UNIVERSITY OF BEIRA INTERIOR

Health Sciences

Assessment of the State of Health by the Measurement of a Set of Biophysiological Signals

Paula Sofia Barros de Sousa

Dissertation to obtain the Degree of Master in

Biomedical Sciences

(2nd cycle of study)

Supervisor: Prof. Dr. Nuno M. Garcia
Co-supervisor: Prof. Dr. Miguel Castelo Branco

Covilhã, October 2010

Acknowledgements

Gratitude is foremost due to my supervisors Prof. Nuno M. Gracia and Prof. Miguel Castelo Branco for their trust, guidance and wisdom. Their ability to find relevant and interesting areas of research is of immense value and importance.

I am truly grateful to my family for their friendship, help and support over the years, especially my mother and father, my sisters and my grandparents.

Last but not least, I am grateful to my friends for being my second family.

Resumo

A dissertação estuda a estimativa do grau de auto-semelhança e da entropia de Shannon de vários sinais reais de electrocardiograma (ECG) obtidos em humanos saudáveis e não saudáveis. O objectivo da dissertação é criar um algoritmo inicial que permita distinguir entre indivíduos saudáveis e não saudáveis e que possa ser usado como base para o estudo de um posterior algoritmo de diagnóstico, necessariamente mais complexo.

Utilizamos um algoritmo novo para estimativa do parâmetro de Hurst baseado no *Embedded Branching Process*, denominado algoritmo *modified Embedded Branching Process*. A entropia foi estimada através da entropia de Shannon. Ambos algoritmos foram aplicados sob a distribuição espacial dos sinais ECG numa forma de janela.

Os sinais estudados foram retirados do website *Physionet*, onde estão diagnosticados como normais ou possuindo uma determinada patologia.

Os resultados apresentados para a estimativa do parâmetro de Hurst permitem confirmar resultados já publicados sobre a auto-semelhança temporal dos sinais ECG, desta vez para a sua distribuição espacial. Também se concluí que os sinais não auto-semelhantes correspondem a indivíduos não saudáveis.

Os resultados obtidos na estimativa da entropia para a distribuição espacial dos sinais de ECG também permitiram uma comparação entre sistemas saudáveis e não saudáveis. Obtiveram-se estimativas de entropia elevadas quer para indivíduos saudáveis quer para indivíduos não saudáveis; no entanto, os indivíduos não saudáveis mostram uma maior variabilidade da entropia de Shannon em relação aos saudáveis.

Palavras-chave

Sinais electrocardiográficos, complexidade biológica, caos, fractal, auto-semelhança, parâmetro de Hurst, *modified Embedded Branching Process*, entropia de Shannon.

Extended Abstract in Portuguese

Introdução

Esta secção apresenta o resumo alargado, em português, da dissertação intitulada "Determinação do Estado de Saúde pela Medição de um Conjunto de Sinais Biofisiológicos" (*Assessment of the State of Health by the Measurement of a Set of Biophysiological Signals*).

Este resumo alargado tem a seguinte organização: em primeiro lugar, são apresentados os objectivos da dissertação e o enquadramento da tese. Seguidamente, é feita uma breve descrição dos algoritmos utilizados assim como referimos os principais resultados. A finalizar são descritas as conclusões deste trabalho e propostas para trabalho futuro. Neste resumo as referências não estão numeradas sequencialmente para manter a coerência da numeração com o texto em inglês.

Descrição do Problema e Objectivos da Dissertação

O problema que define esta pesquisa baseia-se nos seguintes pressupostos:

1. As doenças cardíacas são uma das principais causas de morte e incapacidade da sociedade moderna;
2. Avanços na tecnologia têm melhorado os cuidados de saúde e há uma tendência crescente no desenvolvimento de métodos de diagnóstico;
3. O electrocardiograma (ECG) é o método mais utilizado para observação da actividade eléctrica e do funcionamento do coração;
4. Os dados de ECG são fáceis de obter e de registar;
5. As distribuições espacial e temporal dos sinais ECG variam de indivíduo para indivíduo e entre indivíduos saudáveis e doentes, pelo menos.

Tendo o problema em vista, esta pesquisa aborda as seguintes questões:

1. A importância dos sinais ECG no diagnóstico, uma vez que representam a actividade do coração;
2. O facto de um diagnóstico precoce poder evitar futuras complicações na saúde;
3. A necessidade de estabelecer novos métodos de diagnósticos mais rápidos e simples, de modo a reduzir custos e tempo no diagnóstico, quando comparados aos métodos tradicionais.

O objectivo da pesquisa em que se baseia a dissertação é criar um método baseado na estimativa de alguns parâmetros sobre a distribuição espacial de sinais electrocardiográficos, permitindo que se faça a distinção entre sinais de indivíduos saudáveis e indivíduos que foram diagnosticados com alguma forma de patologia cardíaca.

Enquadramento da Dissertação

O avanço da tecnologia tem trazido melhorias nos cuidados de saúde. Uma área de investigação muito importante nos serviços de saúde é o diagnóstico, com a ajuda do qual é possível detectar precocemente e, assim, tratar doenças, causando menor transtorno para o paciente e poupanças nos sistemas de gestão de saúde. Os investigadores têm desenvolvido novos métodos de diagnóstico, adaptando a tecnologia aos métodos tradicionais.

A pesquisa da dissertação é baseada no facto das doenças cardíacas serem uma das principais causas de morte na sociedade moderna. Para diagnosticar um problema cardíaco, o médico recorre normalmente ao electrocardiograma, porque é fácil de obter e, como representa a actividade eléctrica do coração, permite a detecção pelo médico de um leque alargado de patologias.

Para acompanhar o progresso da tecnologia no diagnóstico de doenças cardíacas, tem-se tornado muito importante criar novos métodos que permitam um diagnóstico mais rápido, preciso e precoce, evitando assim custos materiais e imateriais desnecessários.

Os Algoritmos para a Estimativa do Parâmetro de Hurst e da Entropia

Foram obtidos vários conjuntos de dados de ECG, que correspondem a utilizadores reais aos quais foi diagnosticada uma dada patologia ou atribuída uma classificação de ausência de patologia. Estes sinais estavam armazenados no sítio da Internet *Physionet*, criado e mantido pela Escola Médica de Harvard nos Estados Unidos da América [41]. Os dados de ECG foram tratados através de um algoritmo programado em Java, com o objectivo de estimar os valores de parâmetro de Hurst (*H*) e da entropia da distribuição espacial dos dados. Estes dois parâmetros foram obtidos através do método mEBP e da Entropia de Shannon, respectivamente. A discussão dos resultados baseou-se na comparação de parâmetros obtidos dos traçados dos gráficos e não na análise dos próprios traçados. No entanto, e para facilitar a leitura dos resultados, são apresentados os traçados de alguns indivíduos saudáveis e não saudáveis para ambas as estimativas.

No geral, os indivíduos saudáveis apresentam um valor de *H* muito próximo de 1.0 e variabilidade (desvio padrão) e amplitudes reduzidas, quando comparados com os indivíduos diagnosticados com patologia. Contudo, apenas a estimativa do valor de *H* não garante a correcta classificação dos sinais.

O mEBP é apenas um dos métodos disponíveis para a estimação do parâmetro de Hurst, pelo que, para validar a sua escolha, foram utilizados outros métodos (disponíveis no software Selfis [42]), de modo a verificar a coerência do mEBP com outros métodos. Sabendo da especificidade dos vários estimadores, concluí-se que o estimador mEBP apresenta valores de *H* elevados para os sinais de indivíduos saudáveis, enquanto outros estimadores disponíveis no Selfis não são conclusivos. Mais ainda, a média da variabilidade da estimativa de *H* para indivíduos saudáveis é menor quando usado o mEBP, em comparação com os indivíduos não saudáveis.

Na estimativa da entropia de Shannon foram analisados os mesmos parâmetros estatísticos que na estimativa de *H*. Por o valor médio da entropia não permitir distinguir os indivíduos saudáveis dos não saudáveis, deu-se mais importância ao desvio padrão e amplitude. Para os indivíduos saudáveis, o desvio padrão e amplitude são menores do que os mesmos para os indivíduos não saudáveis, ou seja, a variabilidade é menor.

Neste seguimento, relacionamos os resultados obtidos na estimativa do parâmetro de Hurst com os resultados da estimativa da entropia. Para ser possível estabelecer uma relação, comparamos a variabilidade do parâmetro de Hurst com a variabilidade da entropia, que mostraram ser significantes. A variabilidade do parâmetro de Hurst e da entropia são menores para os indivíduos não saudáveis.

Conclusões e Propostas de Trabalho Futuro

O algoritmo utilizado permitiu estimar o parâmetro de Hurst e a entropia através do método mEBP e da entropia de Shannon, respectivamente. Deste modo, foi possível fazer a distinção entre indivíduos saudáveis e indivíduos aos quais tinha sido diagnosticada uma patologia.

Os resultados mostraram que os sinais para os indivíduos saudáveis obtidos da estimativa do parâmetro de Hurst e da entropia são diferentes da maioria dos sinais de indivíduos doentes. Foram avaliados alguns parâmetros estatísticos como média, desvio padrão e amplitude, que decorrem da análise do traçado dos gráficos obtidos, e concluímos que é possível diferenciar indivíduos saudáveis dos doentes.

Os indivíduos saudáveis apresentam parâmetro de Hurst (*H*) elevado e desvio padrão e amplitude menores que os indivíduos não saudáveis. Um valor de *H* elevado indica elevada

auto-semelhança. Estes resultados permitem confirmar os resultados anteriores apresentados por [13] e [33]. Além disso, os resultados obtidos mostraram que um sinal cuja auto-semelhança é baixa, ou muito variável, pertence a um indivíduo que não é saudável. No entanto, quando usado isoladamente, o parâmetro de Hurst não garante a correcta classificação de saudável ou não saudável.

Os valores médios obtidos para a entropia mostraram-se muito semelhantes entre indivíduos saudáveis e indivíduos não saudáveis. Os valores de desvio padrão e amplitude são muito mais significativos e permitiram uma distinção entre indivíduos saudáveis e não saudáveis, por se apresentarem mais baixos para os indivíduos saudáveis.

Para finalizar, foi possível relacionar a estimativa do parâmetro de Hurst com a estimativa da entropia. Avaliou-se a variabilidade dos valores estimados para os conjuntos de sinais ECG. Os sinais dos indivíduos saudáveis apresentaram baixa variabilidade, indicando que todos os indivíduos saudáveis apresentam coerência entre os valores de parâmetro de Hurst e da entropia de Shannon, quando calculados sobre a sua distribuição espacial.

Abstract

The dissertation studies the estimation of the degree of self-similarity and entropy of Shannon of several real electrocardiography (ECG) signals for healthy and non-healthy humans. The goal of the dissertation is to create a starting point algorithm which allows distinguishing between healthy and non-healthy subjects and can be used as a basis for further study of a diagnosis algorithm, necessarily more complex.

We used a novel Hurst parameter estimation algorithm based on the Embedded Branching Process, termed modified Embedded Branching Process algorithm. The algorithm for estimation of entropy was based on Shannon's entropy. Both algorithms were applied on the spatial distribution of ECG signals in a windowed manner.

The studied signals were retrieved from the Physionet website, where they are diagnosed as normal or as having certain pathologies.

The results presented for the Hurst parameter estimation allow us to confirm the results already published on the temporal self-similarity of ECG signals, this time for its spatial distribution. We also conclude that the non-self similar signals belong to non-healthy subjects.

The results obtained for entropy estimation on the spatial distribution of ECG signals also allowed a comparison between healthy and non-healthy systems. We obtained high entropy estimates both for healthy and non-healthy subjects; nevertheless, non-healthy subjects show higher variability of Shannon's entropy than healthy ones.

Keywords

Electrocardiographic signals, biological complexity, chaos, fractal, self-similarity, Hurst parameter, modified Embedded Branching Process, entropy.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

A	Arrhythmia
ApEn	Approximate Entropy
AV	Atrioventricular
CHF	Congestive Heart Failure
CP	Conditional probability
e	Entropy rate
E	Information-theoretic entropy
EBP	Embedded Branching Process
ECG	Electrocardiogram
E_{KS}	Kolmogorov-Sinai entropy
E_n	Joint entropy
H	Hurst Parameter
IID	Independent Identically Distributed
k_B	Boltzman constant
KS	Kolmogorov-Sinai
mEBP	Modified Embedded Branching Process
MSE	Multiscale Entropy
MVE	Malignant Ventricular Ectopy
NSR	Normal Sinus Rhythm
R/S	Rescaled Statistics
S	Thermodynamic entropy
SA	Sinoatrial
SampEn	Sample Entropy
SD	Standard Deviation
VTA	Ventricular Tachyarrhythmia

xx

Chapter 1. Introduction

In this chapter we describe the dissertation focus and scope, the problem statement and the goals of the research activities.

Here, we mention the advances of technology in health care and its contribution in diagnosis methods improvement; how cardiac problems are today society's leading cause of death and an early diagnosis can prevent severe heart problems. We also refer the goals of our research and its scientific contribution to diagnostic methods

Chapter 1 ends with the description of the organization of the dissertation.

1.1. Dissertation focus and scope

The advance of technology has proved to bring improvements in health services. A very important area of investigation in health services is the diagnosis, focusing on diseases that can be detected as early as possible and treated, with the goal of causing as little disorder as possible to the patient and to maximize the probability of a full recovery, having in view the economic aspects of today's health management systems.

Researchers are developing new methods of diagnosis, adapting technology to traditional methods, and obtaining results that allow making a distinction between young and older subjects and between healthy and unhealthy ones.

Heart diseases are nowadays one of the leading causes of death, in many cases due to unhealthy eating and exercising individual and collective habits.

To diagnose a heart disease or problem, the physician resort usually to the electrocardiogram, because it is easy to obtain and, as it is the recording of the electrical activity of the heart, it allows for the detection of a number of significant diseases, acknowledged by the physician.

This dissertation research is based on the fact that heart diseases are one of the leading causes of death and disability in modern society and, to keep up with the progress of technology in health services, it has become more important to create new methods that allow making a rapid, accurate and early diagnosis, thus avoiding unnecessary costs and waste of time.

Electrocardiographic signals are the way to observe the electrical behavior of the heart, so they are used by the physician when he wants to evaluate whether a patient is healthy or unhealthy, or to determine the cause and origin of the disease or heart problem.

1.2. Problem statement and goals of the research

The statement of the problem that defined this research is based on the following assumptions:

1. Heart disease is one of the leading causes of death and disability in modern society;
2. New advances on technology have improved health services and there is a growing trend in diagnosis methods development;
3. Electrocardiogram (ECG) data is the method used by the physician to observe the electrical activity and proper function of the heart;
4. ECG data is easy to obtain and record;
5. Spatial and temporal distribution of electrocardiographic signals varies from subject to subject and from healthy and non-healthy subjects.

Having this in view, this research addresses the following issues:

1. The importance of electrocardiographic signals in diagnostic, since they represent the electrical activity of one of the most important organs for the proper function of the human body – the heart;
2. The fact that a proper and early diagnosis of heart diseases can avoid further complications in health;
3. The need to establish new faster and easier methods of diagnostic in order to reduce costs and time in the diagnosis for both physicians and patients when compared to traditional methods.

In conclusion, the goal of the research activities was to create a method based on the estimation of some parameters of the spatial distribution of electrocardiographic signals, allowing making the distinction between signals from healthy subjects and non-healthy ones (*i.e.*, subjects who have been diagnosed with some sort of pathology).

1.3. Organization of the dissertation

This dissertation is organized in six chapters. The present chapter (Chapter 1), devoted to the introduction, describes the dissertation focus and scope, the problem statement and the goals of the research activities. It ends with the organization of the dissertation.

Chapter 2 presents the theoretical background. It begins with a brief introduction of Theory of Chaos, and how chaos manifests itself in biological systems It follows the concept of fractality and its inherent properties as self-similarity and fractal dimensionality, followed by the concept of entropy. Fractality is associated with the Hurst parameter since it indicates the degree of self-similarity. There are various Hurst parameter estimators as the modified Embedded Branching Process method which is described in more detail. It is mentioned the meaning of entropy, as a measure of disorder, and how this concept emerged from the Second Law of Thermodynamics to the information-theoretic entropy. Several algorithms to estimate the entropy are enumerated, as Sample Entropy, Approximate Entropy, and Multiscale Entropy.

