Comparison of Feature and Classifier Algorithms for Online Automatic Sleep Staging Based on a Single EEG Signal

Mustafa Radha[1], Gary Garcia-Molina[2], Mannes Poel[3], and Giulio Tononi[4]

Abstract— **Automatic sleep staging on an online basis has recently emerged as a research topic motivated by fundamental sleep research. The aim of this paper is to find optimal signal processing methods and machine learning algorithms to achieve online sleep staging on the basis of a single EEG signal. The classification performance obtained using six different EEG signals and various signal processing feature sets is compared using the kappa statistic which has very recently become popular in sleep staging research. A variable duration of the EEG segment (or epoch) to decide on the sleep stage is also analyzed. Spectral-domain, time-domain, linear, and nonlinear features are compared in terms of performance and two types of machine learning approaches (random forests and support vector machines) are assessed. We have determined that frontal EEG signals, with spectral linear features, epoch durations between 18 and 30 seconds, and a random forest classifier lead to optimal classification performance while ensuring real-time online operation.**

I. INTRODUCTION

Sleep is a state of reversible disconnection from the environment characterized by quiescence and reduced vigilance. Although the precise function of sleep remains to be elucidated, it appears that sleep primarily benefits the brain. Recent research evidence indicates that modulating brain activity patterns during sleep via sensory, magnetic, or electric stimuli at specific sleep stages can be beneficial in a wide range of contexts including memory consolidation [8] and relief from depression [9]. To verify the validity of such interventions in practice requires automated means for online sleep staging.

In conventional sleep staging, experts examine polysomnography (PSG) signals which include electroencephalogram (EEG), electro-oculogram (EOG), and electro-myogram (EMG) to decide on sleep stages on the basis of 30-second long segments (*epochs*). Automated sleep staging algorithms have been primarily developed with the goal of assisting sleep technicians in the manual analysis of sleep recordings in an off-line mode. In this paper we focus on achieving online automatic sleep staging on the basis of a single EEG signal (or channel). For this purpose we consider several alternatives for: a) the single EEG signal (i.e. the electrodes location), b) the signal processing algorithms to extract features that characterize the sleep stages, and c) machine learning methods: random forests and support vector machines.

This paper is organized as follows. Section II summarizes the background information and state-of-the-art. Section III describes this paper dataset and methods. The results are presented in Section IV. Section V concludes the paper.

II. BACKGROUND

Two distinct types of sleep occur in humans: rapid eye movement (REM) sleep, and non-REM (NREM) sleep. Compared to the low voltage, high frequency patterns appearing in the awake EEG, NREM sleep is associated with a synchronized EEG pattern. NREM is subdivided into stages N1, N2, and N3. During REM, the EEG exhibits a pattern similar to that observed during wakefulness [10]. There are several ways in which the overnight EEG can be modeled to find the patterns of different sleep stages. In [4] three families of features are described, being the frequency, temporal and non-linear domain features.

Frequency models describe global trends for the EEG power during sleep in the classical frequency bands: delta, (δ: 0.5-4 Hz), theta (θ: 4-8 Hz), alpha (α: 8-12 Hz), sigma (σ: 11-15 Hz), and beta (β: 15-30 Hz). As sleep deepens the power in the delta and theta bands increase whereas the power in the alpha, sigma, and beta bands follow a quasi-opposite trend. The K-complex band is also popular (K: 0.9-1.1 Hz) and reflects K-complex activity in NREM sleep. The EEG power spectrum density for different sleep stages (including wake) is represented in Fig. 1.

Temporal models assume that the EEG is generated by a generally unknown stochastic process and examine the statistical features of the process over time.

Non-linear models describe the EEG as a non-linear dynamical system. Non-linear EEG features describe the dynamics of the systems [11] (e.g. fractal dimension or entropy). While most non-linear features require phase-space reconstruction, a high-complexity operation, some methods exist that by-pass it and therefore are feasible in real-time applications.

