4ª ed

MIM

EEG Analysis by Compression

Diogo Sato

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

SET | 2011

4ª ed

MIM

EEG Analysis by Compression

Diogo Sato

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

ORIENTADORES:

Luís Filipe Antunes, PhD

Ângela Santos, MSc

OUT|2011 (15/outubro/2011)

Acknowledgements

I would like to thank all colleagues and professors of Medical Informatics Master Course at University of Porto who shared life experiences (aside dinners and soccer matches) and offered support along the academic years, from 2010 to 2011.

Special thanks to Martins Mupueleque, Tatiana Bruna, Rui Mendes, Cátia Silva, Fabio Hedayioglu, Jorge Ribeiro, Rita and Marlon Silva. They're all great friends and important as they are.

As a student, I must praise some teachers and collaborators. To Luís Filipe Antunes, coordinator of this thesis, Pedro Pereira Rodrigues, thesis reviewer, and Angela Santos, epilepsy expert, my special thanks for their primordial roles to keep me on the right way of this thesis' excellence. Also, a due recognition to Ana Paula Tomás and António Porto, both professors from Computer Science Department at University of Porto, with respect to their classes. These masters and collaborators really made the difference and wisely awoken me to the scientific and computer challenges.

I'd like to place on record my acknowledgement to PredialNet ISP company (http://www.predialnet.com.br) for the whole financial support given on those academic years.

Finally, I certainly am thankful to my sweetheart Camila Azevedo, my very very huge family and closer friends, all these people always give me the right directions to be a happy human being.

Abstract

In this thesis we introduce a novel algorithm to inspect and evaluate electroencephalogram exams (EEG). The EEG Segmentation Search (ESS) algorithm is based on the theory of Kolmogorov complexity; in particular we explore this notion over the EEG tracings using compressors as the measurement technique to express the data complexity of its time-series values.

For this work, a total of N=40 exams were used to test the accuracy of the proposed method. This dataset represents EEG records of different patients, all collected at the same place with personal data previously removed to keep patient privacy. These samples were prior classified by an expert as "Normal", "Absence" or "Rolandic". All epileptic exams (N=14) were correctly classified and their seizures crisis, if existed, were noticed by the system using the exact moment (in seconds) of occurrence.

The anonymous dataset used in this study as well as the ESS algorithm can be evaluated online at: http://www.eegonline.org/.

Preamble

This thesis started as an academic attempt to understand the reasons why the clustering method proposed in (Cilibrasi and Vitányi 2005) fits perfectly to the fetal heart rate analysis (Santos, Bernardes et al. 2006), identifying correctly exams which present problematic patterns, but does not present the same performance when we applied on the electroencephalogram (EEG) exams. This new method is based on the formal theory of Kolmogorov complexity, a mathematical rigorous measure of the amount of information in an object. As Kolmogorov complexity is not computable the authors proposed a measure to compute the similarity between two data objects using the 'normalized compression distance'. Our preliminaries' results indicate that this technique based on the compression was not able to express the complexity of EEG data, since this exam constantly present high variability in its values, hence restricting the sensitivity of this method. However, we got interested on this phenomenon, since nowadays there are available many high-quality compressors, and, intuitively, there is no good reason to their algorithms be incapable to 'capture' the drastic numerical randomness difference between the EEG healthy and problematic tracings. This suspicious motivated us to conduct a review on the EEG literature looking for more information about which variables are considered on the examinations analysis, and which automatic techniques' are being developed by researches. Finally, we also checked on the web if would exist any application or service available that could decides seizures presence along EEG exams.

Along the review, we could verify that the EEG analysis is a scientific topic well debated since 1935, when Gibbs, Davis and Lennox studied the brain spike waves on clinical seizures in (Gibbs, Davis et al. 1935). Aside the long time history, this subject has tasting in the last decades the enhancement promoted by the new computational algorithms designed for supporting professionals on the EEG inspection task. In addition, we observed on papers a clue of a hidden battle between two main EEG analysis strategies: On one side, the time and frequency domain approaches, commonly based on statistical measurements, Wavelet and Fourier transforms (FFT), and on the other side, the nonlinear dynamics approach, using many estimated correlations, complexity and entropy measures. However, there is still a lack of real world applications available which can automatic decide if an EEG exam presents seizures patterns or not. These evidences pointed us that the computer analysis of EEG exams remains as an interesting topic and next solutions can take advantage of the pros and cons of these available techniques and can also be concerned about being functional applications. Naturally, this implementation, if possible, would be remarkable to the health business and community.

Contents

Acronyms

ANN – Artificial Neural Networks

AR – Auto Regression

ASCII - American Standard Code for Information Interchange

BFS - breadth first search

BRE - benign rolandic epilepsy

BCECTS - benign childhood epilepsy with centro-temporal spikes

CSV – comma-separated values

ECG - electrocardiogram

EDF – European Data Format

EEG – electroencephalogram examination

EMA – exponential moving average

ESS – EEG Segmentation search

FFT – Fast Fourier transform

SMA – simple moving average

List of Figures

List of Tables

Thesis Structure

The structure of this thesis is as follows:

Chapter 1 introduces the addressed question, motivations, objectives and contributions of this study.

Chapter 2 presents the medical background and current state of art of the proposed subject.

Chapter 3 describes the methodology points: dataset, resources and techniques used.

Chapter 4 shows the results obtained by the tests performed.

Chapter 5 discusses the positive and negative results gathered and the method trade-offs.

Chapter 6 summarizes the contributions, limitations as well as some topics not covered by this thesis but where future research can be conducted.

Chapter 7 lists all scientific papers and references used along this thesis.

Scientific results

1. Presented as master thesis proposal on the 3º Symposium in Medical Informatics at University of Porto, Portugal.
2. Temporally Patent proposal application at INPI/Portugal.

1. Introduction

This chapter introduces the addressed question, motivations, objectives and contributions of this study.

1.1 Research Question

The study of electroencephalography, or recording of electrical activity of the brain, involves more than a century years of research. According to (Swartz and Goldensohn 1998), it dates back to 1875, when Richard Canton discover the existence of electrical activity from the rabbits brain surface, after that, in 1924, Hans Berger succeed in recording the electrical activity of a human brain, using a device that he called electroencephalogram, also known as EEG.

MedlinePlus Medical Encyclopedia defines EEG as a test widely used on medicine to check electrical activity of the brain. (MedlinePlus) It works using between 16 and 25 electrodes plugged on different positions of a patient scalp (Misra and Kalita 2005), all liked to an amplifier and recording machine that reads and converts electrical brain impulses into time-series values that can be showed on a screen as a series of wavy lines or traces, as well as stored in a computer file. EEG provides a set of physiological and pathological information describing the summation of electrical discharges of networks of neurons, which plays an important role in the detection of brain seizures, and also evaluation of epilepsy patients such as determining epileptic zones for presurgical evaluations. (Lehnertz, Mormann et al. 2003; Gandhi, Panigrahi et al. 2010)

In recent years, (Stam 2005) reviewed recent neuroscience researches with new trends defending how electrical brain activity can be interpreted using chaos theory and nonlinear concepts in contrast with the old fashion quantitative (statistics and time-frequency) analysis. Although those works are inspiring the community to publish several computational systems promising accurate approaches to detect seizures, such as (Lerner 1996; Yaylali, Kocak et al. 1996; Celka and Colditz 2002; Altenburg, Vermeulen et al. 2003; Smit, Vermeulen et al. 2004), at this present moment, the EEG analysis still relies on trained neurologists or technicians to inspect the signals recordings visually, seeking for abnormalities expressed on the waves charts of time-series values. Also, this human-dependent procedure is known that can lead to a lot of misdiagnosis and *overreading* problems on benign EEG patterns. In (Hernandez-Frau and Benbadis 2011) the authors mention that the main reason for this sort of fault is the lack of professional experience, summed to the biased reading of chart according to the medical history of patient. Another critical factor of visual examination is the scoring of long terms exams which are stressful and very time-consuming. Thus, in response of these bad issues, the automatic seizures detection procedure becomes a very valuable tool for assisting EEG professionals. (Yuan, Zhou et al. 2011)

1.2 Objectives

This thesis aims to build a feasible web version application to inspect seizures on electroencephalogram exams, which represents an initiative to offer free access to a functional medical tool as a modern manner to stimulate doctors and patients empowerment.

Aside this application, this present work evaluates the nonlinear feature present on EEG analysis by exploiting the Kolmogorov approximation approach proposed by (Cilibrasi and Vitányi 2005) as the main measure to express data complexity present within EEG time-series values, and also discusses the practical pitfalls and trade-offs of this chosen technique.

1.3 Contributions

Since that there is lack of practical implementation available, the great relevance of this thesis does really consists in the development of web free version of an effective algorithm to decide, in a precise manner, the presence of seizures patterns along the EEG exams using their data complexities measures.

This represents a breakthrough for both medical personnel which deals with epilepsy recognition as well as for the information theory research related to the Kolmogorov complexity applicability.

2. Background

Chapter 2 presents the medical scope related and state of the art of the automatic analysis of electroencephalograms.

2.1 What is EEG?

According to MedlinePlus Medical Encyclopedia, electroencephalogram (EEG) is a test widely used on medicine to check electrical activity of the brain. It works using from 16 to 25 electrodes plugged on different positions of patients scalp, all liked to an amplifier and recording machine that converts electrical brain impulses into time-series that can be read on a screen as valued wavy lines chart, as illustrated on Figure 1, as well as stored in a electronic file.

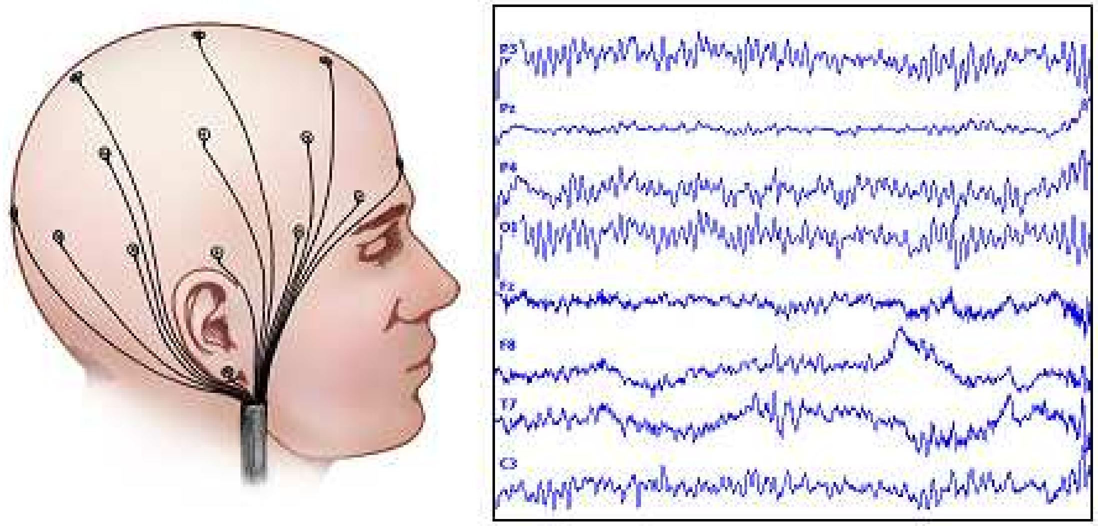

Figure 1: EEG electrodes on patient’s scalp and the waves chart

During the examination usually an expert assists the patient by performing the read of brain signals seeking for abnormalities over the patterns expressed on the channels series which have large use on diagnosis of brain disorders. Along this procedure, the professional pays attention on different kind of variables, the environmental and behavioral ones, such as stress level, drugs use, eyes-blinking and consciousness, and the graphical ones, performing visually analysis of tracings values.

The EEG exam has a special play on medicine, mainly in Neurology, because it is a safe and painless test with a large applicability on the inspection of psychiatry syndromes, drug's effect, brain death, and also as first-line method for the diagnosis of brain tumors, strokes, focal disorders, and in special epileptic seizures. This last better exemplifies the importance of this powerful exam, since that people on this condition present unusual brain electrical activities and doctors can recognize the epileptic patterns along the waves chart.