Chapter 3 presents the biological mechanisms of the heart, enabling the reader to perceive how the electrical signals of the heart are generated and gathered. It is also presented the typical trace for an electrocardiographic (ECG) signal.

In Chapter 4 the methods used are presented. The ECG data used and how and where it was obtained is also discussed. There is also a description of the used algorithms.

Chapter 5 presents the results of the Hurst parameter and entropy estimation for electrocardiographic signals. The discussion of these results follows its presentation.

Chapter 6 presents the main conclusion and some possible future work in continuation of this research.

The reference section presents the detailed information for the references cited in this dissertation. These are numbered sequentially by order of appearance in the English text of the dissertation.

Chapter 2. Background

In this chapter, the theoretical background needed for our research is presented. We present the Theory of Chaos and how chaos manifests in biological systems. Additionally, we relate chaos with homeostasis.

It follows the concept of fractality and its inherent properties. We mention the several Hurst parameter estimators used to estimate self-similarity. We described, in particular, the modified Embedded Branching Process method, to be the method used in our study.

It is mentioned the meaning of entropy as a measure of disorder and how this concept emerged from the Second Law of Thermodynamics to the information-theoretic entropy. Several algorithms to estimate the entropy are also presented.

2.1. The Theory of Chaos

The Theory of Chaos (also known as Chaos Theory) takes its root in the study of nonlinear dynamic systems. Nonlinear dynamic systems have specific properties that mathematicians have studied for more than a century [1]. However, during the last decades, interest in these systems has grown among researchers of different scientific fields, such as physics, chemistry and economics.

In common usage, chaos means a state of disorder; it is the disordered formless matter supposed to have existed before the ordered universe. But the adjective chaotic is defined more precisely in Chaos Theory. The term, originated in the field of nonlinear dynamics, is generally applied in the cases of motions in nonlinear deterministic systems whose time histories have a sensitive dependence on initial conditions [2]. Nonlinear dynamics is concerned with systems whose output is not a linear function of their input [3].

The first studies on the subject were performed by Henri Poincaré, at the end of the 19th Century. He was the pioneer in understanding that chaos is not a synonym of disorder but of complexity pretending to be disorder [4].

The general attributes of chaotic systems are that they appear to have unpredictable behavior; uncertainty grows with time (additionally, there is a loss of information about initial conditions); its behavior is not periodic, even over an arbitrarily long period of time (the state vector of the system never acquires the same value twice); and an underlying fractal structure exists in phase space (which is, for a simple example of pendulum motion, a two-dimensional plot of velocity versus position) [2].

Characteristics of chaotic systems are the following [5]:

   a) The system is dynamical, means that it changes over time;
   b) The behavior of the system is aperiodic and unstable, meaning it does not repeat itself;
   c) Although chaotic behavior is complex, it can have simple causes;
   d) Because the system is nonlinear, it is sensitive to initial conditions;
   e) Because the system is deterministic, chaotic behavior is not random even though its aperiodicity and unpredictability may make it appear to be so. On the other hand, because of the instability, aperiodicity, and sensitivity to initial conditions, the behavior of chaotic systems is not predictable even though it is deterministic; and
   f) Although not included in the above definition, is that of iteration or feedback, in which the output of the system is used as the input in the next calculation.

Systems may display both chaotic and non-chaotic behavior depending on the control parameters used.

Another kind of behavior closely linked to chaos is complexity. A complex system is one in which numerous independent elements continuously interact and spontaneously organize and reorganize themselves into more elaborate structures over time [5].

Complexity is characterized by:

   a) A large number of similar, but independent, elements or agents;
   b) Persistent movement and responses by these elements to other agents;
   c) Adaptiveness, so that the system adjusts to new situations to ensure survival;
   d) Self-organization, in which order is formed spontaneously in the system;
   e) Local rules that apply to each agent; and
   f) Progression in complexity so that over time the system becomes larger and more sophisticated.

As with chaos, the behavior of self-organizing complex systems cannot be predicted and they do not observe the principle of additivity, *i.e.*, their components cannot be divided up and studied in isolation. Complex systems can naturally evolve to a state of self-organized criticality, in which behavior lies at the border between order and disorder. Again, the same system can display order, chaos and self-organizing complexity, depending on the control parameters.

2.1.1. Chaos and Homeostasis

Biological systems combine increasing complexity with self-preservation, acting in a chaotic manner. The process by which living beings operate to reduce variability and to maintain a constancy of internal function is named homeostasis [6].

Physiologic control mechanisms exist from sub-cellular to systemic levels and operate over multiple time scales. Continuous interplay among these different regulatory systems ensures that information is constantly exchanged across all levels of organization and enables an organism to adjust to an ever-changing environment and to perform a variety of activities necessary for survival [7, 8].

One might expect most any physiological variable to return to its normal value after it has been perturbed and to remain steadily at that value until for some reason it is perturbed again. However, the normal heartbeat is not predictably regular.

Table 1. Chaos vs. homeostasis in physiology [6].

Chaos
System does not settle down to constant steady state
Fluctuations arise from internal feedback and do not require external perturbation
Destabilizing factors such as disease or aging usually decrease the degree of complex variability (reduce chaos)

Homeostasis
System will settle down to a steady state (constancy) if perturbed
Fluctuations result from external influences
Destabilizing factors such as disease or aging are anticipated to decrease order (increase chaos)

2.2. Fractality

Inside the Chaos Theory, we find the geometry of things from which scientists intend to explain the strange forms of Nature. In order to do so, the mathematician Benoit Mandelbrot [9] introduced the fractal concept in 1975.

Fractal describes the phenomenon where a certain property of an object – the mathematical object of our interest – is preserved with respect to scaling in space and/or time [10]. The term fractal applies to complex objects, which may be generated by stochastic or nonlinear deterministic mechanisms [11]. The main features of fractal forms are self-similarity, heterogeneity and the absence of a well defined (characteristic) scale of length [3].

Self-similarity means that an object is composed of sub-units on multiple levels that statistically resemble the structure of the whole object [12]. In an idealized model, this property holds on all scales. The real world, however, necessarily imposes upper and lower bounds over which such scale-invariant behavior applies [13].

In this sense, the fractal objects that can not be represented by the Classic or Euclidian Geometry can be represented by iteration or repetition of a basic structure (fractal) [4]. They are not homogeneous. The more closely they are inspected, the more details are revealed [3].

The fractal concept can be applied not just to irregular geometric forms that lack a characteristic (single) scale of length, but also to certain complex processes that lack a single scale of time. Fractal processes generate irregular fluctuations across multiple time scales, analogous to scale-invariant objects that have a branching or wrinkly structure across multiple length scales [13].

2.3. Self-similarity and the Hurst parameter

Self-similarity in data was first observed by Harold E. Hurst, a British scientist, and was first published in 1952, while analyzing fluctuations on the long term capacity of water reservoirs [14].

The Hurst parameter, *H*, also called Hurst exponent, occurs in several areas of applied mathematics, including fractals and spectral analysis, among others. Hurst parameter estimation has been applied in areas ranging from Biophysics to Computer Networking [15]. Although this estimation was originally developed in Hydrology [14], the modern techniques for estimating the Hurst parameter come from fractal mathematics.

The value of the Hurst parameter on a given arbitrarily long data series is an estimation of a probability, as its computation is performed on a limited sub-series of the data [16]. The calculated parameter estimates the degree of self-similarity of a data series, *i.e.*, the degree of persistence of the statistical phenomenon.

The Hurst parameter varies between 0 and 1, and the data is assumed to be self-similar if its corresponding *H* value is bigger than 0.5. Therefore, a value of *H*=0.5 indicates the lack of self-similarity, whereas large values of *H* (close to 1.0) indicate a large degree of self-similarity. The higher the *H* value, the more self-similar is the data. That means that an increasing (or decreasing) trend in the past of a process implies that there is a great probability that an increasing (or decreasing) trend will take place in the future [17].

A continuous-time stochastic process *X(t)* is considered to be statistical self-similar, with parameter *H* (0.5 ≤ *H* ≤ 1), if, for any real, positive *a*, the processes *X(t)* and $a^{-H}X(at)$ have identical finite-dimensional distributions (*i.e.* same statistical properties) for all positive integers *n*:

$$\{X(t_1), X(t_2)...X(t_n)\} \underset{\sim}{D} \{a^{-H}X(at_1), a^{-H}X(at_2)...a^{-H}X(at_n)\} \tag{2.1}$$

The term $\underset{\sim}{D}$ means "asymptotically equal to" in the sense of distribution. Practically, statistical self-similarity means that the following conditions are valid:

- Mean: $E[X(t)] = E[X(at)]/a^H$;
- Variance: $Var[X(t)] = Var[X(at)]/a^H$;
- Autocorrelation function: $R(t,\tau) = R(at,a\tau)/a^{2H}$.

2.4. Hurst parameter estimators

Usually, the degree of self-similarity of a signal is assessed by the estimation of the Hurst parameter.

Various Hurst parameter estimators have been proposed [18]. They can be divided into three basic classes, depending on whether data analysis is done in the time-domain, the frequency-domain or by using wavelets.

(i) Techniques based on data analysis in the time domain;
- R/S-statistic
- Variance-time estimator
- Estimators based on absolute moments
- Fractal dimension estimators
- Estimators based on variance of residuals
- The index of dispersion for counts

(ii) Techniques based on data analysis in the frequency domain;
- Estimators based on the modified periodogram or on the cumulative periodogram
- Daniel PB estimator
- Whittle Maximum Likelihood estimator
- Aggregated Whittle ML estimator
- Local Whittle ML estimator

(iii) Techniques based on wavelet transforms of data.

2.5. The Modified Embedded Branching Process Hurst parameter estimator

The Embedded Branching Process (EBP) was published by Jones and Shen in [19]. As the authors refer, "(...) Self-similar processes appear in telecommunications, finance, medicine and hydrology. (…)". In [20], Jones also acknowledges the existence of a class of self-similar processes, called Embedded Branching Processes, whose crossings of a crossing tree are a simple function of the Hurst parameter. The main characteristic of such processes is that the whole range of values can be fit within a range of crossings, allowing the creation of a crossing tree.

The modified Embedded Branching Process (mEBP) estimator, developed by Hajduczenia *et al*. [21], builds on the definition of the Embedded Branching Process method [20, 22], allowing an estimation on a point by point basis, *i.e.*, suitable to be used in real time operation.

The original EBP estimator calculates the whole crossing tree structure for the given crossing level k at a time, resulting in the retrospective character of the estimation. For the given crossing level k, the crossing points are estimated always for a pair of data points, namely $X_{i-1}(t)$ and $X_i(t)$, then the estimator moves on to another pair of points and so forth. Therefore, at a certain moment of time t_i, it is possible to calculate a full set of crossing tree points for all desired k levels only for $X_i(t)$ and $X_{i-1}(t)$ data points. The whole k level crossing tree structure can be calculated at the given moment of time without the need of processing or storing the whole data trace. Such an approach is equal to vertical rather than horizontal calculation regime and is depicted in Figure 1. Such an approach obviously requires the full set of the examined data, which has to be traversed at various detail levels in search of the possible crossing points.
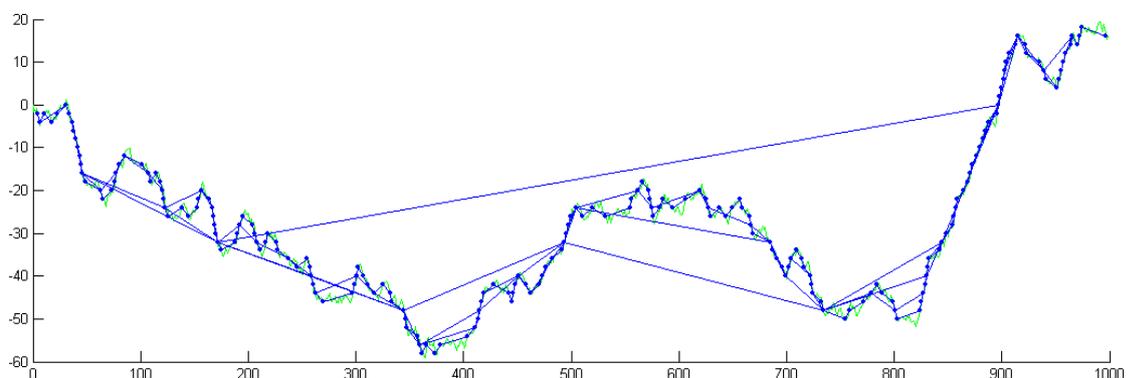
Figure 1. Sample path of a stochastic process and superimposed it three approximations made up of crossings of size 8, 16 and 32 respectively (adapted from [23]).
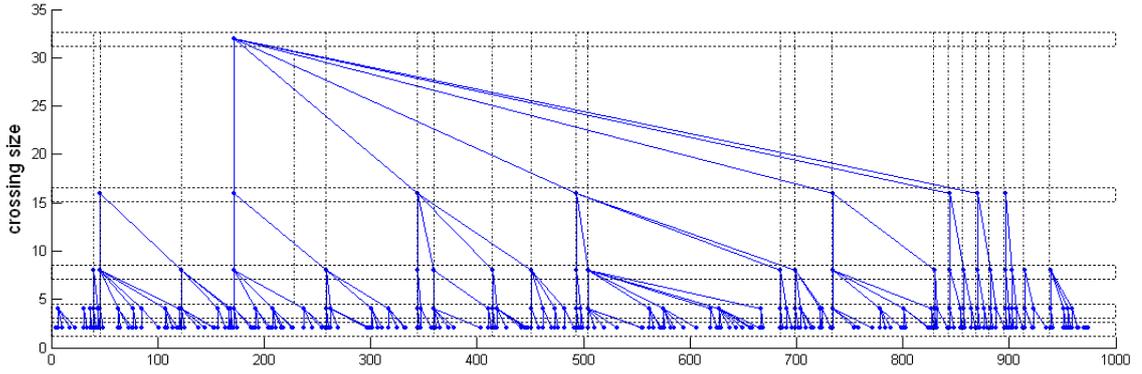
Figure 2. Tree structure of the crossings constructed during operation of the EBP estimator on an arbitrary example of a data trace; a) horizontal calculation regime equal to retrospective Hurst parameter estimation as presented in [22] (dashed line boxes), b) horizontal calculation regime employed in mEBP (dash-dot vertical lines; for clarity, calculation points for levels 3 and above were depicted). The *x*-axis gives the starting point of a crossing and the *y*-axis its size. Crossings are linked if one is a sub-crossing of the other (adapted from [23]).