A. Automated sleep staging using the EEG

Table I shows a summary of results for published research on automatic sleep annotation based on 30 second long epochs. The results are given in Cohen's kappa as it is a metric which is not affected by class imbalance such as

[1] M. Radha is with Philips Group Innovation, 34 High Tech Campus, Eindhoven, AE5656, The Netherlands mustafa.radha at philips.com

[2] G. Garcia-Molina is with Philips Group Innovation, 345 Scarborough Rd., Briarcliff Manor, NY 10510, USA gary.garcia at philips.com

[3] M. Poel is with Human Media Interaction, University of Twente, Drienerlolaan 5, Enschede, 7522 NB, The Netherlands m.poel at utwente.nl

[4] G. Tononi is with Department of Psychiatry, University of Wisconsin, Madison, WI 53719, USA gtononi at wisc.edu at philips.com

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

TABLE I

SUMMARY OF RESULTS FOR AUTOMATED SLEEP STAGING METHODS FOR WHICH κ IS REPORTED.

Signal type	Feature type	Classification method	# of subjects	κ	Reference
2 EOG	Spectral correlation	Rule-based	265	0.63	[1]
6 EEG, 2 EOG, EMG	α, spindle, Slow wave sleep ratio	Rule-based	20	0.79	[2]
Single EEG	α, β, θ, δ, spindle, K-complex	Random forest	16	0.76	[3]
Single EEG	Spectral and non-linear features	SVM ensemble	28	0.86	[4]
ECG	Power Spectrum Density; heart-beat variability	Hidden Markov model	18	0.43	[5]
Single EEG	Power Spectrum Density	Vector Quantization	12	0.64	[6]
Single EEG	Power in δ, θ, α, σ, and β (RMS values)	Gaussian Mixture Model	10	0.63	[7]

Fig. 1. Normalized Welch power spectrum density for Wake, NREM, and REM stges.

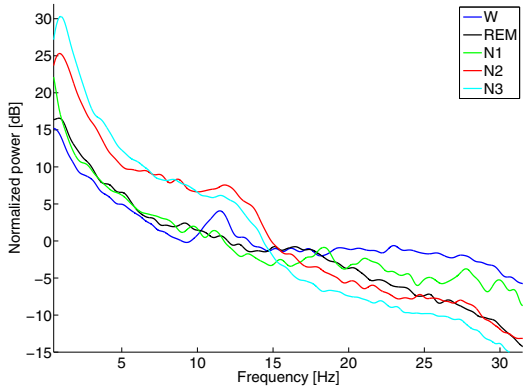


TABLE II

SUMMARY OF EXTRACTED FEATURES

Feature	Symbol
Frequency-domain	
Absolute spectral powers	$\delta_a, \theta_a, \alpha_a, \sigma_a, \beta_a, K_a$
Band power sum	$\delta + \theta$
Relative spectral powers	δ_r, θ_r, α_r, σ_r, β_r, K_r
Power ratios	δ/α, β/δ, θ/α and β/α
Spectral edges	SEF90, SEF95, SC
Spectral peak	sp
Spectral moments	$sm(1)$, $sm(2)$, $sm(3)$, $sm(4)$
Spectral entropy	H_s
Time-domain	
The Hjorth parameters	act, mob, com
Zero-crossing rate	Z_c
75th percentile	75P
Nonlinear	
Mutual information entropy	H_{mi}
Higuchi's fractal dimension	HD
Lempel-Ziv Complexity	LC

with sleep staging. It determines the agreement between annotators while factoring out chance agreement [7]. Kappa values up to 0.2 represent slight agreement while kappa up to 0.4, 0.6, 0.8 and 1 respectively represent fair, moderate, substantial and perfect agreement [12]. Table I shows that single EEG methods can perform just as well as full PSG methods but require more complex feature extraction and classification algorithms. The best performing classification algorithms are the support vector machine ensemble and the random forest.

III. METHODS

The data from 1 night of ten healthy subjects (five female; age 21.9 $\pm$ 0.5 yrs) who participated in a previous study [13] was used in this paper. From this dataset, all the 30-second long epochs of the six bipolar EEG signals (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2, and O1-A2) and the manually annotated hypnogram done by an expert (AASM rules, [10]) are used. The data was band-pass filtered in the frequency band from 0.6 to 27 Hz. The filtered data was down-sampled from 1024 Hz to 64 Hz. Epochs with excessively large-amplitude segments were removed as they are likely to correspond to artifacts.

