: 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–6.

[17] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines, Journal of neuroscience methods 250 (2015) 94–105.

[18] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, Y.-H. Wang, Automatic stage scoring of single-channel sleep eeg by using multiscale entropy and autoregressive models, IEEE Transactions on Instrumentation and Measurement 61 (2012) 1649–1657.

[19] P. Memar, F. Faradji, A novel multi-class eeg-based sleep stage classification system, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (2017) 84–95.

[20] S. S. Prabhu, N. Sinha, Sleep eeg analysis utilizing inter-channel covariance matrices, Biocybernetics and Biomedical Engineering 40 (2020) 527–545.

[21] M. M. Rahman, M. I. H. Bhuiyan, A. R. Hassan, Sleep stage classification using single-channel eog, Computers in biology and medicine 102 (2018) 211–220.

[22] R. Sharma, R. B. Pachori, A. Upadhyay, Automatic sleep stages classification based on iterative filtering of electroencephalogram signals, Neural Computing and Applications 28 (2017) 2959–2978.

[23] M. Sharma, D. Goyal, P. Achuth, U. R. Acharya, An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank, Computers in biology and medicine 98 (2018) 58–75.

[24] A. Supratak, H. Dong, C. Wu, Y. Guo, Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg, IEEE Transactions on Neural Systems and Rehabilitation Engineering 25 (2017) 1998–2008.

[25] C. Sun, J. Fan, C. Chen, W. Li, W. Chen, A two-stage neural network for sleep stage classification based on feature learning, sequence learning, and data augmentation, IEEE Access 7 (2019) 109386–109397.

[26] O. Tsinalis, P. M. Matthews, Y. Guo, Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders, Annals of biomedical engineering 44 (2016) 1587–1597.

[27] O. Tsinalis, P. M. Matthews, Y. Guo, S. Zafeiriou, Automatic sleep stage scoring with single-channel eeg using convolutional neural networks, 2016. arXiv:1610.01683.

[28] K. Šušmáková, A. Krakovská, Discrimination ability of individual measures used in sleep stages classification, Artificial intelligence in medicine 44 (2008) 261–277.

[29] G. Zhu, Y. Li, P. Wen, Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal, IEEE journal of biomed. health informatics 18 (2014) 1813–1821.

[30] F. Ebrahimi, M. Mikaeili, E. Estrada, H. Nazeran, Automatic sleep stage classification based on eeg signals by using neural networks and wavelet packet coefficients, in: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2008, pp. 1151–1154.

[31] A. R. Hassan, M. I. H. Bhuiyan, Automatic sleep scoring using statistical features in the emd domain and ensemble methods, Biocybernetics and Biomedical Engineering 36 (2016) 248–255.

[32] A. R. Hassan, 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 271 (2016) 107–118.

[33] U. R. Acharya, S. Bhat, O. Faust, H. Adeli, E. C.-P. Chua, W. J. E. Lim, J. E. W. Koh, Nonlinear dynamics measures for automated eeg-based sleep stage detection, European neurology 74 (2015) 268–287.

[34] R. Agarwal, T. Takeuchi, S. Laroche, J. Gotman, Detection of rapid-eye movements in sleep studies, IEEE Transactions on biomedical engineering 52 (2005) 1390–1396.

[35] S. A. Imtiaz, E. Rodriguez-Villegas, A low computational cost algorithm for rem sleep detection using single channel eeg, Annals of biomedical engineering 42 (2014) 2344–2359.

[36] H. Phan, Q. Do, T.-L. Do, D.-L. Vu, Metric learning for automatic sleep stage classification, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 5025–5028.

[37] Z. Liu, J. Sun, Y. Zhang, P. Rolfe, Sleep staging from the eeg signal using multi-domain feature extraction, Biomedical Signal Processing and Control 30 (2016) 86–97.

[38] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, Y. Guo, Mixed neural network approach for temporal sleep stage classification, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (2017) 324–333.

[39] M. Ronzhina, O. Janoušek, J. Kolářová, M. Nováková, P. Honzík, I. Provazník, Sleep scoring using artificial neural networks, Sleep medicine reviews 16 (2012) 251–263.

[40] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (2018) 758–769.