In order to calculate the value of the Hurst parameter for the given points, the estimator defined by (2.2) in [22] is employed, though its value is estimated for every single point of the examined data series, resulting in an estimation curve for the whole examined data series rather than a single value with the confidence intervals. This way the evolution curve of the Hurst parameter value may be plotted, allowing for more precise estimation of the true value of the Hurst parameter as well as elimination of numerically unstable solutions.

$$H_n = \log(2)/\log(\hat{\mu}_n): \ \hat{\mu}_n = \sum_{k=1}^{N(n)} Z_k^n / N(n). \tag{2.2}$$

In a general case, estimation of the Hurst parameter (*H*) is based on the 'IID' assumption (Independent Identically Distributed), due to its simplicity and robustness as well as low computational cost and medium implementation complexity. The 'IID' assumption allows for estimation of the data sequence mean and variance, which are recalculated for each newly considered data sample. Recalculation of the sample mean is quite straightforward. The sample mean for the $X_{i-1}(t)$ point is defined as follows:

$$\overline{X}_{i-1} = \frac{1}{i-1}\sum_{m=1}^{i-1} X_m(t), \tag{2.3}$$

while the sample mean value for the $X_i(t)$ point can be estimated in a straightforward manner using the stored previous value of the sample mean as well as the number of data samples used to calculate it:

$$\overline{X}_i = \frac{1}{i}\sum_{m=1}^{i} X_m(t) = \frac{1}{i}\left(\sum_{m=1}^{i-1} X_m(t) + X_i(t)\right), \tag{2.4}$$

which in turns leads to:

$$\overline{X}_i = \left[\overline{X}_i - 1 \cdot (i-1) + X_i(t) \right] / i. \tag{2.5}$$

Similar assumptions and calculation approach results in the estimation of the sample variance without the need to keep track of all the sample values within the examined data sequence. In a general case, the sample variance for the previous data point ($X_{i-1}(t)$) is given by the following equation:

$$\sigma_{i-1}^2 = \frac{1}{i-2} \sum_{m=2}^{i-1} \left(X_m(t) - \overline{X}_{i-1} \right)^2, \tag{2.6}$$

while the sample variance for the current data point ($X_i(t)$) is estimated by (2.7):

$$\sigma_i^2 = \frac{1}{i-1} \sum_{m=1}^{i} \left(X_m(t) - \overline{X}_i \right)^2 = \frac{1}{i-1} \left(\sum_{m=1}^{i-1} \left(X_m(t) - \overline{X}_i \right)^2 + \left(X_i(t) - \overline{X}_i \right)^2 \right). \tag{2.7}$$

Equation (2.7) can be linked with (2.6) as follows:

$$\sigma_1^2 = \frac{1}{i-1} \sum_{m=1}^{i} \left(X_m(t) - \overline{X}_i \right)^2 = \frac{1}{i-1} \left(\sum_{m=1}^{i} X_m^2(t) - 2 \cdot \overline{X}_i \cdot \sum_{m=1}^{i}(t) + \overline{X}_m^2 \right) =$$
$$= \frac{1}{i-1} \left(\sum_{m=1}^{i-1} X_m^2(t) + X_i^2(t) - 2 \cdot \overline{X}_i \cdot \left(\sum_{m=1}^{i-1} X_m(t) + \overline{X}_i(t) \right) + i \cdot \overline{X}_m^2 \right). \tag{2.8}$$

An alternative approach aims at estimating the relation between the current and previous value of the sequence variance in the following form:

$$\sigma_{i-1}^2 = \frac{1}{i-2} \sum_{m=1}^{i-1} \left(X_m(t) - \overline{X}_{i-1} \right)^2 = \frac{1}{i-2} \left(\sum_{m=1}^{i-1} X_m^2(t) - 2 \cdot \overline{X}_{i-1} \cdot \sum_{m=1}^{i-1} X_m(t) + (i-1) \cdot \overline{X}_{i-1}^2 \right). \tag{2.9}$$

Equation (2.9) is used to estimate the aggregated value of squared previous samples:

$$\sum_{m=1}^{i-1} X_m^2(t) = (i-2) \cdot \sigma_{i-1}^2 + 2 \cdot \overline{X}_{i-1} \cdot \sum_{m=1}^{i-1} X_m(t) - (i-1) \cdot \overline{X}_{i-1}^2. \tag{2.10}$$

Next, the current variance value can be further derived to the following form:

$$\sigma_i^2 = \frac{1}{i-1} \left(\sum_{m=1}^{i-1} X_m^2(t) + X_i^2(t) - 2 \cdot \overline{X}_i \cdot \left(\sum_{m=1}^{i-1} X_m(t) + X_i(t) \right) + i \cdot \overline{X}_i^2 \right) =$$
$$= \frac{1}{i-1} \left(\begin{array}{l} (i-2) \cdot \sigma_{i-1}^2 + 2 \cdot \overline{X}_{i-1} \cdot \sum_{m=1}^{i-1} X_m(t) - (i-1) \cdot \overline{X}_{i-1}^2 + X_i^2(t) - \\ -2 \cdot \overline{X}_i \cdot \left(\sum_{m=1}^{i-1} X_m(t) + X_i(t) \right) + i \cdot \overline{X}_i^2 \end{array} \right). \tag{2.11}$$

2.6. Entropy

Entropy is associated with the Second Law of Thermodynamics, first formulated by Sadi Carnot in 1824 [24] and funded with mathematical rigor by Clausius in 1850 [25].

In a general sense, the Second Law formulates that temperature differences between systems in contact with each other tend to equalize and work can be obtained from these non-equilibrium differences. Pressure, density and temperature differences tend to equalize if given the opportunity.

The evolutionary thought, which in Biology meant a move in the direction of order and complexity, in Physics has come to mean a tendency to disorder. First outlined by Carnot, then stated by Clausius, the Second Law brought the entropy concept.

Every increase in entropy is an increase of internal disorder and the maximum entropy corresponds to a total molecular disorder in a system, which manifests itself at the global level of homogenization and equilibrium.

In the 1870s, Boltzmann [26] and Gibbs [27] introduced the statistical probability. It was by Boltzmann kinetic gas work that the concept of entropy was linked with molecular disorder and entropy became a measure of disorder of the movement of molecules of a body. Boltzmann showed the relationship between entropy and probability using the equation in which the entropy of a state is proportional to the logarithm of the probability of that state.

In statistical thermodynamics, the most general formula for the thermodynamic entropy S of a thermodynamic system is the Gibbs entropy:

$$S = -k_B \sum p_i \ln p_i,$$
(2.12)

where k_B is the Boltzmann constant.

Then, an increase in entropy happens when the configurations go from less to more likely, *i.e.,* clutter and disorganization are identified with the highest physical probability in the case of a closed system.

Ilya Prigogine [28], one of the founders of the Chaos Theory, proposed that the information that fights the entropy is represented by the auto organization of the matter and the acquisition of a complex internal order, as it happens in the chaotic systems. In this way, the evolution of life in harmony with the thermodynamics principles can be explained.

There are close parallels between the mathematical expressions for the thermodynamic entropy S of a physical system in the statistical thermodynamics established by Boltzmann and Gibbs and the information-theoretic entropy E of Claude Shannon and Ralph Hartley developed in the 1940s [29].

In information theory, entropy is a measure of the average uncertainty associated with a random variable. The entropy $E(X)$ of a single discrete random variable X is given by Shannon's entropy [29]:

$$E(X) = -\sum_{x_i \in \Theta} p(x_i) \log p(x_i).$$ (2.13)

where X represents a random variable with set of values Θ and probability mass function $p(x_i)=P_r\{X=x_i\}$, $x_i \in \Theta$.

For a time series representing the output of a stochastic process, that is, an indexed sequence of n random variables, $\{X_1,...,X_n\}$, with set of values $\Theta_1,...,\Theta_n$, respectively, the joint entropy is defined as

$$E_n = E(X_1, X_2,...,X_n) = -\sum_{x_1 \in \Theta_1} ... \sum_{x_n \in \Theta_n} p(x_1,...,x_n) \log p(x_i,...,x_n),$$ (2.14)

where $p(x_1,...,x_n) = P_r\{X_1=x_1,...,X_n=x_n\}$ is the joint probability for the n variables $X_1,...,X_n$.

The rate at which the joint entropy grows with n, the entropy rate e, is defined as

$$e = \lim_{n \to \infty} \frac{E_n}{n}.$$ (2.15)

The state of a system at a certain instant, X_n, is partially determined by its history, $X_1, X_2,...,X_{n-1}$. However, each new state carries a certain amount of new information. The mean rate of creation of information, also known as the Kolmogorov-Sinai (KS) Entropy, is a useful parameter to characterize the system dynamics [30]. For deterministic periodic systems, the KS entropy is zero because any state depends only on the initial conditions. In contrast, this entropy measure is maximum for uncorrelated random processes, since each state is totally independent of the previous ones [30].

Considering a D-dimensional dynamical system, which phase space is partitioned into hypercubes of content ε^D and that the state of the system is measured at intervals of time τ, $p(k_1,k_2,...,k_n)$ denote the joint probability that the state of the system is in the hypercube k_1 at $t=\tau$, in the k_2 at $t=2\tau$, and in the hypercube k_n at $t=n\tau$. The Kolmogorov-Sinai entropy is defined as

$$E_{KS} = -\lim_{\tau \to 0} \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n\tau} \sum_{k_1,...,k_n} p(k_1,...,k_n) \log p(k_1,...,k_n) = \lim_{\tau \to 0} \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n\tau} E_n.$$ (2.16)

Then, by the chain rule, it is straightforward to show that

$$E_{KS} = \lim_{\tau \to 0} \lim_{\varepsilon \to 0} \lim_{n \to \infty} (E_{n+1} - E_n).$$ (2.17)

Numerically, only entropies of finite order n can be computed. As soon as n becomes large with respect to the length of a given time series, the entropy E_n is underestimated and decays towards zero. Therefore, the KS entropy for "real-world" time series of finite length cannot usually be estimated with reasonable precision.

Entropy measures, such as the entropy rate and the Kolmogorov complexity, grow monotonically with the degree of randomness. For instance, a randomized time series has higher entropy than the original time series, although the process of generating surrogate data destroys correlations and degrades information content of the original signal [30].

2.7. Entropy-based algorithms

Entropy-based algorithms have been used for measuring the complexity of physiologic time series and proved to be useful in discriminating between healthy and disease states [31, 32], although some results may lead to misleading conclusions [33].

Traditional entropy-based algorithms fail to quantify the complexity of a time series, possibly because they are single-scale based. Time series derived from complex systems are likely to present structures on multiple spatio-temporal scales and time series from simpler systems present single-scaled structures. Therefore, a meaningful measure of complexity should take into account multiple time scales [34].

Healthy systems generate much more complex outputs than diseased ones [33]. Disease systems, when associated with the emergence of more regular behavior, show reduced entropy values compared to the dynamics of free-running healthy systems [30]. Single-scaled entropy measures don't take into account the complex temporal fluctuations inherent in healthy physiological control systems [30].

2.7.1. Approximate Entropy

Pincus [31] developed the Approximate Entropy (ApEn) for the analysis of short and noisy time series and to quantify the concept of changing complexity. ApEn emerged from theory developed in the field of nonlinear dynamics and chaos [32], which is easily applied to physiologic time series [35].

Approximate Entropy is approximately the negative natural logarithm of the conditional probability (CP) of a dataset of length N that having repeated itself within a tolerance r for m points will also repeat itself for $m+1$ points. Approximate Entropy is usually represented by ApEn(m,r,N). Parameter r is commonly expressed as a fraction of the standard deviation of

the data and makes ApEn a scale-invariant measure. A lower value arises from a high probability of repeated template sequences in the data. ApEn is heavily dependent on the record length and is uniformly lower for short records.

Being B the number of matches of length m and A the subset of B that also matches for length m+1, CP=A/B. For ApEn, one calculates -$log(CP)$ for each template and averages these values for all the templates. Since neither A nor B can be 0, CP must be redefined to (1+A)/(1+B), a correction that can be rationalized as allowing templates to match themselves. This is obviously inconsistent with the idea of new information, however, it is a strong source of bias toward CP=1 and ApEn=0, when there are few matches and A and B are small [32].

Both Kolmogorov-Sinai entropy and the related Approximate Entropy parameters depend on an one step difference of the function (*e.g.* H_{n+1} – H_n) and reflect the uncertainty of the next new point, given the past history of the time series [30].

Approximate Entropy was designed to quantify the degree of predictability of a series of data points. Therefore, it is fundamentally a "regularity" statistic, not a direct index of physiologic complexity. Further, ApEn does not probe the nonlinear properties of the signal, nor it does quantify the fractal scaling behavior [11].

2.7.2. Sample Entropy

Sample Entropy (SampEn) quantifies the regularity of a time series. It reflects the conditional probability that two sequences of m consecutive data points that are similar to each other will remain similar when one more consecutive point is included. Being "similar" means that the value of a specific measure of distance is less than r. Therefore, as SampEn is a function of parameters m and r [34] and length of the time series, N, it is represented by SampEn(m,r,N).

Although m and r are critical in determining the outcome of either method for entropy estimation, no guidelines exist for optimizing their values. The various existing rules generally lead to the use of values of r between 0.1 and 0.25 and values of m of 1.0 or 2.0 for data records of length N ranging from 100 to 5 000 data points [32].

While these calculations only aim to estimate, entropy is defined in the limit as m approaches infinite and as r approaches 0. As r increases, the probability of matches tends toward 1.0 and SampEn tends to 0 for all processes, thereby reducing the ability to distinguish any salient features in the dataset; as m decreases, underlying physical processes that are not optimally apparent at smaller values of m may be obscured [32]. SampEn is little affected by loss of more than one third of the data [32].

The accuracy and confidence of the entropy estimate improve as the number of matches of length m and $m+1$ increase. Intuitively, the advantages of SampEn include larger values of A and B and hence more confident estimation of conditional probability. In either method, the number of matches can be increased by choosing small m (short templates) and large r (wide tolerance) [32].

Lower values of SampEn indicate a more self-similar time series.

Approximate Entropy and Sample Entropy derive from formulas suggested to estimate the Kolmogorov Entropy of a process represented by a time series. At their root, each is a measurement of the conditional probability that two vectors that are close to each other for m points will remain close to the next point. ApEn and SampEn are expected to give identical results for uniformly distributed random numbers [35].

In addition to eliminating self-matches, the SampEn algorithm is simpler than the ApEn algorithm, requiring approximately one-half as much time to calculate. SampEn is largely independent of record length and displays relative consistency under circumstances where ApEn does not [35].

There are two major differences between these two statistics. First, SampEn does not count self-matches; entropy is conceived as a measure of the rate of information production and, in this context, comparing data with themselves is meaningless; second, SampEn does not use a template-wise approach when estimating conditional probabilities; to be defined, SampEn requires only that one template find a match of length $m+1$ [35].

2.7.3. Multiscale Entropy

Multiscale Entropy (MSE) technique was introduced by Costa, Goldberger and Peng [30] and it is applicable to the analysis of biologic time series. MSE analysis is a new method for measuring a finite length time series complexity [36].

Given an one-dimensional discrete time series $\{x_1,...,x_2,...,x_N\}$, a consecutive coarse-grained time series is constructed by averaging a successively increasing number of data points in non-overlapping windows (Figure 3). Each element of the coarse-gained time series $\{y^{(\tau)}\}$ is calculated accordingly to the equation:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \qquad (2.18)$$

where τ represents the scale factor and $1 \leq j \leq N/\tau$. For scale 1, the coarse-grained time series is simply the original time series. The length of each coarse-grained time series is equal to the length of the original time series divided by the scale factor τ.
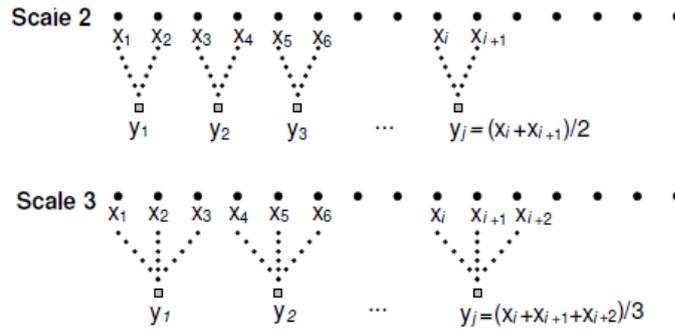
Figure 3. Schematic illustration of the coarse-grained procedure for scales 2 and 3 [33].

Finally, SampEn is calculated for each coarse-grained time series plotted as a function of the scale factor, τ (Figure 4). In the analysis of this plot, it must be taken into account not only the specific values of the entropy but also their dependence on resolution [30] to better characterize a physiologic process.
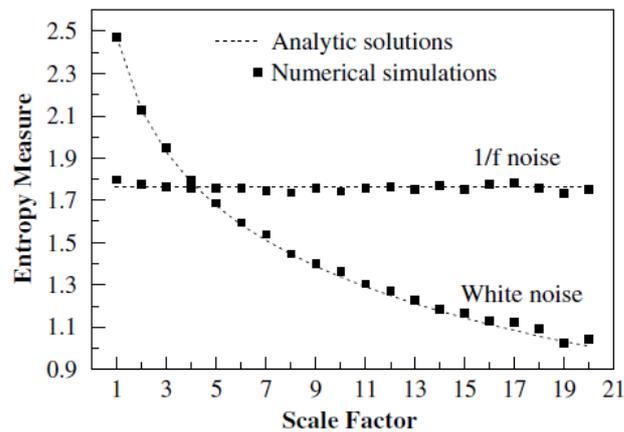
Figure 4. MSE analysis of Gaussian distributed white noise (mean zero, variance one) and 1/f noise. On the y-axis, the value of entropy (SampEn) for the coarse-grained time series is plotted. The scale factor specifies the number of data points averaged to obtain each element of the coarse-grained time series (adapted from [30]).

The Multiscale Entropy method appears to yield a more meaningful approach than conventional entropy measurements. MSE is based on the simple observation that complex physical and biologic systems generally exhibit dynamics that are far from perfect regularity and complete randomness. Instead, complex dynamics typically reveal structure on multiple spatial and temporal scales. These multiscale features, ignored by conventional entropy calculations, are explicitly addressed in the MSE algorithm [30].

Chapter 3. Biological Processes for Generation of Electrocardiogram Signals

This chapter presents the anatomy and the biological mechanisms of the heart. It is explained how electrical impulses spread along the heart and body tissues and how this electricity can be detected by the electrocardiograph electrodes.

