Accepted Manuscript

Noisy EEG signals classification based on entropy metrics. Performance assessment using first and second generation statistics

David Cuesta–Frau, Pau Miró–Martínez, Jorge Jordán Núñez, Sandra Oltra–Crespo, Antonio Molina Picó

PII: S0010-4825(17)30150-6

DOI: 10.1016/j.compbiomed.2017.05.028

Reference: CBM 2683

To appear in: *Computers in Biology and Medicine*

Received Date: 14 February 2017

Revised Date: 5 May 2017

Accepted Date: 28 May 2017

Please cite this article as: D. Cuesta–Frau, P. Miró–Martínez, Jorge.Jordá. Núñez, S. Oltra–Crespo, A. Molina Picó, Noisy EEG signals classification based on entropy metrics. Performance assessment using first and second generation statistics, *Computers in Biology and Medicine* (2017), doi: 10.1016/j.compbiomed.2017.05.028.

Noisy EEG Signals Classification Based on Entropy Metrics. Performance Assessment Using First and Second Generation Statistics

David Cuesta–Frau*, Pau Miró–Martínez, Jorge Jordán Núñez, Sandra Oltra–Crespo, Antonio Molina Picó

D. Cuesta–Frau, S. Oltra–Crespo and A. Molina–Picó are with the Technological Institute of Informatics, Polytechnic University of Valencia, Alcoi Campus, Plaza Ferrandiz y Carbonell 2, Alcoi, Spain

E-mail: dcuesta@disca.upv.es
Pau Miró–Martínez and Jorge Jordán Núñez are with the Department of Statistics, Polytechnic University of Valencia, Alcoi Campus, Alcoi, Spain

Abstract.

This paper evaluates the performance of first generation entropy metrics, featured by the well known and widely used Approximate Entropy (ApEn) and Sample Entropy (SampEn) metrics, and what can be considered an evolution from these, Fuzzy Entropy (FuzzyEn), in the Electroencephalogram (EEG) signal classification context. The study uses the commonest artifacts found in real EEGs, such as white noise, and muscular, cardiac, and ocular artifacts. Using two different sets of publicly available EEG records, and a realistic range of amplitudes for interfering artifacts, this work optimises and assesses the robustness of these metrics against artifacts in class segmentation terms probability. The results show that the qualitative behaviour of the two datasets is similar, with SampEn and FuzzyEn performing the best, and the noise and muscular artifacts are the most confounding factors. On the contrary, there is a wide variability as regards initialization parameters. The poor performance achieved by ApEn suggests that this metric should not be used in these contexts.

Keywords: Electroencephalograms , Signal Classification , Approximate Entropy , Sample Entropy , Fuzzy Entropy , EEG Artifacts

1. Introduction

Electroencephalography is a very important medical monitoring technique based on recording and analysing the brain's electrical activity. These recordings are termed electroencephalograms (EEGs), and are usually obtained non invasive by placing electrodes on the surface of scalps. The resulting time series can then be used to study the electrical activity of different brain regions and their correlation with clinical variables [1]. This analysis, performed by skilled operators using classical signal processing algorithms, was successfully used to assess a multitude of brain disorders, damage or processes.

For example, the authors in [2] propose a method based on the EEG power spectrum to estimate users' level of alertness while they performed critical tasks. Similarly, [3] report a method to classify states of fatigue and alertness while driving. Another field of extensive research is the assessment of sleep or anesthesia depth. In Rodriguez et al. [4], the authors describe an unsupervised sleep stages classification method based on pattern recognition techniques and a feature optimisation algorithm. EEG has also been used to evaluate the brain function after a stroke. The study [5] proposes a dense–array EEG to capture stroke effects, with a high correlation with the NIH stroke scale by partial least squares modelling. EEG and different types of dementia form another very active field of research. In [6], the authors carried out a meta–analysis based on 4157 papers to assess the correlation between abnormal EEGs and early–onset dementia (EOD). A clear relationship was found and demonstrated the capability of EEG to become a reliable tool for EOD diagnosis and prognosis. EEG analysis and processing can also contribute significantly to diagnosing and managing epilepsy [7] with a number of specific applications, such as seizure type determination or identification of epileptogenic regions, among many more.

However, not all the information provided by EEGs can be directly extracted because some information may be buried far down in the dynamics of the time series itself. In order to place this information within reach of the understanding of physicians, it is necessary to implement advanced mathematical methods and algorithms that extract additional subclinical information efficiently and expeditiously [8]. In line with this, one of the most successful groups of tools is the time series entropy estimation methods.

A diverse varied collection of these methods has been proposed in the last few decades, including Approximate Entropy, Sample Entropy, Fuzzy Entropy, Lempel–Ziv complexity, Permutation Entropy, Distribution Entropy, Renyi Entropy, Detrended Fluctuation Analysis, and some others, with a broad range of capabilities and applications in mainly economy and medicine. Specifically, in the field of EEG processing, two of the most widely used and successful entropy estimators are Approximate Entropy (ApEn) [9] and Sample Entropy (SampEn) [10], with hundreds of studies in the scientific literature.

ApEn quantifies the similarity probability of patterns of length m and $m + 1$.

Unlike other previous non linear methods, ApEn has demonstrated its robustness against noise and its capability to detect complexity changes using finite size datasets, and has provided at least 1000 data values whenever available [9]. By using similarity threshold r, defined as a fraction of the standard deviation of the input data, ApEn is also scale-independent. ApEn has been used to find EEG differences in schizophrenia patients [11], with lower entropy values obtained for these patients, or in comatose patients [12]. A significant number of studies has assessed anaesthesia depth where ApEn was the chosen tool, e.g., the work described in [13]. In that study, the ApEn metrics was able to track EEG changes in different anesthesia stages. Other research works have focused on measuring the effects of specific treatments or therapies on a range of neurological conditions through quantifiable changes in EEG. For example, the authors in [14] investigated the effect of current stimulation on aphasic patients. EEG changes due to aging or sleep have also been assessed using ApEn, as in [15], where ApEn was able to distinguish consciousness levels, and to find differences between age groups.

SampEn is a similar statistic. It also measures the probability of subsequences being close at two lengths m and $m+1$ within tolerance r. However, SampEn does not include self–comparisons and exhibits greater consistency than ApEn [16]. The algorithm to compute SampEn is also faster than that of ApEn, but its execution time is still $O(N^2)$, with N being the length of the time series [17]. SampEn has not yet been used as extensively as ApEn as this was proposed later, but it is quickly catching up given its better performance. The scope of application is very similar to that of ApEn. So, there are works that have studied EEG differences between control subjects and individuals with traumatic brain injury [18]. Sleep stages have also been classified using SampEn, as in [19, 20]. Alzheimer screening using EEG and SampEn is another promising area of research with already significant results [21].

ApEn and SampEn are very successful data entropy estimators, but they also have their weaknesses. As stated above, ApEn is biased since it includes self–matches in the count, and SampEn requires a relatively large r to find similar subsequences and to avoid the log(0) problem (Table 1). They are also very sensitive to input parameters m, r, and N. More recently, an evolution of these metrics, Fuzzy entropy (FuzzyEn), has been proposed to mitigate these problems [22]. FuzzyEn is based on a continuous function to compute the dissimilarity between two zero–mean subsequences and, consequently, it is more stable in noise and parameter initialisation terms. This metrics is still scarcely used in EEG studies, but it is expected to replace ApEn and SampEn because of its excellent stability, mainly when applied to noisy or short records. At present, very few studies have already demonstrated its capability to detect epileptic seizures [23], EEG abnormalities in Alzheimer's disease [24], or in recognizing wake or sleep stages [25, 26].

ApEn and SampEn have played, or are playing, a very important role in unveiling hidden information in EEGs, and will still be used for some time unless a more efficient metrics, such as FuzzyEn, completely replaces these older methods. To distinguish between these two generations of metrics, those initially proposed, even decades ago, and

those proposed less than 5 years ago as an evolution or improvement of the initial ones, we coined the terms first- and second-generation metrics, which will be used throughout this paper.

Signal classification efficiency is often assessed in relation to more robustness against difficult processing conditions: class separability, initialisation dependence, data size or noise. This paper focuses specifically on the effect on entropy metrics of EEG signals noise. Biomedical records are often corrupted with artifacts and noise, and EEGs are no exception. In general, biomedical record interferences can be of a physiological (EEGs are corrupted with data from other biosignals) or technical (EEGs are corrupted with noise generated by acquisition or other nearby systems) origin, with a myriad of methods to remove, or at least, reduce these artifacts proposed in the scientific literature [27, 28, 29, 30]. However, this is not always possible: signal and artifacts overlap in time and/or frequency domains (they cannot be removed without degrading the underlying valid signal), there is a high computational cost or complexity of the required algorithms, and the parameter optimization needs of filtering or cancelling methods cannot be addressed due to lack of time or resources.

As a result, a certain level of interference should be expected in any EEG signal, and the methods applied must therefore be robust against it. The present study addresses this issue by assessing of the performance of the above cited methods, ApEn, SampEn, and FuzzyEn, in the noisy EEG signal classification context. Specifically, we analyse the influence of the commonest physiological artifacts in EEG records: ocular artifacts [31], cardiac artifacts [32] and muscular artifacts [33]. The study also includes technical artifacts, such as noise and spikes [34]. The objective of the study is to improve the understanding of the metrics' behaviour under real conditions, and to provide practical advice about optimal performance.

The methodology employed is based on quantitative research. The analysis involves the collection of labelled EEG data, considered as the ground truth, since they do not contain artifacts (intra-cranial visually inspected EEGs), and apply a correlational research to find differences among the three entropy metrics studied (ApEn, SampEn, and FuzzyEn), based on a statistical treatment. The ultimate goal is to support or refute the robustness against artifacts hypothesis of each one of the metrics.

2. Materials and methods

2.1. Entropy metrics

The three entropy metrics chosen for this study are ApEn, SampEn, and FuzzyEn. ApEn and SampEn are undoubtedly the two most widely used indices for entropy estimations in physiological time series. FuzzyEn is an evolution of these two, where pattern dissimilarity computation has been improved by applying the fuzzy membership function concept instead of the Heaviside step function [35].

For a sequence $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ of size N, these metrics are mathematically

defined as described in Table 1:

Table 1. Mathematical definition of ApEn, SampEn, and FuzzyEn ($\mu(d,r)$:Fuzzy membership function).

	ApEn(m,r,N)	SampEn(m,r,N)	FuzzyEn(m,r,N)
1) Create a set of subsequences of length m	$\mathbf{x}_i = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}$ $i = 1, \ldots, N-m+1$	$\mathbf{x}_i = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}$ $i = 1, \ldots, N-m+1$	$\mathbf{y}_i = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}$ $\overline{y}_i$ =mean$(\mathbf{y}_i)$ $\mathbf{x}_i = \{x_i - \overline{y}_i, x_{i+1} - \overline{y}_i, \ldots, x_{i+m-1} - \overline{y}_i\}$ $i = 1, \ldots, N-m+1$
2) Dissimilarity computation	$d_{ij} = \max(\lvert x_{i+k} - x_{j+k}\rvert)$, $0 \le k \le m-1$	$d_{ij} = \max(\lvert x_{i+k} - x_{j+k}\rvert)$, $0 \le k \le m-1, j \ne i$	$d_{ij} = \max(\lvert x_{i+k} - x_{j+k}\rvert)$, $D_{ij} = \mu(d_{ij}, r), 0 \le k \le m-1, j \ne i$
3) Count matches	$B_i(r)$ no. of j so that $d[X_m(i), X_m(j)] \le r$ $A_i(r)$ no. of j so that $d[X_{m+1}(i), X_{m+1}(j)] \le r$ $(1 \le j \le N-m+1)$	$B_i(r)$ no. of j so that $d[X_m(i), X_m(j)] \le r$ $A_i(r)$ no. of j so that $d[X_{m+1}(i), X_{m+1}(j)] \le r$ $(1 \le j \le N-m, j \ne i)$	$\phi_i^m(r) = \frac{1}{N-m-1} \sum_{j=1, j\ne i}^{N-m} D_{ij}^m$
4) Statistics	$B_i^m(r) = \frac{1}{N-m+1} B_i(r)$ $A_i^m(r) = \frac{1}{N-m} A_i(r)$ $\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log B_i^m(r)$ $\phi^{m+1}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \log A_i^m(r)$	$B_i^m(r) = \frac{1}{N-m-1} B_i(r)$ $B^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r)$ $A_i^m(r) = \frac{1}{N-m-1} A_i(r)$ $A^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} A_i^m(r)$	$\varphi^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \phi_i^m(r)$ $\varphi^{m+1}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \phi_i^{m+1}(r)$
5) Result	$\mathrm{ApEn}(m,r) = \lim_{N\to\infty} \left[\phi^m(r) - \phi^{m+1}(r)\right]$ $\mathrm{ApEn}(m,r,N) = \left[\phi^m(r) - \phi^{m+1}(r)\right]$	$\mathrm{SampEn}(m,r) = \lim_{N\to\infty} \left(-\log\left[\frac{A^m(r)}{B^m(r)}\right]\right)$ $\mathrm{SampEn}(m,r,N) = -\log\left[\frac{A^m(r)}{B^m(r)}\right]$	$\mathrm{FuzzyEn}(m,r) = \lim_{N\to\infty} \left[\log\varphi^m(r) - \log\varphi^{m+1}(r)\right]$ $\mathrm{FuzzyEn}(m,r,N) = \left[\log\varphi^m(r) - \log\varphi^{m+1}(r)\right]$

The three metrics are computed similarly. First, the entire time data series under study is decomposed into subsequences of length m. Then dissimilarity is computed between subsequence x_i and another x_j one. While ApEn allows the case $i = j$ (self–matches), SampEn and FuzzyEn avoid this bias by setting $i \ne j$. Specifically, FuzzyEn removes each subsequence mean before computing this dissimilarity. Next the matches between subsequences are counted. This is an integer number for ApEn and SampEn, whereas it is the average of distances for all the neighboring vectors for FuzzyEn. Finally, the statistics for lengths m and $m+1$ are obtained, from which the final metrics result can be calculated. The computational cost of ApEn and SampEn is $O(N_2)$ [17], but it is $O(N_3)$ for FuzzyEn, because all the values in the subsequences have to be compared.