[41] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. Hu, M. De Vos, Multichannel sleep stage classification and transfer learning using convolutional neural networks, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 171–174.

[42] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, M. De Vos, Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 1452–1455.

[43] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, M. De Vos, Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging, IEEE Transactions on Neural Systems and Rehabilitation Engineering 27 (2019) 400–410.

[44] N. Banluesombatkul, P. Ouppaphan, P. Leelaarporn, P. Lakhan, B. Chaitusaney, N. Jaimchariya, E. Chuangsuwanich, W. Chen, H. Phan, N. Dilokthanakul, et al., Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning, IEEE Journal of Biomedical and Health Informatics (2020).

[45] J. Zhang, R. Yao, W. Ge, J. Gao, Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel eeg, Computer methods and programs in biomedicine 183 (2020) 105089.

[46] H. Kim, S. Choi, Automatic sleep stage classification using eeg and emg signal, in: 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), IEEE, 2018, pp. 207–212.

[47] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220.

[48] S. Devuyst, The dreams databases and assessment algorithm, 2005. URL: https://doi.org/10.5281/zenodo.2650142. doi:10.5281/zenodo.2650142.

[49] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, S. Moslehpour, Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation, Entropy 18 (2016) 272.

[50] E. Olbrich, P. Achermann, P. Meier, Dynamics of human sleep eeg, Neurocomputing 52 (2003) 857–862.

[51] A. Accardo, M. Affinito, M. Carrozzi, F. Bouquet, Use of the fractal dimension for the analysis of electroencephalographic time series, Biological cybernetics 77 (1997) 339–350.

[52] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D: Nonlinear Phenomena 31 (1988) 277–283.

[53] Q. Wei, Q. Liu, S.-Z. Fan, C.-W. Lu, T.-Y. Lin, M. F. Abbod, J.-S. Shieh, Analysis of eeg via multivariate empirical mode decomposition for depth of anesthesia based on sample entropy, Entropy 15 (2013) 3458–3470.

[54] X. Chen, I. C. Solomon, K. H. Chon, Comparison of the use of approximate entropy and sample entropy: applications to neural respiratory signal, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 4212–4215.

[55] G. J. Jiang, S.-Z. Fan, M. F. Abbod, H.-H. Huang, J.-Y. Lan, F.-F. Tsai, H.-C. Chang, Y.-W. Yang, F.-L. Chuang, Y.-F. Chiu, et al., Sample entropy analysis of eeg signals via artificial neural networks to model patients' consciousness level based on anesthesiologists experience, BioMed research international (2015).

[56] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Transactions on information theory 22 (1976) 75–81.

[57] X.-S. Zhang, R. J. Roy, E. W. Jensen, Eeg complexity as a measure of depth of anesthesia for patients, IEEE Transactions on Biomedical Engineering 48 (2001) 1424–1433.

[58] R. Morales, T. Di Matteo, R. Gramatica, T. Aste, Dynamical generalized hurst exponent as a tool to monitor unstable periods in financial time series, Physica A: statistical mechanics and its applications 391 (2012) 3180–3189.

[59] H. Namazi, V. V. Kulish, J. Hussaini, J. Hussaini, A. Delaviz, F. Delaviz, S. Habibi, S. Ramezanpoor, A signal processing based analysis and prediction of seizure onset in patients with epilepsy, Oncotarget 7 (2016) 342.

[60] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, J. Rolink, Sleep stage classification with ecg and respiratory effort, Physiological measurement 36 (2015) 2027.

[61] Y. Shen, E. Olbrich, P. Achermann, P. Meier, Dimensional complexity and spectral properties of the human sleep eeg, Clinical Neurophysiology 114 (2003) 199–209.

[62] L. Xu, P. C. Ivanov, K. Hu, Z. Chen, A. Carbone, H. E. Stanley, Quantifying signals with power-law correlations: A comparative study of detrended fluctuation analysis and detrended moving average techniques, Physical Review E 71 (2005) 051101.

[63] B. Hjorth, Time domain descriptors and their relation to a particular model for generation of eeg activity, CEAN-Computerized EEG analysis (1975) 3–8.

[64] W. H. Kruskal, W. A. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American statistical Association 47 (1952) 583–621.