It is presented a typical trace for an electrocardiographic signal and its characteristic waves. It is also explained the meaning of each of these waves in the interpretation of the ECG signals and its contribution to the diagnosis.

3.1. The heart's anatomy

The human heart consists of two ventricles that work as pumping chambers and two atria as filling chambers. The heart is a pulsatile pump, operating via muscular contraction of both the ventricles and atria and producing positive displacement of blood through the circulatory system. The right ventricle pumps blood into the pulmonary circulation, where blood becomes oxygenated and the left ventricle pumps into the systemic circulation that allows oxygenated blood to reach tissues throughout the body, where oxygen is transported to those tissues. Figure 5 depicts the heart with all four chambers.

3.2. Cardiac electrophysiology

Electrical impulses from the heart muscle cause the heart to contract. These impulses are created by the movement of ions across membranes in a fashion called depolarization of cells. This depolarization is propagated along pathways as the ion transport continues. This electrical signal begins in the sinoatrial (SA) node, located at the top of the right atrium. An electrical impulse from the SA node travels through the muscle fibers of the atria and ventricles, causing them to contract. The initiating electrical signal from the SA node travels down a preferred pathway of specialized conducting myocardial tissue toward the atrioventricular (AV) node, which is the electrical connecting point from the atria to the ventricles. After a slight transmission delay (designed to allow the atria to contract and fill the ventricles before the ventricles contract), the specialized conduction pathway continues into the ventricles. There are two bundles of fibers, called the left and right bundle branches

of His, culminating in purkinje fibers, that promote rapid transmission of electrical current through both ventricles. The electrical pathways are shown in Figure 5.
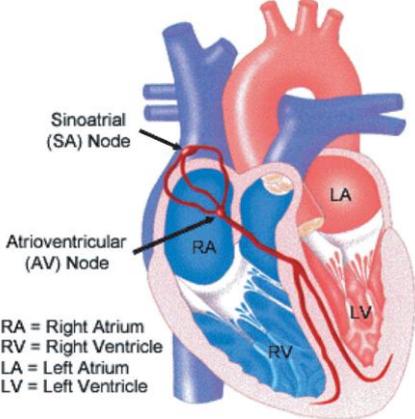
Figure 5. The heart's anatomy and electrical excitatory pathways (adapted from [37]).

3.3. The electrocardiogram

As the heart undergoes depolarization and repolarization, the electrical currents generated do not only spread within the heart but also throughout the body. This electrical activity generated by the heart can be measured by an array of electrodes placed on the body surface. Recording electrodes are used to detect such electrical potentials. These signals are filtered, amplified and recorded as a measure of the underlying cardiac electrical activity. The record that results from this procedure is termed an electrocardiogram (ECG). A typical ECG tracing is shown in Figure 6.
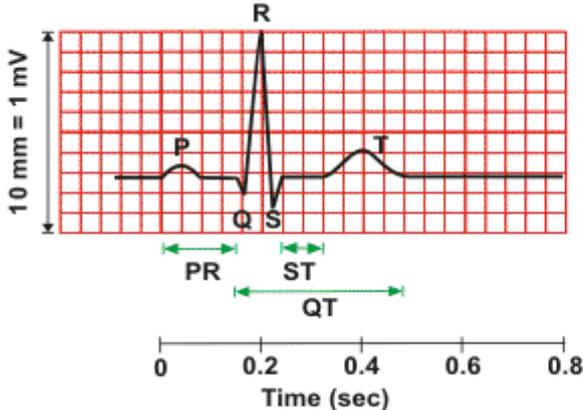
Figure 6. A typical ECG trace showing P, QRS and T waves and typical duration intervals (adapted from [38]).

Contrary to the predictions of homeostasis, the output of healthy human individuals' heartbeat, even under resting conditions, is neither constant or strictly periodic [6]. The human heart displays an extraordinarily large range of complex rhythms, in both health and

disease [39]. For example, heart rate varies in response to factors as emotional stress, exercise, postural changes and respiration.

3.3.1. Electrocardiogram waves

The normal ECG consists of three basic features, described bellow: a P wave, a QRS complex and a T wave.

**P wave:** The P wave represents the wave of depolarization that spreads from the SA node throughout the atria and is usually 80-100 ms in duration. The brief isoelectric period after the P wave represents the time in which the impulse is travelling within the AV node and the bundle of His.

QRS complex: The QRS complex represents ventricular depolarization. The QRS complex has a short duration which indicates that ventricular depolarization normally occurs very rapidly. If the QRS complex is prolonged (> 100 ms), conduction is impaired within the ventricles. The shape of the QRS complex in Figure 6 is idealized. In fact, the shape changes depending on which recording electrodes are being used. The shape will also change when there is abnormal conduction of electrical impulses within the ventricles.

T wave: The T wave represents ventricular repolarization. Conduction of the repolarization wave is slower than the wave of depolarization. Sometimes a small positive U wave may be seen following the T wave. The U wave, not shown in Figure 6, represents the last remnants of ventricular repolarization. Inverted or prominent U waves indicate underlying pathology or conditions affecting repolarization.

ST segment: The ST segment is the isoelectric period that follows the QRS. It is the time at which the entire ventricle is depolarized and roughly corresponds to the plateau phase of the ventricular action potential. The ST segment is important in the diagnosis of ventricular ischemia or hypoxia because, under those conditions, the ST segment can become either depressed or elevated.

Q-T interval: The Q-T interval represents the time for both ventricular depolarization and repolarization to occur and, therefore, roughly estimates the duration of an average ventricular action potential. At high heart rates, ventricular action potentials shorten in duration, which decreases the Q-T interval. Because prolonged Q-T intervals can be diagnostic for susceptibility to certain types of tachyarrhythmias, it is important to determine if a given Q-T interval is excessively long. There is no distinctly visible wave representing atrial repolarization in the

ECG because it occurs during ventricular depolarization. Because the wave of atrial repolarization is relatively small in amplitude, it is masked by the much larger ventricular-generated QRS complex.

3.3.2. Electrocardiogram leads

By convention, electrodes are placed on each arm and leg, and six electrodes are placed at defined locations on the chest. These electrode leads are connected to a device that measures potential differences between selected electrodes to produce the characteristic ECG tracings.

Some of the ECG leads are bipolar leads that utilize a single positive and a single negative electrode between which electrical potentials are measured. Unipolar leads have a single positive recording electrode and utilize a combination of the other electrodes to serve as a composite negative electrode. Normally, when an ECG is recorded, all leads are recorded simultaneously, giving rise to what is called a 12-lead ECG.

There are three types of ECG leads:

- Limb Leads (Bipolar),
- Augmented Limb Leads (Unipolar) and
- Chest Leads (Unipolar).

Bipolar recordings use standard limb lead configurations depicted in Figure 7. By convention, lead I, II and III measure the potential difference between the two arms, between the left leg and the right arm and between left leg and the left arm, respectively.

In addition to the three bipolar limb leads described above, there are three augmented unipolar limb leads. These are termed unipolar leads because there is a single positive electrode that is referenced against a combination of the other limb electrodes. The positive electrodes for these augmented leads are located on the left arm (aV_L), the right arm (aV_R) and the left leg (aV_F). In practice, these are the same electrodes used for leads I, II and III.

The three augmented leads, along with the three standard bipolar limb leads, are depicted, as shown in Figure 7, using the axial reference system. The aV_L lead is at -30° relative to the lead I axis, aV_R is at -150° and aV_F is at +90°.

In addition to the three standard limb leads and the three augmented limb leads that view the electrical activity of the heart from the frontal plane, there are six pre-cordial, unipolar chest leads. This configuration places six positive electrodes on the surface of the chest over

different regions of the heart in order to record electrical activity in a plane perpendicular to the frontal plane. These six leads are named V_1 - V_6.
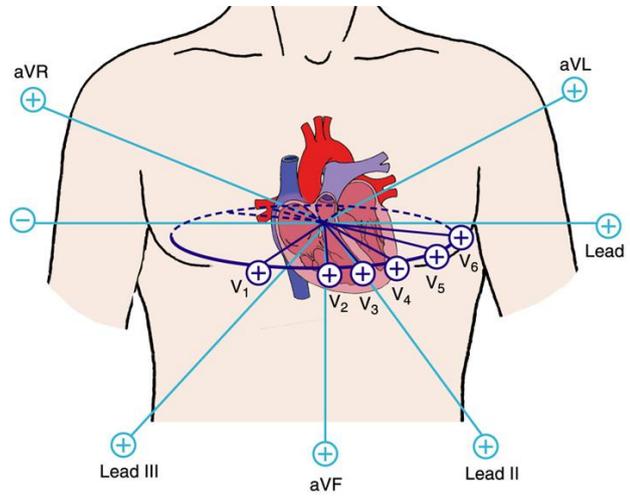

Figure 7. Electrocardiographic views of the heart [40].

Chapter 4. Methods

This chapter presents the data and the algorithms under research. The source of the ECG data is presented, along with its characteristics. There is also a description of the used algorithms for the Hurst parameter and entropy estimation.

4.1. Analyzed electrocardiogram data

The electrocardiogram (ECG) data used to test the Hurst parameter and entropy estimator algorithm was downloaded from the Physionet website [41]. In Physionet, the data is classified according to the pathology of the subject from whom the data was recorded. A set of 14 recordings was selected from subjects who have a Normal Sinus Rhythm (NSR) or who have been diagnosed with a pathology, such as Arrhythmia (A), Congestive Heart Failure (CHF), Malignant Ventricular Ectopy (MVE) and Ventricular Tachyarrhythmia (VTA).

These data files were labeled including the prefixes *nsr*, *a*, *chf*, *mve* and *vta*, respectively. Thus, for example, the file *chf07* refers to the subject number 7 who has been diagnosed with Congestive Heart Failure. The samples were chosen randomly for the *a* and *nsr* file sets and sequentially for the *chf*, *mve* and *vta* file sets. If a downloaded file has more than one data series, resulting from various recordings in the same file, it was considered only one record. In the *a* files, the two data series result from different leads (V1 and II). In this case we only considered lead II. As mentioned in Chapter 3, lead II is a limb lead and corresponds to the potential difference between the right arm and the left leg. Moreover, the waves are clearer, namely the QRS complex that presents itself positive with high amplitude.

Although we didn't use the same lead on all ECG files, the outcome measure is not influenced by such, as it is only affected by the spatial distribution of the ECG plot points. The plots for healthy subjects have a different spatial distribution than unhealthy subjects, independently the ECG lead, as present in this chapter.

According to the information available in Physionet, the sampling frequency of the signals is as follows: 128 Hz for Normal Sinus Rhythm signals, 360 Hz for Arrhythmia signals, and 250 Hz for the other signals. Yet, and without loss of generalization, because of the self-similar nature of the data [13], we considered the signals independently of their sampling frequency, *i.e.*, each data point is taken only by its nominal value disregarding the moment in time it was captured. This approach allows the estimation of the degree of self-similarity and the entropy of the ECG data in its spatial distribution.

The signals in Physionet website were exported in a plain text format, although there were other possible formats. The plain text format allows the sample to be manipulated by the Hurst and entropy estimation tools used.

The length of each sample may be different in each data file. Arrhythmia files have each 21 600 samples, while the Congestive Heart Failure files have 15 000 samples each. The smallest sample can be found in Normal Sinus Rhythm set, with a little less than 8 000 samples. The other analyzed sets have 15 000 samples each.

The analyzed files, in a total of 70 samples, are shown in Table 2.

Table 2. Analyzed data files.

Arrhythmia	Congestive Heart Failure	Normal Sinus Rhythm	Malignant Ventricular Ectopy	Ventricular Tachyarrhythmia
a_103_mlii	chf01	nsr_16265	mve418	vta01
a_106_mlii	chf02	nsr_16272	mve419	vta02
a_108_mlii	chf03	nsr_16273	mve420	vta03
a_111_mlii	chf04	nsr_16420	mve421	vta04
a_115_mlii	chf05	nsr_16483	mve422	vta05
a_119_mlii	chf06	nsr_16539	mve423	vta06
a_124_mlii	chf07	nsr_16773	mve424	vta07
a_203_mlii	chf08	nsr_16786	mve425	vta08
a_207_mlii	chf09	nsr_16795	mve426	vta09
a_209_mlii	chf10	nsr_17052	mve427	vta10
a_213_mlii	chf11	nsr_17453	mve428	vta11
a_217_mlii	chf12	nsr_18177	mve429	vta12
a_222_mlii	chf13	nsr_18184	mve430	vta13
a_228_mlii	chf14	nsr_19830	mve602	vta14

4.2. The algorithm

The above mentioned files (ECG data) were processed by an application written in Java, with the main goal of estimating the Hurst parameter and the entropy of the data in a windowed manner. The source code for this application is available in [42].

There are several algorithms that can be used to estimate the Hurst parameter. The comparison of efficiency for ECG signals of these algorithms and the one selected in this research is presented in Chapter 5.

The Hurst parameter is estimated by the modified Embedded Branching Process and the entropy is estimated by Shannon's entropy, both described in Chapter 2. These algorithms estimate these two parameters in a windowed manner on the absolute values of recorded and

previously diagnosed ECG signals downloaded from Physionet. For clarity of display, some charts were clipped to show only the first 10 000 data points.

Both estimators are evaluated in a windowed manner, because it was proven that this allowed a better understanding of the nature of the signal, *i.e.*, it allows for a finer assessment of the degree of complexity involved in the several phases the signals passes through. Or in other words, the goal here was to evaluate how much complexity was involved in the generation of the signal, in particular, in each different stage of the PQRST complex. Following previous results, it was clear that the higher the complexity of the signal, the more healthy the subject. Nevertheless, the signal in its full extension is complex enough to dissipate any minor variation in the underlying complexity of the physiological processes. Concluding, it became clear that a certain degree of history was needed, but not all of the signal's history, as this would result in the assessment of the totality of the mechanisms involved in the process.

The size of the sliding window was defined for values ranging from 50 to 1 000 samples (Figure 8, Figure 9, Figure 10, Figure 11 and Figure 12, respectively) and the observed results showed that the larger the size of the window, the smaller the variability both of the estimated Hurst parameter and entropy value.

For the Hurst parameter estimation, the window size was defined as 100 samples, as the variability of the parameter is an important parameter to assess minor changes in the nature of the signal. A window of 100 samples corresponds to 0.781 seconds for Normal Sinus Rhythm files, 0.278 seconds for Arrhythmia files, and 0.400 seconds for Congestive Heart Failure, Malignant Ventricular Ectopy and Ventricular Tachyarrhythmia files.
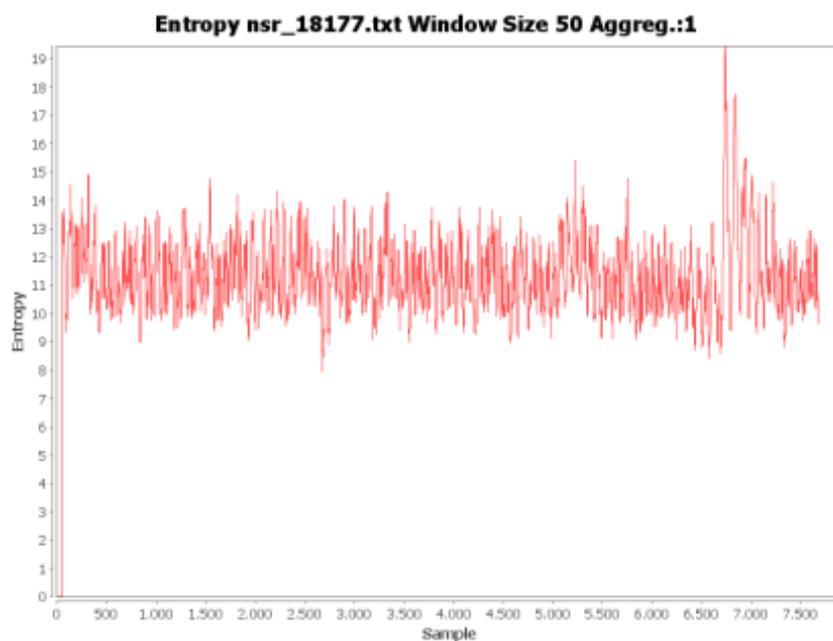

Figure 8. Evolution of entropy estimation in time for the nsr_18177 healthy subject, considering moving windows of 50 samples.