2.2. Experimental dataset

The experimental dataset was composed of the real EEG records obtained from different databases so as to ensure a rich varied set of features and properties. In addition, they do not contain significant acquisition artifacts to not interfere with the analysis since they were manually inspected to ensure that they were artifact-free [36]. The chosen databases were:

- The Bonn database [37]. This database is composed of 500 records from five different classes (100 records each). Sets A and B correspond to the surface EEG recordings of healthy subjects. Volunteers were awaken in a relaxed state, with their the eyes open (set A) or closed (set B). The surface electrodes were placed according to the standard 10–20 system [37]. Sets C, D, and E correspond to epileptic patients, obtained using intracranial electrodes placed as described in [37].

Set E contains seizure activity, whereas sets C and D contain only seizure-free activity. Each record contains 4096 samples, and a sampling rate of 173.61 Hz was used (23.6s duration). All the signals in this database were used in the experiments. An example of records of each class is shown in Figure 1.

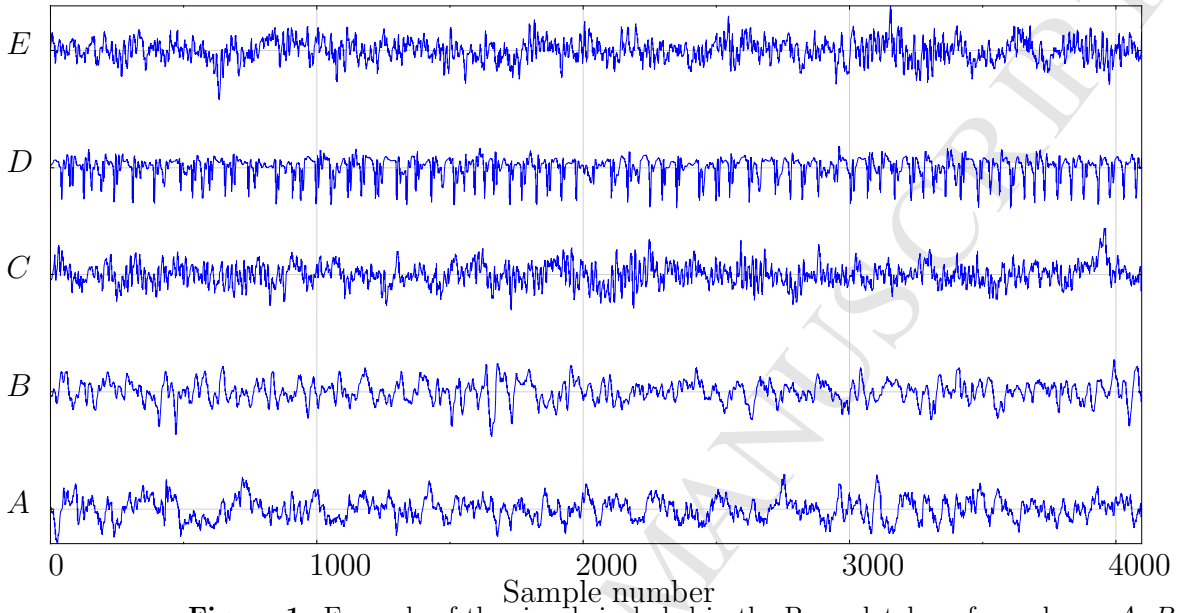

Figure 1. Example of the signals included in the Bonn database from classes A, B, C, D, and E. The 500 records are composed of 4096 samples obtained at 173.61 Hz (duration of 23.6s).

- The Bern–Barcelona database [36]. This database is composed of 3750 intracranial records from two classes. A surface electrode located between positions Fz and Pz was used as a reference. Set F corresponds to focal signals and set N to non focal records. Each series contains one pair of simultaneously recorded EEG signals (F_1, F_2 and N_1, N_2). Each record contains 10240 samples, and a sampling rate of 512 Hz was used (20s duration). Only a subset of 50 records per class and per pair was included in the experiments, which is also available at the database site (200 records). An example of the records of each class is shown in Figure 2.

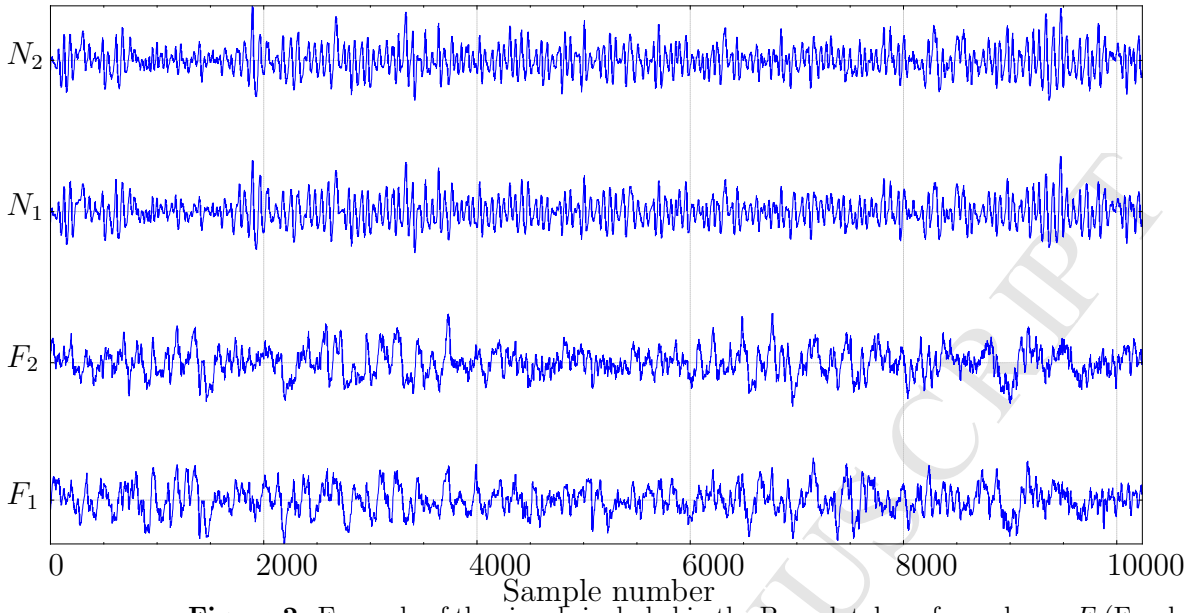

Figure 2. Example of the signals included in the Bern database from classes F (Focal, F_1 and F_2 pair) and N (Non Focal, N_1 and N_2 pair). Only 50 records per class and per pair of the 3750 records were used in the experiments (200 in all). Each record is composed of 10240 samples, obtained at 512 Hz (duration of 20s).

These databases correspond to intra–cranial EEGs actually. These signals usually exhibit a very low level of noise compared to extra–cranial EEGs, and therefore they can be considered as the ground–truth for the experiments, avoiding introducing bias to the results. In addition, the datasets have been classified successfully in other works [38, 39]. Thus, we simulate real extra-cranial EEGs by adding noise to initially clearly separable intra–cranial EEGs.

The noisy observations were obtained by linearly superimposing the synthetic artifacts to an otherwise pure, noise–free EEG signal. The resulting time series was normalised before computing the entropy metrics (zero mean and unit variance). The level of interference was in accordance with the type of artifact and with what occurs in a real clinical setting [27]. The signal to noise ratio (SNR) was 26dB, 20dB, 16dB, 12dB, and 10dB for noise, spikes, muscular, and cardiac artifacts, respectively, and 15dB, 9dB, 6dB, 4dB, and 2dB for ocular artifacts as their amplitude is usually larger. These SNR levels were chosen visually to resemble real cases. The length of all the records involved in the experiment was set at $N = 1000$ samples (the first 1000 values), which is long enough to ensure good entropy estimations [40]. The details of the employed artifacts are described bellow:

- White noise. This synthetic artifact was generated by a Gaussian random process. It accounts for possible sources in real environments, such as thermal noise or electro–magnetic noise.
- Spikes. Spikes were synthetically generated as described in [41]. These interferences can be of a technological (sensor movement, electrical interferences) or physiological (mainly eye blinks) origin. The probability of appearance was kept relatively low

(0.005), as expected in a real case. Duration was set at 1 sample. Only amplitude varied [41].

- Muscular artifacts. Muscular artifacts were drawn from a long real electromyogram (EMG) signal downloaded from PhysioNet [42] (`https://www.physionet.org/physiobank/database/emgdb/`), and corresponds to a patient with myopathy. Data were acquired at 50KHz and then downsampled to 4KHz. For each run, an EMG epoch of length N was extracted from the entire record by commencing at a random sample. These artifacts account for muscular activity during EEG recording.
- Ocular artifacts. These artifacts were obtained similarly to that of the EMG artifacts. From a real long electrooculogram (EOG) record processed with the EYE–EEG extension [43] (`http://www2.hu-berlin.de/eyetracking-EEG`), random segments were cut out for each experiment run. This record also includes high frequency artifacts, and white noise. This interference mainly causes the EEG baseline to drift, and can be of greater amplitude than that of the underlying EEG signal [44].
- Cardiac artifacts. Synthetic electrocardiogram (ECG) records were generated as described in [45] (`https://www.physionet.org/physiotools/ecgsyn/`). The average heart rate was set at 60 bpm, and amplitude was kept lower than that of the EEG following the above cited SNR levels. No additional noise was added.

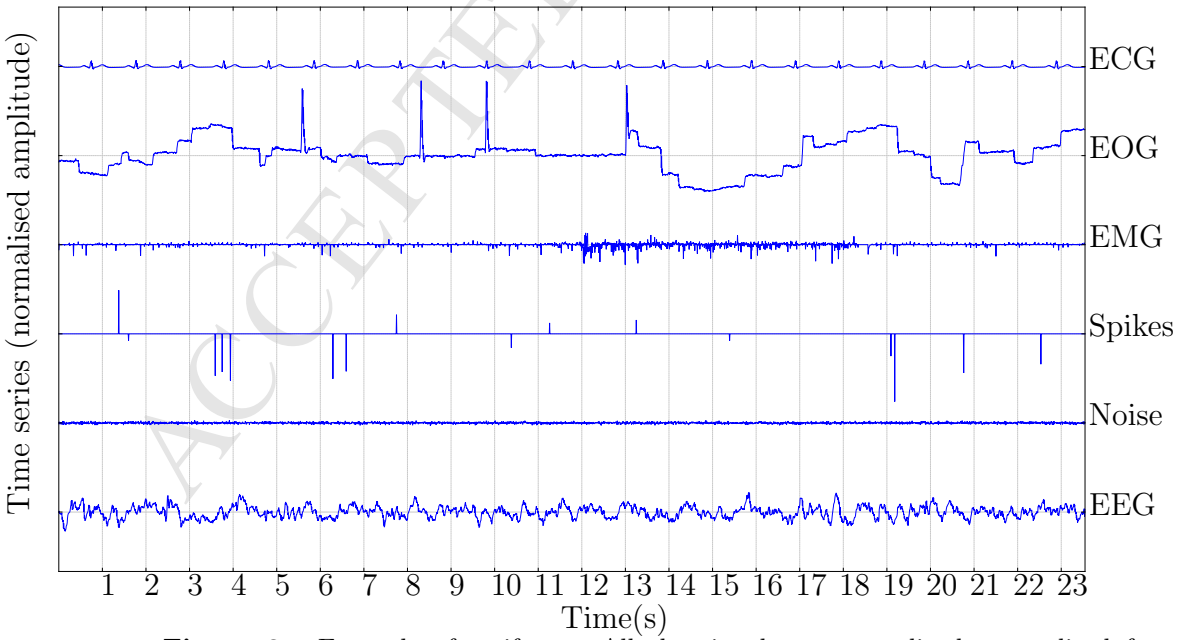

Figure 3. Example of artifacts. All the signals were amplitude-normalized for visualisation purposes. EOG, EMG, and spike artifacts may vary depending on the point from which they were extracted from the original record (EOG and EMG), or on the results of the Bernouilli process than sets occurrence and amplitude of spikes.

These artifacts were resampled to match the sampling frequency of the underlying

EEG signal. Due to the random nature of such artifacts, and their non stationarity (except the white noise and cardiac artifacts), each EEG record was corrupted in a slightly different manner which improved the generality of the results. An exemplary set of artifacts is shown in Figure 3, where this non stationarity can be easily observed for EOG, EMG, and spike artifacts.

2.3. Entropy metric parameter selection

The three entropy metrics require their input parameters m, r, and N to be initialized. As stated above, N was set to 1000 for all the experiments. This length significantly lowers the computational cost of the experiments, $O(N^2)$ or even $O(N^3)$, but preserves the stability and validity of the results. This value is in accordance with the suggestions made in [46, 47, 10, 40] ($N \geq 10^m$), and it is well above the minimum length required in other cases [48, 49, 50].

General recommendations exist for the other parameters: e.g., $m = 1$ or $m = 2$, and r in the $[0.2, 0.3]$ range [51]. Specifically for FuzzyEn, the recommendation for the membership function is to be continuous and convex [22]. Methods for the automatic selection of these parameters have also been proposed [52], but no general consensus about what method is best for each scenario has yet been reached.

We chose to find the optimal parameter configuration by maximizing the probability of class separation of the experimental dataset by minimizing the probability of equal EEG class means (null hypothesis) using the Student's t–test (when no artifact was present in the EEGs, a baseline case). A range of parameter values in the vicinity of the recommended ones was analyzed. For the m parameter, we studied the $p-$values obtained using $m = 1, 2, 3$. For r, performance was assessed using the values from 0.15 to 0.3 in steps of 0.05. For FuzzyEn, the chosen membership function was the exponential function, $\mu(d_{ij}, r) = \exp(-(d_{ij}/r)^q)$, as in many other works [22, 53]. In this case, there is an additional parameter to set, q. We attempted values 1, 2, 3 and 4 for q. Other membership functions, such as that described in [54], were also tested, but their performance was clearly lower (an equal means hypothesis accepted in more cases).