A. Feature extraction

Per epoch, 34 features (frequency, time and non-linear types) were extracted (see Table II). Most features are taken from [4] where the origin and equations of each feature are given.

1) Frequency domain features: Frequency-based features were derived from the Welch [14] estimate of the power spectrum density function (PSD).

Band powers Absolute spectral band powers were extracted as the integral of the PSD for these bands. These are noted as $\delta_a, \theta_a, \alpha_a, \sigma_a,$, β_a and K_a. Furthermore the band-power sum $\delta + \theta$ was included as an indication of low frequency activity. Relative features of frequency ($\delta_r, \theta_r, \alpha_r, \sigma_r$, β_r and K_r) were also obtained by dividing the absolute power in a band by the total spectral power. Finally, ratios between different band powers were obtained: δ/α, β/δ, θ/α, and β/α. It was shown that such ratios strongly correlate with sleep stages [7].

Spectral edge as the frequency for which the power obtained by integrating the PSD from 0 to that frequency is equal to a given fraction r of the total power. The spectral edge frequencies SEF90 ($r = 0.9$), SEF95 ($r = 0.95$) and the spectral centroid SC ($r = 0.5$) were obtained.

Spectral peak (sp) is defined as the frequency at which the PSD is the highest. This feature is used in [15], [4]. This is written as sp.

Spectral moments of different orders $m = 1$ to $m = 4$ were estimated according to the equations in [16]. These features quantify high frequency momentum.

Spectral entropy (H_s) characterizes the non-uniformity of the PSD. It is equal to the negative value of the sum of spectral powers multiplied by their natural logarithms over all frequency bins.

2) Time domain features: To extract these features, the EEG is considered as a time series with samples $\{x[1], ..., x[|x|]\}$ where $|x|$ is the total number of data points. μ_x is the average of the data points, and std_x the standard deviation.

The **Hjorth parameters** are three classifical time features of the EEG. The activity (act) is defined as the standard deviation of the epoch. Mobility (mob) is the squared ratio of the activity of the derivative of the epoch to the activity of the epoch. Complexity (com) is the ratio of the mobility of the derivative of the epoch to the epoch itself.

The **zero-crossing rate** (Z_c) counts the number of crossings around the mean within a 3-second long moving average. This quantity is used in [4].

The **75th percentile** (75P) is the amplitude value that is larger or equal than 25% of the samples in the epoch. This percentile was used in [4] for sleep staging.

3) Non-linear features: Similarly to the time-domain features, an epoch is considered as a time series.

The **mutual information entropy** quantifies the unpredictability of the time series. This quantity is referred to as H_{mi} and is calculated as the negative of the sum of all points in the time series multiplied by their relative chance of occuring in the series (i.e. their a-priori probabilities).

The **Higuchi fractal dimension** [17] (HD) is a fast approximation of the fractal dimension, a powerful feature for seperating deep sleep [4]. The detailed equations to derive HD are described in [18].

The **Lempel-Ziv complexity** [4] (LC) provides a measure of complexity in the signal which can be calculated by setting a power threshold and counting each crossing of the threshold after a sub-sequence of consecutive values below or above the threshold. The optimized threshold [4] for sleep stage classification is $1.24 * \mu_x$.

B. Classification methods

Both the random forests (RF) [19] and the support vector machine (SVM) ensemble [4] were considered as classification algorithms. The Weka software [20] implementation of these algorithms was used. The RF decides on the sleep stage based on a weighted vote of ten decision trees with randomized inputs. The number of features to be used in random selection is $log(\#_F)+1$ where $\#_F$ is the number of total features being used for classification. These parameters are suggested in [3]. The second classifier, the SVM, is in principle a binary classifier. To cope with multiclass problems such as in sleep staging, ensembles of SVM's are employed with a voting strategy. There are two widely-used ensemble layouts being the one-versus-all (1vA) and one-versus-one (1v1) voting layouts [4] and they will both be tested here. Each SVM is trained using a polynomial kernel (with parameter set to 1). The soft margin parameter was set to 1, the epsilon parameter to 10^{-12}, and the tolerance parameter to 0.001.