Entropy nsr_18177.txt Window Size 100 Aggreg.:1
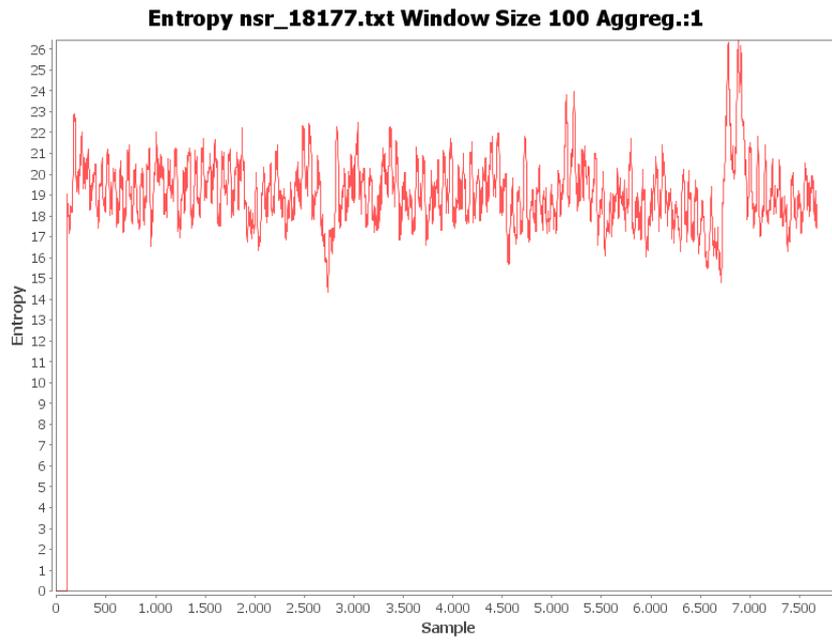
Figure 9. Evolution of entropy estimation in time for the nsr_18177 healthy subject, considering moving windows of 100 samples.


Entropy nsr_18177.txt Window Size 200 Aggreg.:1
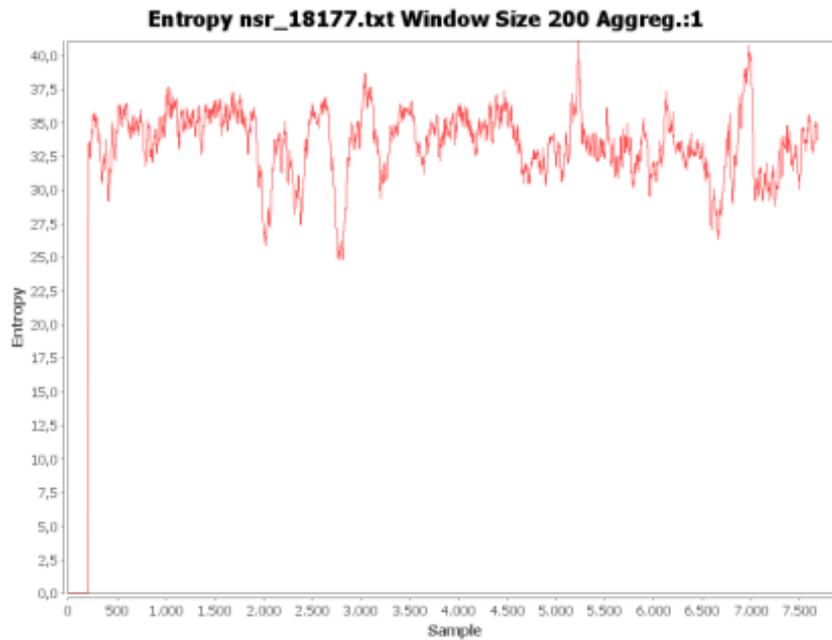
Figure 10. Evolution of entropy estimation in time for the nsr_18177 healthy subject, considering moving windows of 200 samples.
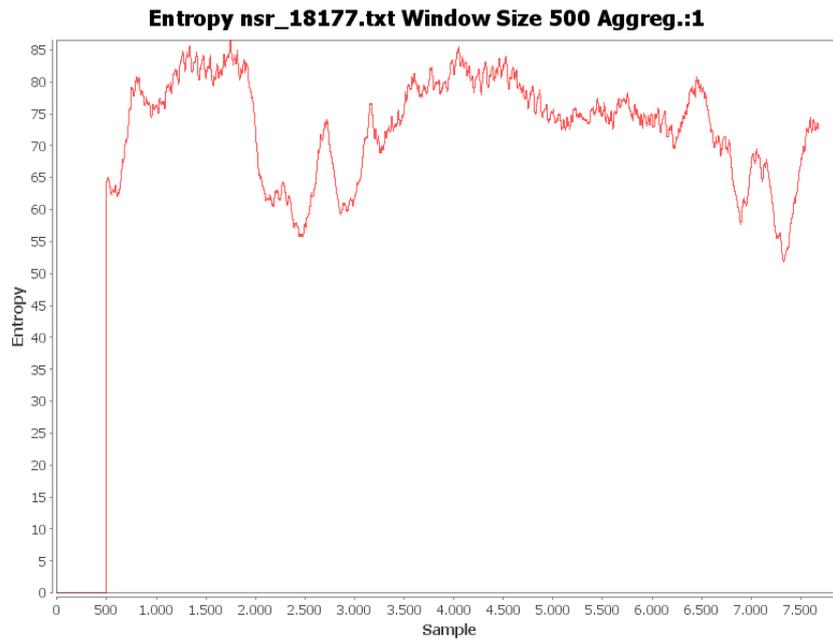
Figure 11. Evolution of entropy estimation in time for the nsr_18177 healthy subject, considering moving windows of 500 samples.
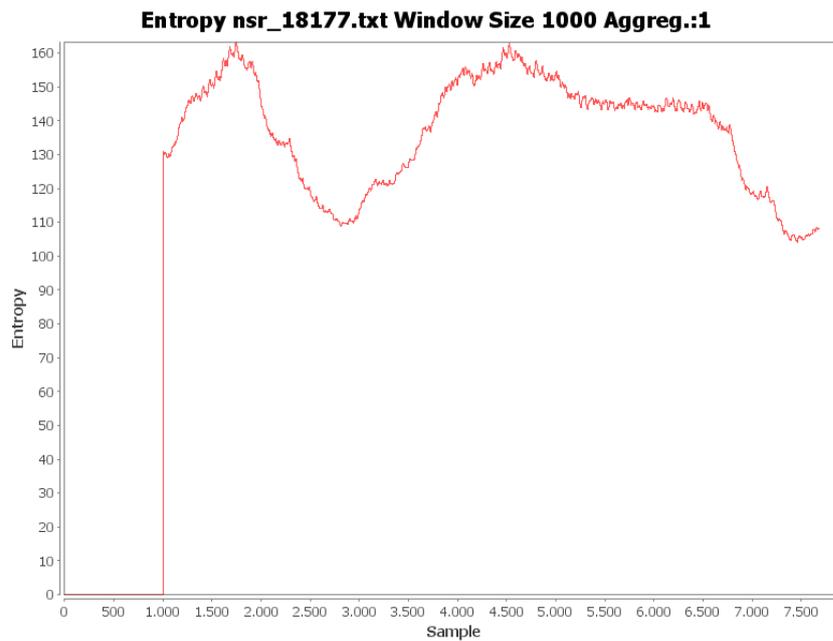
Figure 12. Evolution of entropy estimation in time for the nsr_18177 healthy subject, considering moving windows of 1000 samples.

For the entropy estimation, the window size used was also 100 samples, in line with the previously presented arguments.

The Hurst parameter and entropy estimation results obtained with this algorithm are shown in Chapter 5.

Chapter 5. Results and Discussion

In this chapter, it is presented the results of the Hurst parameter and entropy estimation for electrocardiographic signals. The discussion of these results is done as they are presented. The comparison of results obtained for the estimation of the Hurst parameter using the mEBP and other methods is also presented.

5.1. The Hurst parameter estimation

5.1.1. The mEBP Algorithm

In the following figures, the results for the Hurst parameter (H) estimation are presented. The x-axis represents the number of sample points and the y-axis represents the estimated value for the Hurst parameter, considering a sliding window of 100 samples.

Our aim is not comparing the obtained plots from healthy and non-healthy subjects, but evaluating parameters, as mean value and standard deviation of the Hurst parameter estimate, which allow us to differentiate them.

Figure 13 and Figure 14 show the H estimation for two healthy subjects. One can see that, in both cases, the mean value for the estimated H series is very close to 0.95, with a small variability (standard deviation in the order of hundredths) (Table 3). The oscillation of the H values is associated with the window size.

Figure 15 and Figure 16 show the plots for subjects who have been diagnosed with Arrhythmia. It is clear that these plots don't follow the same distribution as the previously shown in Figure 13 and Figure 14. In Figure 15 and Figure 16, two main characteristics are observed: firstly, the overall mean value is very close to 0.6; secondly, the variability of the values is very large, compared with values from NSR signals (Table 3). They show amplitude approximate to the unit, sometimes it covers all the range of values. The rhythm that can be detected in these signals is related to the heart beats.

Figure 17 and Figure 18 show the plots for subjects diagnosed with Congestive Heart Failure. Also here, the plots don't follow the same distribution of the plots obtained for healthy subjects. Moreover, the variability and mean value of the CHF plots are quite similar to those of the plots obtained for subjects with Arrhythmia, as they show high mean value and small variability.