It is noteworthy that not all input classes are separable, even without artifacts, and such cases were not taken into account; e.g., for the Bern–Barcelona database, it is obviously impossible to discern between records within the same pair (F_1 and F_2, and N_1 and $N2$, cases 01 and 23 of the experiments, respectively). For the Bonn database, it is also impossible to find differences between records of healthy subjects with their eyes open or closed [37] (A and B, case 01 of the experiments).

Table 2 shows some of the parameter optimisation stage results. As stated above, some class combinations are impossible to distinguish because they are conceptually and analytically too similar, as other researchers also found [37]. Such results are also included in Table 2 to illustrate their consistency, but were omitted in the experiments that used artifacts. There are other input parameter values that yield worse class

separability, mainly when $m = 1$, and these combinations were avoided in the final tests. The optimal configuration is selected from the parameter settings that reject the equal mean hypothesis in all the separable cases. Some combinations yield negligible differences in $p-$values, as shown in Table 2 (mainly for Bern–Barcelona database). Those with a higher greater $p-$value are chosen.

Table 2. Parameter optimization results with $q = 3$ for FuzzyEn. An accepted hypothesis (equal means between classes) is featured by $p-$values in bold.

	Bern–Barcelona			Bonn		
	$m = 1, r = 0.15$	$m = 2, r = 0.25$	$m = 3, r = 0.3$	$m = 1, r = 0.15$	$m = 2, r = 0.25$	$m = 3, r = 0.3$
ApEn	$\mathbf{p_{01} = 0.379732}$	$\mathbf{p_{01} = 0.358934}$	$\mathbf{p_{01} = 0.400093}$	$\mathbf{p_{01} = 0.336989}$	$\mathbf{p_{01} = 0.051005}$	$\mathbf{p_{01} = 0.035396}$
	$p_{02} = 0.000087$	$p_{02} = 0.000076$	$p_{02} = 0.000090$	$p_{02} = 0.000038$	$p_{02} = 0.000026$	$p_{02} = 0.000025$
	$p_{03} = 0.000064$	$p_{03} = 0.000056$	$p_{03} = 0.000062$	$p_{03} = 0.000027$	$\mathbf{p_{03} = 0.212683}$	$\mathbf{p_{03} = 0.214318}$
	$p_{12} = 0.000515$	$p_{12} = 0.000541$	$p_{12} = 0.000547$	$p_{04} = 0.000030$	$p_{04} = 0.000025$	$p_{04} = 0.000026$
	$p_{13} = 0.000234$	$p_{13} = 0.000206$	$p_{13} = 0.000228$	$p_{12} = 0.000032$	$p_{12} = 0.000025$	$p_{12} = 0.000026$
	$\mathbf{p_{23} = 0.865867}$	$\mathbf{p_{23} = 0.830141}$	$\mathbf{p_{23} = 0.872516}$	$p_{13} = 0.000025$	$p_{13} = 0.000199$	$p_{13} = 0.000303$
				$p_{14} = 0.000026$	$p_{14} = 0.000026$	$p_{14} = 0.000025$
				$p_{23} = 0.000033$	$p_{23} = 0.000026$	$p_{23} = 0.000026$
				$\mathbf{p_{24} = 0.016003}$	$p_{24} = 0.000027$	$p_{24} = 0.000027$
				$p_{34} = 0.000026$	$p_{34} = 0.000026$	$p_{34} = 0.000025$
SampEn	$\mathbf{p_{01} = 0.371318}$	$\mathbf{p_{01} = 0.370324}$	$\mathbf{p_{01} = 0.398312}$	$\mathbf{p_{01} = 0.149373}$	$\mathbf{p_{01} = 0.036537}$	$\mathbf{p_{01} = 0.024077}$
	$p_{02} = 0.000067$	$p_{02} = 0.000059$	$p_{02} = 0.000630$	$p_{02} = 0.000036$	$p_{02} = 0.000026$	$p_{02} = 0.000025$
	$p_{03} = 0.000057$	$p_{03} = 0.000053$	$p_{03} = 0.000055$	$\mathbf{p_{03} = 0.382615}$	$p_{03} = 0.000196$	$p_{03} = 0.000057$
	$p_{12} = 0.000282$	$p_{12} = 0.000213$	$p_{12} = 0.000202$	$p_{04} = 0.000029$	$p_{04} = 0.000025$	$p_{04} = 0.000025$
	$p_{13} = 0.000149$	$p_{13} = 0.000107$	$p_{13} = 0.000113$	$p_{12} = 0.000031$	$p_{12} = 0.000025$	$p_{12} = 0.000026$
	$\mathbf{p_{23} = 0.882336}$	$\mathbf{p_{23} = 0.871806}$	$\mathbf{p_{23} = 0.908126}$	$\mathbf{p_{13} = 0.543533}$	$p_{13} = 0.000025$	$p_{13} = 0.000026$
				$p_{14} = 0.000026$	$p_{14} = 0.000026$	$p_{14} = 0.000026$
				$p_{23} = 0.000033$	$p_{23} = 0.000026$	$p_{23} = 0.000028$
				$\mathbf{p_{24} = 0.036623}$	$p_{24} = 0.000055$	$p_{24} = 0.000031$
				$p_{34} = 0.000026$	$p_{34} = 0.000026$	$p_{34} = 0.000028$
FuzzyEn	$\mathbf{p_{01} = 0.952577}$	$\mathbf{p_{01} = 0.502495}$	$\mathbf{p_{01} = 0.338560}$	$\mathbf{p_{01} = 0.402402}$	$\mathbf{p_{01} = 0.660507}$	$\mathbf{p_{01} = 0.875549}$
	$p_{02} = 0.009812$	$p_{02} = 0.000768$	$p_{02} = 0.000129$	$p_{02} = 0.000034$	$p_{02} = 0.000025$	$p_{02} = 0.000029$
	$p_{03} = 0.008163$	$p_{03} = 0.000533$	$p_{03} = 0.000097$	$p_{03} = 0.000038$	$p_{03} = 0.000028$	$p_{03} = 0.000026$
	$p_{12} = 0.006139$	$p_{12} = 0.002257$	$p_{12} = 0.000697$	$p_{04} = 0.000041$	$p_{04} = 0.000031$	$p_{04} = 0.000026$
	$p_{13} = 0.005240$	$p_{13} = 0.001534$	$p_{13} = 0.000387$	$p_{12} = 0.000030$	$p_{12} = 0.000025$	$p_{12} = 0.000028$
	$\mathbf{p_{23} = 0.894929}$	$\mathbf{p_{23} = 0.891299}$	$\mathbf{p_{23} = 0.843148}$	$p_{13} = 0.000035$	$p_{13} = 0.000027$	$p_{13} = 0.000026$
				$p_{14} = 0.000039$	$p_{14} = 0.000030$	$p_{14} = 0.000026$
				$p_{23} = 0.000027$	$p_{23} = 0.000027$	$p_{23} = 0.000032$
				$\mathbf{p_{24} = 0.033899}$	$p_{24} = 0.004689$	$p_{24} = 0.000079$
				$p_{34} = 0.000031$	$p_{34} = 0.000026$	$p_{34} = 0.000025$

After analyzing the $p-$values obtained using all these parameter configurations, the initialization parameters chosen for each experimental dataset were:

- Bonn database. The optimal parameter configuration found for ApEn and SampEn was the same: $m = 3$ and $r = 0.15$. The FuzzyEn optimal parameters were $m = 3$, $r = 0.3$, and $q = 4$. In this case, suboptimal configurations led to equal means acceptance in some class combinations (e.g., $p_{03} = 0.382615$, instead of $p_{03} = 0.00003$ in the optimal case). This occurred mainly for low m values, almost independently of the r values, for the three metrics. This was more severe with

FuzzyEn, with hypotheses rejected for $m = 1$ and $m = 2$ (classes not considered different).

- Bern–Barcelona database. Both ApEn and SampEn performed best for $m = 2$ and $r = 0.3$, and FuzzyEn with $m = 3$, $r = 0.15$, and $q = 1$. Differences were small, with ApEn and SampEn achieving the full rejection of all the cases (equal means rejected), but the optimal combination yielded higher probabilities and achieved the same rejection threshold (e.g., $p_{12} = 0.000282$ against $p_{12} = 0.000235$ in the optimal case). However for FuzzyEn, any combination that included a value $m < 3$, with $q = 4$, caused the hypothesis test to fail in some class comparisons. This suggests that FuzzyEn is very sensitive to the m parameter within this EEG analysis framework.

As a preliminary conclusion of this study, it seems that the values of $m = 1$ should be ruled out, with $m = 3$ being the most robust assumption as a general rule. There is wider variability for r, depending on the experimental set, and no recommendation can be made. As stated above, FuzzyEn seems the most parameter-sensitive metrics, conversely to what other researchers found [22], but in different contexts. No additional parameter values were studied since full separability (with the above-stayed exceptions) was already achieved with the proposed optimal parameter configurations.

3. Results

The separability of all the classes from the two datasets was assessed using the optimal parameter configuration described in the previous section. Classes were numbered as follows: 0 (records of type A), 1 (records of type B), 2 (records of type C), 3 (records of type D), and 4 (records of type E) for the Bonn database, and 0 (F_1), 1 (F_2), 2 (N_1), and 3 (N_2) for the Bern–Barcelona database. All the entropy means $\bar{\lambda}$ for all the classes compared on a one to one basis using a Student's t–test. The hypothesis was the equality of means $H_0 : \bar{\lambda}_i = \bar{\lambda}_j$ (null hypothesis), where $\bar{\lambda}_i$ is the average of the corresponding entropy statistic for class i. The $p-$value related to the consistency of the hypothesis of two classes i and j having the same mean was termed p_{ij}. The threshold for rejecting the null hypothesis was set at $\alpha = 0.01$. A smaller $p-$value rejects H_0 and, therefore, accepts the alternative hypothesis $H_1 : \bar{\lambda}_i \neq \bar{\lambda}_j$ [55]. In other words, if $p_{ij} < \alpha$, then we can consider it more likely that the means are different and, therefore, the classes can be more easily distinguished analytically. This experimental setting has been used in similar works, e.g. [56, 46]. No assumption about the probability distribution of the data was necessary as the mean was the only focus of the analysis with sample sizes of at least 50 [57].

3.1. Bern–Barcelona database

This section describes the results obtained using the Bern–Barcelona database. The baseline results (input signals without artifacts) are included only in the case of Gaussian

noise (Table 3) since they are the same for the other types of artifacts. The $p-$values for different levels of artifacts are shown in Tables 3-6.

Table 3. The results for the Bern–Barcelona database using different levels of Gaussian noise. The results for non separable classes (p_{01} and p_{23}) are not included.An accepted hypothesis is featured by $p-$values in bold.

	No artifact	**SNR**$(26dB)$	**SNR**$(20dB)$	**SNR**$(16dB)$	**SNR**$(12dB)$	**SNR**$(10dB)$
ApEn	$p_{02} = 0.000076$	$p_{02} = 0.000092$	$p_{02} = 0.000205$	$p_{02} = 0.001354$	$p_{02} = 0.002550$	$\mathbf{p_{02} = 0.011902}$
	$p_{03} = 0.000058$	$p_{03} = 0.000058$	$p_{03} = 0.000068$	$p_{03} = 0.000083$	$p_{03} = 0.000189$	$p_{03} = 0.000486$
	$p_{12} = 0.000538$	$p_{12} = 0.000778$	$p_{12} = 0.002861$	$\mathbf{p_{12} = 0.020972}$	$\mathbf{p_{12} = 0.032507}$	$\mathbf{p_{12} = 0.106199}$
	$p_{13} = 0.000214$	$p_{13} = 0.000212$	$p_{13} = 0.000539$	$p_{13} = 0.001152$	$p_{13} = 0.003253$	$p_{13} = 0.009348$
SampEn	$p_{02} = 0.000061$	$p_{02} = 0.000068$	$p_{02} = 0.000133$	$p_{02} = 0.000724$	$p_{02} = 0.002410$	$\mathbf{p_{02} = 0.015954}$
	$p_{03} = 0.000054$	$p_{03} = 0.000054$	$p_{03} = 0.000062$	$p_{03} = 0.000065$	$p_{03} = 0.000180$	$p_{03} = 0.000617$
	$p_{12} = 0.000235$	$p_{12} = 0.000351$	$p_{12} = 0.001504$	$\mathbf{p_{12} = 0.012484}$	$\mathbf{p_{12} = 0.034411}$	$\mathbf{p_{12} = 0.085657}$
	$p_{13} = 0.000117$	$p_{13} = 0.000116$	$p_{13} = 0.000357$	$p_{13} = 0.000639$	$p_{13} = 0.003717$	$p_{13} = 0.006086$
FuzzyEn	$p_{02} = 0.000061$	$p_{02} = 0.000068$	$p_{02} = 0.000090$	$p_{02} = 0.000320$	$p_{02} = 0.000585$	$p_{02} = 0.003072$
	$p_{03} = 0.000053$	$p_{03} = 0.000054$	$p_{03} = 0.000057$	$p_{03} = 0.000060$	$p_{03} = 0.000082$	$p_{03} = 0.000101$
	$p_{12} = 0.000211$	$p_{12} = 0.000342$	$p_{12} = 0.000781$	$p_{12} = 0.005157$	$p_{12} = 0.007527$	$\mathbf{p_{12} = 0.027387}$
	$p_{13} = 0.000096$	$p_{13} = 0.000111$	$p_{13} = 0.000194$	$p_{13} = 0.000390$	$p_{13} = 0.000738$	$p_{13} = 0.001026$

It can be noted from the $p-$values shown in Table 3 that ApEn and SampEn are very sensitive to presence of noise in EEG records. Even for levels that are barely discernible visually (16dB), they fail to provide a robust metrics capable of maximizing the separation between the means of most classes (except case 03, classes F_1 and N_2, and case 13, classes F_2 and N_2). FuzzyEn appears more robust against white noise because it does not fail until level 10dB, and also to a lesser extent (case 12). A visual example of the white noise impact on EEG records is shown in Figure 4 for all the studied levels.