C. Evaluation methods

The following procedure is applied to determine the feature subset, channel, epoch duration, and algorithm that leads to the best classification performance.

1) **Feature ranking**. All the features are first ranked according to their predictive power per channel. The ReliefF attribute evaluation method [21] is used.
2) **Channel, feature and epoch duration comparison**. The leave-one-subject-out cross-validation (L1SOCV) method (consisting in training with the data of all the subjects but one and testing on the remaining subject) is applied using an SVM 1vA ensemble for different: (a) channels, (b) subsets of top features, and (c) epoch durations to determine the combination that leads to the maximum performance.
3) **Algorithm selection**. Finally the different algorithms are tested using the best channel and feature subset to select the optimal classification algorithm. The classification performance is compared in terms of kappa.

TABLE III
TOP-10 FEATURES OVER ALL CHANNELS SELECTED USING THE RELIEFF [21] FEATURE EVALUATOR.

Rank	Feature	Count	Rank	Feature	Count
1	δ_r	6	6	θ/α	5
2	θ_r	6	7	H_{mi}	5
3	Z_c	6	8	σ_r	4
4	SEF95	6	9	LC	4
5	K_r	5	10	SEF90	4

IV. RESULTS AND DISCUSSION

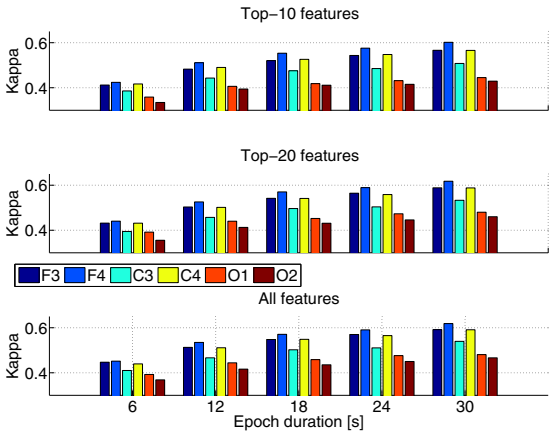
A. Feature ranking

As described in Section III-C, the ReliefF [21] feature ranking method was applied on all the data for each channel in a 10-fold cross-validation manner and rankings were averaged over each fold. Table III reports the top ten most occurring features over all channels, with the count indicating the number of channels in which they occur.

The high ranking of the δ, θ and K_r reveal the importance of the lower frequencies in sleep staging. The fact that relative band powers were chosen over absolute features of band power provides evidence that the PSD's total amplitude is specific to the context (e.g. subject or cycle) and simple normalization will improve performance. Other strong spectral predictors are the σ_r and θ/α features. The σ band reflects spindle activity and therefore can distinguish N2 from the other NREM stages. The θ/α ratio is powerful because it contrasts high frequency against low frequency. The SEF95 and SEF90 features are also strong features from the spectral domain.

As for the non-linear features, the mutual information entropy and Lempel-Ziv complexity features rank high in Table III yet they appear less important than linear features. The zero-crossing rate, while similar to Lempel-Ziv complexity, is ranked higher than it. This hints towards the validity of the moving average model which is assumed by the zero-crossing rate (but not by the Lempel-Ziv method). The zero-crossing rate was originally applied for the detection of spindles [22] and therefore serves as a

Fig. 2. Classification performance characterized by the kappa statistic for each epoch duration, EEG signal, and three feature sets (top-10, top-20, and the whole set).



similar function as the σ band.

B. EEG signal and feature selection

In an L1SOCV manner the κ performance was obtained for varying numbers of top features for each of the 6 EEG signals and for a variety of epoch lengths. The results are visualized in Figure 2.

As expected, larger epoch size and a higher amount of features always increase the performance. For practical purposes and taking into account that real-time operation is sought, the top-20 features can be used since adding more features only results in marginal increases. Frontal channels F3 and F4 lead to the best κ when compared to central (C3, C4) and occipital (O1, O2) channels. Signals originating at right sites (F4 and C4) have slightly better kappa values than their left counterparts (F3 and C3 respectively). This lateralization trend is not present on occipital channels. The F4-A1 signal has the best results.