Figure 19 and Figure 20 show the plots for subjects who have been diagnosed with Malignant Ventricular Ectopy. Again, these are clearly distinguishable from the plots from healthy subjects, as they present mean values of 0.6579 and 0.6910, respectively, and variability of 0.1225 and 0.1563.
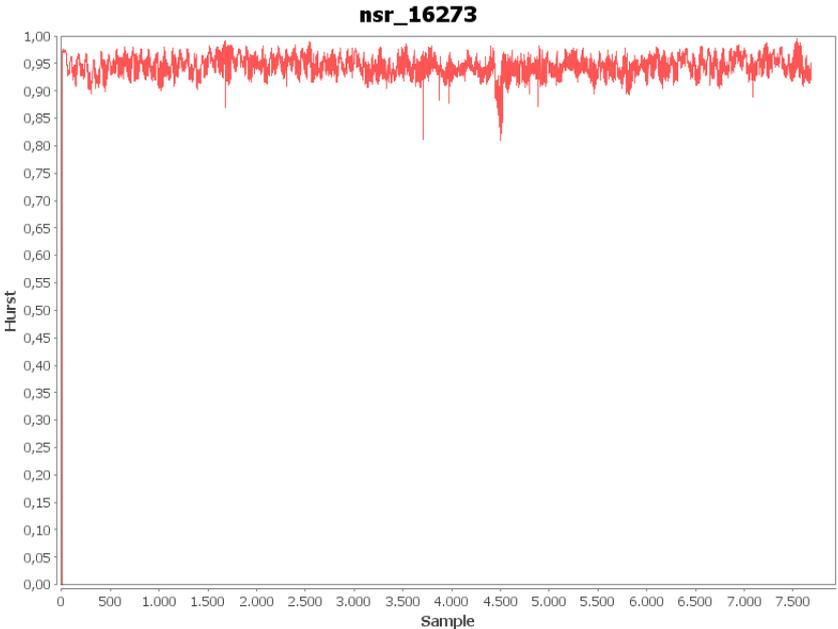
Figure 13. Evolution of the Hurst parameter estimation for the nsr_16273 healthy subject (mean value: 0.9450; standard deviation: 0.0183; amplitude: 0.1862), considering a moving window of 100 samples.
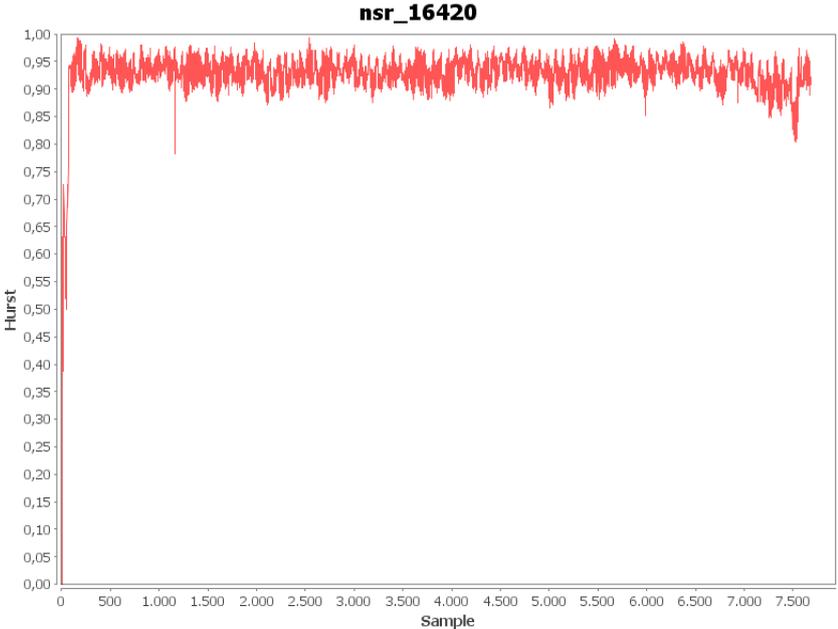
Figure 14. Evolution of the Hurst parameter estimation for the nsr_16420 healthy subject (mean value: 0.9308; standard deviation: 0.0257; amplitude: 0.4051), considering a moving window of 100 samples.
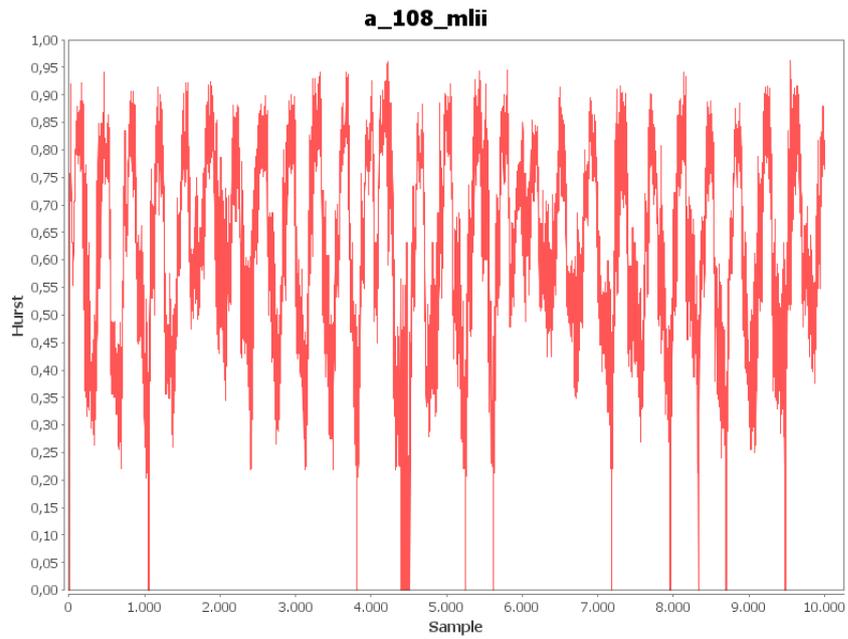
Figure 15. Evolution of the Hurst parameter estimation for the a_108 non-healthy subject (mean value: 0.6112; standard deviation: 0.1812; amplitude: 0.9620), considering a moving window of 100 samples.
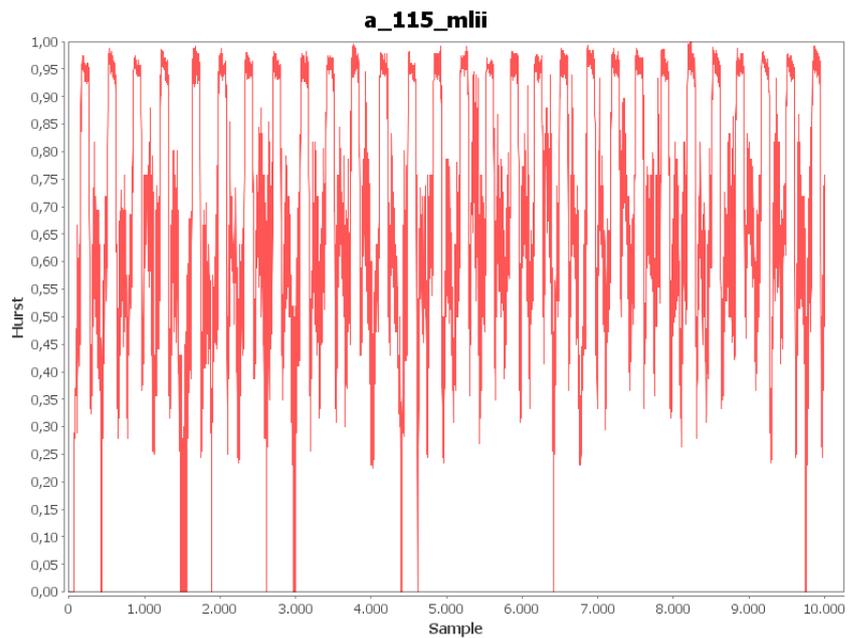
Figure 16. Evolution of the Hurst parameter estimation for the a_115 non-healthy subject (mean value: 0.6860; standard deviation: 0.2212; amplitude: 1.000), considering a moving window of 100 samples.
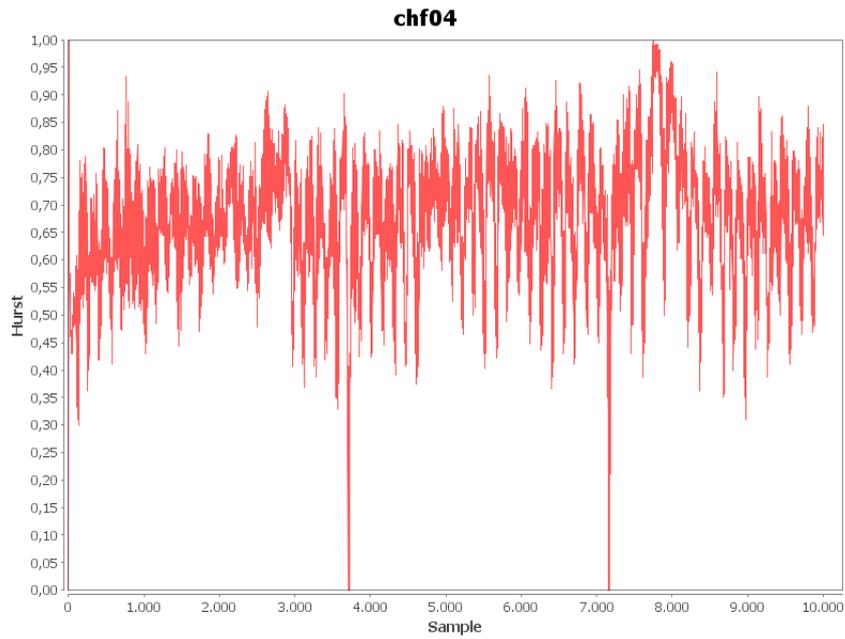
Figure 17. Evolution of the Hurst parameter estimation for the chf04 non-healthy subject (mean value: 0.6743; standard deviation: 0.1107; amplitude: 1.000), considering a moving window of 100 samples.
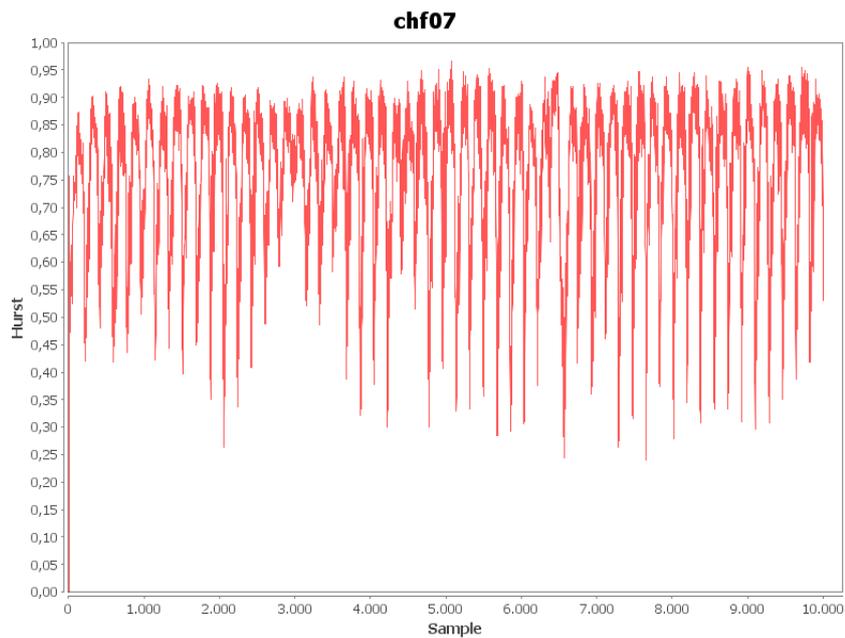
Figure 18. Evolution of the Hurst parameter estimation for the chf07 non-healthy subject (mean value: 0.7555; standard deviation: 0.1373; amplitude: 0.7274), considering a moving window of 100 samples.
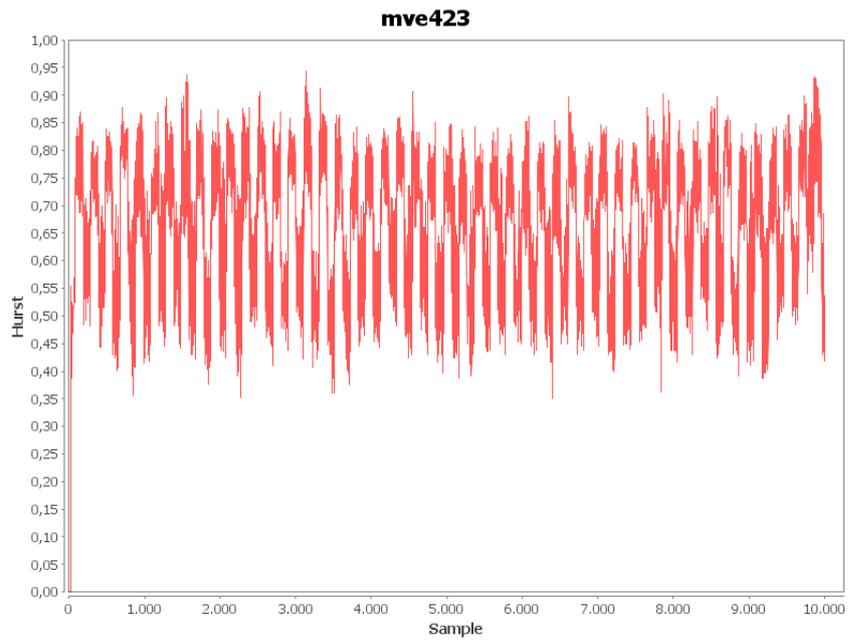
Figure 19. Evolution of the Hurst parameter estimation for the mve423 non-healthy subject (mean value: 0.6579; standard deviation: 0.1225; amplitude: 0.9433), considering a moving window of 100 samples.
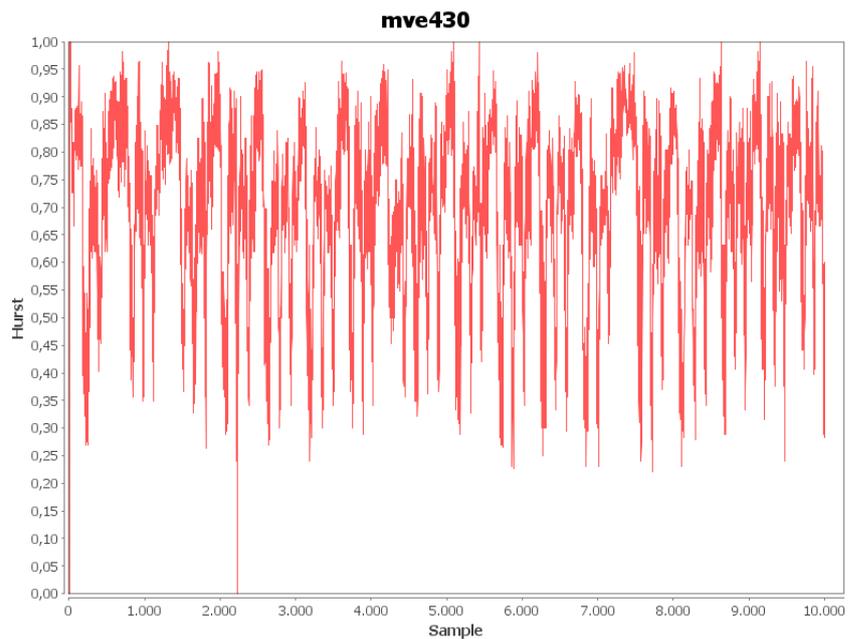

Figure 20. Evolution of the Hurst parameter estimation for the mve430 non-healthy subject (mean value: 0.8089; standard deviation: 0.1269; amplitude: 0.9926), considering a moving window of 100 samples.

Figure 21 and Figure 22 show the plots for subjects diagnosed with Ventricular Tachyarrhythmia. Once again, it is visible a high variability of the plot, and a mean value not close to the values expected in the estimation for healthy subjects, as we can see in Table 3.
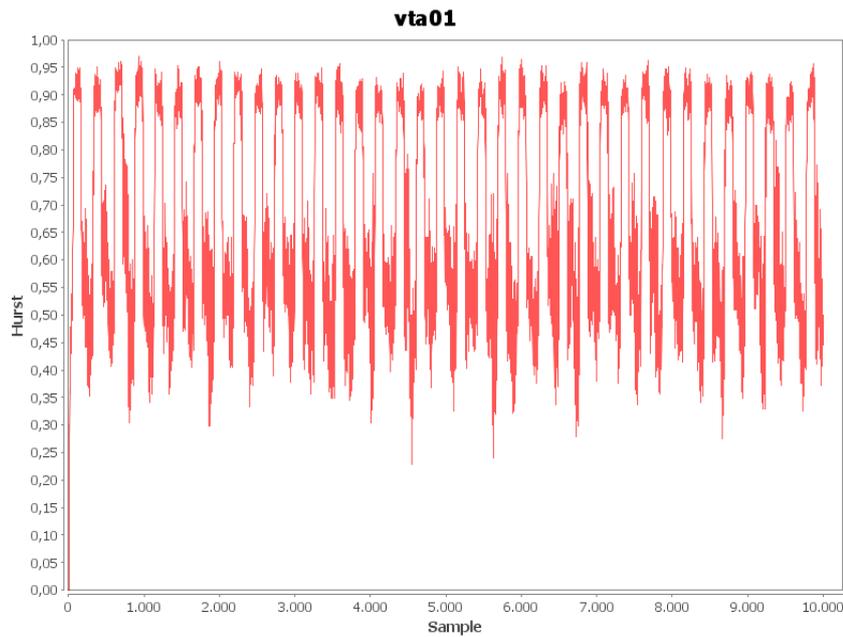
Figure 21. Evolution of the Hurst parameter estimation for the vta01 non-healthy subject (mean value: 0.6749; standard deviation: 0.1867; amplitude: 0.7399), considering a moving window of 100 samples.
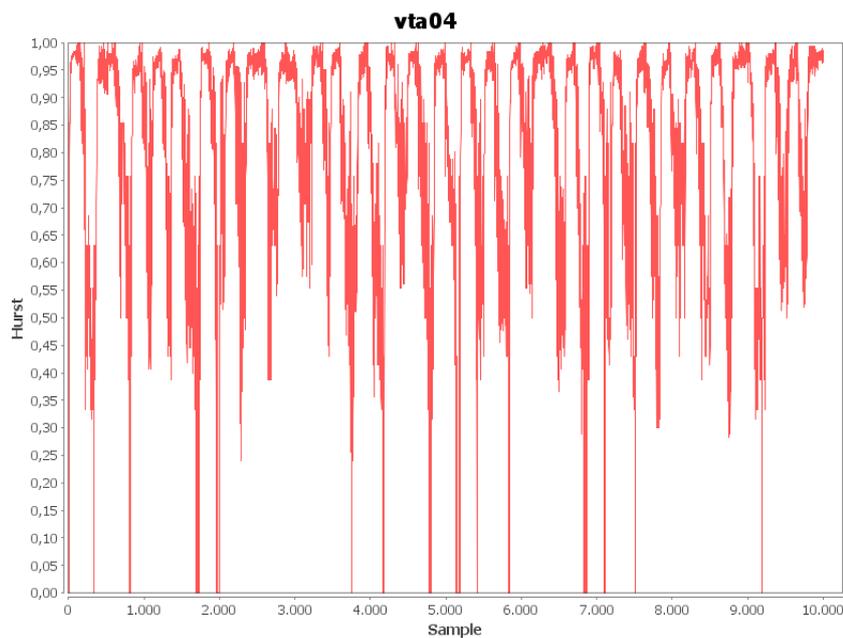
Figure 22. Evolution of the Hurst parameter estimation for the vta04 non-healthy subject (mean value: 0.8053; standard deviation: 0.2046; amplitude: 1.000), considering a moving window of 100 samples.

In general, all the charts for healthy subjects show a mean estimation of the Hurst parameter very close to 1.0, and a very small variability (standard deviation) and amplitude (Table 4) when compared to plots representing several types of pathologies. A larger value of *H* means that the data series is more-self-similar, and then ECG waveforms from healthy subjects are more self-similar than waveforms from unhealthy ones.

36

Table 3. Statistical data from the Hurst parameter estimation (values from the plots of Figures 10 - 19).

	Mean value	Standard deviation	Amplitude
nsr_16273	0.9450	0.0183	0.1862
nsr_16420	0.9308	0.0257	0.4051
a_108_mlii	0.6112	0.1812	0.9620
a_115_mlii	0.6860	0.2212	1.000
chf04	0.6743	0.1107	1.000
chf07	0.7555	0.1373	0.7274
mve423	0.6579	0.1225	0.9433
mve430	0.6910	0.1563	1.000
vta01	0.6749	0.1867	0.7399
vta04	0.8053	0.2046	1.000

Table 4. Statistical data from the Hurst parameter estimation of healthy and non-healthy subjects (mean values from 14 recordings).

	Mean ± SD	Amplitude
Normal Sinus Rhythm	0.9067 ± 0.0394	0.3260
Arrhythmia	0.7184 ± 0.1973	0.8866
Congestive Heart Failure	0.8109 ± 0.1015	0.5988
Malignant Ventricular Ectopy	0.7243 ± 0.1233	0.9815
Ventricular Tachyarrhythmia	0.9139 ± 0.0636	0.4466
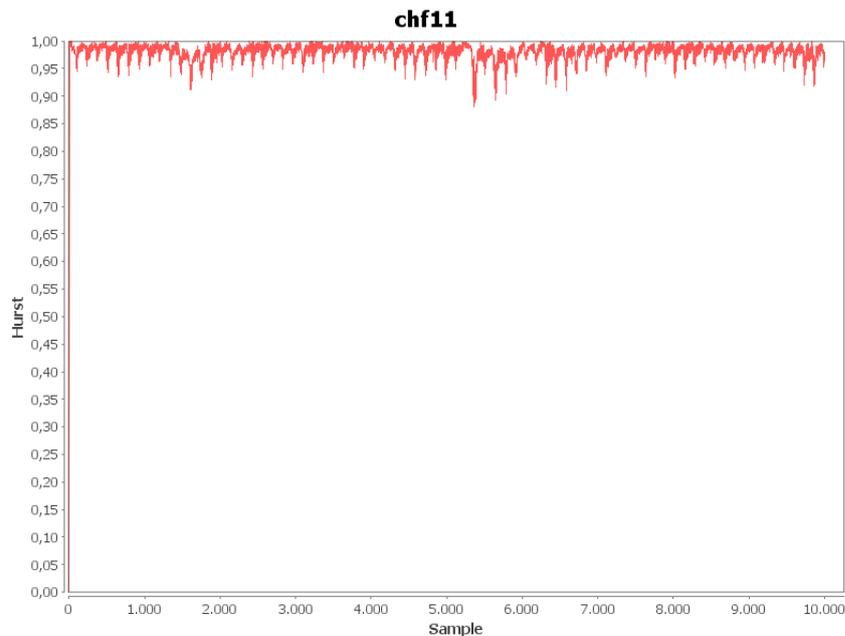
Figure 23. Evolution of the Hurst parameter estimation for the chf11 non-healthy subjects, considering a moving window of 100 samples.

Yet, some charts for subjects diagnosed as non-healthy, such as, *e.g.*, the chart for subject chf11 (Figure 23), show a plot that has the characteristics of the plot for a healthy subject (mean value of 0.9809; standard deviation of 0.0130; and amplitude of 0.1197). Thus this is the case of a false-healthy classification, *i.e.*, the case where the estimation of the Hurst parameter alone does not guarantee the affirmative classification of the healthy or unhealthy condition of the subject.