[65] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, Rusboost: A hybrid approach to alleviating class imbalance, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40 (2009) 185–197.

[66] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (2020) 1–13.

[67] D. Jiang, Y.-n. Lu, M. Yu, W. Yuanyuan, Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement, Expert Systems with Applications 121 (2019) 188–203.

[68] O. Yildirim, U. B. Baloglu, U. R. Acharya, A deep learning model for automated sleep stages classification using psg signals, International journal of environmental research and public health 16 (2019) 599.

[69] N. Seliya, T. M. Khoshgoftaar, J. Van Hulse, A study on the relationships of classifier performance metrics, in: 2009 21st IEEE international conference on tools with artificial intelligence, IEEE, 2009, pp. 59–66.

[70] H. Ghimatgar, K. Kazemi, M. S. Helfroush, A. Aarabi, An automatic single-channel eeg-based sleep stage scoring method based on hidden markov model, Journal of neuroscience methods 324 (2019) 108320.

[71] C. Vural, M. Yildiz, Determination of sleep stage separation ability of features extracted from eeg signals using principle component analysis, Journal of medical systems 34 (2010) 83–89.

[72] T. Zhu, W. Luo, F. Yu, Convolution-and attention-based neural network for automated sleep stage classification, International Journal of Environmental Research and Public Health 17 (2020) 4152.

[73] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-S. Cheng, A rule-based automatic sleep staging method, Journal of neuroscience methods 205 (2012) 169–176.

[74] B. Yang, X. Zhu, Y. Liu, H. Liu, A single-channel eeg based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden markov model, Biomedical Signal Processing and Control 68 (2021) 102581.

[75] H. Shen, M. Xu, A. Guez, A. Li, F. Ran, An accurate sleep stages classification method based on state space model, IEEE Access 7 (2019) 125268–125279.

[76] H. Shen, F. Ran, M. Xu, A. Guez, A. Li, A. Guo, An automatic sleep stage classification algorithm using improved model based essence features, Sensors 20 (2020) 4677.

[77] S.-L. Himanen, J. Hasan, Limitations of rechtschaffen and kales, Sleep medicine reviews 4 (2000) 149–167.

[78] P. Jadhav, G. Rajguru, D. Datta, S. Mukhopadhyay, Automatic sleep stage classification using time–frequency images of cwt and transfer learning using convolution neural network, Biocybernetics and Biomedical Engineering 40 (2020) 494–504.

[79] E. Khalili, B. M. Asl, Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel eeg, Computer Methods and Programs in Biomedicine 204 (2021) 106063.

[80] T. L. da Silveira, A. J. Kozakevicius, C. R. Rodrigues, Single-channel eeg sleep stage classification based on a streamlined set of statistical features in wavelet domain, Medical & biological engineering & computing 55 (2017) 343–352.

[81] S. Mousavi, F. Afghah, U. R. Acharya, Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach, PloS one 14 (2019) e0216456.

[82] R. Jain, A. G. Ramakrishnan, Electrophysiological and neuroimaging studies–during resting state and sensory stimulation in disorders of consciousness: a review, Frontiers in Neuroscience 14 (2020) 987.

[83] J. A. Christensen, M. Nikolic, S. C. Warby, H. Koch, M. Zoetmulder, R. Frandsen, K. K. Moghadam, H. B. Sorensen, E. Mignot, P. J. Jennum, Sleep spindle alterations in patients with parkinson's disease, Frontiers in human neuroscience 9 (2015) 233.

[84] F. Ferrarelli, G. Tononi, The thalamic reticular nucleus and schizophrenia, Schizophrenia bulletin 37 (2011) 306–315.

[85] V. Latreille, J. Carrier, M. Lafortune, R. B. Postuma, J.-A. Bertrand, M. Panisset, S. Chouinard, J.-F. Gagnon, Sleep spindles in parkinson's disease may predict the development of dementia, Neurobiology of aging 36 (2015) 1083–1090.

[86] D. Riemann, M. Berger, U. Voderholzer, Sleep and depression—results from psychobiological studies: an overview, Biological psychology 57 (2001) 67–103.

[87] M. J. Sateia, International classification of sleep disorders, Chest 146 (2014) 1387–1394.