Table IV lists the top-20 features for F4-A1. These features can be efficiently estimated in real-time. Most of these features result from linear spectral analysis of the signal. For the F4-A1 signal and the top-20 features a drop in kappa (difference = 0.17) occurs when 6-second long epochs (kappa=0.44; moderate agreement) are considered instead of 30-second long epochs (kappa=0.61; substantial agreement). Using 12 (kappa=0.53) or 18 (kappa=0.57) second long epochs can however ensure real-time operation with quasi-substantial agreement with professionally annotated data. In the context of this paper in which real-time sleep staging is considered for the purposes of applying an external intervention to influence sleep, it is assumed that being able to determine the sleep stage on the basis of shorter epochs is preferable for the timely intervention. This type of assumption applies in the memory consolidation enhancement cases considered in [8], [23].

TABLE IV

TOP 20 FEATURES FOR THE F4-A1 CHANNEL.

Rank	Feature	Rank	Feature	Rank	Feature
1	θ/α	8	SEF95	15	mob
2	K_r	9	σ_a	16	β_r
3	δ_r	10	LC	17	α_r
4	θ_r	11	θ_a	18	δ_a
5	H_{mi}	12	SEF90	19	θ_a
6	σ_r	13	HD	20	SC
7	C	14	K_a		

TABLE V

CLASS-BY-CLASS PRECISION AND RECALL, ACCURACY AND COHEN'S KAPPA PER ALGORITHM.

		SVM 1vA	SVM 1v1	RF
Precision	W	0.86	0.75	0.78
	REM	0.56	0.58	0.69
	N1	na[1]	0.18	0.52
	N2	0.86	0.85	0.85
	N3	0.32	0.82	0.83
Recall	W	0.51	0.71	0.73
	REM	0.55	0.79	0.70
	N1	0.00	0.00	0.31
	N2	0.83	0.88	0.91
	N3	0.70	0.70	0.73
Accuracy		0.69	0.77	0.80
Cohen's κ		0.49 $\pm$0.06	0.61 $\pm$0.06	0.66 $\pm$0.15

C. Classification algorithms

The classification algorithms from Section III-B were applied on the data from the signal (F4-A1) and its top 20 features. For each algorithm, the kappa statistic and accuracy were obtained from an L1SOCV procedure. In addition, sleep stage specific precision and recall values were estimated. All statistics are reported in table V.

The RF classifier has the highest average kappa but also the largest variance across subjects. The one-versus-all SVM ensemble led to the lowest performance. The SVM 1v1 ensemble performs better than the 1vA. Yet, this SVM approach did not perform well for the transitional N1 sleep stage.

The RF works by modeling decision rules and therefore resembles the AASM methodology of sleep staging. Yet it is more prone to overfitting than the statistical SVM and therefore will perform very well on "average" subjects while outliers will not work well. Limiting the tree depth might lower the standard deviation of the performance. The SVM approaches seem to cope poorly with the N1 sleep stage. Using more sophisticated kernel functions such as the Gaussian radial basis function may increase the SVM performance.

Yet, all 3 algorithms perform reasonably well given the very limited amount of data used. The studies shown in table I all use much bigger data sets. A future step is to test the selected classification algorithm with a larger data-set.

V. CONCLUSIONS

In this paper, the focus was on developing an automated sleep staging online algorithm using a single EEG signal. We have compared the classification performance in terms

of the kappa statistic for six EEG signal candidates, 5 epoch durations, and different types of signal features (time and frequency domain; linear and non-linear). In addition, we have also considered two of the best machine learning algorithms foudn in literature (support vector machine and random forest).

We have found that the EEG signals leading to the best classification performance are the ones corresponding to frontal channels. While the performance increases with the epoch duration a good compromise was found with an epoch duration of 18 seconds which can ensure online operation with a reasonable performance (kappa=0.57 quasi-substantial agreement with a professionally annotated hypnogram).

We have found that spectral linear features lead to higher performance than temporal and nonlinear features. Using relative frequency band powers (which are obtained by dividing by the total power or the power in another band) lead to higher performance than absolute ones. This is due to the within-subject normalization effect that is introduced by the ratio calculation.

The random forest had the highest performance (average kappa across subjects=0.66 for 30-second long epochs), higher than that of the SVM.