2.2 Seizures and epilepsies

From the basics, a seizure is a surge of brain electrical activity that suddenly occurs and usually affects, for a short period of time, the sensibility and actions of a person who suffers this episode. There are many kinds of seizures that can be classified by neurologists according to the brain area affected, the primary generalized seizures or partial seizures, respectively. Although seizures do not represent a disease, in fact, they are a symptom of many brain disorders, such is the epilepsy. (Epilepsy.com 2008)

On the other hand, epilepsy is a neurological condition, also known as a seizure disorder that usually is diagnosed after a person has had at two or more seizures episodes distinct from a known medical condition like alcohol withdrawal or extremely hypoglycemic episode. This disorder is also called a syndrome since it is defined by a set of characteristic features that commonly occur together. There are different types of epileptic syndromes, which can be described by the types and frequencies of seizures, the patient age and background, the part of the brain involved, certain patterns on the EEG and others disorders in addition to the symptoms. This factors analysis has a special relevance on the patient's treatment, since doctors can prescript the most helpful medications according to the characterized syndrome type. (Epilepsy.com 2007)

In this study we are interested in two particular childhood seizures types, the absence seizures, present on both case of Childhood/Juvenile Absence Epilepsy and Juvenile Myoclonic Epilepsy, and the partial seizures on the rolandic fissure area, present on the Benign Rolandic Epilepsy, respectively. They have a special relevance because the EEG examination is highly recommended for these disorders identification, since the others brain scans, such as CT and MRI, present normal results. (Epilepsy.com 2007; Epilepsy.com 2008; Epilepsy.com 2009; Epilepsy.com 2011)

2.2.1 Absence seizures

According to (Temkin 1971) absence seizures are a type of generalized seizures, which were first studied in the beginning of the 18th century by Poupart and followed by Tissot, who used the term petit access to describe this phenomenon. The term absence was introduced years later by Calmeil, and Gibbs, Davis, and Lennox described it as the association of impaired consciousness and 3 Hz spike-and-slow-wave complexes on electroencephalograms (EEGs). (Gibbs, Davis et al. 1935)

The absence seizures can be present in both idiopathic and symptomatic generalized epilepsies. (Epilepsia 1989) Among the idiopathic ones, they are seen in cases of childhood (pyknolepsy) and juvenile absence epilepsy, and juvenile myoclonic epilepsy (also called as petit mal). (Benbadis and Berkovic 2006) When occur in these conditions, they are known as typical absence seizures and are frequently identified through an EEG examination, associated with sudden start-termination and seizure duration commonly less than 10 seconds with 3-4 Hz spike-and-slow-wave complexes. (Gibbs, Davis et al. 1935; Epilepsy.com 2007)

2.2.2 Rolandic fissure seizures

The benign rolandic epilepsy (BRE), also called as benign childhood epilepsy with centro-temporal spikes (BCECTS), is an epileptic syndrome which involves partial seizures in a specific brain region named Rolandic fissure or *central sulcus*, as shown in Figure 2.

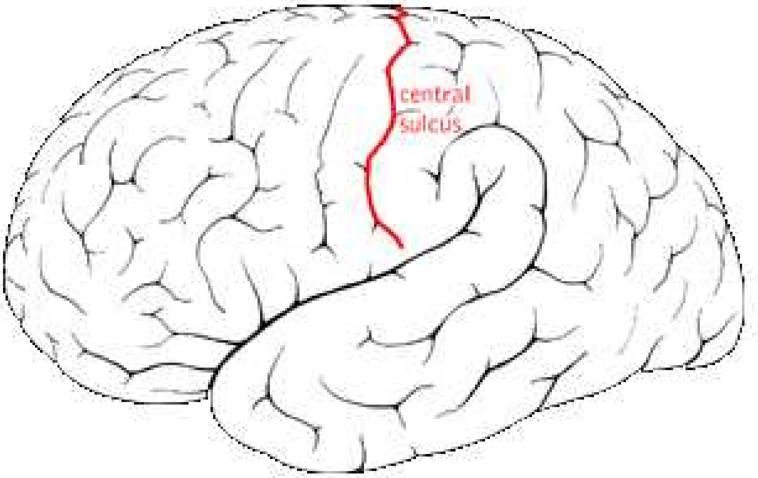

Figure 2: Rolandic fissure localization

It is one of the most common types of epilepsy occurring in children, special in the ages range from 3 to 13 years, with the major incidence occurring between the 7-8th year. (Kriz and Gazdik 1978) Usually these children have epilepsy in their family background and the seizures commonly occur at night time while sleeping. Also, these kids can express symptoms such as difficulty on speak, twitching or stiffness (clonic or tonic, respectively) movements on the face, arms and legs. Although the episodes of seizures occur, those cases have good outcomes, due to almost children with BRE do not present any loss of intelligence and will outgrow the disorder on puberty. (Lerman and Kivity 1975; Epilepsy-Action 2010)

In the BRE scenario, the EEG is also a key role for the right diagnosis, as others exams, such as neurological examination and MRI, do not offer any clues to doctors about this kind of brain disorder. (Epilepsy.com 2007)

2.3 Review of EEG Analysis

Since Berger's invent, the EEG subject has been improved along the years, and no less than 10 years later, in 1935, three scientists, namely Gibbs, Davis and Lennox, started the study of the interictal brain spikes on clinical absence seizures cases. (Gibbs, Davis et al. 1935) Although it was the preliminary shot on the clinical usage of EEG, that work contributed a lot on the understanding of brain system by inspecting the relationship between EEG signals and brain disorders. Also that, their approach was already be in accordance to the principles of modern neuroscience, published years later by Kendal, which states that the main task of neuroscience is to explain behavior in terms of brain activity. (Kandel, Schwartz et al. 2000)

Since the EEG analysis has become a very default and important procedure with large medical use, the researches communities have identified an opportunity to use the computational capability to extract rich and more meaningful information from the EEG recordings, offering several analysis tools to health professionals. This movement can be exemplified by several statistical, time and frequency based techniques. In 1970, (Hjorth 1970) published an article introducing a method to characterize an EEG trace through 'descriptive parameters' based on statistics derived from its power spectrum. Fifteen years later, (Penttilä, Partanen et al. 1985) studied the different stages of Alzheimer's disease by using the fast Fourier transform (FFT) as frequency analysis. Since the 90's, others studies initiatives have introduced different techniques in contrast to the standard FFT time-series analysis. In (Pardey, Robertst et al. 1996), the authors indicate the use of auto regressive (AR) modeling for spectral estimation. Also, (Kalayci and Ozdamar 1995) applied artificial neural network (ANN) on the EEG spikes detection. Finally, many researches appeared confirming the better fitting of the wavelet transform (WT) along the EEG signals analysis. (Akin 2002; Rosso, Martín et al. 2006; Ocak 2009)

In parallel, a secondary research movement has emerged in more recent years, putting the nonlinear dynamics analysis as a hot topic into the EEG scenario. As (Stam 2005) indicates, the application of this approach to electroencephalography has helped the development of new nonlinear measures which are suitable to be applied on the description of noisy EEG data. One of the most important contribute to this understanding is the attempt to solve the challenge of "inscrutable oscillation of the EEG" mentioned by (Jones 1999). Many researches dealt with this problem using measures based on notions of dimensional correlations, as discussed in (Aftanas, Koshkarov et al. 1994) and (Besthorn, Sattel et al. 1995), Lyapunov exponent in (Fell, Röschke et al. 1993; Kantz 1994; Aftanas, Lotovaa et al. 1997), and data entropy in (Dünki 1991; Bruhn, Röpcke et al. 1993; Dhamala, Pagnoni et al. 2002). However, these efforts have not being consolidated into any standard technique, leaving open gaps in both theoretical understanding and proper interpretation of these tools.

Faced this unstable scenario, where the scientific community still does not have a concrete proof of how exactly the nonlinear dynamics of EEG reflects the human brain behavior, (Stam 2005) suggests that the development of nonlinear analysis tools will be the future of this research area.

Therefore, this study will be attempt to demonstrate how apply the nonlinear concept in a practical way into the EEG analysis, through the development of a new clinical application to identify abnormal brain activities on EEG series data, using data complexity as main technique.

2.4 Kolmogorov Complexity Applications

In recent years, the nonlinear dynamic analysis had introduced many measures in attempt to undercover more relevant information intrinsic into EEG data series. One of them is called Kolmogorov complexity, which can be referred as the absolute quantification of the amount of information contained in a finite data object. (Cilibrasi and Vitányi 2005)

In Information Theory, the computational complexity of an object x is given by the minimum amount of bit required for its description. This measure, also called the Kolmogorov complexity $K(x)$ or the minimal algorithmic description of x, aims to express the compressibility or randomness of information that specifies a particular object. (Cilibrasi and Vitányi 2005) In computing, $K(x)$ represents the smallest binary program that, without any additional input, would result in the output x. Extending this formulation, the Kolmogorov complexity of an object x given an input y, can be expressed as $K(x|y)$.

Although this measure still remains not computable, i.e. without an exact algorithm to calculate this function, in recent research published in 2005, (Cilibrasi and Vitányi 2005) suggests that the Kolmogorov complexity measure

$K(x)$ could be approximated using real compressors $C(x)$, such as gzip, bzip2 and others. In this manner, the authors also introduce the *normalized compression distance* (NCD) as a similarity function to compute the distance between two objects in a simple strategy, which do not requires any parameter settings neither any prior knowledge about the nature of information data, being only restricted by their own description of objects.

Since then, the Kolmogorov complexity approximation effectiveness has been well tested by many studies, through the NCD function, in different domains of information. For instance, (Li, Badger et al. 2001) uses it to construct the phylogeny tree based on mitochondrial genomes, (Li, Chen et al. 2004) uses this technique to compute the similarity distance between 50 Euro-Asian languages, (Chen, Francia et al. 2004) detects plagiarism in student programming assignments, (Cilibrasi, Vitányi et al. 2004) clusters different styles of music and (Wehner 2007) classifies worms on the Internet data traffic. And, in a special case, (Santos, Bernardes et al. 2006) made a successful experiment of this method for the detection of abnormalities on fetal heart rate tracings. It was a successful case of the use of this remarkable approach with a medical purpose.

Hence, this present work contributes to this research line by testing the Kolmogorov complexity applicability into another medical field. In this study the data complexity approximation using compressors is exploited for the analysis of seizures into EEG data recordings.

3. Materials and Methods

Chapter 3 describes the medical dataset, computational resources and also explains the designed techniques.

3.1 Dataset

A total of 40 EEG tests were collected from children patients' scalp. All tests are electronic files with personal information excluded previously. From this set, 28 files represent complete long-term exams samples obtained using the European Data Format[1] (EDF), and others 12 are in CSV format, representing small samples with only 5 minutes period data. Each file received a sequential number, from 1 to 40, concatenated with their file extensions, respectively.

The exams were prior classified by an expert, using the default visual analysis as technique. On this set, 14 tests presented epileptic wave-forms and 26 were classified as normal samples, i.e. with no seizures crisis identified. Within those epileptic exams, 5 have absence seizures patterns and 9 rolandic epilepsy patterns; however we lack the annotations when these seizures exactly have occurred along the examinations.

---

[1] Specification at: http://www.edfplus.info/specs/edf.html

Table 1: EEG dataset description

Exam files	**Classification**
28 EDF (complete exams)	3 with absence crisis 9 with rolandic crisis 14 normal
12 CSV (5 minutes exams)	2 with absence crisis 10 normal
N=40	**14 epileptic/ 26 normal**

Unfortunately, some samples presented no recording values in some specifics EEG channels. To overcome these setbacks and keep consistency of input data, the removal of those deaf channels from the analysis process was decided. Thus, a free software named "edf2ascii" was adapted to convert the EDF files into free-text files (see Annex I) containing only values of 19 channels, as depicted on Figure 3, namely, F3, F4, F7, F8, Fz, FT9, FT10, T7, T8, P3, P4, P7, P8, Pz, C3, C4, Cz, O1 and O2.

