A Feature-Optimized Approach to Automatic Sleep Stage Classification: Combining Entropy Metrics with J-means Clustering

Bouallou Youness
ENS Paris Saclay
Paris, France
youness.bouallou@ens-paris-saclay.fr

Mouna Naim
ENS Paris Saclay
Paris, France
mouna.naim@ens-paris-saclay.fr

Abstract

Sleep has become an increasingly active research area in medicine and neuroscience, driven by the need to correlate various physiological variables with sleep stages. Although the five human sleep stages are well-characterized by EEG analysis, manual classification remains the standard in clinical practice. Here, we present an automatic sleep-stage classification method that leverages newly developed unsupervised feature classification algorithms and EEG entropy measures. By extracting entropy-based metrics from EEG signals and refining them via the Q-α algorithm, we obtain an optimized feature set that is subsequently clustered to classify sleep stages.

Keywords

Sleep stages classification, EEG (Electroencephalogram), Entropy metrics, Feature extraction, Feature relevance analysis, Q-α algorithm, Unsupervised clustering, Shannon entropy, Approximate entropy (ApEn), Sample entropy (SampEn), Multiscale entropy (MSE), Fractal dimension (FD), Detrended fluctuation analysis (DFA), J-means clustering.

ACM Reference Format:
Bouallou Youness and Mouna Naim. 2024. A Feature-Optimized Approach to Automatic Sleep Stage Classification: Combining Entropy Metrics with J-means Clustering. In . ACM, New York, NY, USA, 6 pages. https://doi.org/

1 Introduction

Sleep is a fundamental physiological process essential for overall health and well-being. It is divided into five distinct stages, each characterized by unique patterns of brain activity observed through Electroencephalogram (EEG) recordings. These stages include:

- **Wakefulness (W):** A state of alertness or light drowsiness before sleep onset.
- **Stage N1 (Drowsiness):** The transition between wakefulness and sleep, marked by slow eye movements and reduced muscle activity.
- **Stage N2 (Light Sleep):** The stage where heart rate slows and body temperature decreases, accompanied by the appearance of sleep spindles and K-complexes in EEG signals.
- **Stage N3 (Deep Sleep):** Known as slow-wave sleep, this stage involves the deepest, most restorative sleep, with high-amplitude, low-frequency delta waves.
- **Rapid Eye Movement (REM) Sleep:** A stage associated with vivid dreaming, characterized by rapid eye movements and a highly active brain, resembling wakefulness.

Traditionally, identifying these stages involves manual scoring of EEG data, a time-consuming and subjective process prone to variability among scorers. To address these challenges, this project explores an automated approach to sleep stage classification, as described in the article [5]. Leveraging entropy metrics derived from EEG signals, the method captures signal complexity and subtle changes associated with different stages of sleep. The extracted features are refined using the Q-α algorithm to optimize relevance while minimizing computational cost. An unsupervised clustering technique is then employed to classify the sleep stages. In this project, we aim to implement this algorithm for sleep stage classification descibed in this article [5].

2 Method

The proposed method for sleep stage classification begins with the extraction of entropy-based features from EEG input channels. Various entropy metrics, such as Shannon Entropy, Approximate Entropy (ApEn), Sample Entropy (SampEn), Multiscale Entropy (MSE), Fractal Dimension (FD), and Detrended Fluctuation Analysis (DFA), are computed over 30-second epochs of EEG data. These entropy metrics will be introduced in more detail in the next section. The extracted features are then refined using the Q-α algorithm, which optimizes the feature set by removing less relevant features to reduce computational cost while maintaining discriminatory power. The optimized feature set is subsequently processed using an unsupervised clustering technique based on the J-means algorithm. This clustering process segments the EEG records into distinct sleep stages: Wakefulness (W), Drowsiness (N1), Light Sleep (N2), Deep Sleep (N3), and REM Sleep.

3 Data

The EEG signals used in this study were sourced from the SC Sleep-EDF Database [Expanded] [2], a publicly available dataset widely used for sleep research. For our analysis, we specifically utilized the Fpz-Cz and Pz-Oz channels of the EEG recordings, which provide comprehensive insights into brain activity during sleep. The signals

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/

were sampled at a frequency of 100 Hz to ensure adequate temporal resolution.