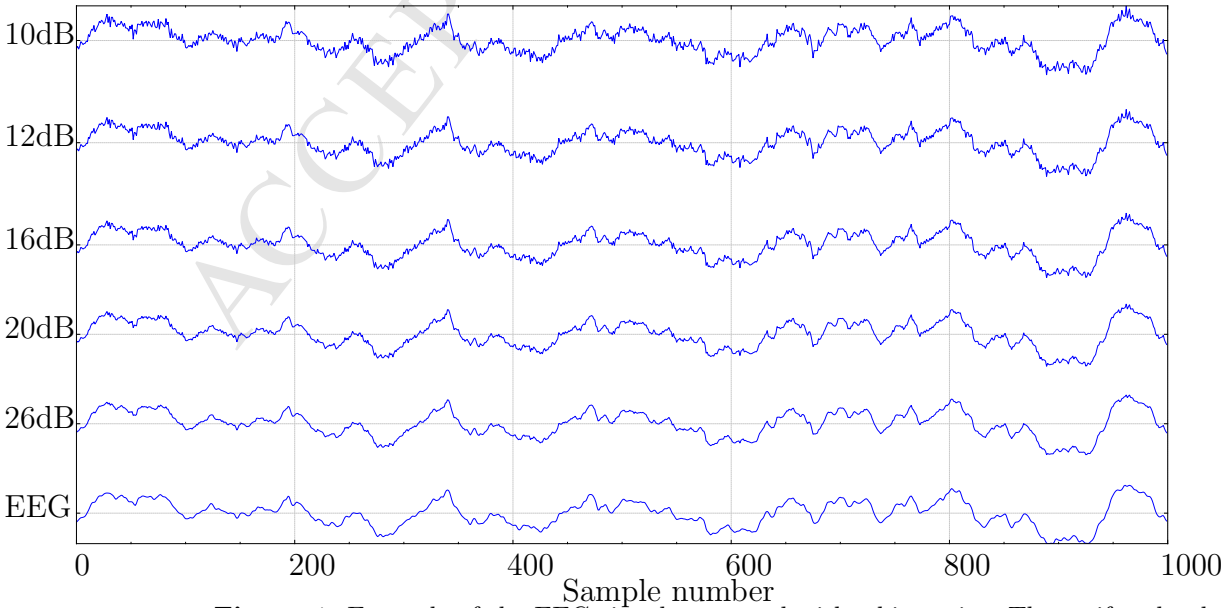

Figure 4. Example of the EEG signal corrupted with white noise. The artifact level lowers from top to bottom. The length of signals is 1000 samples, approximately 2s.

The Student's t–test is very useful for assessing the separability of the EEG signal groups in different scenarios, which is the main objective of the present paper. The entropy values obtained would be the features used by a classifier. However, this test does not quantify the correct classification rate that can be achieved or the optimal entropy threshold that should be used.

As stated above, data size N may also influence the results of the studied entropy metrics. Although the validity of the value employed, $N = 1000$, is justified in Section 2.3, and despite the fact that it is beyond the scope of the study to test a wide range of N values, Table 4 shows the results for SNR($10dB$), and $N = 1000, 2000, 3000$ and 4000.

Table 4. The results for Bern–Barcelona database using Gaussian noise (SNR($10dB$)) and different lengths (N). The results for non separable classes (p_{01} and p_{23}) are not included. An accepted hypothesis is featured by $p-$values in bold.

	$N = 1000$	$N = 2000$	$N = 3000$	$N = 4000$
ApEn	$\mathbf{p_{02} = 0.011902}$	$p_{02} = 0.009883$	$\mathbf{p_{02} = 0.018458}$	$\mathbf{p_{02} = 0.024699}$
	$p_{03} = 0.000486$	$p_{03} = 0.000922$	$p_{03} = 0.001628$	$p_{03} = 0.001784$
	$\mathbf{p_{12} = 0.106199}$	$\mathbf{p_{12} = 0.035246}$	$\mathbf{p_{12} = 0.047285}$	$\mathbf{p_{12} = 0.240016}$
	$p_{13} = 0.009348$	$p_{13} = 0.003384$	$p_{13} = 0.004749$	$\mathbf{p_{13} = 0.037036}$
SampEn	$\mathbf{p_{02} = 0.015954}$	$\mathbf{p_{02} = 0.010447}$	$\mathbf{p_{02} = 0.021456}$	$\mathbf{p_{02} = 0.026582}$
	$p_{03} = 0.000617$	$p_{03} = 0.001145$	$p_{03} = 0.002227$	$p_{03} = 0.002083$
	$\mathbf{p_{12} = 0.085657}$	$\mathbf{p_{12} = 0.026117}$	$\mathbf{p_{12} = 0.048137}$	$\mathbf{p_{12} = 0.245755}$
	$p_{13} = 0.006086$	$p_{13} = 0.002632$	$p_{13} = 0.005552$	$\mathbf{p_{13} = 0.040598}$
FuzzyEn	$p_{02} = 0.003072$	$p_{02} = 0.002202$	$p_{02} = 0.004631$	$p_{02} = 0.004806$
	$p_{03} = 0.000101$	$p_{03} = 0.000200$	$p_{03} = 0.000344$	$p_{03} = 0.000248$
	$\mathbf{p_{12} = 0.027387}$	$p_{12} = 0.005592$	$\mathbf{p_{12} = 0.010827}$	$\mathbf{p_{12} = 0.087149}$
	$p_{13} = 0.001026$	$p_{13} = 0.000391$	$p_{13} = 0.000781$	$p_{13} = 0.009276$

The experiment was repeated using spike artifacts. However, these results are not included because means were assumed different in all cases (all rejected hypotheses, $p_{ij} < \alpha, \forall i, j$ considered). With a probability of 0.005, and a duration of 1 sample, spikes did not seem to significantly impact the matches count and, therefore, impacted the entropy metrics [41]. Figure 5 shows an example of an EEG record corrupted with synthetic spikes. Further information about the influence of spikes on entropy metrics can be found in [41].

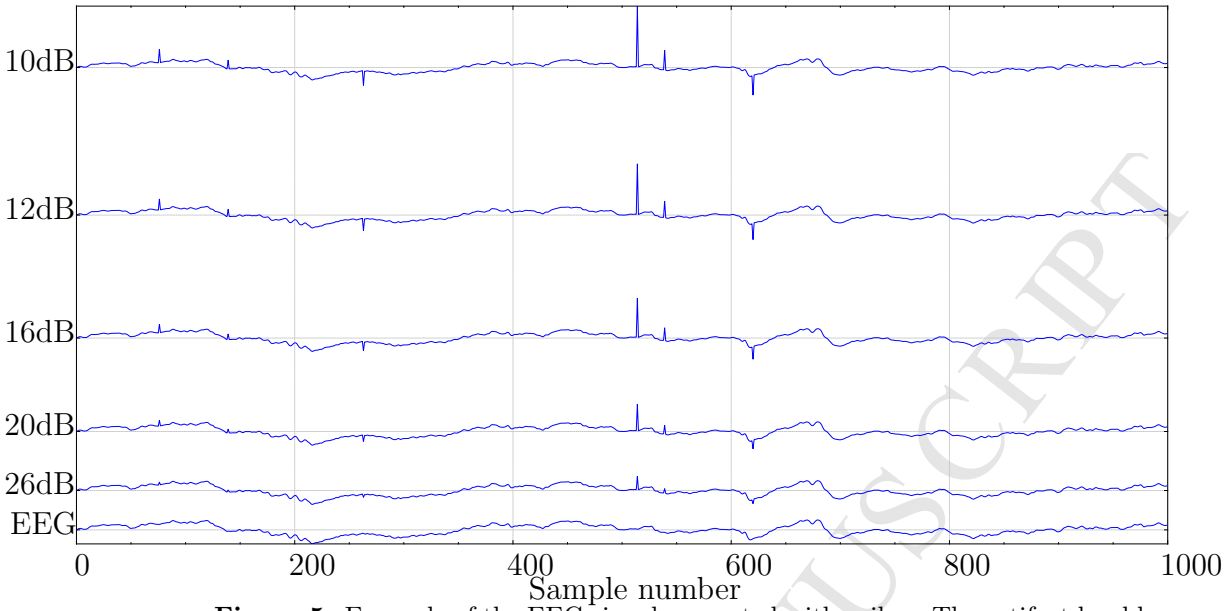

Figure 5. Example of the EEG signal corrupted with spikes. The artifact level lowers from top to bottom. The length of signals is 1000 samples, approximately 2s.

The results for muscular artifacts are shown in Table 5, and a visual example of the resulting corrupted EEG record is depicted in Figure 6. In this case, the metrics performance is quite poor, with ApEn failing at 16dB SNR, but with complete statistical inability to discern between means at 10dB. SampEn performance is only slightly better, with the first pair of means considered equal at 12dB, and two at 10dB. FuzzyEn is once again the most robust metric as it provides a full H_0 rejection in all cases.

Table 5. The results for the Bern–Barcelona database using EMGs as artifacts. The results for non separable classes (p_{01} and p_{23}) are not included. An accepted hypothesis is featured by $p-$values in bold.

	SNR$(26dB)$	**SNR**$(20dB)$	**SNR**$(16dB)$	**SNR**$(12dB)$	**SNR**$(10dB)$
ApEn	$p_{02} = 0.000129$	$p_{02} = 0.000559$	$p_{02} = 0.002662$	$p_{02} = 0.007750$	$\mathbf{p_{02} = 0.011586}$
	$p_{03} = 0.000061$	$p_{03} = 0.000168$	$p_{03} = 0.001351$	$p_{03} = 0.005115$	$\mathbf{p_{03} = 0.011076}$
	$p_{12} = 0.001198$	$p_{12} = 0.004739$	$\mathbf{p_{12} = 0.014735}$	$\mathbf{p_{12} = 0.034762}$	$\mathbf{p_{12} = 0.052257}$
	$p_{13} = 0.000309$	$p_{13} = 0.001603$	$p_{13} = 0.008845$	$\mathbf{p_{13} = 0.025999}$	$\mathbf{p_{13} = 0.053218}$
SampEn	$p_{02} = 0.000074$	$p_{02} = 0.000161$	$p_{02} = 0.000617$	$p_{02} = 0.001913$	$p_{02} = 0.004284$
	$p_{03} = 0.000053$	$p_{03} = 0.000064$	$p_{03} = 0.000214$	$p_{03} = 0.001021$	$p_{03} = 0.003749$
	$p_{12} = 0.000464$	$p_{12} = 0.001447$	$p_{12} = 0.004338$	$\mathbf{p_{12} = 0.010343}$	$\mathbf{p_{12} = 0.017941}$
	$p_{13} = 0.000138$	$p_{13} = 0.000365$	$p_{13} = 0.001780$	$p_{13} = 0.006668$	$\mathbf{p_{13} = 0.017464}$
FuzzyEn	$p_{02} = 0.000081$	$p_{02} = 0.000162$	$p_{02} = 0.000450$	$p_{02} = 0.001184$	$p_{02} = 0.002615$
	$p_{03} = 0.000054$	$p_{03} = 0.000064$	$p_{03} = 0.000129$	$p_{03} = 0.000394$	$p_{03} = 0.001160$
	$p_{12} = 0.000365$	$p_{12} = 0.000930$	$p_{12} = 0.002286$	$p_{12} = 0.004739$	$p_{12} = 0.008379$
	$p_{13} = 0.000102$	$p_{13} = 0.000216$	$p_{13} = 0.000641$	$p_{13} = 0.001785$	$p_{13} = 0.004185$

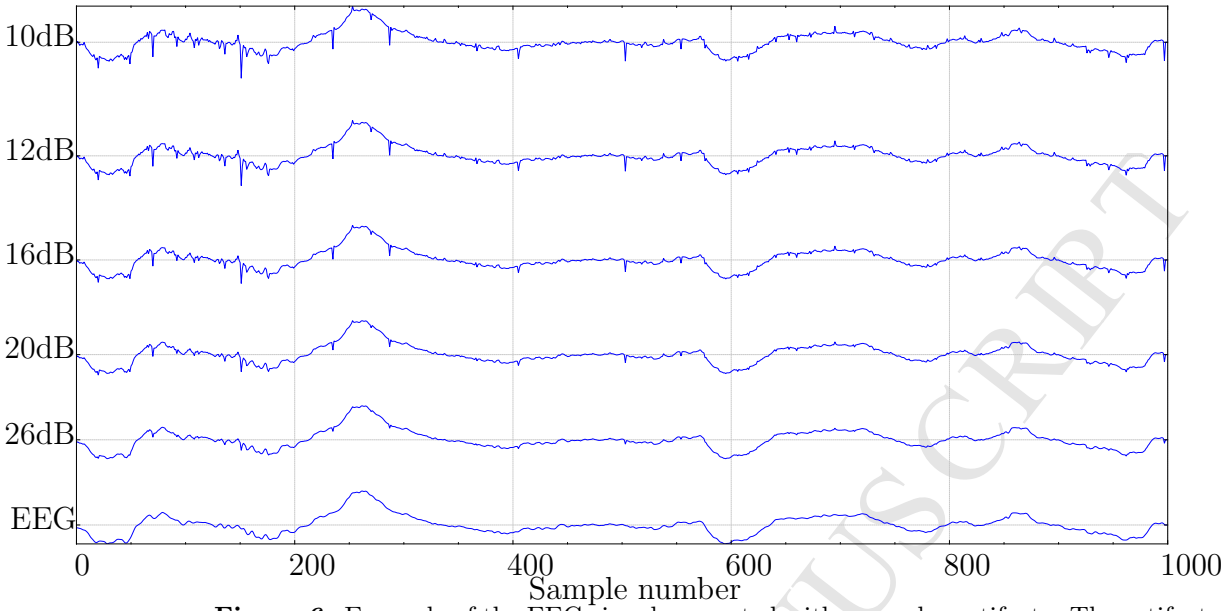

Figure 6. Example of the EEG signal corrupted with muscular artifacts. The artifact level lowers from top to bottom. The length of signals is 1000 samples, approximately 2s.

The results for the ocular artifacts are shown in Table 6, and a visual example of the resulting corrupted EEG record is depicted in Figure 7. No class combination exist of the separable classes where the hypothesis is accepted, but the quantitative results are included for comparative purposes as these artifacts have no influence on the Bonn database case.