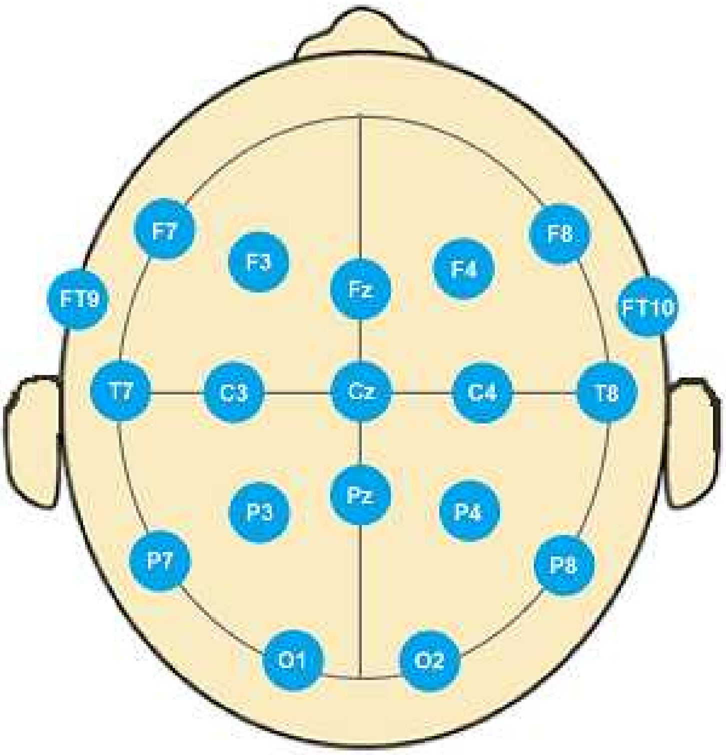

Figure 3: Electrodes signals captured for the EEG analysis

3.2 Computational Resources

The follow IT resources supported the development and execution of this project. A commercial computer machine with Pentium Intel Core 2 Duo processor and 2 GB of RAM was used to install a free Ubuntu 10.10 operating system running the Apache2 and MySQL services for hosting the algorithm and the web page files necessaries to perform the EEG analysis remotely. Due to this online feature, the web pages were written as PHP scripts to support the callbacks to the main program written in GNU C language. Also, two Linux command line programs were used: the *split* command line, necessary to divide the EEG time series data into smaller period files, and the *bzip2* data compressor, as the Kolmogorov complexity function approximation.

3.3 EEG Analysis

The development of the proposed techniques involved three moments, according to the needs, challenges and knowledge gain of each phase.

3.3.1 Detecting Spikes

In a first moment, the data analysis began in an attempt to better understand the transition of EEG rhythmic activity from normal to abnormal behavior. For this goal, we would like to observe the difference between the periodic oscillations on both cases, i.e. count the spikes on these time periods.

The counting of periodic oscillations reduces to the task of detecting local maximum within a temporal series. Although this subject is widely discussed on science with several exemplifications such as (Kolawole 2003; Sanei and Chambers 2007; Kim, Yu et al. 2009; MathWorks 2011), there is no standardization on any approaches and the effectiveness of these algorithms commonly relies on the nature of application as well as on the fine tuning of parameters.

Motivated by these factors, a new spike detection algorithm was designed to better fit our data set domain. The proposed technique consisted in a comparison of instant values with their respective mean values, which involved the computation of three simple moving averages (SMA) with only the last 2, 32 and 64 sample values[2], representing the mean voltage values of the past data periods of 8ms[3], 125ms and 250ms, respectively, and also an exponential moving average (EMA) was used with the smoothing factor equals

---

[2] Values considering a recording rate of 256 samples per second

[3] The last 8ms values before the current value

0.9. Then, for each value point t in an EEG channel series, the algorithm evaluates the following condition sentences:

```
if (not Peak[t-1]) and SMA(t-1,2) < SMA(t,64) < Val[t] then
    if SMA(t-1,2) < SMA(t,32) < Val[t] or SMA(t-1,2) < EMA(t,0.9) < Val[t] then
        Peak[t] = true
```

Figure 4: Pseudo-code for the spikes detection conditions

In Figure 5 is exemplified the running of spike detection algorithm in a small sample values of a unique EEG recording channel.

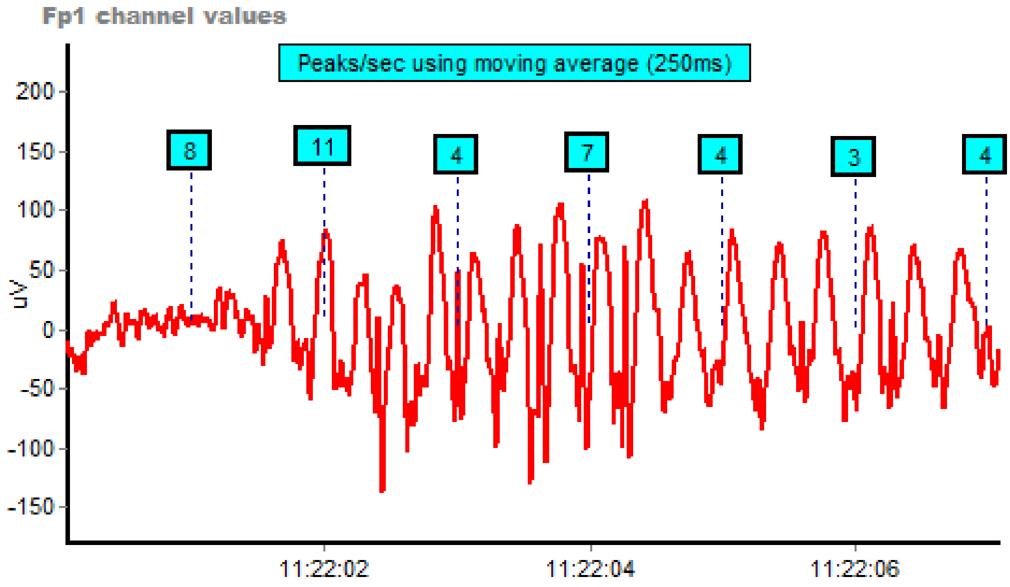

Figure 5: A tiny execution (7 seconds) of spikes detection algorithm

After counting spikes, the EEG channels could be transliterated into universally recognized Greek mnemonics, representing the frequency bands according to the number of spikes detected in each second values. The following Table 1 describes each representation:

Table 2: Representation of Greeks mnemonics

Frequency bands (peaks/sec)	Greek mnemonic
up to 4	(D) Delta
5 – 7	(T) Theta
8 – 12	(A) Alpha
13 – 29	(B) Beta
30 – 100	(G) Gamma

Even being a very promising approach, this procedure designed, however, requires long time consumption due to the inspection of each time series values of 19 EEG channels involved in the calculation of statistics necessaries to detect and count spikes. Hence, this method proved to be too slow for the online inspection of common long term EEG exams.

3.3.2 Extracting Complexity Measures from Exams

Actually, the previous undesired issue was relevant to motivate the next development step, which involves the data complexity analysis using the Kolmogorov approximation by compressors. This one, when available, could offer a great advantage of data inspection in a global manner, without the necessity of visitation of each value points.

As mentioned earlier, the proposed complexity measure can express, instead of counting spikes, the randomness of EEG data involved if exploited using its compressibility notion. So, this notion can be directly extracted from the Kolmogorov approximation C(x) by:

$$\mathrm{Z}(x) = 1 - \frac{|C(x)|}{|x|}$$

Where $|x|$ and $|C(x)|$ are the representation in bits of the length and the compressed version length of object x, respectively.

Assuming that the compression function always return a smaller or equal version of x, then the value of $Z(x)$ can be interpreted as the compression rate of it. Given this, how much higher this measure is, the more the object data are compressible, and vice versa. Hence, we intuitively expect that the disturbed (or maybe epileptic) EEG files may present their Z-values lower than the regular (and healthy) ones. Thus, doing a straight application of this notion in every EEG sample, it could be very useful to signal different wave patterns between the exams in a comparative analysis.

Although this second attempt overcame the long processing time problem, when applied on a test bed of 10 EEG files it did not reach the goal of signaling abnormalities, given the fact that the measured compression rate of exams, in both normal and epileptic cases, almost belong to the same range of Z-values, as detailed on Table 2.

Table 3: Compression rate of a test bed of 10 EEG files

Exam	**Patterns**	**Z-value**
1.edf	Absence	0,398
2.edf	Absence	0,410
9.edf	Rolandic	0,428
13.edf	Rolandic	0,377
14.edf	Rolandic	0,387
17.edf	Normal	0,305
19.edf	Normal	0,395
20.edf	Normal	0,446
23.edf	Normal	0,367
25.edf	Normal	0,414
N=40	**4 epileptic/ 26 normal**	**$\overline{Z}$=0,393**

This was a really bad issue, since the approach turned innocuous to evaluate the exams healthiness. In addition, we believed that this behavior occurred due to the nature of brain data domain that typically present a high variability inherited on the voltage values of the EEG recording channels.

3.3.3 Complexity from EEG Segments

Driven by this second pitfall, a third attempt was conducted not exploring the disagreement between EEG Z-values but using the own information contained in the interested period analysis. In this manner, we promote a divide and conquer[4] strategy design by splitting a unique exam into smaller time periods (segments) and computing their own compressibility; it could be possible to categorize these small chunks by observing the discrepancy of their compression rates with the chunks set mean Z-value using a dispersion measurement, such as the standard deviation.

This simple change on the second approach shown to be very useful on the identification of high complexity periods along the EEG test, and also has promoted the independence of the analysis process, being performed with the own data contained in an unique exam file without the necessity of comparisons against others tests. It becomes notable that the method could serves as an indicator of regions with possible presence of abnormal and high disturbed EEG tracings.

According to the preliminary results, the periods indicated do really present high randomness patterns on their data values, hence they produce a very disturbed wave lines charts, however, some of them were false positive cases, i.e. they do not present seizures discharges patterns, but only noise data. This wasn't an unexpected behavior since the EEG recording is known to be very susceptible to many sources of noise, which represent challenges in the interpretation of this kind of exam.

---

[4] A divide and conquer strategy works by recursively breaking down a problem into smaller sub-problems of the same (or related) type, until these become simple enough to be solved directly.

This present work does not intend to discuss which standard technique to deal with noise data fits better, despite (Repovš 2010) discusses several known strategies available, both at the time of EEG recording as well as during a preprocessing phase. However, as the computation of Kolmogorov complexity uses the raw values of EEG channels and as we already developed a proper procedure for counting spikes in the beginning of this study, hence, we decided to make some little adjustments on this previous method by transforming it in a parametric function to produce filtering measurements which encompass a bypass filter implementation to count the EEG channels which present voltage values exceeding a specific range given, and also a calculation of the mean standard deviation of these channels series involved.

In fact, the main algorithm runs this filter procedure on the outlier segments selected. On this post phase, are calculated the filtering measures and, then, the result set is filtered by excluding the outliers which not attend some particular conditions defined. To exemplify what kind of conditions were built, we empirically observed that most false positives cases have presented their mean standard deviation values less than 70 points, also the seizures in our dataset commonly occurs when presenting at least 15 EEG channels with values exceeding the voltage range of [-170uV, +170uV] and most of them really present 3-4 HZ waveforms (or Delta, using the Greek nomenclature).

3.4 ESS Algorithm

In this section we present the main algorithm description as the result of the EEG analysis development.

As mentioned, the third method proposed inspects an EEG test in two complementary steps. As the first step, it runs a modified version of breadth first search (BFS) algorithm, nicknamed as “EEG Segmentation Search” (ESS), to locate the periods with high complexity values. For this task, it uses the following set of parameters as input:

- F: the exam file containing the EEG channels values;
- HZ: the record sampling rate;
- WND: the window split wanted (greater or equal than HZ);

In this phase, the algorithm separates all period sample values contained in F, the root node, into smaller splits FS containing WND samples each one. They represent the sequential time segments children containing exact (WND/HZ) seconds values each one. After the split, the method observes computes the discrepancy of the **Z**-values of each child node in comparison with the set mean **Z**-value, classifying as outlier those which present a high deviation from that expected compressibility value.