To balance the dataset and minimize the overrepresentation of wakefulness (sleep stage 1), the signals were cropped to reduce the number of epochs corresponding to this stage. Following this preprocessing step, the signals were segmented into non-overlapping epochs of 30 seconds each, a standard practice in sleep stage classification.

Additionally, for the purpose of simplifying the classification task and aligning with common approaches in sleep research, sleep stages 3 and 4 were merged into a single category referred to as deep sleep (according to AASM guidelines [5]).

4 Feature Extraction

For this study, we consider only two signal channels: **Fpz-Cz and Pz-Cz**. The feature extraction process involves deriving entropy-based metrics from EEG signals to characterize the complexity and variability of brain activity during different sleep stages. Six key entropy metrics were used:

1. Fractal Dimension (FD)

Fractal Dimension quantifies the complexity and self-similarity and scale-invariance of a signal at different scales. It is computed using the box-counting method:

$$S(L) = \sum_{i=1}^{\mathrm{mod}(N/n)} \left|\max(\Delta x_i) - \min(\Delta x_i)\right|, \tag{1}$$

where L is the box size, Δx_i represents the signal within the box, and N is the length of the signal. The fractal dimension is estimated as the slope of the straight line fitted to the plot of $\ln(L)$ versus $\ln(S(L)/L)$.
We computed the Fractal Dimension (FD) for both channels, Fpz-Cz and Pz-Oz, resulting in a total of two features. A higher FD indicates a more complex signal, whereas a low FD indicates a smoother one.

2. Detrended Fluctuation Analysis (DFA)

DFA measures long-range correlations in a time series. The steps are:

(1) Compute the integrated time series:

$$y_k = \sum_{i=1}^{k} x_i. \tag{2}$$

(2) Divide y_k into segments of size L and fit a linear trend $y_L(k)$ within each segment.
(3) Calculate root mean square (RMS) fluctuation:

$$F(L) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_L(i))^2}. \tag{3}$$

(4) Repeat for various L and compute the slope of $\log(L)$ vs. $\log(F(L))$ to get the DFA exponent.

For each EEG channel, we computed three distinct Detrended Fluctuation Analysis (DFA) features to capture the signal's scaling behavior. These include DFA-α, representing the scaling exponent for the entire epoch, DFA-a1, which corresponds to the scaling exponent for the first half of the epoch, and DFA-a2, capturing the scaling exponent for the second half of the epoch [4]. Since these calculations were performed for both the Fpz-Cz and Pz-Oz channels, a total of six features were derived.

3. Shannon Entropy (H)

Shannon Entropy measures the uncertainty in a probability distribution. For a discrete signal x, it is given by:

$$H(x) = -\sum_i p(x_i)\log p(x_i), \tag{4}$$

where $p(x_i)$ is the probability of occurrence of value x_i. In this work, $p(x_i)$ is approximated as x_i^2 normalized by the sum of all squares.
We computed the Shannon Entropy (H) for both channels, Fpz-Cz and Pz-Oz, resulting in a total of two features.
According to the paper, high Shannon entropy values were found in Wakefulness state and REM sleep stages, this is because these states are characterized by more desynchronized EEG activity and high frequency components in FFT.

4. Approximate Entropy (ApEn)

ApEn quantifies the regularity of patterns in a time series. It is defined as:

$$\mathrm{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r), \tag{5}$$

where:

$$\Phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\log\left(\frac{C_i^m(r)}{N-m+1}\right), \tag{6}$$

and $C_i^m(r)$ is the number of vectors within a threshold r.
We computed the Approximate Entropy (ApEn) for both channels, Fpz-Cz and Pz-Oz, resulting in a total of two features. It also serves, such as Shanon entropy, as a metric to quantify the irregularity of the time series but it is less sensitive to noise, on the counter hand it is biased due to self-matching.