Table 6. The results for the Bern–Barcelona database using EOGs as artifacts. The results for non separable classes (p_{01} and p_{23}) are not included.

	SNR$(15dB)$	**SNR**$(9dB)$	**SNR**$(6dB)$	**SNR**$(4dB)$	**SNR**$(2dB)$
ApEn	$p_{02} = 0.000076$	$p_{02} = 0.000081$	$p_{02} = 0.000096$	$p_{02} = 0.000141$	$p_{02} = 0.000232$
	$p_{03} = 0.000059$	$p_{03} = 0.000063$	$p_{03} = 0.000068$	$p_{03} = 0.000081$	$p_{03} = 0.000106$
	$p_{12} = 0.000487$	$p_{12} = 0.000430$	$p_{12} = 0.000425$	$p_{12} = 0.000512$	$p_{12} = 0.000606$
	$p_{13} = 0.000210$	$p_{13} = 0.000209$	$p_{13} = 0.000200$	$p_{13} = 0.000203$	$p_{13} = 0.000214$
SampEn	$p_{02} = 0.000061$	$p_{02} = 0.000063$	$p_{02} = 0.000070$	$p_{02} = 0.000076$	$p_{02} = 0.000102$
	$p_{03} = 0.000055$	$p_{03} = 0.000056$	$p_{03} = 0.000058$	$p_{03} = 0.000062$	$p_{03} = 0.000072$
	$p_{12} = 0.000207$	$p_{12} = 0.000200$	$p_{12} = 0.000227$	$p_{12} = 0.000233$	$p_{12} = 0.000237$
	$p_{13} = 0.000116$	$p_{13} = 0.000111$	$p_{13} = 0.000107$	$p_{13} = 0.000113$	$p_{13} = 0.000117$
FuzzyEn	$p_{02} = 0.000061$	$p_{02} = 0.000063$	$p_{02} = 0.000068$	$p_{02} = 0.000078$	$p_{02} = 0.000101$
	$p_{03} = 0.000054$	$p_{03} = 0.000055$	$p_{03} = 0.000057$	$p_{03} = 0.000062$	$p_{03} = 0.000073$
	$p_{12} = 0.000201$	$p_{12} = 0.000185$	$p_{12} = 0.000179$	$p_{12} = 0.000183$	$p_{12} = 0.000198$
	$p_{13} = 0.000098$	$p_{13} = 0.000094$	$p_{13} = 0.000093$	$p_{13} = 0.000098$	$p_{13} = 0.000109$

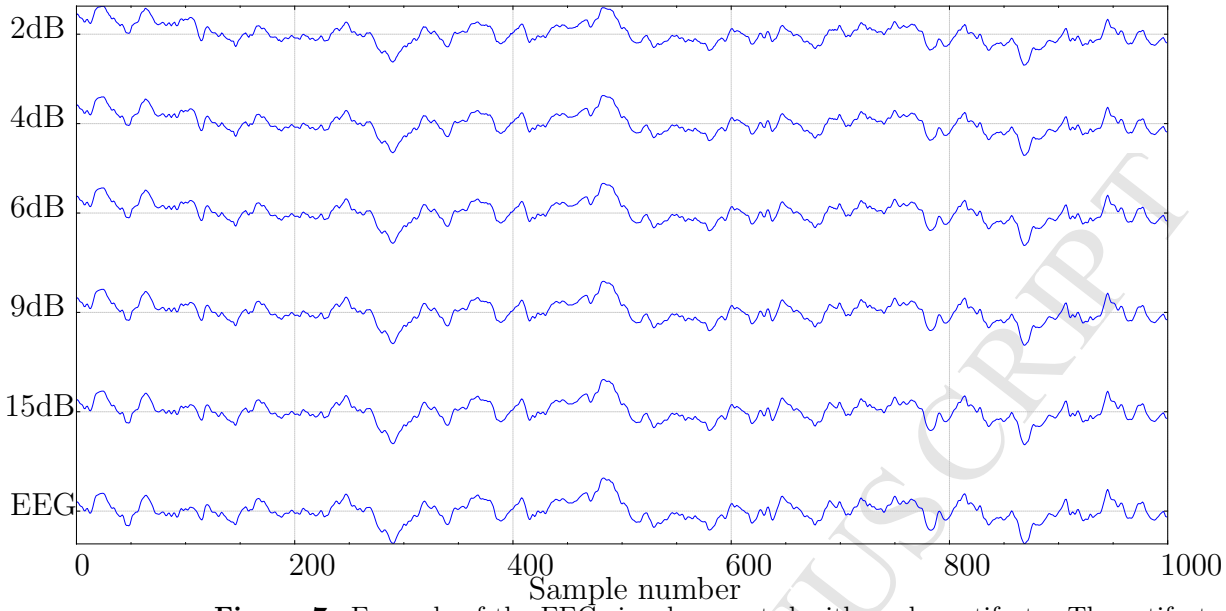

Figure 7. Example of the EEG signal corrupted with ocular artifacts. The artifact level lowers from top to bottom. The length of signals is 1000 samples, approximately 2s.

For the spikes case, cardiac artifacts do not significantly influence the separability of the means. Therefore, the numerical results are not included ($p_{ij} < \alpha, \forall i, j$ considered). Figure 8 shows an example of an EEG record corrupted with an underlying ECG signal.

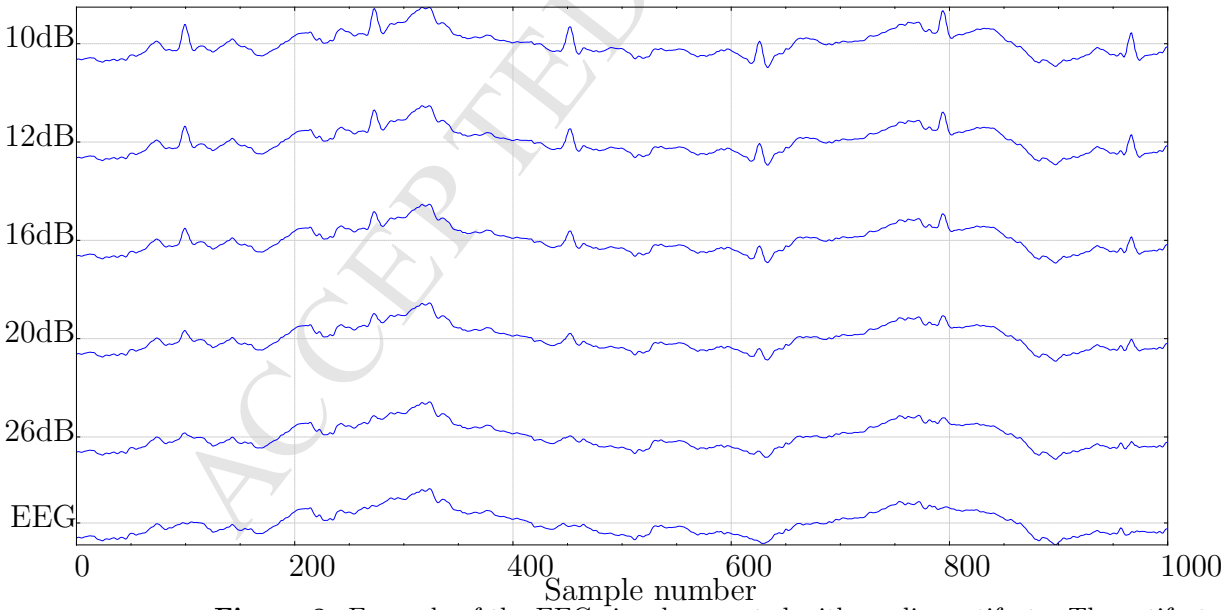

Figure 8. Example of the EEG signal corrupted with cardiac artifacts. The artifact level lowers from top to bottom. The length of signals is 1000 samples, approximately 2s.

3.2. Bonn database

This section describes the results achieved using the Bonn database. The baseline results (input signals without artifacts) are included only for Gaussian noise (Table 7) as they

are the same for the other artifact types. The $p-$values for the different levels of artifacts are shown in Tables 7-10. Since the visual examples of each artifact are practically the same for both databases, they are included only in the previous subsection.

As for the Bern–Barcelona database, the EEG records in the Bonn database also seem quite sensitive to presence of noise. With ApEn, means are considered equal very early, at 26dB, and performance worsens significantly at 10dB. SampEn holds until 12dB, and also fails at 10dB, and FuzzyEn enables all the means to be considered statistically different.

Table 7. The results for the Bonn database using different levels of Gaussian noise. Class 0 corresponds to set A, class 1 to B, 2 to C, 3 to D, and 4 to E. Case 01 (AB) is not included because it is impossible to separate the two classes. An accepted hypothesis is featured by $p-$values in bold.

	No artifact	**SNR**$(26dB)$	**SNR**$(20dB)$	**SNR**$(16dB)$	**SNR**$(12dB)$	**SNR**$(10dB)$
ApEn	$p_{02} = 0.000039$	$p_{02} = 0.000047$	$p_{02} = 0.000032$	$p_{02} = 0.000027$	$p_{02} = 0.000026$	$p_{02} = 0.000028$
	$p_{03} = 0.000031$	$p_{03} = 0.000027$	$p_{03} = 0.000025$	$p_{03} = 0.000032$	$p_{03} = 0.000025$	$\mathbf{p_{03} = 0.261416}$
	$p_{04} = 0.000039$	$p_{04} = 0.000035$	$p_{04} = 0.002387$	$p_{04} = 0.000036$	$p_{04} = 0.000027$	$p_{04} = 0.000025$
	$p_{12} = 0.000036$	$\mathbf{p_{12} = 0.163427}$	$p_{12} = 0.000025$	$p_{12} = 0.000027$	$p_{12} = 0.000025$	$p_{12} = 0.000026$
	$p_{13} = 0.000026$	$p_{13} = 0.000025$	$p_{13} = 0.000033$	$p_{13} = 0.000032$	$p_{13} = 0.000025$	$\mathbf{p_{13} = 0.013195}$
	$p_{14} = 0.000033$	$p_{14} = 0.000027$	$p_{14} = 0.000034$	$p_{14} = 0.000036$	$p_{14} = 0.000028$	$p_{14} = 0.000026$
	$p_{23} = 0.000030$	$p_{23} = 0.000035$	$p_{23} = 0.000034$	$p_{23} = 0.000027$	$p_{23} = 0.000026$	$p_{23} = 0.000028$
	$p_{24} = 0.000025$	$p_{24} = 0.000028$	$p_{24} = 0.000196$	$p_{24} = 0.001881$	$p_{24} = 0.007121$	$\mathbf{p_{24} = 0.012558}$
	$p_{34} = 0.000030$	$p_{34} = 0.000029$	$p_{34} = 0.000025$	$\mathbf{p_{34} = 0.018446}$	$p_{34} = 0.000027$	$p_{34} = 0.000025$
SampEn	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000026$	$p_{02} = 0.000025$	$p_{02} = 0.000025$
	$p_{03} = 0.000030$	$p_{03} = 0.000026$	$p_{03} = 0.000025$	$p_{03} = 0.000025$	$p_{03} = 0.000026$	$p_{03} = 0.000028$
	$p_{04} = 0.000026$	$p_{04} = 0.000026$	$p_{04} = 0.000025$	$p_{04} = 0.000025$	$p_{04} = 0.000026$	$p_{04} = 0.000025$
	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$
	$p_{13} = 0.000028$	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000027$	$p_{13} = 0.000027$	$p_{13} = 0.000029$
	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000026$
	$p_{23} = 0.000030$	$p_{23} = 0.000027$	$p_{23} = 0.000025$	$p_{23} = 0.000027$	$p_{23} = 0.000025$	$p_{23} = 0.000029$
	$p_{24} = 0.000026$	$p_{24} = 0.000031$	$p_{24} = 0.000046$	$p_{24} = 0.000613$	$\mathbf{p_{24} = 0.023369}$	$\mathbf{p_{24} = 0.037315}$
	$p_{34} = 0.000028$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000026$	$p_{34} = 0.000028$	$p_{34} = 0.000027$
FuzzyEn	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$
	$p_{03} = 0.000027$	$p_{03} = 0.000027$	$p_{03} = 0.000026$	$p_{03} = 0.000026$	$p_{03} = 0.000026$	$p_{03} = 0.000026$
	$p_{04} = 0.000027$	$p_{04} = 0.000027$	$p_{04} = 0.000027$	$p_{04} = 0.000027$	$p_{04} = 0.000026$	$p_{04} = 0.000026$
	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$
	$p_{13} = 0.000027$	$p_{13} = 0.000027$	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000026$
	$p_{14} = 0.000027$	$p_{14} = 0.000027$	$p_{14} = 0.000027$	$p_{14} = 0.000027$	$p_{14} = 0.000026$	$p_{14} = 0.000026$
	$p_{23} = 0.000032$	$p_{23} = 0.000031$	$p_{23} = 0.000031$	$p_{23} = 0.000030$	$p_{23} = 0.000030$	$p_{23} = 0.000030$
	$p_{24} = 0.000039$	$p_{24} = 0.000056$	$p_{24} = 0.000206$	$p_{24} = 0.000491$	$p_{24} = 0.000933$	$p_{24} = 0.002410$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$

Figure 9 offers a ROC curve for the case linked to p_{13} when SNR$= 10dB$, with $p_{13} = 0.013195$ for ApEn, $p_{13} = 0.000029$ for SampEn, and $p_{13} = 0.000026$ for FuzzyEn. The ROC curve shows how FuzzyEn would achieve the highest correct classification ratio, followed by SampEn and then by ApEn. This is numerically supported by the Area Under Curve (AUC) value, which is 0.7796, 0.7308, and 0.5773, respectively. This scheme could be replicated in any other case where separability between two classes is required in quantitative terms, and a classifier should be implemented based on the thresholds obtained in the corresponding ROC curve using the entropy results as input features. For instance, using the threshold obtained from the optimal point in the ROC curve (minimum distance to point $(0, 1)$), the classification results in this case are 75.56 % TP (True Positives) and 73.64 % TN (True Negatives) for FuzzyEn, 90.91 % TP and

27.84 TN % for SampEn, and 57.61 % TP and 56.48 % TN for SampEn. Figure 10 depicts the same $p13$ case when no noise is present. In this case the AUC is 0.9039 for FuzzyEn, 0.8959 for SampEn, and 0.885 for ApEn.