Once classified, the algorithm enqueues these (WND/HZ)-seconds segments using HZ as the next window split. Hence, on the following recursive cycles, these outliers will be subdivided into subsequent splits of 1 second values each (with HZ samples only), and the function may check the specifics seconds which are very discrepant, in terms of compressibility, from their related neighbors’ in set. So, assuming that the studied seizures are commonly associated with short period duration and sudden start-termination, the algorithm classifies as “abnormal” the instant times which present randomness in their values so far distant from the expected healthy ones.

```
function CompressionRate(FS, Zip)
    sum = 0, sum2 = 0
    for each f in FS do
        f.compressed_size = Zip(f)
        f.z_value = 1 - (f.compressed_size / f.size)
        sum = sum + f.z_value
        sum2 = sum2 + f.z_value * f.z_value
    FS.z_mean = sum / FS.num_files
    FS.z_deviation = (sum2 - sum * FS. z_mean) / FS.num_files
return (FS)

function ESS(F, HZ, WND)
    Results = [ ],Q = [ ]
    Enqueue(Q, [F, WND])
    while Q is not empty do
        v = Dequeue(Q)
        FS = Split(v.period, v.samples)
        CompresionRate(FS, Bzip2)
        for each f in FS do
            f.distance = (f.entropy - FS. z_mean) / FS. z_deviation
            if (not -1 < f.distance < 1) and v.samples > HZ then
                Enqueue(Q, [f, HZ])
            else if (not -1 < f.distance < 1)
                Enqueue(Results, f)
return (Results)
```

Figure 6: Pseudo-code of EEG Segmentation Search

3.5 Algorithm Analysis

In this section, we develop the performance analysis of the ESS algorithm using the big O notation for a better comprehension of the execution time expectation of this proposed method.

As showed earlier in Figure 6, the pseudo-code of ESS algorithm implements a modified version of the Breadth First Search technique, basically differing by maximum search depth and the calling functions, *Split* and *CompressionRate*, respectively. Thus, our performance study will be based on some characteristics of this well known BFS analysis.

Let assume that the BFS, in the worst scenario, visits b nodes (branching factor) in a constant time $O(1)$ in each tree search level, and suppose its tree has a maximum depth of d, then it would finish its search function in $1 + b + b^2 \ldots + b^d$ steps, which is $O(b^d)$. However, in the ESS case the time complexity differs a little, given that it performs only a two level search along the file F data, with the branching factor in each level depending on the splitting parameters given, WND and HZ, respectively, as shown in Table 4. In addition, as depicted in Figure 7, the *Split* function is called on every parent node, and the compression *Zip* function, used to compute the Kolmogorov complexity, is called over the children nodes set at the *CompressionRate* procedure.

Table 4: ESS tree level analysis

Level	**Parent nodes**	**Parent length**	**Branching factor**	**Children length**
1	1	F	F/WND	WND
2	F/WND	F/WND	F/WND * WND/HZ	HZ

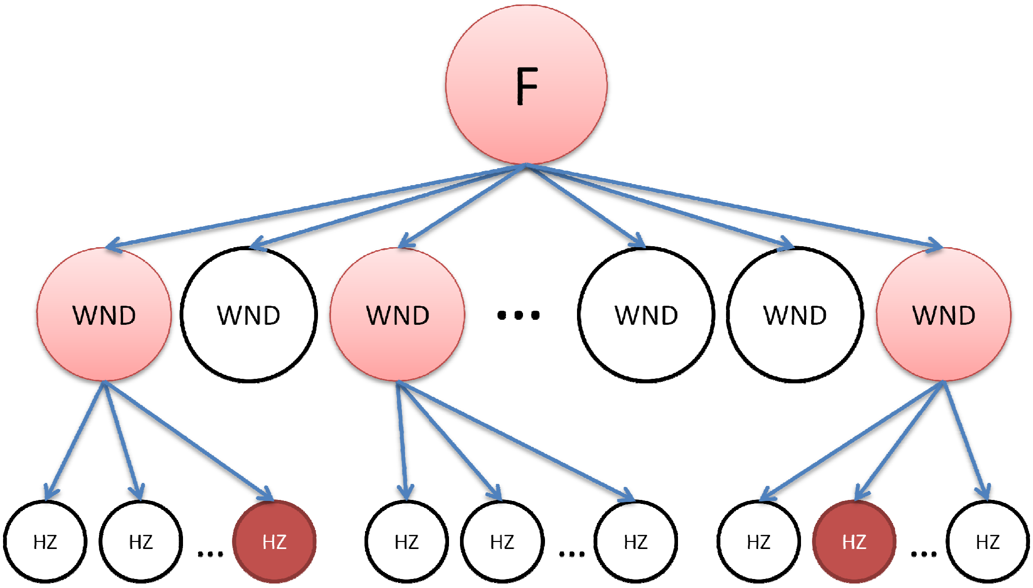

Figure 7: Execution tree of ESS algorithm

Hence, the number of nodes and the spent time could be summarized as in Table 5:

Table 5: ESS times in the worst case scenario

	Nodes	Time
Parents	1 + F/WND	1 * *Split*(F) + F/WND * *Split*(F/WND)
Children	F/WND + F/HZ	F/WND * *Zip*(WND) + F/HZ * *Zip*(HZ)

Note that the times above can be simplified by considering that both *Split* and *Zip* functions are linear to their given input length, hence:

- Parents time = *Split*(F) + *Split*(F) = 2**Split*(F) which is *O*(F)
- Children time = *Zip*(F) + *Zip*(F) = 2**Zip*(F) which is *O*(F)

Thus,

- Time Complexity = *O*(F) + *O*(F) which is *O*(F)

So, this result indicates that the ESS algorithm time is limited by, or grows asymptotically no faster than, a linear function according to the EEG samples length presented.

4. Results

The results described in Table 6 represent the experiment of all samples of our dataset using the third algorithm analysis explained on the previous section. The table contains information about exams duration, processing time, classification and the number of dangerous instants identified.

Due to the high number of outliers checked in some exams, the reader can observe at the last column that their timestamps were joined into groups of intervals with at maximum 10 seconds each, and also an auxiliary tag were used informing how many outliers does the interval contains. For instance, the "00:22:34 (3)" expresses that 3 critical patterns were found along the time span of [00:22:34, 00:22:44), but the first occurred at 00:22:34 moment.

Table 6: Results of EEG analysis using ESS algorithm

Exam	Length	Analysis Time	Patterns	Outliers	Timestamps
1.edf	00:04:05	00:00:03	Absence	4	00:00:34 (2) 00:01:00 (1) 00:03:53 (1)
2.edf	00:50:35	00:00:30	Absence	33	00:05:58 (1) 00:07:39 (1) 00:08:23 (1) 00:09:55 (2) 00:11:31 (1) 00:11:47 (1) 00:11:55 (1) 00:13:12 (1) 00:18:07 (1) 00:20:14 (1) 00:23:16 (2)

					00:23:26 (1) 00:24:38 (1) 00:26:47 (2) 00:28:33 (1) 00:29:31 (1) 00:34:43 (1) 00:35:01 (2) 00:36:47 (1) 00:37:29 (1) 00:38:48 (1) 00:42:02 (1) 00:45:00 (2) 00:47:41 (1) 00:48:44 (1) 00:49:22 (1) 00:50:15 (2)
3.edf	00:31:18	00:00:18	Absence	7	00:00:19 (3) 00:07:53 (2) 00:09:19 (2)
4.csv	00:05:00	00:00:02	Absence	1	00:00:57 (1)
5.csv	00:05:00	00:00:02	Absence	1	00:00:40 (1)
6.edf	01:03:10	00:00:37	Rolandic	5	00:45:51 (1) 00:54:59 (1) 00:55:38 (1) 00:56:59 (1) 01:02:37 (1)
7.edf	00:36:55	00:00:27	Rolandic	31	00:14:29 (1) 00:14:56 (2) 00:16:31 (2) 00:16:59 (1) 00:19:32 (1) 00:21:36 (1) 00:22:01 (1) 00:22:17 (1) 00:22:47 (2) 00:23:18 (1) 00:24:18 (2) 00:25:20 (2) 00:25:44 (1) 00:25:58 (1) 00:26:13 (1) 00:26:23 (1) 00:26:35 (2) 00:26:58 (3) 00:27:38 (3) 00:28:25 (1)

					00:29:01 (1)
8.edf	00:47:54	00:00:30	Rolandic	10	00:08:20 (2) 00:10:38 (1) 00:11:43 (1) 00:13:37 (1) 00:16:05 (1) 00:18:05 (1) 00:25:26 (1) 00:32:39 (1) 00:38:20 (1)
9.edf	00:40:58	00:00:24	Rolandic	1	00:17:27 (1)
10.edf	00:50:26	00:00:35	Rolandic	4	00:33:00 (1) 00:35:38 (1) 00:39:44 (1) 00:40:54 (1)
11.edf	00:59:02	00:00:36	Rolandic	2	00:47:42 (1) 00:50:15 (1)
12.edf	00:43:47	00:00:28	Rolandic	8	00:24:11 (1) 00:27:52 (1) 00:29:02 (1) 00:29:18 (1) 00:30:29 (1) 00:34:13 (1) 00:35:33 (2)
13.edf	02:06:43	00:01:24	Rolandic	3	00:17:50 (1) 00:40:23 (1) 01:48:42 (1)
14.edf	01:01:44	00:00:40	Rolandic	1	00:40:24 (1)
15.edf	00:15:54	00:00:10	Normal	1	00:08:59 (1)
16.edf	00:24:53	00:00:15	Normal	0	
17.edf	00:39:04	00:00:26	Normal	0	
18.edf	01:08:02	00:00:47	Normal	23	00:00:33 (2) 00:00:49 (2) 00:01:24 (2) 00:30:43 (2) 00:31:12 (1) 00:32:48 (1) 00:43:35 (1) 00:46:58 (1) 00:47:24 (1) 00:47:46 (3) 00:48:01 (1) 00:49:12 (2) 00:52:42 (2)

					00:53:25 (1) 00:58:14 (1)
19.edf	00:33:13	00:00:15	Normal	0	
20.edf	00:56:14	00:00:36	Normal	1	00:40:58 (1)
21.edf	00:51:26	00:00:35	Normal	0	
22.edf	01:18:36	00:00:52	Normal	0	
23.edf	00:21:50	00:00:16	Normal	2	00:05:44 (1) 00:14:16 (1)
24.edf	00:40:42	00:00:25	Normal	0	
25.edf	00:27:09	00:00:17	Normal	0	
26.edf	00:20:32	00:00:13	Normal	0	
27.edf	00:26:52	00:00:18	Normal	0	
28.edf	01:01:10	00:00:39	Normal	0	
29.edf	00:31:21	00:00:20	Normal	0	
30.edf	00:37:09	00:00:24	Normal	0	
31.csv	00:05:00	00:00:01	Normal	0	
32.csv	00:05:00	00:00:01	Normal	0	
33.csv	00:05:00	00:00:02	Normal	0	
34.csv	00:05:00	00:00:01	Normal	0	
35.csv	00:05:00	00:00:01	Normal	0	
36.csv	00:05:00	00:00:01	Normal	0	
37.csv	00:05:00	00:00:01	Normal	0	
38.csv	00:05:00	00:00:01	Normal	0	
39.csv	00:05:00	00:00:01	Normal	0	
40.csv	00:05:00	00:00:01	Normal	0	
N=40	**21:50:44**	**00:13:35**	**14 epileptic**	**138**	**18 checked as with seizures**

The result table figures out that a total of 138 outlier periods distributed along 18 exam files were suggested as contained seizures crisis. All 14 samples which were previously classified by doctors as containing epileptic patterns were right classified, and only 4 misclassifications occurred from the normal exams set. The Table 6 also summarizes the total amount of examination and processing time taken, 21:50:44 and 00:13:25, respectively.