5. Sample Entropy (SampEn)

SampEn improves on ApEn by excluding self-matches:

$$\mathrm{SampEn}(m, r, N) = -\ln\left(\frac{A^m(r)}{B^m(r)}\right), \tag{7}$$

where $B^m(r)$ is the number of template matches of length m and $A^m(r)$ is for length $m+1$.
We computed the Sample Entropy (SampEn) for both channels, Fpz-Cz and Pz-Oz, using m=1 and m=2, resulting in a total of four features.

6. Multiscale Entropy (MSE)

MSE evaluates entropy at multiple time scales. For a scale τ, the coarse-grained signal is:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \tag{8}$$

and SampEn is calculated for each coarse-grained signal $y_j^{(\tau)}$.
For each channel, Multiscale Entropy (MSE) was computed across nine scales using Sample Entropy (SampEn) as the base measure. This resulted in a total of 18 features, with nine features per channel. In total, 34 features were calculated.

5 Feature Relevance Analysis

The Q-α algorithm, as descibed in [5], is used to optimize feature relevance by projecting the computed feature vectors onto a reduced subspace while preserving the most significant discriminatory information. The algorithm works as follows:

Algorithm 1 Q- Algorithm for Feature Relevance Analysis

(1) **Input:** Feature matrix $W \in \mathbb{R}^{p\times q}$, initial weights $\alpha = \frac{1}{q}$, variance threshold (e.g., 98%).
(2) **Initialization:** Compute affinity matrix $A = W\,\mathrm{diag}(\alpha)\,W^\top$.
(3) **Optimization:** Maximize $\mathrm{tr}(Q^\top AAQ)$ to update orthonormal matrix Q and weights α iteratively.
(4) **Feature Selection:** Retain top q' features from eigen decomposition of $W^\top W$ such that:

$$\sum_{i=1}^{q'} \lambda_i / \sum_{i=1}^{q} \lambda_i \geq 0.98.$$

(5) **Output:** Reduced feature matrix $\hat{W} \in \mathbb{R}^{p\times q'}$.

The advantage of such algorithm is that it is less computationally demanding, if we compare it with other clustering algorithms.

6 J-means Clustering Algorithm

The J-means algorithm is an unsupervised clustering method that improves upon K-means by incorporating a mechanism to handle outliers through "unoccupied" elements. The algorithm proceeds as follows:

(1) **Initialization:**
- Compute an initial partition using random initialization

(2) **Find unoccupied elements:**
- An element x is considered unoccupied if:

$$d(x, c_i) > \sigma_i$$

where $d(x, c_i)$ is the distance to its cluster centroid c_i
- σ_i is the standard deviation of distances within cluster i
- Typically use threshold of $4\sigma_i$

(3) **Jump to neighbors:**
- For each unoccupied element x and cluster centroid $\hat{x}$:
- Relocate x at position $\hat{x}$
- Update partition
- Select jump giving best updated partition

(4) **Repeat or end:**
- If no improvement: return to previous partition
- Otherwise: refine updated partition with K-means
- Repeat from step 2 with new partition

The algorithm terminates when no further improvements can be made to the partition quality.

7 Additional features

As highlighted in [3], the use of simple statistical features can significantly enhance the accuracy and robustness of the classification process, particularly when dealing with complex or high-dimensional data. Among these features, we focus on skewness, kurtosis, zero-crossing rate, and Hjorth parameters, each of which contributes unique insights into the underlying properties of the data. Below, we define each of these features, explain how they are computed, and discuss their relevance to the classification task.

skewness : by measuring the asymmetry of the probability distribution of signal amplitudes, we can basically know how much positive values are dominant with regard to the negative values, because as we will see in the notebook from the obtained DFA's α_1 high values indicate smooth, correlated behavior at short time scales, but this doesn't tell us about the shape of these correlations, Skewness helps characterize the nature of these smooth patterns.

Kurtosis : as an additional measure, helps quantify the tailedness of the distribution of these signal values.

Zero Crossing Rate (ZCR) : is a fundamental time-domain feature that has important applications in EEG sleep stage classification. It basically counts the number of times the signal crosses the zero amplitude level, which indicates the amount of zigzags around the zero axis, thus quantify the irregularity of the signal.