REFERENCES

[1] J. Virkkala, J. Hasan, A. Varri, S.-L. Himanen, and K. Muller, "Automatic sleep stage classification using two-channel electro-oculography," *Journal of neuroscience methods*, vol. 166, no. 1, pp. 109–115, 2007.

[2] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, and Y.-S. Cheng, "A rule-based automatic sleep staging method," *Journal of neuroscience methods*, vol. 205, no. 1, pp. 169–176, 2012.

[3] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and 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*, 2011.

[4] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in biology and medicine*, 2012.

[5] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Staging based on signals acquired. through bed sensor," *IEEE Trans Biomed Eng CHAPTER 7. BIBLIOGRAPHY*, vol. 149, 2009.

[6] I. J. Berezhnoy, G.-J. D. Vries, T. Weysen, J. Dimov, and G. Garcia-Molina, "Towards Unobtrusive Automated Sleep Stage: Polysomnography using electrodes on the face," in *HEALTHINF 2012; Int. Conf. on Health Informatics*, 2012.

[7] G. Garcia-Molina, M. Bellesi, S. Pastoor, S. Pfundtner, B. A. Riedner, and G. Tononi, "Online Single EEG Channel Based Automatic Sleep Staging," in *Engineering Psychology and Cognitive Ergonomics. Applications and Services*, D. Harris, Ed. Springer Berlin Heidelberg, 2013, pp. 333–342. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-39354-9_36#

[8] L. Marshall, H. Helgadóttir, M. Mölle, and J. Born, "Boosting slow oscillations during sleep potentiates memory," *Nature*, vol. 444, no. 7119, pp. 610–613, 2006.

[9] E. C. Landsness, M. R. Goldstein, M. J. Peterson, G. Tononi, and R. M. Benca, "Antidepressant effects of selective slow wave sleep deprivation in major depression: a high-density EEG investigation," *Journal of psychiatric research*, vol. 45, no. 8, pp. 1019–1026, 2011.

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

[11] D. H. Campos, "NON LINEAR TIME SERIES ANALYSIS OF THE EEG DURING SLEEP," *Revista de la Academia colombiana de ciencias exactas, físicas y naturales*, vol. 20, no. 78, p. 491, 1996.

[12] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/843571

[13] B. K. Hulse, E. C. Landsness, S. Sarasso, F. Ferrarelli, J. J. Guokas, T. Wanger, and G. Tononi, "A postsleep decline in auditory evoked potential amplitude reflects sleep homeostasis," *Clinical Neurophysiology*, vol. 122, no. 8, pp. 1549–1555, 2011.

[14] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, no. 2, pp. 70–73, 1967.

[15] D. Jordan, G. Stockmanns, E. F. Kochs, and G. Schneider, "Median frequency revisited: an approach to improve a classic spectral electroencephalographic parameter for the separation of consciousness from unconsciousness," *Anesthesiology*, vol. 107, no. 3, pp. 397–405, 2007.

[16] J. Fell, J. Röschke, K. Mann, and C. Schäffner, "Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures," *Electroencephalography and clinical neurophysiology*, vol. 98, no. 5, pp. 401–410, 1996.

[17] T. Higuchi, "Relationship between the fractal dimension and the power law index for a time series: a numerical investigation," *Physica D: Nonlinear Phenomena*, vol. 46, no. 2, pp. 254–264, 1990.

[18] W. Klonowski, E. Olejarczyk, R. Stepien, P. Jalowiecki, and R. Rudner, "Monitoring the depth of anaesthesia using fractal complexity method," *Complexus mundi. Emergent patterns in nature*, pp. 333–342, 2006.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Exploration*, vol. 11, no. 1, 2009.

[21] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.

[22] A. Krakovská and K. Mezeiová, "Automatic sleep scoring: A search for an optimal combination of measures," *Artificial Intelligence in Medicine*, vol. 53, no. 1, pp. 25–33, 2011.

[23] L. Marshall, M. Mölle, M. Hallschmid, and J. Born, "Transcranial direct current stimulation during sleep improves declarative memory," *The Journal of neuroscience*, vol. 24, no. 44, pp. 9985–9992, 2004.