5. Discussion

Chapter 5 discusses the results gathered arguing the method trade-offs and also check its achieved performance.

Is very important to recall that we did not have access to the information about when the seizure epochs exactly start and end over the epileptic exams. Due to this lack, hence even the system right classifying the samples, it is possible that some of them could be wrongly checked due to artifacts presence. Then, the numerical performance discussed along this chapter will be driven by the exams' classification, not by the seizures characterization

5.1 Why have seizures been found?

The results in Table 6 show that the ESS algorithm has obtained great responses in identifying exams with presence of seizures. From the set classified as containing these patterns, 100% (N=14) of samples were correctly marked and their seizures timestamps highlighted.

To illustrate this behavior, we glimpse at the 10 seconds snapshot containing the first two outlier points from the EEG sample "1.edf", respectively detected at 00:00:34 and 00:00:36, and print the compression rate values for the whole second points.

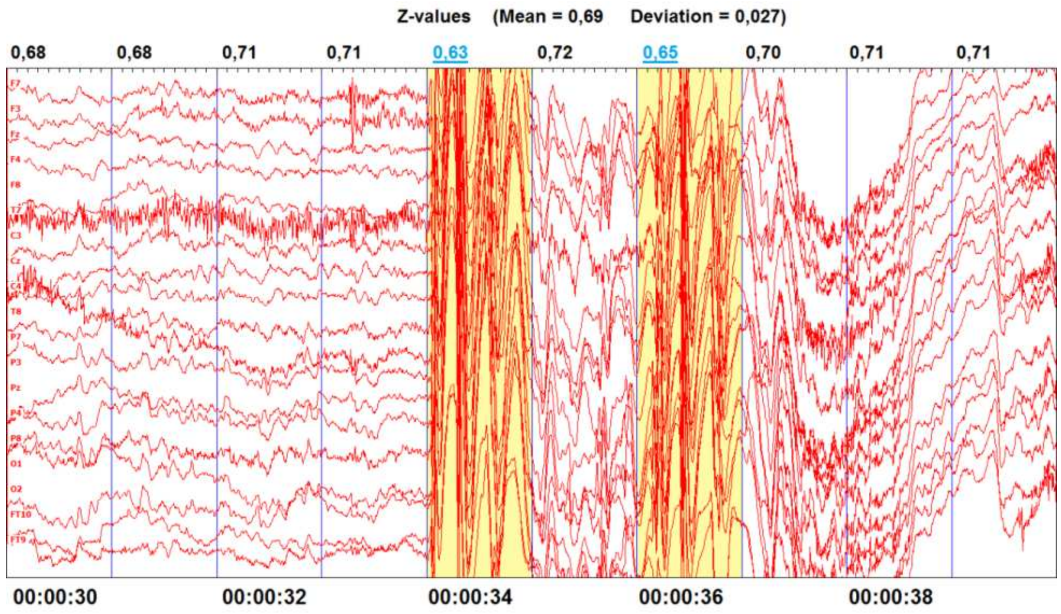

Figure 8: Seizures detected in a 10 second snapshot of 1.edf sample

As shown in the Figure 8, the two seizures points identified had low **Z**-values scores when compared with their neighbors' seconds. This outcome was not a surprise, since that the ESS algorithm was designed to take advantage from the expected gap between the seizures and non-seizures data complexities, respectively. There is no doubt that this existing feature could be only noticed by the algorithm due to the promoted divide and conquer strategy, which splits and computes the Kolmogorov complexities along smaller period segments of EEG data. In fact, we could see how crucial was the evolution of our methodology steps, since the second promoted strategy in section 3.3.2 had been surpassed by the ESS algorithm.

5.2 False positive and negative cases

In medical testing, the false positive and false negative studies are very significant issues. In our scenario we can define a false positive result when the EEG analysis of a healthy exam incorrectly reports that it contains epileptic wave-forms in its tracings. On the other hand, a false negative case occurs when there are no dangerous signals identified on a real epileptic exam. Thus, according to these definitions, the results shown in Table 6 confirms that the analysis method achieved a false negative rate of 0% (N=0), i.e. none epileptic files were classified as containing normal patterns, and also, as mentioned earlier, just 10% of files (N=4) were false positive cases.

Taking the assumption that patients usually submit an EEG test when their doctors want to confirm some brain malfunctioning suspicious, then the false negative rate equals 0% is much more relevant than the false positives cases checked. In addition, as the proposed method alerts the exact timestamps when seizures were considered, the health professional could easily check and filter those “false alarms”.

Even that, we conducted a visual inspection on each bad outcome to understand which conditions lead the algorithm to signal erroneously the normal patterns as outliers. Doing this revision was revealed that those false positive cases pointed can really be considered as outlier periods, given that some of them present hard noisily or artifacts data, as exemplified on Figures 9-10, and others present very disturbed values on almost all channels, as shown on Figures 11-13. These cases may occur due to interferences along the recording procedure or maybe could represent unnoticed seizures. Nevertheless, both cases might be proper evaluated as they may contain relevant data patterns.