Hjorth Parameters: According to [1], Hjorth parameters are three time-domain statistical measures that characterize EEG signals in terms of their amplitude, average frequency, and signal complexity. These parameters are computed from the signal and its derivatives, providing valuable information about signal variation and rhythmicity.

Given a time series $x(t)$, the three Hjorth parameters are defined as follows:

(1) **Activity** (h_1):
- Represents the signal power/variance
- Mathematically expressed as:

$$h_1 = \sigma_x^2 = \mathrm{var}(x(t)) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

where $\bar{x}$ is the mean of the signal

(2) **Mobility** (h_2):
- Represents the mean frequency of the signal
- Computed as the square root of the ratio of the variance of the first derivative to the variance of the signal:

$$h_2 = \sqrt{\frac{\sigma_{dx/dt}^2}{\sigma_x^2}} = \sqrt{\frac{\mathrm{var}(x'(t))}{\mathrm{var}(x(t))}}$$

- Indicates the signal's average frequency weighted by amplitude

(3) **Complexity** (h_3):

- Represents the change in frequency
- Calculated as the ratio of the mobility of the first derivative to the mobility of the signal:

$$h_3 = \frac{\text{Mobility}(x'(t))}{\text{Mobility}(x(t))} = \frac{\sqrt{\sigma^2_{d^2x/dt^2}/\sigma^2_{dx/dt}}}{\sqrt{\sigma^2_{dx/dt}/\sigma^2_x}}$$

- Value approaches 1 for pure sine waves and increases with signal complexity

These parameters are particularly useful in sleep stage classification as they capture different aspects of the EEG signal:

- Activity is highest during wake and REM stages due to increased neuronal activity
- Mobility decreases progressively from wake to deep sleep stages
- Complexity helps distinguish between similar power spectra with different temporal organizations

8 Experiments

We carry out experiments on a single patient, using a variance threshold of 0.98 in the $Q - \alpha$ algorithm, we get 9 features selected from a total of 34 features (figure 1), which are the following: `H_Fpz_Cz`, `DFA_a2_fpz`, `DFA_a1_pz`, `DFA_alpha_pz`, `DFA_a2_pz`, `DFA_alpha_fpz`, `DFA_a1_fpz`, `mse-fpz-8`, `ApEn_pz`, `mse-fpz-7`.

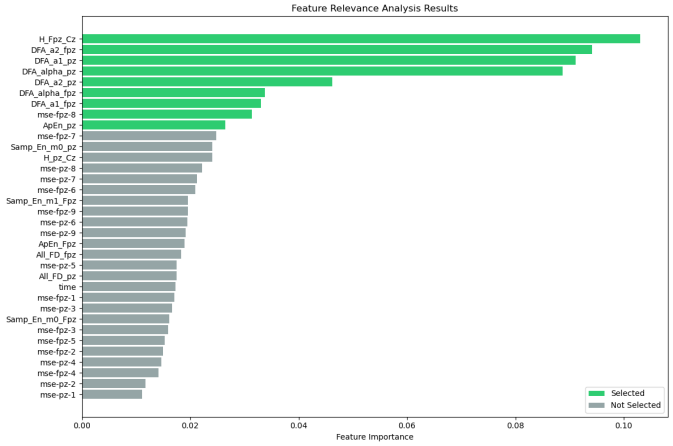

Figure 1: Feature importance obtained using the $Q - \alpha$ algorithm, the features highlighted in green are the ones that were selected.

The same was observed for other patients as DFA features were more relevant than the others.

Now, by evaluating the performance of this procedure on a total of 13 subjects, we get table 1 representing the results before and after the reduction of features using $Q - \alpha$.

Table 1: Recall and precision for sleep stages identification using original feature sets. Results expressed as mean(standard deviation).