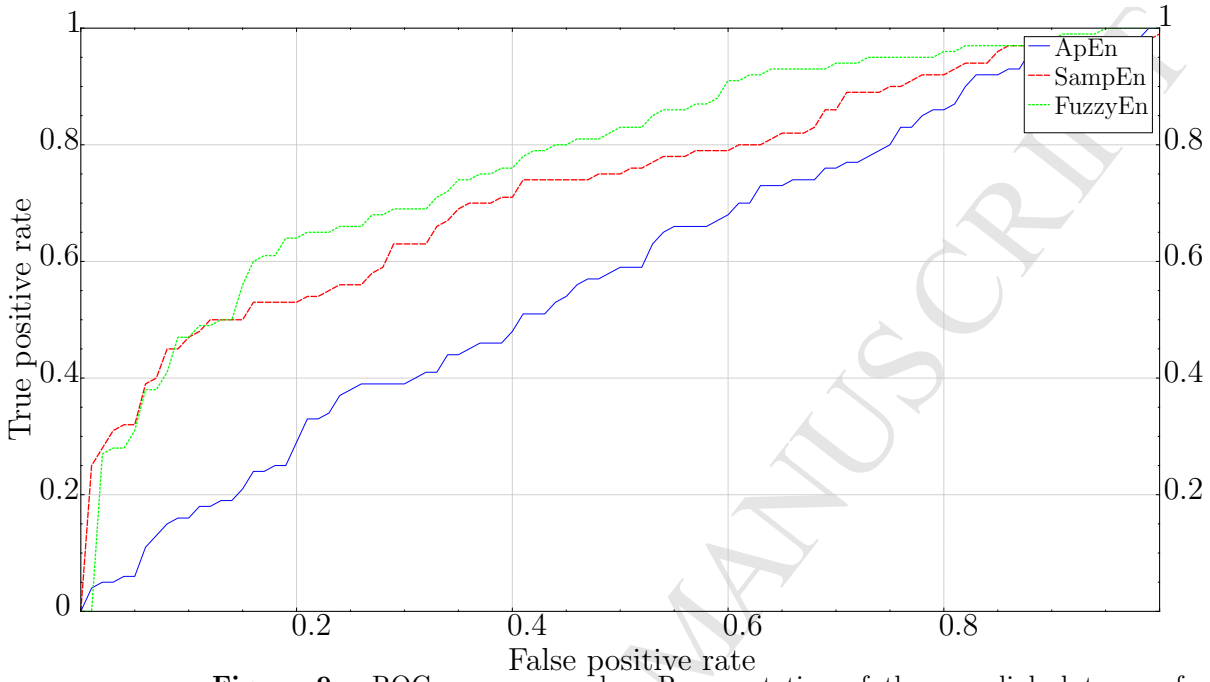

Figure 9. ROC curve example. Representation of the case linked to p_{13} for SNR= $10dB$. The curves for ApEn, SampEn and FuzzyEn are included. Higher detection accuracy corresponds to FuzzyEn.

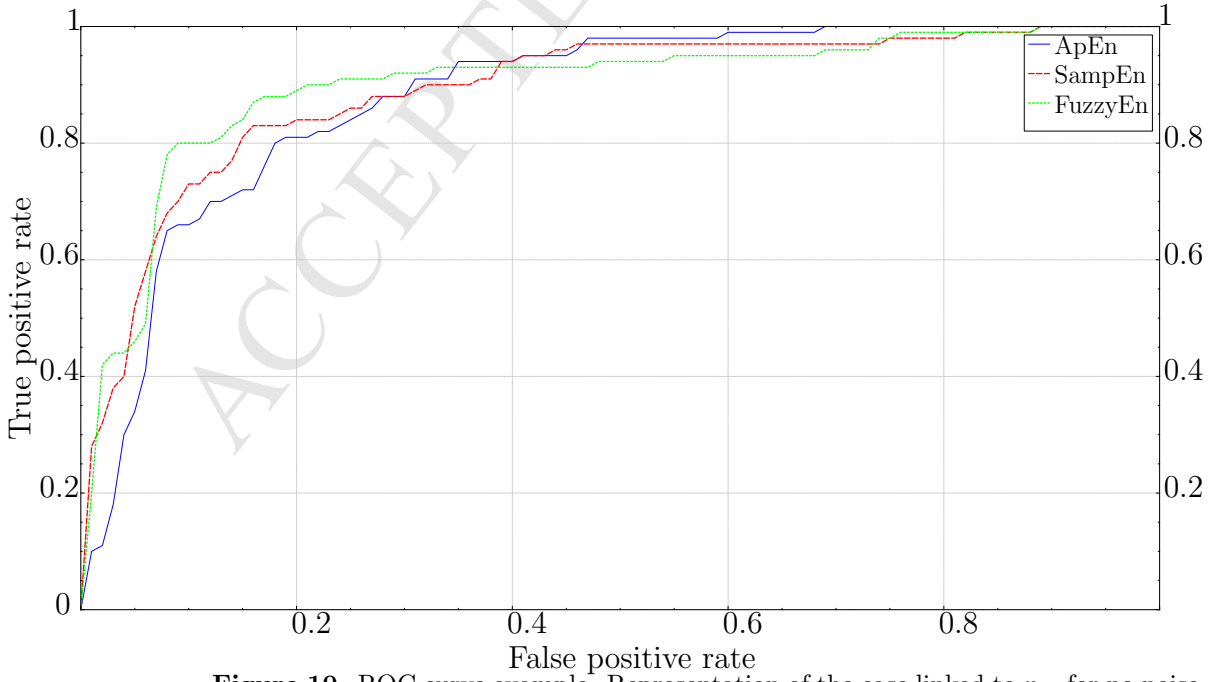

Figure 10. ROC curve example. Representation of the case linked to p_{13} for no noise. The curves for ApEn, SampEn and FuzzyEn are included. Higher detection accuracy corresponds to FuzzyEn.

Table 8 shows the results obtained for $N = 1000, 2000, 3000$ and 4000 using Gaussian noise at SNR($10dB$). The objective of this table is to ensure that the signal classification performance of the three entropy metrics under study is similar, regardless of the length of records.

Table 8. The results for the Bonn database using Gaussian noise (SNR($10dB$)) and different lengths (N). Class 0 corresponds to set A, class 1 to B, 2 to C, 3 to D, and 4 to E. Case 01 (AB) is not included as it is impossible to separate the two classes. An accepted hypothesis is featured by $p-$values in bold.

	$N = 1000$	$N = 2000$	$N = 3000$	$N = 4000$
ApEn	$p_{02} = 0.000028$	$p_{02} = 0.000026$	$p_{02} = 0.000025$	$p_{02} = 0.000025$
	$\mathbf{p_{03} = 0.261416}$	$p_{03} = 0.000418$	$p_{03} = 0.000027$	$p_{03} = 0.000029$
	$p_{04} = 0.000025$	$p_{04} = 0.000026$	$p_{04} = 0.000027$	$p_{04} = 0.000027$
	$p_{12} = 0.000026$	$p_{12} = 0.000025$	$p_{12} = 0.000026$	$p_{12} = 0.000028$
	$\mathbf{p_{13} = 0.013195}$	$p_{13} = 0.000030$	$p_{13} = 0.000031$	$p_{13} = 0.000037$
	$p_{14} = 0.000026$	$p_{14} = 0.000028$	$p_{14} = 0.000032$	$p_{14} = 0.000036$
	$p_{23} = 0.000028$	$p_{23} = 0.000027$	$p_{23} = 0.000028$	$\mathbf{p_{23} = 0.061182}$
	$\mathbf{p_{24} = 0.012558}$	$p_{24} = 0.001144$	$p_{24} = 0.003696$	$p_{24} = 0.000212$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000501$	$\mathbf{p_{34} = 0.147776}$
SampEn	$p_{02} = 0.000025$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000028$
	$p_{03} = 0.000028$	$p_{03} = 0.000026$	$p_{03} = 0.000028$	$p_{03} = 0.000027$
	$p_{04} = 0.000025$	$p_{04} = 0.000026$	$p_{04} = 0.000027$	$p_{04} = 0.000027$
	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000026$	$p_{12} = 0.000026$
	$p_{13} = 0.000029$	$p_{13} = 0.000029$	$p_{13} = 0.000029$	$p_{13} = 0.000029$
	$p_{14} = 0.000026$	$p_{14} = 0.000025$	$p_{14} = 0.000026$	$p_{14} = 0.000025$
	$p_{23} = 0.000029$	$p_{23} = 0.000031$	$p_{23} = 0.000033$	$p_{23} = 0.000032$
	$\mathbf{p_{24} = 0.037315}$	$\mathbf{p_{24} = 0.165770}$	$\mathbf{p_{24} = 0.475905}$	$\mathbf{p_{24} = 0.506105}$
	$p_{34} = 0.000027$	$p_{34} = 0.000030$	$p_{34} = 0.000032$	$p_{34} = 0.000032$
FuzzyEn	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000028$
	$p_{03} = 0.000026$	$p_{03} = 0.000064$	$p_{03} = 0.000092$	$p_{03} = 0.001015$
	$p_{04} = 0.000026$	$p_{04} = 0.000025$	$p_{04} = 0.000025$	$p_{04} = 0.000025$
	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$
	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000027$	$p_{13} = 0.000169$
	$p_{14} = 0.000026$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000026$
	$p_{23} = 0.000030$	$p_{23} = 0.000030$	$p_{23} = 0.000030$	$p_{23} = 0.000030$
	$p_{24} = 0.002410$	$p_{24} = 0.000159$	$p_{24} = 0.000199$	$p_{24} = 0.000145$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$

The spikes case is also consistent for the two datasets. Even at 10dB, no hypothesis is accepted for any combination of classes. It is arguably possible that for a lower SNR, the equal means hypothesis will eventually be accepted. However, such a low SNR does not fall in line with what happens in a real case. Classes A and B are not separable in any case. They are too similar in entropy terms even in their original form, and without artifacts. Although class D also comes close to classes A and B when ApEn is used, the equal means hypothesis is not analytically accepted.

The results for muscular artifacts are shown in Table 9. In this case, the performance of the metrics is not as bad as for the Bern–Barcelona database, except for ApEn, which still fails in many comparisons, even at 26dB. SampEn and FuzzyEn perform much better, and no hypothesis is accepted in terms of equal means between classes.

Table 9. The results for the Bonn database using EMGs as artifacts. Class 0 corresponds to set A, class 1 to B, 2 to C, 3 to D, and 4 to E. Case 01 (AB) is not included as it is impossible to separate the two classes. An accepted hypothesis is featured by $p-$values in bold.

	SNR(26dB)	**SNR**(20dB)	**SNR**(16dB)	**SNR**(12dB)	**SNR**(10dB)
ApEn	$p_{02} = 0.000040$	$\mathbf{p_{02} = 0.023474}$	$\mathbf{p_{02} = 0.819953}$	$\mathbf{p_{02} = 0.024763}$	$p_{02} = 0.000553$
	$p_{03} = 0.000028$	$p_{03} = 0.000026$	$p_{03} = 0.000025$	$p_{03} = 0.000025$	$p_{03} = 0.000025$
	$p_{04} = 0.000037$	$p_{04} = 0.000032$	$p_{04} = 0.000625$	$\mathbf{p_{04} = 0.049641}$	$\mathbf{p_{04} = 0.374565}$
	$\mathbf{p_{12} = 0.051204}$	$\mathbf{p_{12} = 0.163194}$	$p_{12} = 0.000030$	$p_{12} = 0.000025$	$p_{12} = 0.000025$
	$p_{13} = 0.000025$	$p_{13} = 0.000026$	$p_{13} = 0.000028$	$p_{13} = 0.000028$	$p_{13} = 0.000028$
	$p_{14} = 0.000031$	$p_{14} = 0.002240$	$\mathbf{p_{14} = 0.532845}$	$\mathbf{p_{14} = 0.327657}$	$\mathbf{p_{14} = 0.093265}$
	$p_{23} = 0.000033$	$p_{23} = 0.000036$	$p_{23} = 0.000034$	$p_{23} = 0.000029$	$p_{23} = 0.000027$
	$p_{24} = 0.000026$	$p_{24} = 0.000029$	$p_{24} = 0.000029$	$p_{24} = 0.000032$	$p_{24} = 0.000062$
	$p_{34} = 0.000030$	$p_{34} = 0.000028$	$p_{34} = 0.000027$	$p_{34} = 0.000025$	$p_{34} = 0.000025$
SampEn	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$
	$p_{03} = 0.000027$	$p_{03} = 0.000026$	$p_{03} = 0.000025$	$p_{03} = 0.000025$	$p_{03} = 0.000027$
	$p_{04} = 0.000026$	$p_{04} = 0.000026$	$p_{04} = 0.000025$	$p_{04} = 0.000026$	$p_{04} = 0.000026$
	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$
	$p_{13} = 0.000026$	$p_{13} = 0.000025$	$p_{13} = 0.000025$	$p_{13} = 0.000025$	$p_{13} = 0.000025$
	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$
	$p_{23} = 0.000028$	$p_{23} = 0.000026$	$p_{23} = 0.000025$	$p_{23} = 0.000025$	$p_{23} = 0.000025$
	$p_{24} = 0.000028$	$p_{24} = 0.000032$	$p_{24} = 0.000054$	$p_{24} = 0.000104$	$p_{24} = 0.000197$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000026$	$p_{34} = 0.000026$
FuzzyEn	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$	$p_{02} = 0.000027$
	$p_{03} = 0.000027$	$p_{03} = 0.000027$	$p_{03} = 0.000026$	$p_{03} = 0.000026$	$p_{03} = 0.000026$
	$p_{04} = 0.000027$	$p_{04} = 0.000027$	$p_{04} = 0.000026$	$p_{04} = 0.000026$	$p_{04} = 0.000025$
	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000027$	$p_{12} = 0.000026$
	$p_{13} = 0.000027$	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000026$	$p_{13} = 0.000026$
	$p_{14} = 0.000027$	$p_{14} = 0.000027$	$p_{14} = 0.000026$	$p_{14} = 0.000026$	$p_{14} = 0.000026$
	$p_{23} = 0.000031$	$p_{23} = 0.000031$	$p_{23} = 0.000030$	$p_{23} = 0.000029$	$p_{23} = 0.000029$
	$p_{24} = 0.000054$	$p_{24} = 0.000120$	$p_{24} = 0.000280$	$p_{24} = 0.000635$	$p_{24} = 0.001158$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$

The results for ocular artifacts are shown in Table 10. Unlike what happens with the ocular artifacts in the Bern–Barcelona database, in this case two of the metrics fail at some point. ApEn and FuzzyEn fail at 10dB, whereas SampEn is the most robust metric in this case, with no test in which H_0 is accepted. Performance degradation is not generally as severe as for muscular artifacts, but is still measurable.