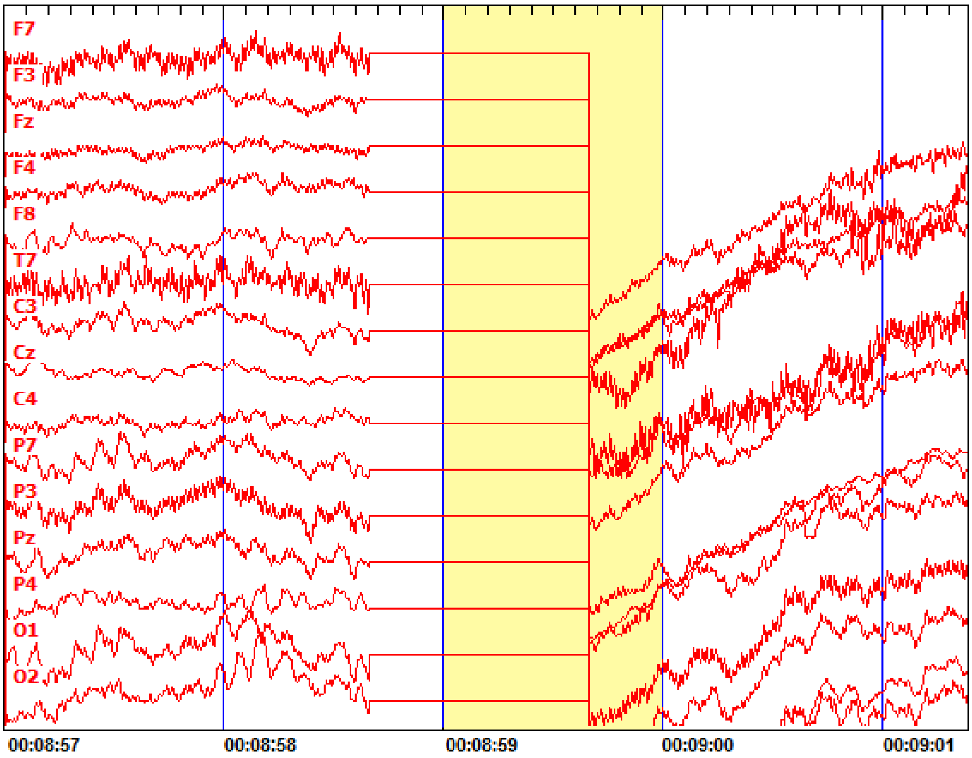

Figure 9: False positive outlier of sample 15.edf

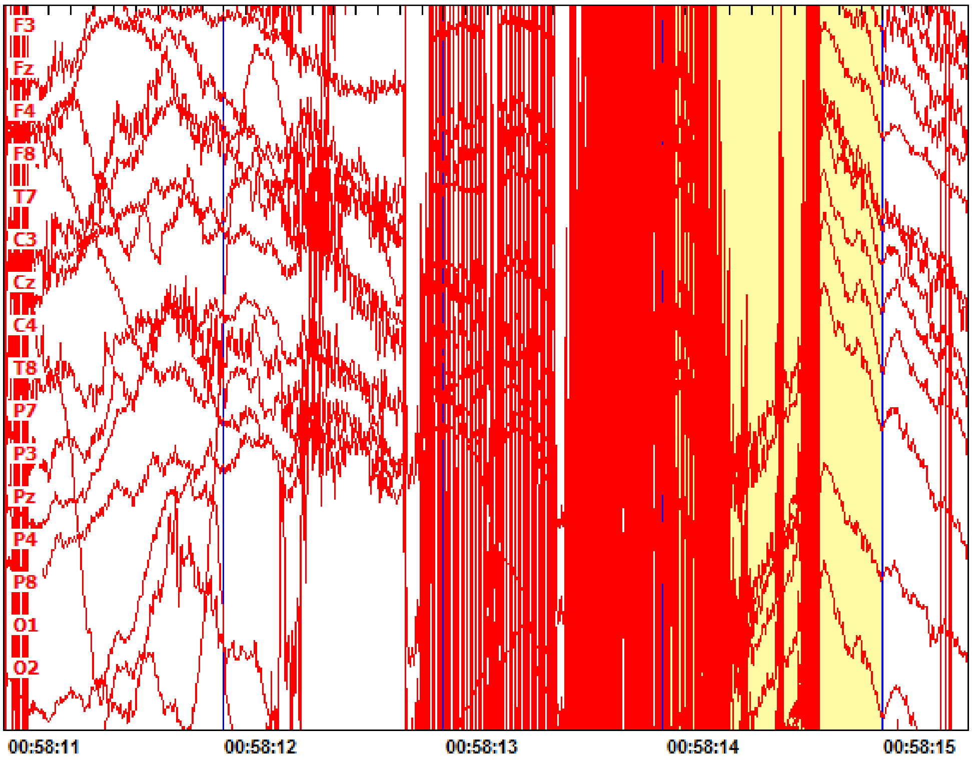

Figure 10: False positive outlier of sample 18.edf

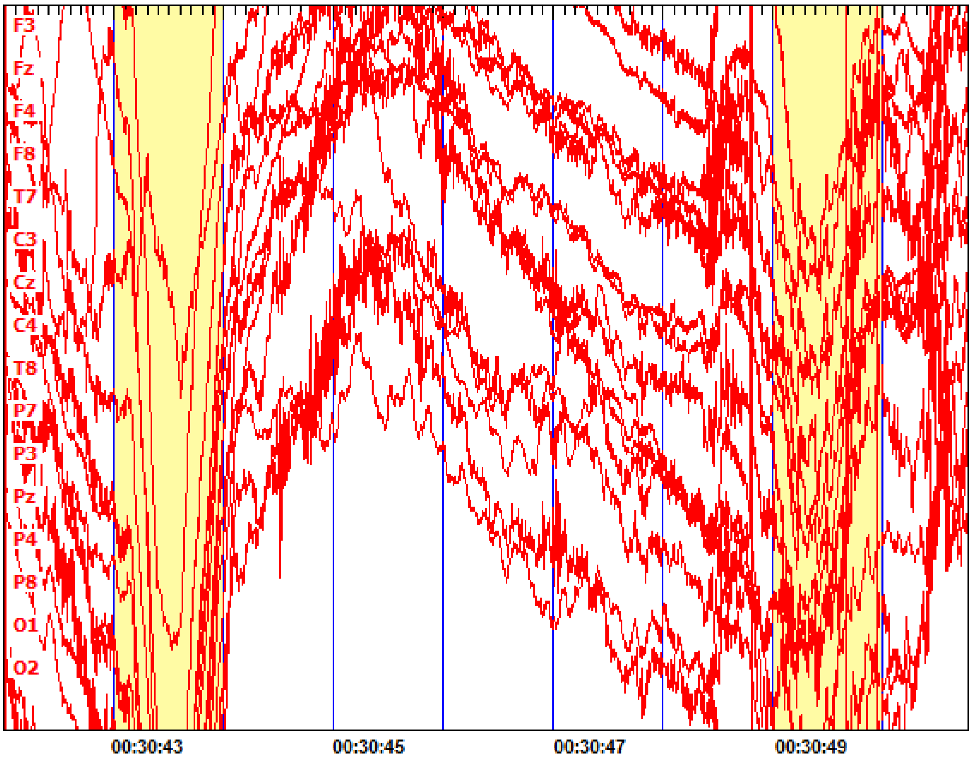

Figure 11: Dubious false positive outlier of sample 18.edf

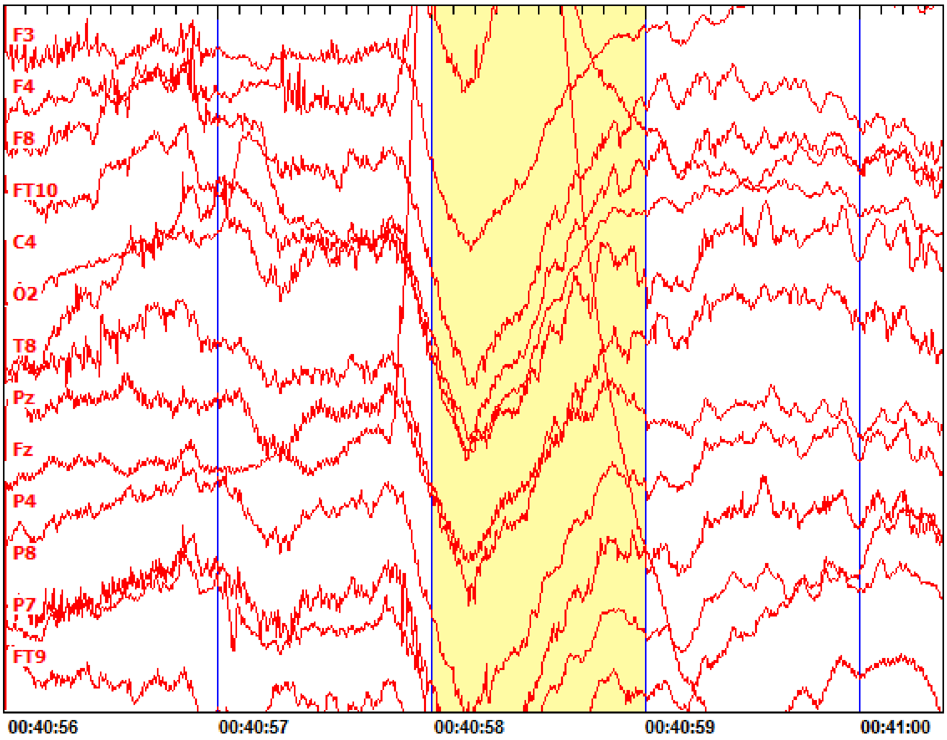

Figure 12: Dubious false positive outlier of sample 20.edf

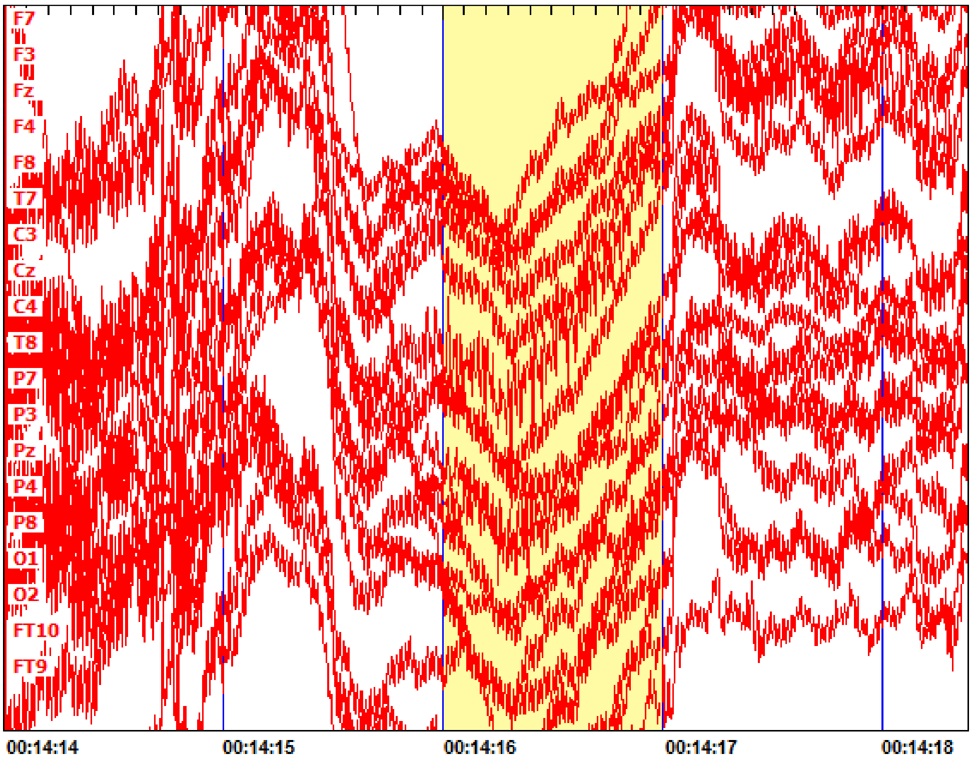

Figure 13: Dubious false positive outlier of sample 23.edf

5.3 The speedup of EEG analysis

Another very interesting measure provided by the outcomes is about the time spent on each exam analysis. A brief comparison between the second and third columns from Table 6 reveals that the ESS algorithm could perform very quick evaluations on all samples contained in our study dataset. The summary line of this same table shows that almost 22 hours of EEG examination have been analyzed in just 13:35 minutes. So, if we take in account a tangible assumption that usually an EEG expert pays attention on the wave lines chart along the whole examination procedure, then that summarized times represent an economy rate of almost 99% of analysis time spent.

At a first glance, this large saving could be a remarkable number, however this good performance could be algorithmically explained by two arguments. First, as the method only investigates the windowed segments which are considered outliers, then it does not spend time searching seizures on those periods which have low complexity, and probably also do not contain abnormalities on their wave patterns. Hence, this designed strategy may provide a large number of jumps over the unnecessary data periods. Finally, as a second argument, we already showed theoretically that the ESS algorithm, in the worst case, runs in linear time according to the input EEG data length, as discussed on the section 3.5. To confirm this last argument, we plot the practical dispersion chart of the processing time spent in comparison with the samples periods length, as shown in Figure 14.

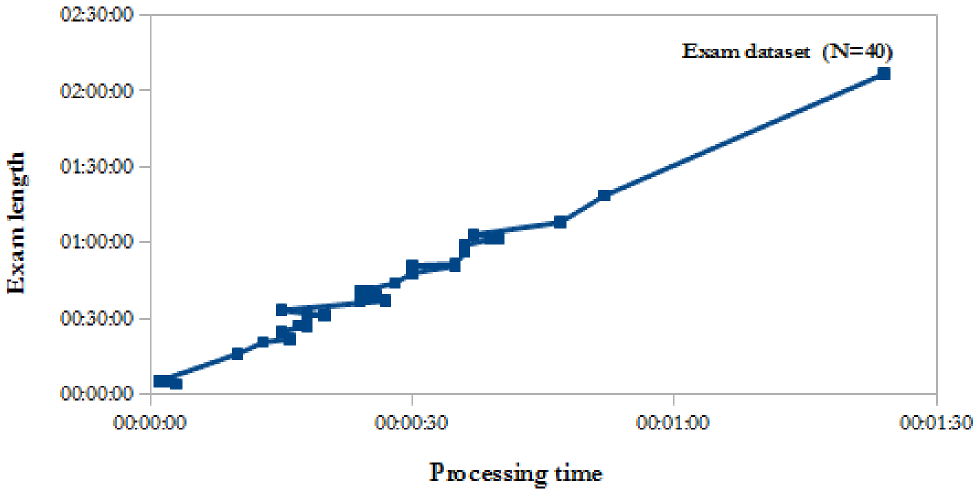

Figure 14: Dispersion plot of ESS time versus exams length

The depicted line above confirms the prevision of our time complexity analysis study, suggesting, in a practical way, that the ESS algorithm is quick enough to be a functional and real-world tool to assists the doctors and health professionals on the EEG analysis.

6. Conclusions

In this chapter are summarized the contributions, limitations as well as some recommendation not covered by this thesis but where future research can be conducted.

6.1 The contributions

As reviewed by this study, the automatic seizures detection over the electroencephalogram (EEG) tests still remains as a desirable tool promise, without an effective algorithm available as a real-world application to decide, in a precise manner, the presence of these dangerous patterns along the exam tracings.

Due to this practical lack, this thesis builds a functional method for the electroencephalogram analysis for childhood absence and rolandic seizures with its implementation published as a free web service. Also, as a side effect of this study, the concept of Kolmogorov complexity, from the algorithmic information theory, is visited and its trade-offs discussed through the usage of a Unix-like compressor, *bzip2*, as the approximation framework to quantify the minimal description of the EEG time-series values.

While this complexity notion has been successfully applied in many scientific experiments aiming patterns recognition, however we could identify a pitfall in its direct applicability over the EEG dataset, once that it became clear how complex this medical data domain is, as containing a numerous set of recording channels series as well as high variance, sometimes due to noise presence, on their values. From this finding, we conducted the development of a divide-and-conquer algorithm by splitting the exams periods into smaller intervals. This change just boosted the pattern recognition capability of seizure crisis detection using the data complexity values.

We propose the ESS algorithm method which achieved results that could classify the whole set of exams (N=14) with seizures patterns present. Also, the few false positive cases (N=4) were discussed through a viewpoint that showed how hard have been the interpretations of these outlier periods. Nevertheless, the proposed method proved that it could produce an economy rate of almost 99% in the time spent on the examinations analysis. Thus, the results presented by this thesis do represent successful cases of EEG analysis made by a really fast computer algorithm.

6.2 Limitations

Is very important to mention that the automatic EEG analysis proposed by this thesis does not mean or intend eliminate the fundamental relevance of the health professional evaluation about the exams results. In the contrary; the development of this computer tool sought the reinforcement of doctors and patients empowerment as a main goal. Thus, this is a synergy between technology and medicine.

Along this study several preliminary assumptions have been considered regarding on the EEG features related. As the base principle, was taken in account that the data complexity of channels series tends to be higher along seizures epochs when compared against the normal activity. At a first glance, this seems to be a controversial assumption according to the nonlinear systems studies reviewed by (Stam 2005), which reveals the 'loss of complexity' phenomenon along epileptic crisis. However, this loss refers to the reduction of the dimension of a dynamic system attractor which represents the extensiveness of the geometric subspace of the total state space to where these systems converge. Thus, this kind of nonlinear analysis deals with 'states' in 'state space' and not with amplitudes or frequencies of values, as we does.

Besides to this theoretical distinction, some other clarifications should be mentioned. From the background section, we took in account that the absence and rolandic epilepsy seizures are commonly associated with short period duration to develop the algorithm strategy. Also, in the practical methodology was considered that the compressor used, *bzip2*, always returns a smaller version of the original EEG sample file. Finally, on the results analysis, we made three considerations: first, that due to the lack of seizures annotations were considered only the exams prior classification made by an EEG expert for the analysis of true positives, false positives and false negatives cases, also was taken into consideration that usually the patient does an electroencephalogram test when his/her doctor recommends this kind of examination, and finally, the EEG expert takes almost the whole procedure time to inspect the waves chart.

6.3 What's next?

From the several topics involving the scope presented by this thesis, we suggest three main points as promising next steps to be expanded in a future research.

First, the ESS algorithm seems to be very adequate for the inspection of dangerous periods along electroencephalogram tests. Although it has resulted positive responses, seizures epochs should be annotated firstly then compared to check the real performance of ESS algorithm, since that almost all false positive cases signaled artifacts presence, which can be avoided by investing in a better pre or post filtering procedure, however probably have produced biased results on our discussion analysis.

Also, as it was designed to check the complexity differences between segmented time-series data, its implementation has a great potential to be successful applied on different data domains. Still on the medicine field, an arising promise with diagnosis purposes can be the application on the electrocardiograms (ECG) tracings analysis, by tuning the ESS algorithm parameters.