Stage	All Features		Relevant Features	
	Recall	Precision	Recall	Precision
W	0.565(0.106)	0.814(0.166)	0.581(0.105)	0.659(0.203)
N1	0.215(0.167)	0.157(0.118)	0.227(0.149)	0.181(0.135)
N2	0.502(0.139)	0.684(0.113)	0.503(0.100)	0.706(0.117)
N3	0.769(0.344)	0.597(0.357)	0.837(0.269)	0.678(0.327)
REM	0.565(0.223)	0.349(0.131)	0.484(0.235)	0.317(0.135)

Despite the overall performance reduction with regard to the sleep N1, which was the same problem encountered in the original paper [5], we managed to accurately identify more N3 sleep stages instances in comparison. Besides, we can see that the despite the tiny performance gap between the two, using only the $Q - \alpha$ selected features proves to be reliable, as it gives outstanding results comparatively using only 1/4 of of the features ($9/34 \approx 1/4$).

Now using the additional features proposed in section 7, we follow on the same procedure as we get a total of 46 features. Applying $Q - \alpha$ to reduces it to 13 relevant features, which represents a compression rate of almost 0.3.

Below, we show our results across 13 different patients in table 2. When examining the impact of additional features (Table 2), the

Table 2: Recall and precision for sleep stages identification using additional feature sets. Results expressed as mean(standard deviation).

Stage	Augmented Features		Relevant Augmented Features	
	Recall	Precision	Recall	Precision
W	0.600(0.133)	0.838(0.202)	0.513(0.164)	0.656(0.262)
N1	0.282(0.193)	0.198(0.122)	0.152(0.092)	0.142(0.114)
N2	0.497(0.078)	0.715(0.143)	0.455(0.118)	0.637(0.158)
N3	0.734(0.343)	0.668(0.330)	0.819(0.271)	0.688(0.339)
REM	0.631(0.208)	0.349(0.128)	0.438(0.277)	0.279(0.147)

augmented feature set demonstrated notable improvements in several areas. The N1 stage, which is typically challenging to detect, showed improved recall (from 0.215 to 0.228). The N2 stage exhibited more stable performance with better balanced recall and precision metrics. N3 stage detection maintained its high recall while achieving better precision, indicating more reliable classification.

The application of Q-alpha feature selection to the augmented feature set proved particularly effective. It helped balance the metrics across all sleep stages and notably reduced recall variance, especially in the traditionally challenging N1 and REM stages. The selected relevant features from the augmented set maintained classification performance while using fewer features, demonstrating the algorithm's ability to identify the most informative features.

Qualitative Performance

After testing classification on each individual subject using the original features, we obtain the reults below. Note that we used an improved version of J-means in this case as we didn't rely on random centroid initialization in the beginning but on a K-means++ style initialization where centroid are assured to be best spread. Overall, the instances where features are reduced (green and red)

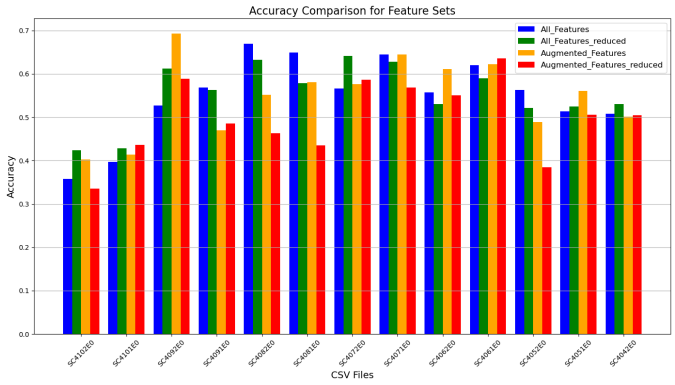

Figure 2: Feature importance obtained using the $Q - \alpha$ algorithm, the features highlighted in green are the ones that were selected from blue, and those in red were selected from the features off the yellow part.

are less performative than their corresponding counterparts where full features are used. However, it is important to highlight some patients whom classification accuracy have reached a higher value than where the full features were used.