Cardiac artifacts do not significantly influence the separability of the means and, therefore, the numerical results are not included ($p_{ij} < \alpha, \forall i, j$ considered), as for the previous database.

Table 10. The results for the Bonn database using EOGs as artifacts. Class 0 corresponds to set A, class 1 to B, 2 to C, 3 to D, and 4 to E. Case 01 (AB) is not included as it is impossible to separate the two classes. An accepted hypothesis is featured by $p-$values in bold.

	SNR$(26dB)$	**SNR**$(20dB)$	**SNR**$(16dB)$	**SNR**$(12dB)$	**SNR**$(10dB)$
ApEn	$p_{02} = 0.000040$	$p_{02} = 0.000039$	$p_{02} = 0.000039$	$p_{02} = 0.000037$	$p_{02} = 0.000035$
	$p_{03} = 0.000032$	$p_{03} = 0.000035$	$p_{03} = 0.000219$	$p_{03} = 0.005070$	$\mathbf{p_{03} = 0.099963}$
	$p_{04} = 0.000041$	$p_{04} = 0.000041$	$p_{04} = 0.000041$	$p_{04} = 0.000040$	$p_{04} = 0.000038$
	$p_{12} = 0.000036$	$p_{12} = 0.000034$	$p_{12} = 0.000034$	$p_{12} = 0.000034$	$p_{12} = 0.000032$
	$p_{13} = 0.000027$	$p_{13} = 0.000028$	$p_{13} = 0.000028$	$p_{13} = 0.000029$	$p_{13} = 0.000030$
	$p_{14} = 0.000036$	$p_{14} = 0.000037$	$p_{14} = 0.000037$	$p_{14} = 0.000037$	$p_{14} = 0.000036$
	$p_{23} = 0.000029$	$p_{23} = 0.000028$	$p_{23} = 0.000027$	$p_{23} = 0.000027$	$p_{23} = 0.000026$
	$p_{24} = 0.000025$	$p_{24} = 0.000026$	$p_{24} = 0.000026$	$p_{24} = 0.000026$	$p_{24} = 0.000026$
	$p_{34} = 0.000030$	$p_{34} = 0.000030$	$p_{34} = 0.000030$	$p_{34} = 0.000030$	$p_{34} = 0.000028$
SampEn	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$	$p_{02} = 0.000025$
	$p_{03} = 0.000030$	$p_{03} = 0.000030$	$p_{03} = 0.000033$	$p_{03} = 0.000145$	$p_{03} = 0.004188$
	$p_{04} = 0.000026$	$p_{04} = 0.000025$	$p_{04} = 0.000025$	$p_{04} = 0.000025$	$p_{04} = 0.000025$
	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$	$p_{12} = 0.000025$
	$p_{13} = 0.000028$	$p_{13} = 0.000028$	$p_{13} = 0.000028$	$p_{13} = 0.000028$	$p_{13} = 0.000028$
	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$	$p_{14} = 0.000025$
	$p_{23} = 0.000030$	$p_{23} = 0.000030$	$p_{23} = 0.000029$	$p_{23} = 0.000028$	$p_{23} = 0.000029$
	$p_{24} = 0.000026$	$p_{24} = 0.000032$	$p_{24} = 0.000036$	$p_{24} = 0.000045$	$p_{24} = 0.000269$
	$p_{34} = 0.000028$	$p_{34} = 0.000028$	$p_{34} = 0.000028$	$p_{34} = 0.000028$	$p_{34} = 0.000028$
FuzzyEn	$p_{02} = 0.000027$	$p_{02} = 0.000026$	$p_{02} = 0.000025$	$p_{02} = 0.000026$	$p_{02} = 0.000028$
	$p_{03} = 0.000027$	$p_{03} = 0.000027$	$p_{03} = 0.000028$	$p_{03} = 0.000028$	$p_{03} = 0.000028$
	$p_{04} = 0.000027$	$p_{04} = 0.000027$	$p_{04} = 0.000028$	$p_{04} = 0.000030$	$p_{04} = 0.000032$
	$p_{12} = 0.000027$	$p_{12} = 0.000026$	$p_{12} = 0.000025$	$p_{12} = 0.000026$	$p_{12} = 0.000028$
	$p_{13} = 0.000027$	$p_{13} = 0.000027$	$p_{13} = 0.000027$	$p_{13} = 0.000027$	$p_{13} = 0.000028$
	$p_{14} = 0.000027$	$p_{14} = 0.000027$	$p_{14} = 0.000028$	$p_{14} = 0.000029$	$p_{14} = 0.000031$
	$p_{23} = 0.000031$	$p_{23} = 0.000030$	$p_{23} = 0.000027$	$p_{23} = 0.000026$	$p_{23} = 0.000025$
	$p_{24} = 0.000043$	$p_{24} = 0.000060$	$p_{24} = 0.000252$	$p_{24} = 0.002246$	$\mathbf{p_{24} = 0.014164}$
	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000025$	$p_{34} = 0.000030$	$p_{34} = 0.000715$

4. Discussion

The goal of this study was to find out the best entropy metrics and parameter configuration for noisy EEG records employed in signal classification applications. The results for the two databases exhibit the same trend, with noise and muscular artifacts yielding the lowest rejection levels (the same mean accepted, the same class assumed), whereas spikes and cardiac artifacts appear to not influence the separability of classes. Performance was assessed in terms of equal means hypothesis acceptance or rejection.

First, the parameter initialization analysis confirmed what has been found in many scientific works [35]: the m, r, N and q values may significantly influence the results obtained using these entropy metrics. The influence of N was minimized using a value, 1000, that meets the well–known requirement of $N \geq 10^m$ [10] and other similar recommendations [16]. Obviously, other N values would certainly change the quantitative entropy results, as specifically shown in Tables 4 and 8. However, the qualitative results remain the same; i.e., FuzzyEn performs best, whereas ApEn is the metrics with more cases of equal means acceptance. $N = 1000$ keeps the computational burden relatively low, and uniformizes the length of the two datasets. It is stressed that in real clinical settings, it is not always possible to acquire very long time series, and the search for entropy metrics that performs well for very short biosignals is an ongoing

research line [58].

The rest of the parameters were studied using different combinations, which exceeded the usually recommended ranges. In this case, they were heuristically chosen to maximize the probability of class separability (equal means hypothesis rejection) when no artifact was present. The m parameter varied from 1 to 3, r from 0.15 to 0.30, and q from 1 to 4. The test with no artifacts was repeated for each case, and the $p-$value was computed. The results of this analysis, some of which are shown in Table 2, reflect the fact that the two input datasets are very likely to be separable when no artifact is present (mainly the Bern database) and the specific values were picked from these cases. Although differences are minor (e.g., $p_{03} = 0.000064$ for $m = 1, r = 0.15$, and $p_{03} = 0.000056$ for $m = 2, r = 0.25$, for the first database), there is a clear trend where p_{ij} decreases as m increases in almost each case.

There is wide variability between datasets in terms of optimal input parameters. For the Bern database, ApEn($m = 2$, $r = 0.3$), SampEn($m = 2$, $r = 0.3$), FuzzyEn($m = 3, r = 0.15, q = 1$) vs. ApEn($m = 3, r = 0.15$), SampEn($m = 3, r = 0.15$), and FuzzyEn($m = 3, r = 0.3, q = 4$) for the Bonn database. In fact, if these parameter sets were swapped between the two databases, there would be baseline cases where the equal means hypothesis would be accepted, or even the influence of spikes and/or cardiac artifacts would become significant. Although FuzzyEn seems more stable, mainly as regards m, it is not as stable as claimed in other contexts [22]. Consequently, special care must be taken to appropriately select these parameters, and even dependency on disturbance type can be arguably assumed. In other words, the EEG classification using any of these metrics requires prior class knowledge, supervision and customization to ensure optimal results.

For the two employed datasets, the results show that noise and muscular artifacts have the strongest influence on the class separability of the input data (Tables 3, 7, 5 and 9), with rejections found even at 26dB, specifically for ApEn. Isolated spikes and cardiac artifacts do not seem to significantly degrade the segmentation capabilities of the studied metrics, with no acceptance found for all the studied cases. Ocular artifacts fall in–between these two extreme cases, with a minimal, but measurable, influence (Tables 10 and 6), that starts later at 10dB. It is also important to note that acceptance does not only depend on the SNR level since some artifacts, especially the EMG artifacts, are clearly non stationary. Since EMG and EOG epochs are randomly chosen for the experiments, their influence may vary depending on their spike distribution (as illustrated by the changes in the EMG signal at time 12s in Figure 3). Lack of consistency for ApEn also becomes apparent in some cases, where means are considered equal at some SNR levels, but are considered different at a lower SNR level (Table 7).

The changes in the $p-$values with artifacts are due to changes in the pattern that matches the ratios of the metrics; e.g., regardless of their amplitude, isolated spikes, only represent an extremely minor variation in the number of subsequences that match/do not match. Consequently, the ratio is almost the same, as is the entropy metrics, and the equal means hypothesis is rejected. The situation is similar with cardiac artifacts.

Only QRS complexes have a significant amplitude, but are spaced time series as spikes are, and are even more regular, so their influence on the dissimilarity computation is minimal. Conversely, more evenly time distributed artifacts, such as noise or EMGs, introduce variations into almost every signal sample, with a more significant variation of the pattern matches count. Thus the ratio is very likely to be altered, as is the entropy estimation. As a result, the distribution of entropy values notably varies, with completely different $p-$values and a masking of the groups' boundaries.

Although all the metrics provide full separability for the baseline case (no artifact) in terms of statistically significant difference in means, ApEn is very sensitive to presence of outliers, even for a high SNR like 26 dB. Its performance degrades rapidly with a drop in SNR (Tables 5 and 9). FuzzyEn appears to be the most robust metrics, but in one case (Table 10), SampEn outperforms FuzzyEn. This situation may suggest that a more crispy dissimilarity function would be preferable for these cases, in contrast to what is suggested in [54].

5. Conclusions

We studied the performance of ApEn, SampEn, and FuzzyEn metrics in the noisy EEG classification context. It was based on an equal means hypothesis test, and other performance influencing factors, such as parameter configuration, were removed by manual optimisation. The results demonstrate that the ApEn and SampEn metrics are sensitive to the artifacts commonly found in EEG records, mainly white Gaussian noise and muscular artifacts. Even with the barely visible artifacts in the EEG, the signal classification can significantly alter. These and other artifacts can be minimized using the myriad of methods proposed in the literature [27, 34], but this is not always possible, and special care has to be taken when deciding on the final configuration of the metrics to employ.

The selection of input parameters r, m, N, and q is also critical. With low m values, the performance of the three metrics is very poor. The r parameter seems more stable. The size of the data, N, provided it is large enough to ensure a reliable estimation of the number of matches, does not influence the results that much. We recommend using at least $N = 1000$, which is in accordance with the scientific literature and provides reliable results for larger values, e.g., $2000, 3000$ and 4000, but with a much lower computational cost. The q parameter, in conjunction with the membership function, also plays a key role, with variations within the range $[1, 4]$.

FuzzyEn achieves the best results. However, this metrics is not as robust to parameters as usually claimed [22]. In addition to m and r, the fuzzy membership function, and the q parameter, also have to be defined, and they also greatly influence the accuracy of the results [35]. As a general rule, parameter m should be initially set at 3. The main weakness of this metrics is its computational cost. As all the comparisons made between subsequence samples have to be computed, the algorithm burden is $O(N^3)$ instead of $O(N^2)$. This may become a serious problem for large databases or very long

records, and researchers should prioritize the optimization of this algorithm, as with ApEn or SampEn [17], given its superior performance.

In summary, we conclude that broadband artifacts, such as white noise or EMG interference, are the most influential artifacts in EEG records when processed using the entropy measures studied herein. Regardless of their amplitude, other more infrequent artifacts, like spikes, do not significantly modify entropy results, no the classification statistics. Therefore, researchers or medical technology manufacturers will have to better implement artifact removal methods and more robust entropy estimators to protect their studies or systems against misleading results if white noise–like outliers enter EEG acquisition systems. If complete broadband artifact removal can not be ensured, then FuzzyEn seems the most robust metrics for EEG classification if the configuration parameters are properly chosen. However, finding the optimal parameter configuration when no prior knowledge of classes is available can be difficult, and unsupervised parameter optimization methods should be investigated. These parameters could be optimized in each particular case, and similarly to that proposed for SampEn in [59], provided a normalization scheme takes place to make all the results comparable. In any case, we recommend not using ApEn, but to replace it with FuzzyEn or, at least with SampEn, if the computational cost is an issue.

References

[1] C. Babiloni, V. Pizzella, C. D. Gratta, A. Ferretti, G. L. Romani, Chapter 5 fundamentals of electroencefalography, magnetoencefalography, and functional magnetic resonance imaging, Vol. 86 of International Review of Neurobiology, Academic Press, 2009, pp. 67–80.

[2] T.-P. Jung, S. Makeig, M. Stensmo, T. J. Sejnowski, Estimating alertness from the EEG power spectrum, IEEE transactions on bio-medical engineering 44 (1) (1997) 60–9. doi:10.1109/10.553713.