Finally, the extension of the test dataset could promote the effectiveness of our algorithm in a large scale. Some suggestions in favor to this increment can be through agreements between hospitals and universities as well as through the dissemination of the available online version at http://www.eegonline.org website for neurologists and EEG experts.

7. References

Aftanas, L., V. Koshkarov, et al. (1994). "Dimensional analysis of human EEG during experimental affective experience." International Journal of Psychophysiology **18**(1): 67-70.

Aftanas, L., N. Lotovaa, et al. (1997). "Non-linear analysis of emotion EEG: calculation of Kolmogorov entropy and the principal Lyapunov exponent." Neuroscience Letters **226**(1): 13-16.

Akin, M. (2002). "Comparison of Wavelet Transform and FFT Methods in the Analysis of EEG Signals." Journal of Medical Systems **26**(3): 241-247.

Altenburg, J., R. Vermeulen, et al. (2003). "Seizure detection in the neonatal EEG with synchronization likelihood." Clinical Neurophysiology **114**(1): 50-55.

Benbadis, S. and S. Berkovic (2006). Absence Seizure. Treatment of Epilepsy: Principles & Practice. E. Wyllie, Lippincott Williams & Wilkins**:** 305-315.

Besthorn, C., H. Sattel, et al. (1995). "Parameters of EEG dimensional complexity in Alzheimer's disease." Electroencephalography and clinical Neurophysiology **95**(2): 84-89.

Bruhn, J., H. Röpcke, et al. (1993). "Electroencephalogram approximate entropy correctly classifies the occurrence of burst suppression pattern as increasing anesthetic drug effect." Anesthesiology **93**(4): 981-985.

Celka, P. and P. Colditz (2002). "Time-varying statistical dimension analysis with application to newborn scalp EEG seizure signals." Medical Engineering & Physics **24**(1): 1-8.

Chen, X., B. Francia, et al. (2004). "Shared information and program plagiarism detection." IEEE Transactions on Information Theory **50**(7): 1545-1551.

Cilibrasi, R. and P. Vitányi (2005). "Clustering by Compression." IEEE Transactions on Information Theory **51**(4): 1523-1545.

Cilibrasi, R., P. Vitányi, et al. (2004). "Algorithmic clustering of music based on string compression." Computer Music Journal **28**(4): 49-67.

Dhamala, M., G. Pagnoni, et al. (2002). "Measurements of brain activity complexity for varying mental loads." Physical Review E - Statistical, Nonlinear and Soft Matter Physics **65**(4).

Dünki, R. (1991). "The estimation of the Kolmogorov entropy from a time series and its limitations when performed on EEG." Bulletin of mathematical biology **55**(5): 665-678.

Epilepsia (1989). "Proposal for revised classification of epilepsies and epileptic syndromes, Commission on Classification and Terminology of the International League Against Epilepsy." Epilepsia **30**(4): 389-399.

Epilepsy-Action. (2010). "Benign rolandic epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.org.uk/info/syndromes/benign-rolandic-epilepsy.

Epilepsy.com. (2007). "Absence Seizures." Retrieved January 24th, 2011, from http://www.epilepsy.com/EPILEPSY/seizure_absence.

Epilepsy.com. (2007). "Benign Rolandic Epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.com/epilepsy/epilepsy_benignrolandic.

Epilepsy.com. (2007). "Types of Epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.com/EPILEPSY/TYPES_EPILEPSY.

Epilepsy.com. (2008). "Childhood Absence Epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.com/epilepsy/epilepsy_childhoodabsence.

Epilepsy.com. (2008). "Types of Seizures." Retrieved March 1st, 2011, from http://www.epilepsy.com/EPILEPSY/TYPES_SEIZURES.

Epilepsy.com. (2009). "Juvenile Absence Epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.com/epilepsy/epilepsy_juvenileabsence.

Epilepsy.com. (2011). "Juvenile Myoclonic Epilepsy." Retrieved March 1st, 2011, from http://www.epilepsy.com/epilepsy/newsletter/May11_OnMyMind.

Fell, J., J. Röschke, et al. (1993). "Deterministic chaos and the first positive Lyapunov exponent: a nonlinear analysis of the human electroencephalogram during sleep." Biological Cybernetics **69**(2): 139-146.

Gandhi, T., B. Panigrahi, et al. (2010). "Expert model for detection of epileptic activity in EEG signature." Expert Systems with Applications: An International Journal **37**(4): 3513-3520.

Gibbs, F., H. Davis, et al. (1935). "The electroencephalogram in epilepsy and in conditions of impaired consciousness." Arch Neurol Psychiatry **34**(6): 1133-1148.

Hernandez-Frau, P. and S. Benbadis (2011). "Pearls & Oy-sters: Errors in EEG interpretations : What is misinterpreted besides temporal sharp transients?" Neurology **76**(12): 57-59.

Hjorth, B. (1970). "EEG analysis based on time domain properties." Electroencephalography and Clinical Neurophysiology **29**(3): 306-310

Jones, S. (1999). "Book review." Clinical Neurophysiology **110**: 1669.

Kalayci, T. and O. Ozdamar (1995). "Wavelet preprocessing for automated neural network detection of EEG spikes." IEEE Engineering in Medicine and Biology Magazine **14**(2): 160-166.

Kandel, E., J. Schwartz, et al. (2000). Principles of Neural Science. New York, McGraw-Hill.

Kantz, H. (1994). "A robust method to estimate the maximal Lyapunov exponent of a time series." Physics Letters A **185**(1): 77-87.

Kim, J., S. Yu, et al. (2009). "A robust peak detection method for RNA structure inference by high-throughput contact mapping." Bioinformatics **25**(9): 1137-1144.

Kolawole, M. (2003). Radar Systems, Peak Detection and Tracking, Newnes.

Kriz, M. and M. Gazdik (1978). "Epilepsy with centrotemporal (Rolandic) spikes. A peculiar seizure disorder of childhood." Neurologia i neurochirurgia polska **12**(4): 413-419.

Lehnertz, K., F. Mormann, et al. (2003). Seizure prediction by nonlinear EEG analysis. IEEE Engineering in Medicine and Biology Magazine. **22:** 57-63.

Lerman, P. and S. Kivity (1975). "Benign focal epilepsy of childhood-a follow up study of 100 recovered patients." Archive of Neurology **32**(4): 261-264.

Lerner, D. (1996). " Monitoring changing dynamics with correlation integrals: case study of an epileptic seizure." Physica D: Nonlinear Phenomena **97**(4): 563-576.

Li, M., J. Badger, et al. (2001). "An information-based sequence distance and its application to whole mitochondrial genome phylogeny." Bioinformatics **17**(2): 149-154.

Li, M., X. Chen, et al. (2004). "The similarity metric." IEEE Transactions on Information Theory **50**(12): 3250-3264.

MathWorks. (2011). "Signal Processing Toolbox." Retrieved February 2nd, 2011, from http://www.mathworks.com/products/signal/.

MedlinePlus. "Medical Encyclopedia - EEG." Retrieved July 5th, 2010, from http://www.nlm.nih.gov/medlineplus/ency/article/003931.htm.

Misra, U. and J. Kalita (2005). Clinical Electroencephalography. Elsevier, Reed Elsevier India Private Ltd**:** 130.

Ocak, H. (2009). "Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy." Expert Systems with Applications: An International Journal **36**(2): 2027–2036.

Pardey, J., S. Robertst, et al. (1996). "A Review of Parametric Modelling Techniques for EEG Analysis." Medical Engineering & Physics **18**(1): 2-11.

Penttilä, M., J. Partanen, et al. (1985). "Quantitative analysis of occipital EEG in different stages of Alzheimer's disease." Electroencephalography and Clinical Neurophysiology **60**(1): 1-6.

Repovš, G. (2010). "Dealing with Noise in EEG Recording and Data Analysis " Informatica Medica Slovenica **15**(1): 18-25.

Rosso, O., M. Martín, et al. (2006). "EEG analysis using wavelet-based information tools." Journal of Neuroscience Methods **153**(2): 163-182.

Sanei, S. and J. Chambers (2007). EEG signal processing, Wiley-Interscience.

Santos, C. C., J. Bernardes, et al. (2006). Clustering Fetal Heart Rate Tracings by Compression. 19th IEEE International Symposium on Computer-Based Medical Systems**:** 685-690.

Smit, L., R. Vermeulen, et al. (2004). "Neonatal seizure monitoring using non-linear EEG analysis." Neuropediatrics **35**(6): 329-335.

Stam, C. (2005). "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field." Clinical Neurophysiology **116**(10): 2266–2301.

Swartz, B. and E. Goldensohn (1998). "Timeline of the history of EEG and associated fields." Electroencephalography and clinical Neurophysiology **106**(2): 173-176.

Temkin, O. (1971). The Falling Sickness, The Johns Hopkins University Press.

Wehner, S. (2007). "Analysing Worms and Network Traffic Using Compression." Journal of Computer Security **15**(3): 12.

Yaylali, I., H. Kocak, et al. (1996). "Detection of seizures from small samples using nonlinear dynamic system theory." IEEE Transactions on Biomedical Engineering **43**(7): 743-751.

Yuan, Q., W. Zhou, et al. (2011). "Epileptic EEG classification based on extreme learning machine and nonlinear features." Epilepsy Research **96**(1-2): 29-38.