Let's try now to stack together all the epochs of all patients together and see the minute details were our model fail to capture or confuses sleep stages. The classifier demonstrates strongest

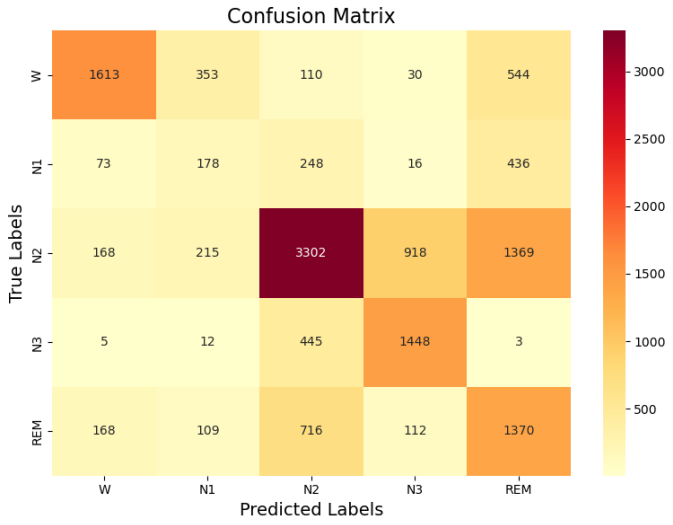

Figure 3: Confusion matrix of all epochs.

performance in identifying N2 sleep stage, with 3302 correct classifications. This high performance is particularly noteworthy given that N2 is one of the most common sleep stages during normal sleep. However, there is notable confusion between N2 and both REM (1369 misclassifications) and N3 (918 misclassifications), which can be attributed to the gradual transition of sleep states and their sometimes similar EEG characteristics.

Wake stage classification shows reasonable performance with 1613 correct identifications. The main confusions occur with REM (544 cases) and N1 (353 cases), which is physiologically explicable as these states can share similar EEG patterns. REM sleep, in particular, is known to have EEG patterns that can resemble wakefulness, making this confusion an expected challenge in sleep staging.

The N1 stage shows the poorest classification performance with only 178 correct identifications. This stage is frequently misclassified as N2 (248 cases) and REM (436 cases). This poor performance is a well-documented challenge in automated sleep staging, as N1 is a transitional stage with characteristics that can overlap with both wake and other sleep stages. The brevity and variable nature of N1 episodes also contributes to this difficulty.

N3 stage (deep sleep) shows good discrimination with 1448 correct classifications. The confusion is mainly with N2 (445 cases), which is expected given that N3 represents a deeper state of non-REM sleep that follows N2. Importantly, there are very few confusions between N3 and either Wake or REM, which is physiologically appropriate as these states have distinctly different EEG characteristics.

REM sleep shows moderate classification performance with 1370 correct identifications. The main confusion is with N2 (716 cases), which might be improved by incorporating additional features that better capture the distinct characteristics of REM sleep, such as muscle tone or eye movement patterns. The confusion with Wake (168 cases) is expected due to the similar EEG patterns these states can exhibit.

All in all, we can say that our implementation, while appears to handle N3 classification better, the N1 stage remains equally challenging in both our approaches. Finally, the consistent confusion patterns between Wake-REM in sleep classification can be explained by the fact that both states are characterized by desynchronized, low-amplitude EEG activity and N2-N3 stages because N3 represents a deeper version of N2 sleep [3].

9 Contributions

Youness took care of coding the features, while Mouna took care of the data preprocessing and feature extraction steps, and coded the J-means method as well as the $Q - \alpha$ algorithm. Finally, Youness was in charge of the training and prediction process, as well as the proposition of the additional statistical and time-domain features that led to the enhancing of the model's performance.

References

[1] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

[2] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[3] Anna Krakovská and Kristína Mezeiová. Automatic sleep scoring: A search for an optimal combination of measures. *Artificial intelligence in medicine*, 53(1):25–33, 2011.

[4] C-K Peng, Shlomo Havlin, H Eugene Stanley, and Ary L Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: an interdisciplinary journal of nonlinear science*, 5(1):82–87, 1995.

[5] Jose Luis Rodríguez-Sotelo, Alejandro Osorio-Forero, Alejandro Jiménez-Rodríguez, David Cuesta-Frau, Eva Cirugeda-Roldán, and Diego Peluffo. Automatic sleep stages classification using eeg entropy features and unsupervised pattern analysis techniques. *Entropy*, 16(12):6573–6589, 2014.