[3] R. Chai, Y. Tran, G. R. Naik, T. N. Nguyen, S. Ling, A. Craig, H. T. Nguyen, Classification of EEG based-mental fatigue using principal component analysis and bayesian neural network, in: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2016, Orlando, FL, USA, August 16-20, 2016, 2016, pp. 4654–4657.

[4] J. L. Rodríguez-Sotelo, A. Osorio-Forero, A. Jiménez-Rodríguez, D. Cuesta-Frau, E. Cirugeda-Roldán, D. Peluffo, Automatic sleep stages classification using eeg entropy features and unsupervised pattern analysis techniques, Entropy 16 (12) (2014) 6573.

[5] J. Wu, R. Srinivasan, E. Burke Quinlan, A. Solodkin, S. L. Small, S. C. Cramer, Utility of eeg measures of brain function in patients with acute stroke, Journal of Neurophysiology 115 (5) (2016) 2399–2405. doi:10.1152/jn.00978.2015.

[6] C. Micanovic, S. Pal, The diagnostic utility of eeg in early-onset dementia: a systematic review of the literature with narrative analysis, Journal of Neural Transmission 121 (1) (2014) 59–69.

[7] S. J. M. Smith, Eeg in the diagnosis, classification, and management of patients with epilepsy, Journal of Neurology, Neurosurgery and Psychiatry 76 (suppl 2) (2005) ii2–ii7.

[8] D. P. Subha, P. K. Joseph, R. Acharya U, C. M. Lim, Eeg signal analysis: A survey, Journal of Medical Systems 34 (2) (2010) 195–212.

[9] S. Pincus, I. Gladstone, R. Ehrenkranz, A regularity statistic for medical data analysis, J. of Clin. Monit. and Comput. 7 (4) (1991) 335–345.

[10] J. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am J Physiol Heart Circ Physiol 278 (6) (2000) H2039–2049.

[11] S. A. Akar, S. Kara, F. Latifoglu, V. Bilgiç, Analysis of the complexity measures in the eeg of schizophrenia patients, International Journal of Neural Systems 26 (02) (2016) 1650008.

[12] L. Mesin, P. Costa, Prognostic value of eeg indexes for the glasgow outcome scale of comatose patients in the acute phase, Journal of Clinical Monitoring and Computing 28 (4) (2014) 377–385.

[13] Z. Liang, Y. Wang, X. Sun, D. Li, L. J. Voss, J. W. Sleigh, S. Hagihira, X. Li, Eeg entropy measures in anesthesia, Frontiers in Computational Neuroscience 9 (2015) 16.

[14] D. Wu, J. Wang, Y. Yuan, Effects of transcranial direct current stimulation on naming and cortical excitability in stroke patients with aphasia, Neuroscience Letters 589 (2015) 115 – 120.

[15] G. Lee, S. Fattinger, A.-L. Mouthon, Q. Noirhomme, R. Huber, Electroencephalogram approximate entropy influenced by both age and sleep, Frontiers in Neuroinformatics 7 (2013) 33.

[16] D. E. Lake, J. S. Richman, M. P. Griffin, J. R. Moorman, Sample entropy analysis of neonatal heart rate variability, American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 283 (3) (2002) R789–R797. doi:10.1152/ajpregu.00069.2002.

[17] Y. Jiang, D. Mao, Y. Xu, A fast algorithm for computing sample entropy, Advances in Adaptive Data Analysis 03 (01n02) (2011) 167–186.

[18] C. Gómez, J. Poza, M. T. Gutiérrez, E. Prada, N. Mendoza, R. Hornero, Characterization of {EEG} patterns in brain-injured subjects and controls after a snoezelen® intervention, Computer Methods and Programs in Biomedicine 136 (2016) 1 – 9. doi:http://dx.doi.org/10.1016/j.cmpb.2016.08.008.

[19] H. Li, C. Peng, D. Ye, A study of sleep staging based on a sample entropy analysis of electroencephalogram, Bio-Medical Materials and Engineering 26 (s1) (2015) S1149–S1156.

[20] M. O. Mendez, I. Chouvarda, A. Alba, A. M. Bianchi, A. Grassi, E. Arce-Santana, G. Milioli, M. G. Terzano, L. Parrino, Analysis of a-phase transitions during the cyclic alternating pattern under normal sleep, Medical & Biological Engineering & Computing 54 (1) (2016) 133–148.

[21] J. Solé-Casals, F.-B. Vialatte, Towards semi-automatic artifact rejection for the improvement of alzheimer's disease screening from eeg signals, Sensors 15 (8) (2015) 17963.

[22] W. Chen, J. Zhuang, W. Yu, Z. Wang, Measuring complexity using fuzzyen, apen, and sampen, Medical Engineering and Physics 31 (1) (2009) 61 – 68. doi:http://dx.doi.org/10.1016/j.medengphy.2008.04.005.

[23] J. Xiang, C. Li, H. Li, R. Cao, B. Wang, X. Han, J. Chen, The detection of epileptic seizure signals based on fuzzy entropy, Journal of Neuroscience Methods 243 (2015) 18 – 25.

[24] Y. Cao, L. Cai, J. Wang, R. Wang, H. Yu, Y. Cao, J. Liu, Characterization of complexity in the electroencephalograph activity of alzheimer's disease based on fuzzy entropy, Chaos 25 (8).

[25] N. Sriraam, B. R. Purnima, U. M. Krishnaswamy, Comparative study of fuzzy entropy with relative spike amplitude features for recognizing wake–sleep stage 1 eegs, International Journal of Biomedical and Clinical Engineering (IJBCE) 2 (4) (2015) 12–25. doi:10.4018/IJBCE.2015070102.

[26] A. U. Rajendra, B. Shreya, F. Oliver, A. Hojjat, C. E. Chern-Pin, L. W. Jie, K. J. En, Nonlinear dynamics measures for automated eeg-based sleep stage detection, Eur. Neurol. 74 (2015) 268–287.

[27] J. A. Urigüen, B. Garcia-Zapirain, Eeg artifact removal—state-of-the-art and guidelines, Journal of Neural Engineering 12 (3) (2015) 031001.

[28] S. Bhardwaj, P. Jadhav, B. Adapa, A. Acharyya, G. R. Naik, Online and automated reliable system design to remove blink and muscle artefact in eeg, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 6784–6787. doi:10.1109/EMBC.2015.7319951.

[29] R. A. Feis, S. M. Smith, N. Filippini, G. Douaud, E. G. P. Dopper, V. Heise, A. J. Trachtenberg,

J. C. van Swieten, M. A. van Buchem, S. A. R. B. Rombouts, C. E. Mackay, Ica-based artifact removal diminishes scan site differences in multi-center resting-state fmri, Frontiers in neuroscience 9 (2015) 395. doi:10.3389/fnins.2015.00395.

[30] P. N. Jadhav, D. Shanamugan, A. Chourasia, A. R. Ghole, A. A. Acharyya, G. Naik, Automated detection and correction of eye blink and muscular artefacts in eeg signal for analysis of autism spectrum disorder, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 1881–1884. doi:10.1109/EMBC.2014.6943977.

[31] W. Zhou, J. Gotman, Automatic removal of eye movement artifacts from the eeg using ica and the dipole model, Progress in Natural Science 19 (9) (2009) 1165–1170.

[32] S. Devuyst, T. Dutoit, P. Stenuit, M. Kerkhofs, E. Stanus, Cancelling ecg artifacts in eeg using a modified independent component analysis approach, EURASIP Journal on Advances in Signal Processing 2008 (1) (2008) 747325. doi:10.1155/2008/747325.

[33] S. Muthukumaraswamy, High-frequency brain activity and muscle artifacts in meg/eeg: A review and recommendations, Frontiers in Human Neuroscience 7 (2013) 138. doi:10.3389/fnhum.2013.00138.

[34] K. T. Sweeney, T. E. Ward, S. F. McLoone, Artifact removal in physiological signals–practices and possibilities, IEEE Transactions on Information Technology in Biomedicine 16 (3) (2012) 488–500. doi:10.1109/TITB.2012.2188536.

[35] L. Zhao, S. Wei, C. Zhang, Y. Zhang, X. Jiang, F. Liu, C. Liu, Determination of sample entropy and fuzzy measure entropy parameters for distinguishing congestive heart failure from normal sinus rhythm subjects, Entropy 17 (9) (2015) 6270–6288. doi:10.3390/e17096270.

[36] R. G. Andrzejak, K. Schindler, C. Rummel, Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients, Phys. Rev. E 86 (2012) 046206. doi:10.1103/PhysRevE.86.046206.
URL http://link.aps.org/doi/10.1103/PhysRevE.86.046206

[37] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C. E. Elger, Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E 64 (2001) 061907.

[38] N. Kannathal, M. L. Choo, U. R. Acharya, P. Sadasivan, Entropies for detection of epilepsy in {EEG}, Computer Methods and Programs in Biomedicine 80 (3) (2005) 187 – 194.

[39] S. Deivasigamani, C. Senthilpari, W. H. Yong, Classification of focal and nonfocal eeg signals using anfis classifier for epilepsy detection, Int. J. Imaging Syst. Technol. 26 (4) (2016) 277–283.

[40] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, N. Stergiou, The appropriate use of approximate entropy and sample entropy with short data sets, Annals of Biomedical Engineering 41 (2) (2013) 349–365. doi:10.1007/s10439-012-0668-3.

[41] A. Molina-Picó, D. Cuesta-Frau, M. Aboy, C. Crespo, P. Miró-Martínez, S. Oltra-Crespo, Comparative study of approximate entropy and sample entropy robustness to spikes, Artif. Intell. Med. 53 (2) (2011) 97–106.

[42] A. L. Goldberger, L. A. N. 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 (23) (2000 (June 13)) e215–e220.

[43] O. Dimigen, W. Sommer, A. Hohlfeld, A. M. Jacobs, R. Kliegl, Coregistration of eye movements and eeg in natural reading: Analyses and review, Journal of Experimental Psychology: General 140 (4) (2011) 552–572.

[44] R. Croft, R. Barry, Removal of ocular artifact from the eeg: a review, Neurophysiologie Clinique/Clinical Neurophysiology 30 (1) (2000) 5 – 19.

[45] P. E. McSharry, G. D. Clifford, L. Tarassenko, L. A. Smith, A dynamical model for generating synthetic electrocardiogram signals, IEEE Transactions on Biomedical Engineering 50 (3) (2003) 289–294. doi:10.1109/TBME.2003.808805.

[46] D. Abásolo, R. Hornero, P. Espino, D. Álvarez, J. Poza, Entropy analysis of the eeg background

activity in alzheimer's disease patients, Physiological Measurement 27 (3) (2006) 241.

[47] M. O. Sokunbi, Sample entropy reveals high discriminative power between young and elderly adults in short fmri data sets, Frontiers in Neuroinformatics 8 (2014) 69. doi:10.3389/fninf.2014.00069.

[48] K. Balasubramanian, N. Nagaraj, Aging and cardiovascular complexity: effect of the length of rr tachograms, PeerJ 4 (2016) e2755. doi:10.7717/peerj.2755.

[49] R. Hornero, M. Aboy, D. Abasolo, J. McNames, B. Goldstein, Interpretation of approximate entropy: analysis of intracranial pressure approximate entropy during acute intracranial hypertension, IEEE Transactions on Biomedical Engineering 52 (10) (2005) 1671–1680. doi:10.1109/TBME.2005.855722.

[50] S.-D. Wu, C.-W. Wu, S.-G. Lin, K.-Y. Lee, C.-K. Peng, Analysis of complex time series using refined composite multiscale entropy, Physics Letters A 378 (20) (2014) 1369 – 1374.

[51] C. C. Mayer, M. Bachler, M. Hörtenhuber, C. Stocker, A. Holzinger, S. Wassertheurer, Selection of entropy-measure parameters for knowledge discovery in heart rate variability data, BMC Bioinformatics 15 (6) (2014) S2. doi:10.1186/1471-2105-15-S6-S2.

[52] S. Lu, X. Chen, J. K. Kanters, I. C. Solomon, K. H. Chon, Automatic selection of the threshold value r for approximate entropy, IEEE Transactions on Biomedical Engineering 55 (8) (2008) 1966–1972. doi:10.1109/TBME.2008.919870.

[53] C. Liu, K. Li, L. Zhao, F. Liu, D. Zheng, C. Liu, S. Liu, Analysis of heart rate variability using fuzzy measure entropy, Comput. Biol. Med. 43 (2) (2013) 100–108. doi:10.1016/j.compbiomed.2012.11.005.

[54] L. Ji, P. Li, K. Li, X. Wang, C. Liu, Analysis of short-term heart rate and diastolic period variability using a refined fuzzy entropy method, BioMedical Engineering OnLine 14 (1) (2015) 64. doi:10.1186/s12938-015-0063-z.

[55] R. Heijungs, P. J. Henriksson, J. B. Guinée, Measures of difference and significance in the era of computer simulations, meta-analysis, and big data, Entropy 18 (10) (2016) 361. doi:10.3390/e18100361.

[56] K. Balasubramanian, N. Nagaraj, Aging and cardiovascular complexity: effect of the length of RR tachograms, PeerJ e2755 (4) (2016) 1–18.

[57] T. Lumley, P. Diehr, S. Emerson, L. Chen, The importance of the normality assumption in large public health data sets., Annual review of public health 23 (1) (2002) 151–169.

[58] D. E. Lake, J. R. Moorman, Accurate estimation of entropy in very short physiological time series: The problem of atrial fibrillation detection in implanted ventricular devices, American Journal of Physiology - Heart and Circulatory Physiologydoi:10.1152/ajpheart.00561.2010.

[59] D. E. Lake, J. R. Moorman, Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices, American Journal of Physiology - Heart and Circulatory Physiology 300 (1) (2011) H319–H325. doi:10.1152/ajpheart.00561.2010.

- Muscular artifacts are the most influencing artifacts in EEG records in terms of entropy calculation.

- Approximate Entropy is very sensitive to the presence of outliers and should not be used in this context.

- Fuzzy Entropy is the most robust entropy metric against the usual EEG signal artifacts.

- There is a great input parameter variability and each case should be configured independently.

- No need to process EEG records longer than 1000 samples.