5.1.2. Comparison of efficiency between Hurst parameter estimation methods

Using statistical analysis, the results obtained with the mEBP estimator confirm the existence of differences between healthy and non-healthy systems. However, it is necessary to make a comparison with other Hurst parameter estimation methods and verify if the results are consistent within and across the various estimators.

For this study we used the Selfis software [43]. Selfis estimates the Hurst parameter through several techniques. Among the available algorithms, Aggregate Variance, Rescaled Statistics (R/S), Periodogram and Absolute Methods were chosen. This *H* estimation is obtained for the whole signal, unlike mEBP in which we used a sliding window of 100 samples. Thus, to allow the comparison, we also determined *H* for the entire signal using mEBP. The Hurst parameter mean values for 14 subjects of each diagnosis set are shown in Table 5.

Table 5. Mean values for mEBP and Selfis Hurst parameter estimators for healthy and non-healthy subjects.

	mEBP	Selfis			
		Aggregate variance	R/S	Periodogram	Absolute Moments
NSR	0,920	0,486	0,516	0,361	0,554
A	0,840	0,731	0,488	1,094	0,783
CHF	0,854	0,729	0,477	1,109	0,783
MVE	0,771	0,806	0,565	1,169	0,846
VTA	0,948	0,491	0,433	1,177	0,506

At first, it appears that the mEBP estimator is the one that presents the largest *H* estimation of all the estimators for Normal Sinus Rhythm signals, contrary to Selfis estimators that show higher values of *H* for non-healthy subjects than healthy subjects. However, the values for standard deviation for NSR (averaged over all the 14 subjects) are lower for mEBP (Table 6), comparatively to non-healthy systems, meaning the values are less variable, as we can see in Figure 24.

Table 6. Standard deviation values fro mEBP and Selfis Hurst parameter estimators for healthy and non-healthy subjects.

	mEBP	Selfis			
		Aggregate variance	R/S	Periodogram	Absolute Moments
NSR	0,039	0,114	0,268	0,283	0,105
A	0,086	0,078	0,125	0,327	0,084
CHF	0,078	0,099	0,070	0,222	0,122
MVE	0,102	0,154	0,095	0,307	0,148
VTA	0,044	0,163	0,217	0,166	0,158

It is known that different Hurst parameter estimators are more efficient in certain types of signals [44]. According to the measurements we present here, we can conclude that mEBP is better suited for ECG signals and thus allows for higher efficiency regarding its application as component of a diagnosis algorithm based on these signals.
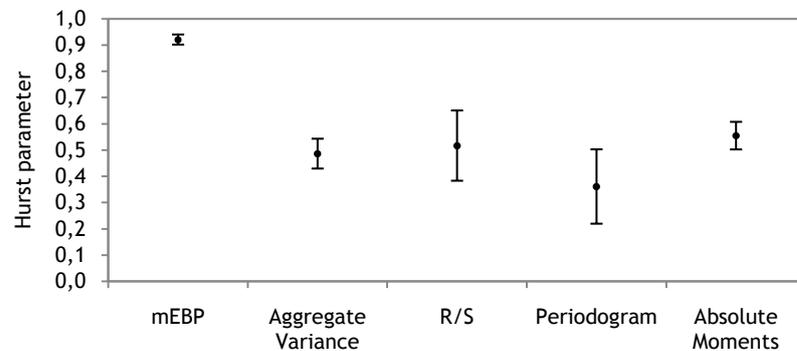

Figure 24. Hurst parameter estimation for NSR signals from the several methods used: mean values and standard deviation.

5.2. The entropy estimation

The initial question on the estimation of the entropy of ECG signals in a windowed manner was to select the size of the window. The chosen size was 100 samples and the conclusion was reached after the calculation of several statistics for windows of various sizes (50, 100, 200, 500 and 1000 sample windows). The statistics obtained allowed the comparison of the estimated entropy for the abovementioned window sizes and, finally, supported the choice of 100 samples as the appropriate window size.

Figure 25, Figure 26 and Figure 27 show the median, standard deviation and amplitude of the entropy estimation and we can see that these values for all window sizes remain in the same ratio, *i.e.*, the larger the window size the larger the median, standard deviation and amplitude. In the figures in Chapter 4, we can see some differences in signal distribution

between the plots obtained with different window sizes, as oscillation and variation over time. For a window of 100 samples, the standard deviation is lower, which means that the statistical dispersion of values of the median or mean value is lower.

With our algorithm we intend to analyze statistics such as median and standard deviation, instead of construe the plot. Then, for this reason, and those mentioned above, a window of 100 samples is adequate.
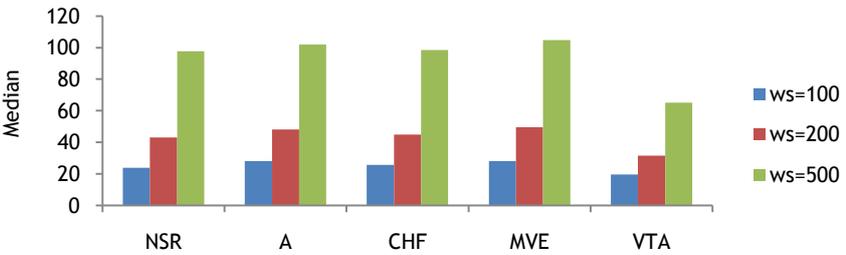
Figure 25. Median values of the entropy estimation, considering window sizes of 100, 200 and 500 samples, respectively (mean values for 14 subjects set).
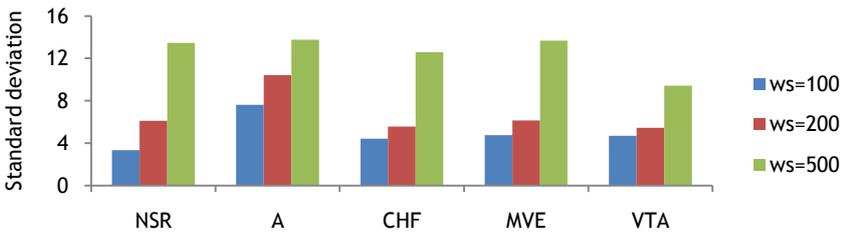
Figure 26. Standard deviation of the entropy estimation, considering window sizes of 100, 200 and 500 samples (mean values for 14 subjects set).
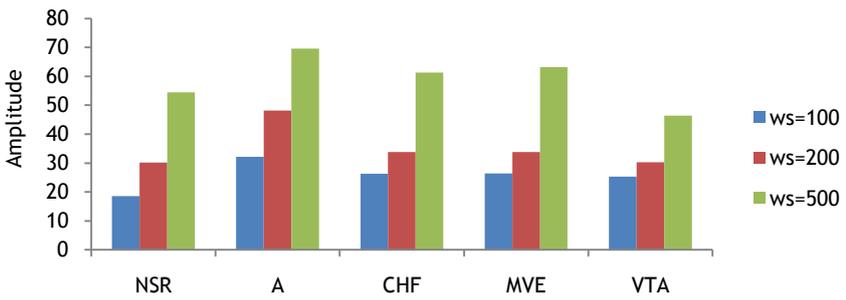

Figure 27. Amplitude of the entropy estimation, considering window sizes of 100, 200 and 500 samples (mean values for 14 subjects set).

5.2.1. Shannon's Entropy Algorithm

The following figures present the obtained results for the entropy estimation of ECG signal from healthy and non-healthy subjects. The *x*-axis represents the number of sample points and the *y*-axis represents the estimated value for the entropy, considering a sliding window of 100 samples.

Figure 28 and Figure 29 show the entropy estimation for two healthy subjects. Although the average value of the entropy plot of both subjects does not allow drawing conclusions, the standard deviation (1.50 and 2.32, respectively) and amplitude present low values (12.10 and 14.93, respectively).

Figure 30 and Figure 31 show the plots for windowed entropy estimation for subjects who have been diagnosed with Arrhythmia. They show a higher standard deviation (around 9.51 and 10.67, respectively) and amplitude (34.79 and 38.75, respectively), when compared to Normal Sinus Rhythm plots, regardless of its inconstant mean value.
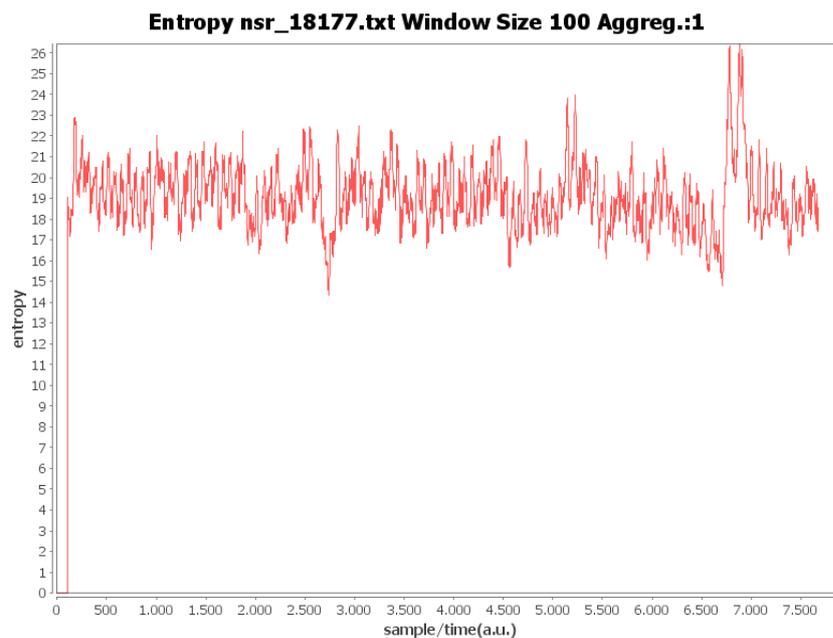
Figure 28. Evolution of entropy estimation in time for the nsr_18177 healthy subject (mean value: 19.0781; standard deviation: 1.5048; amplitude: 12.0967), considering a moving window of 100 samples.
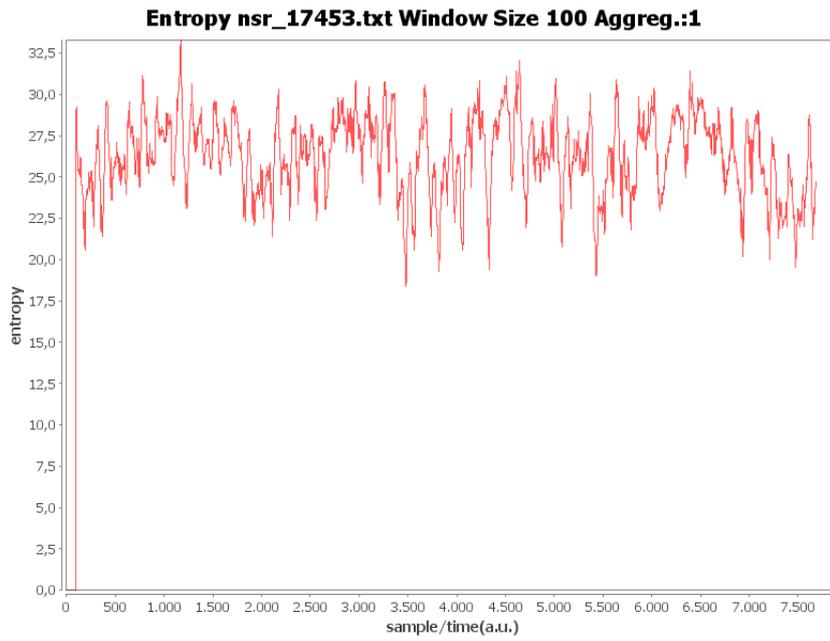
Figure 29. Evolution of entropy estimation in time for the nsr_17453 healthy subject (mean value: 26.2895; standard deviation: 2.3194; amplitude: 14.9281), considering a moving window of 100 samples.
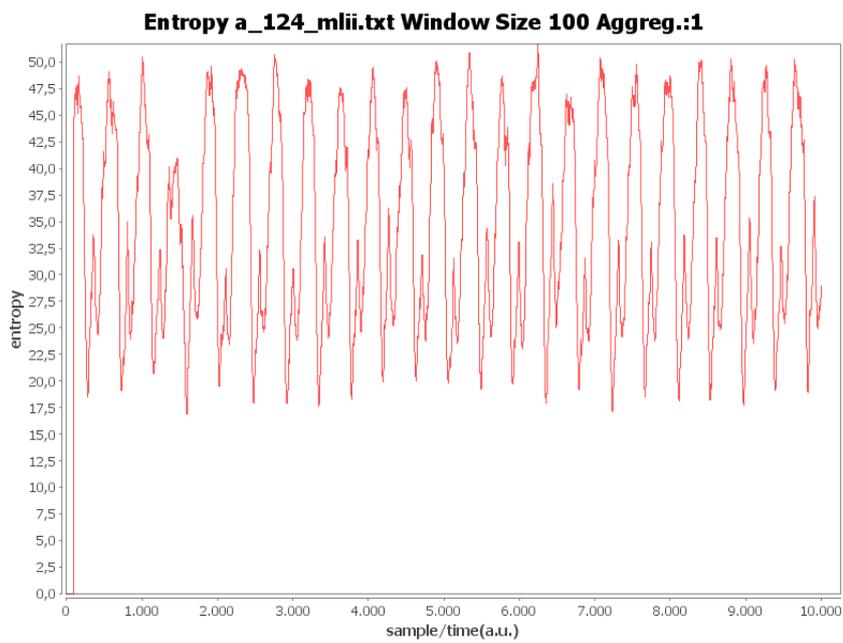
Figure 30. Evolution of entropy estimation in time for the a_124_mlii non-healthy subject (mean value: 34.4393; standard deviation: 9.5079; amplitude: 34.7918), considering a moving window of 100 samples.

Entropy a_217_mlii.txt Window Size 100 Aggreg.:1
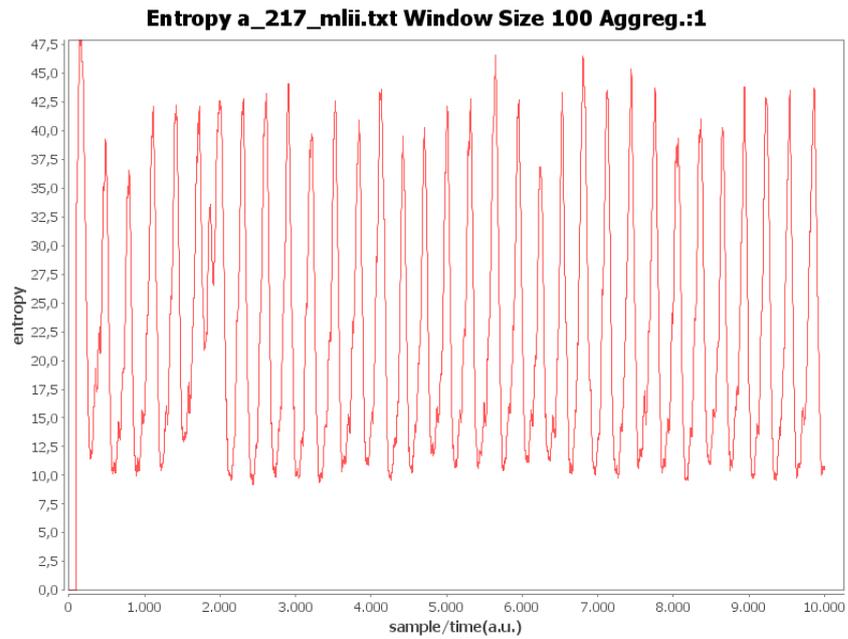
Figure 31. Evolution entropy estimation in time for the a_217_mlii non-healthy subject (mean value: 22.4929; standard deviation: 10.6714; amplitude: 38.7535), considering a moving window of 100 samples.

Figure 32 and Figure 33 show the plots for entropy estimation for subjects who have been diagnosed with Congestive Heart Failure. They show higher standard deviation (5.88 and 7.37, respectively) and amplitude (33.74 and 26.11, respectively), when compared to the plots from healthy subjects.