8. Annexes

8.1 Annex 1

```
/*
*******************************************************************
* edf2ascii converter
* Author: Teunis van Beelen <teuniz@gmail.com>
* Original source code at: http://www.teuniz.net/edf2ascii/
*******************************************************************
* Modified by Diogo Sato <mululo@gmail.com>
* Last change on 10/feb/2011
* 1. disabled outputs files: annotations, header and signals
* 2. 19 signals captured: F3, F4, F7, F8, Fz, FT9, FT10, T7, T8,
*                         P3, P4, P7, P8, Pz, C3, C4, Cz, O1, O2
* 3. all time-series values were truncated with no decimal places.
* 4. replaced ',' to '\t' in output file values
* 5. columns label were stripped out
********************************************************************/

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <locale.h>

struct edfparamblock{
         int smp_per_record,smp_written,dig_min,dig_max;
         int offset, buf_offset;
         double phys_min,phys_max,time_step,sense;
       } *edfparam;

int main(int argc, char *argv[]) {
  FILE *inputfile=NULL, *outputfile=NULL;
  const  char  interested_signals[19*17+1]  =  "F3                   ,F4
,F7                  ,F8                  ,Fz                  ,FT9
,FT10                ,T7                  ,T8                   ,P3
,P4                   ,P7                  ,P8                   ,Pz
,C3                   ,C4                  ,Cz                   ,O1
,O2              ,";
  const char *fileName;

  int i, j, k, p, r, m, n,
      pathlen, fname_len, signals, datarecords, datarecordswritten,
      recordsize, recordfull, edf=0, bdf=0, edfplus=0, bdfplus=0,
      annot_ch[256], nr_annot_chns, skip, max, onset, duration,
```

```
      zero, max_tal_ln, samplesize;

  char signal_label[40], path[512], ascii_path[512],
      *edf_hdr=NULL, *scratchpad=NULL, *cnv_buf=NULL,
      *time_in_txt=NULL, *duration_in_txt=NULL;
  unsigned char *excluded_signals=NULL;

  double data_record_duration, elapsedtime, time_tmp,
         d_tmp, value_tmp=0.0;

  union {
          unsigned int one;
          signed int one_signed;
          unsigned short two[2];
          signed short two_signed[2];
          unsigned char four[4];
        } var;

  setlocale(LC_ALL, "C");
  if(argc!=2) return(1);
  strcpy(path, argv[1]);
  strcpy(ascii_path, argv[1]);
  pathlen = strlen(path);
  if(pathlen<5) {
    printf("Error, filename is too short.\n");
    return(1);
  }
  scratchpad = (char *)malloc(128);
  if(scratchpad==NULL) {
    printf("Malloc error! (scratchpad).\n");
    return(1);
  }
  fname_len = 0;
  for(i=pathlen; i>0; i--) {
       if((path[i-1]=='/')||(path[i-1]=='\\'))  break;
       fname_len++;
  }
  fileName = path + pathlen - fname_len;
  for(i=0; fileName[i]!=0; i++);
  if(i==0) {
    printf("Error, filename is too short.\n");
    return(1);
  }
  i -= 4;
  if((strcmp((const char *)fileName + i, ".edf")) &&
     (strcmp((const char *)fileName + i, ".EDF")) &&
     (strcmp((const char *)fileName + i, ".bdf")) &&
     (strcmp((const char *)fileName + i, ".BDF")))
  {
    printf("Error, wrong filename extension.\n");
    return(1);
  }
  if((!strcmp((const char *)fileName + i, ".edf")) ||
     (!strcmp((const char *)fileName + i, ".EDF"))) {
    edf = 1;
    samplesize = 2;
  }
  else {
    bdf = 1;
```

```
    samplesize = 3;
  }
/***************** check header ******************************/
  inputfile = fopen(path, "rb");
  if(inputfile==NULL) {
    printf("Error, can not open file %s for reading\n", path);
    return(1);
  }
  if(fseek(inputfile, 0xfc, SEEK_SET)) {
    printf("Error, reading file %s\n", path);
    goto _free_resources;
  }
  if(fread(scratchpad, 4, 1, inputfile)!=1) {
    printf("Error, reading file %s\n", path);
    goto _free_resources;
  }
  scratchpad[4] = 0;
  signals = atoi(scratchpad);
  if((signals<1)||(signals>256)) {
    printf("Error, number of signals in header is %i\n", signals);
    goto _free_resources;
  }
  edf_hdr = (char *)malloc((signals + 1) * 256);
  if(edf_hdr==NULL) {
    printf("Malloc error! (edf_hdr)\n");
    goto _free_resources;
  }
  rewind(inputfile);
  if(fread(edf_hdr, (signals + 1) * 256, 1, inputfile)!=1) {
    printf("Error, reading file %s\n", path);
    goto _free_resources;
  }
  for(i=0; i<((signals+1)*256); i++)
    if(edf_hdr[i]==',') edf_hdr[i] = '\'';

  if(edf) {
    if(strncmp(edf_hdr, "0       ", 8)) {
      printf("Error, EDF-header has unknown version\n");
      goto _free_resources;
    }
  } else if(bdf) {
    if(strncmp(edf_hdr + 1, "BIOSEMI", 7) || (edf_hdr[0] != -1)) {
      printf("Error, BDF-header has unknown version\n");
      goto _free_resources;
    }
  }

  strncpy(scratchpad, edf_hdr + 0xec, 8);
  scratchpad[8] = 0;
  datarecords = atoi(scratchpad);
  if(datarecords<1) {
    printf("Error,  number  of  datarecords  in  header  is  %i\n",
datarecords);
    goto _free_resources;
  }
  strncpy(scratchpad, edf_hdr + 0xf4, 8);
  scratchpad[8] = 0;
  data_record_duration = atof(scratchpad);
  nr_annot_chns = 0;
```

```
  if((strncmp(edf_hdr + 0xc0, "EDF+C     ", 10))
    &&(strncmp(edf_hdr + 0xc0, "EDF+D     ", 10))) {
    edfplus = 0;
  } else {
    edfplus = 1;
    for(i=0; i<signals; i++) {
      if(!(strncmp(edf_hdr + 256 + i * 16, "EDF Annotations ", 16))) {
        annot_ch[nr_annot_chns] = i;
        nr_annot_chns++;
        if(nr_annot_chns>255)  break;
      }
    }
    if(!nr_annot_chns) {
      printf("Error, file EDF+ but no annotation signal.\n");
      goto _free_resources;
    }
  }
  if((strncmp(edf_hdr + 0xc0, "BDF+C     ", 10))
   &&(strncmp(edf_hdr + 0xc0, "BDF+D     ", 10))) {
    bdfplus = 0;
  } else {
    bdfplus = 1;
    for(i=0; i<signals; i++) {
      if(!(strncmp(edf_hdr + 256 + i * 16, "BDF Annotations ", 16))) {
        annot_ch[nr_annot_chns] = i;
        nr_annot_chns++;
        if(nr_annot_chns>255)  break;
      }
    }
    if(!nr_annot_chns) {
      printf("Error, file BDF+ but no annotation signal.\n");
      goto _free_resources;
   }
  }
  edfparam = (struct edfparamblock *)malloc(signals * sizeof(struct
edfparamblock));
  if(edfparam==NULL) {
    printf("Malloc error! (edfparam)\n");
    goto _free_resources;
  }

  recordsize = 0;
  for(i=0; i<signals; i++) {
    strncpy(scratchpad, edf_hdr + 256 + signals * 216 + i * 8, 8);
    scratchpad[8] = 0;
    edfparam[i].smp_per_record = atoi(scratchpad);
    edfparam[i].buf_offset = recordsize;
    recordsize += edfparam[i].smp_per_record;
    strncpy(scratchpad, edf_hdr + 256 + signals * 104 + i * 8, 8);
    scratchpad[8] = 0;
    edfparam[i].phys_min = atof(scratchpad);
    strncpy(scratchpad, edf_hdr + 256 + signals * 112 + i * 8, 8);
    scratchpad[8] = 0;
    edfparam[i].phys_max = atof(scratchpad);
    strncpy(scratchpad, edf_hdr + 256 + signals * 120 + i * 8, 8);
    scratchpad[8] = 0;
    edfparam[i].dig_min = atoi(scratchpad);
    strncpy(scratchpad, edf_hdr + 256 + signals * 128 + i * 8, 8);
    scratchpad[8] = 0;
```

```
    edfparam[i].dig_max = atoi(scratchpad);
    edfparam[i].time_step        =        data_record_duration        /
edfparam[i].smp_per_record;
    edfparam[i].sense = (edfparam[i].phys_max - edfparam[i].phys_min)
/ (edfparam[i].dig_max - edfparam[i].dig_min);
    edfparam[i].offset  =  edfparam[i].phys_max  /  edfparam[i].sense  -
edfparam[i].dig_max;
  }

  cnv_buf = (char *)malloc(recordsize * samplesize);
  if(cnv_buf==NULL) {
    printf("Malloc error! (cnv_buf)\n");
    goto _free_resources;
  }
  free(scratchpad);
  scratchpad = NULL;
  max_tal_ln = 0;
  for(r=0; r<nr_annot_chns; r++) {
    if(max_tal_ln<edfparam[annot_ch[r]].smp_per_record   *   samplesize)
max_tal_ln = edfparam[annot_ch[r]].smp_per_record * samplesize;
  }
  if(max_tal_ln<128)  max_tal_ln = 128;
  scratchpad = (char *)malloc(max_tal_ln + 3);
  if(scratchpad==NULL) {
    printf("Malloc error! (scratchpad)\n");
    goto _free_resources;
  }
  duration_in_txt = (char *)malloc(max_tal_ln + 3);
  if(duration_in_txt==NULL) {
    printf("Malloc error! (duration_in_txt)\n");
    goto _free_resources;
  }
  time_in_txt = (char *)malloc(max_tal_ln + 3);
  if(time_in_txt==NULL) {
    printf("Malloc error! (time_in_txt)\n");
    goto _free_resources;
  }
  excluded_signals = (unsigned char *) malloc( signals+1 );
  if (excluded_signals==NULL) {
    printf("Malloc error! (excluded_signals)\n");
    goto _free_resources;
  }
  for(i=0; i<signals; i++) {
    excluded_signals[i] = 0x00;
    if(edfplus || bdfplus) {
      skip = 0;
      for(j=0; j<nr_annot_chns; j++)
        if(i==annot_ch[j])  skip = 1;
      if(skip) {
        excluded_signals[i] = 0x01;
        continue;
      }
    }
    sprintf(signal_label, "%.16s,", edf_hdr + 256 + i * 16);
    if (strstr(interested_signals,signal_label)==NULL)
      excluded_signals[i] = 0x01;
  }
/**************** write data ****************************/
  ascii_path[pathlen-4] = 0;
```

```
  outputfile = fopen(ascii_path, "wb");
  if(outputfile==NULL) {
    printf("Error, can’t open file %s for writing\n", ascii_path);
    goto _free_resources;
  }
  if(fseek(inputfile, (signals + 1) * 256, SEEK_SET)) {
    printf("Error when reading inputfile\n");
    goto _free_resources;
  }
/*****************           start           data           conversion
******************************/
  datarecordswritten = 0;
  for(i=0; i<datarecords; i++) {
    for(j=0; j<signals; j++)  edfparam[j].smp_written = 0;
    if(fread(cnv_buf, recordsize * samplesize, 1, inputfile)!=1) {
      printf("Error when reading inputfile during conversion\n");
      goto _free_resources;
    }
    do {
      time_tmp = 10000000000.0;
      for(j=0; j<signals; j++) {
        if(edfplus || bdfplus) {
          skip = 0;
          for(p=0; p<nr_annot_chns; p++) {
            if(j==annot_ch[p]) {
              skip = 1;
              break;
            }
          }
          if(skip) continue;
        }
        if (excluded_signals[j]==0x01) continue;
        d_tmp = edfparam[j].smp_written * edfparam[j].time_step;
        if(d_tmp<time_tmp) time_tmp = d_tmp;
      }

      for(j=0; j<signals; j++) {
        if(edfplus || bdfplus) {
          skip = 0;
          for(p=0; p<nr_annot_chns; p++) {
            if(j==annot_ch[p]) {
              skip = 1;
              break;
            }
          }
          if(skip) continue;
        }
        if (excluded_signals[j]==0x01) continue;
        d_tmp = edfparam[j].smp_written * edfparam[j].time_step;
        if((d_tmp<(time_tmp+0.00000000000001))
        &&(d_tmp>(time_tmp-0.00000000000001))
        &&(edfparam[j].smp_written<edfparam[j].smp_per_record)) {
          if(edf)
            value_tmp           =           ((*(((signed           short
*)cnv_buf)+edfparam[j].buf_offset+edfparam[j].smp_written))           +
edfparam[j].offset) * edfparam[j].sense;
          else if (bdf) {
            var.two[0]    =    *((unsigned    short    *)(cnv_buf    +
((edfparam[j].buf_offset + edfparam[j].smp_written) * 3)));
```

```
      var.four[2] = *(cnv_buf + ((edfparam[j].buf_offset +
edfparam[j].smp_written) * 3) + 2);
      if(var.four[2]&0x80) {
       var.four[3] = 0xff;
      }
      else {
       var.four[3] = 0x00;
      }
      value_tmp = (var.one_signed + edfparam[j].offset) *
edfparam[j].sense;
     }
     fprintf(outputfile, "%.0f\t", value_tmp);
     edfparam[j].smp_written++;
    }
    else fputc('\t', outputfile);
   }
   if(fputc('\n', outputfile)==EOF) {
    printf("Error when writing to outputfile during
conversion\n");
    goto _free_resources;
   }
   recordfull = 1;
   for(j=0; j<signals; j++) {
    if(edfparam[j].smp_written<edfparam[j].smp_per_record) {
     if(edfplus || bdfplus) {
      skip = 0;
      for(p=0; p<nr_annot_chns; p++) {
       if(j==annot_ch[p]) {
        skip = 1;
        break;
       }
      }
      if(skip) continue;
     }
     if (excluded_signals[j]==0x01) continue;
       recordfull = 0;
     break;
    }
   }
  }
  while(!recordfull);
  datarecordswritten++;
 }

 _free_resources:
 if (inputfile) fclose(inputfile);
 if (outputfile) fclose(outputfile);
 if (edf_hdr) free(edf_hdr);
 if (edfparam) free(edfparam);
 if (cnv_buf) free(cnv_buf);
 if (time_in_txt) free(time_in_txt);
 if (duration_in_txt) free(duration_in_txt);
 if (scratchpad) free(scratchpad);
 if (excluded_signals) free(excluded_signals);
 return(0);
}
```