Entropy chf03.txt Window Size 100 Aggreg.:1
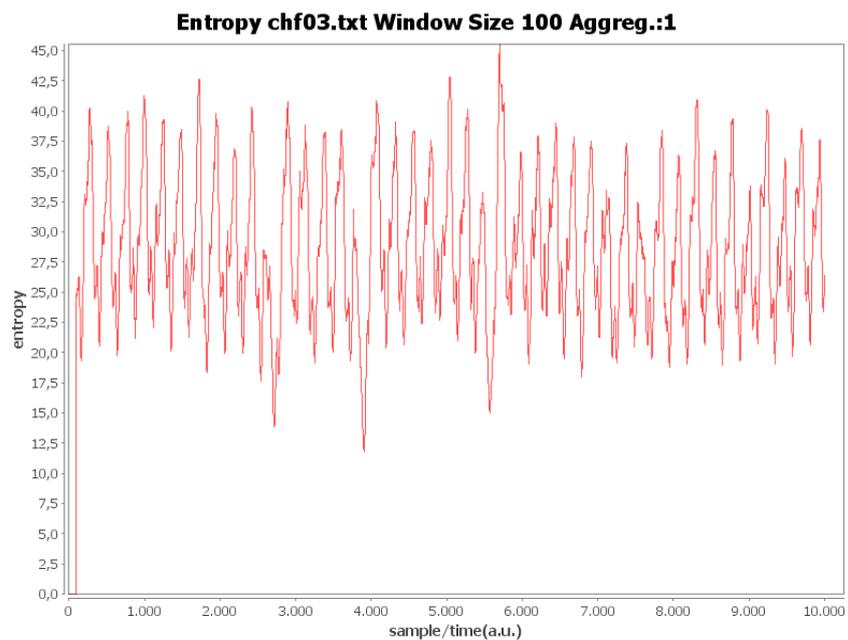
Figure 32. Evolution of entropy estimation in time for the chf03 non-healthy subject (mean value: 28.5977; standard deviation: 5.8816; amplitude: 33.7376), considering a moving window of 100 samples.

Entropy chf08.txt Window Size 100 Aggreg.:1
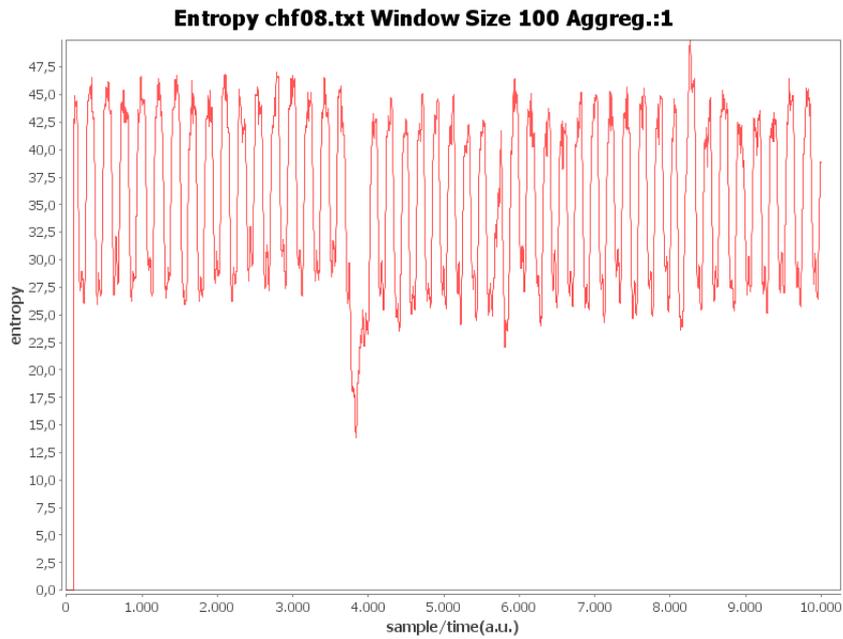
Figure 33. Evolution of entropy estimation in time for the chf08 non-healthy subject (mean value: 35.3840; standard deviation: 7.3688; amplitude: 26.1115), considering a moving window of 100 samples.

The same happens to the entropy estimation of subjects who have been diagnosed with Malignant Ventricular Ectopy, as shown in Figure 34 and Figure 35. They show higher values for entropy standard deviation (6.17 and 6.17, respectively) and amplitude (31.61 and 35.43, respectively) than Normal Sinus Rhythm plots.

Entropy mve422.txt Window Size 100 Aggreg.:1
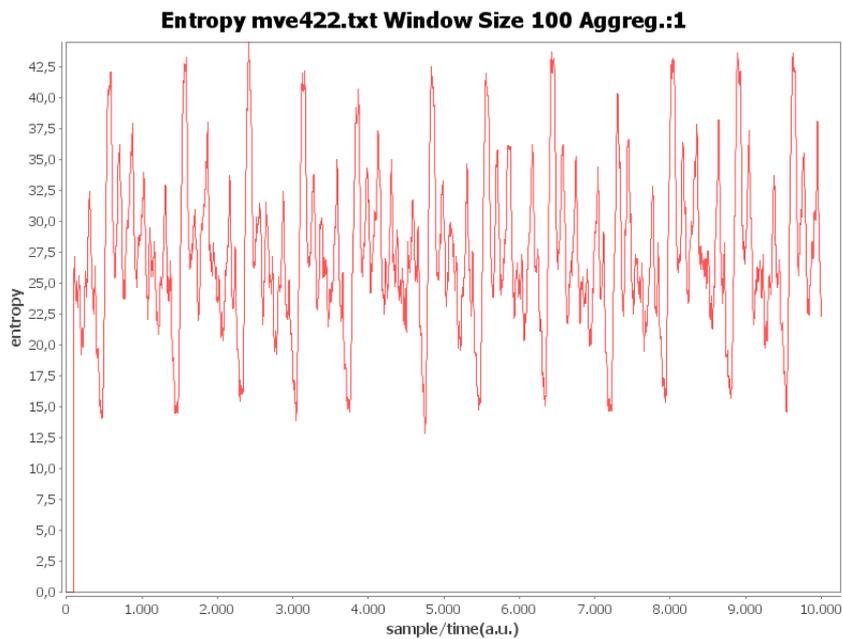
Figure 34. Evolution of entropy estimation in time for the mve422 non-healthy subject (mean value: 27.2105; standard deviation: 6.1683; amplitude: 31.6056), considering a moving window of 100 samples.

44

Figure 36 and Figure 37 show the plots for entropy estimation for subjects who have been diagnosed with Ventricular Tachyarrhythmia. They show a mean value lower than Normal Sinus Rhythm entropy, but its standard deviation and amplitude are higher than Normal Sinus Rhythm, still lower than the other pathologies.


Entropy mve602.txt Window Size 100 Aggreg.:1
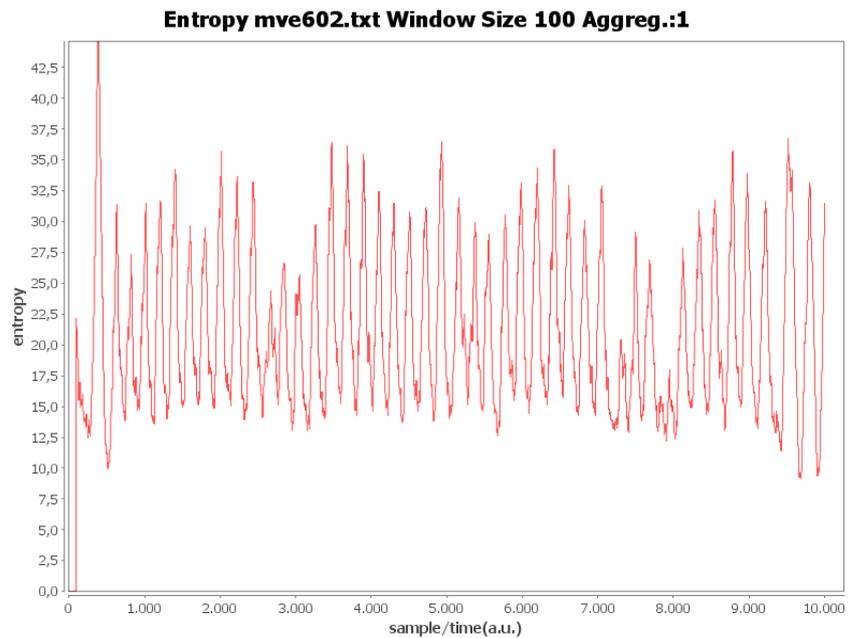
Figure 35. Evolution of entropy estimation in time for the mve602 non-healthy subject (mean value: 20.9588; standard deviation: 6.1706; amplitude: 35.4296), considering a moving window of 100 samples.


Entropy vta_cu05.txt Window Size 100 Aggreg.:1
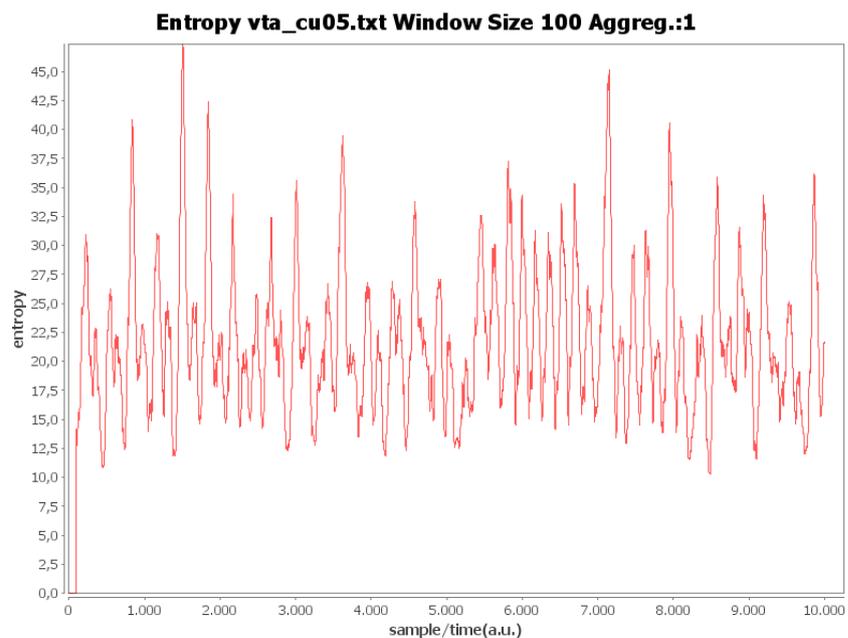
Figure 36. Evolution of entropy estimation in time for the vta05 non-healthy subject (mean value: 21.6539; standard deviation: 6.2162; amplitude: 37.1397), considering a moving window of 100 samples.
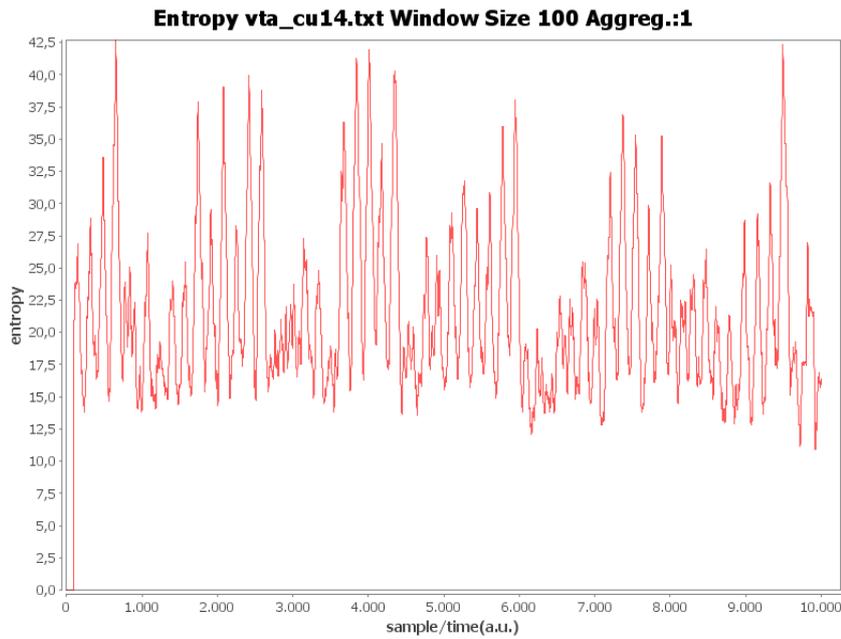
Figure 37. Evolution of entropy estimation in time for the vta14 non-healthy subject (mean value: 21.3250; standard deviation: 5.9727; amplitude: 31.7597), considering a moving window of 100 samples.

A quick glance at these plots suggests that there are differences in the nature of signals: the plots from non-healthy subjects show a different pattern than the plots from healthy ones.

The statistical data obtained from the present plots can be analyzed in Table 7.

Table 7. Statistical data from the Hurst parameter estimation (values from the plots of Figures 28 - 37).

	Median value	Standard deviation	Amplitude
nsr_17453	26.2895	2.3194	14.9281
nsr_18177	19.0781	1.5048	12.0967
a_124_mlii	34.4393	9.5079	34.7918
a_217_mlii	22.4929	10.6714	38.7535
chf03	28.5977	5.8816	33.7376
chf08	35.3840	7.3688	26.1115
mve422	27.2105	6.1683	31.6056
mve602	20.9588	6.1706	35.4296
vta05	21.6539	6.2162	37.1397
vta14	21.3250	5.9727	31.7597

By determining the entropy of ECG signals, we obtain different values for healthy subjects and for subjects who have been diagnosed with a pathology (Table 8). The mean value of the entropy alone does not allow distinguishing between healthy and non-healthy subjects. On the other hand, the standard deviation and amplitude are more reliable because, although the average value differ from one system to another, these are more significant, because

46

they show a clearer separation between healthy and non-healthy subjects, having in view its possible use in a diagnose algorithm. For healthy subjects the standard deviation is lower, at least, in two units, comparatively to non-healthy subjects, and amplitude, at least, in seven units.

Table 8. Statistical data from entropy estimation of healthy and non-healthy subjects (mean values from 14 recordings).

	Median	Standard Deviation	Amplitude
Normal Sinus Rhythm	25,6733	2,5980	16,8121
Arrhythmia	28,2004	7,6049	32,1163
Congestive Heart Failure	25,6227	4,4276	26,3662
Malignant Ventricular Ectopy	28,1852	4,7410	26,4017
Ventricular Tachyarrhythmia	19,5404	4,7023	25,2660

The entropy median, standard deviation and amplitude values for healthy and non-healthy subjects are very close to each other and also variable from subject to subject, as shown in Figure 38, Figure 39 and Figure 40. The major differentiations are in the standard deviation (variability) and amplitude, which are higher for non-healthy subjects.
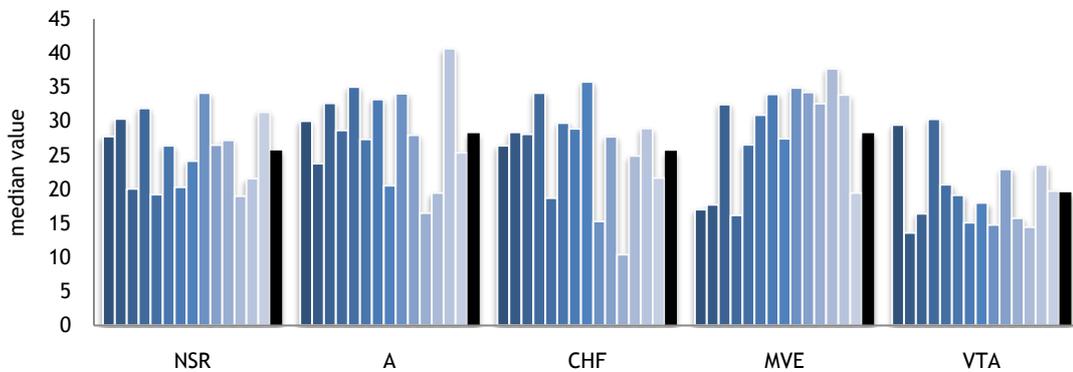
Figure 38. Estimation of the mean value of entropy for 14 subjects of each type of classification (the black bar represents the mean value for each set of values).

Analyzing Figure 39, it appears that the plot for Normal Sinus Rhythm (NSR) presents lower and more constant values than all pathologic subjects. In general, values for NSR entropy standard deviation are low, however some subjects may be classified as non-healthy, *e.g.*, subjects 11, 12 and 13, as they present the higher value for entropy standard deviation in all 14 subjects.

The plot bars for Arrhythmia show higher and more variable values. As for NSR, there are subjects who have a standard deviation outlying its mean value and thus may be considered

as healthy, as subject 11. Thus, the entropy value alone cannot be used to reach for conclusive results.
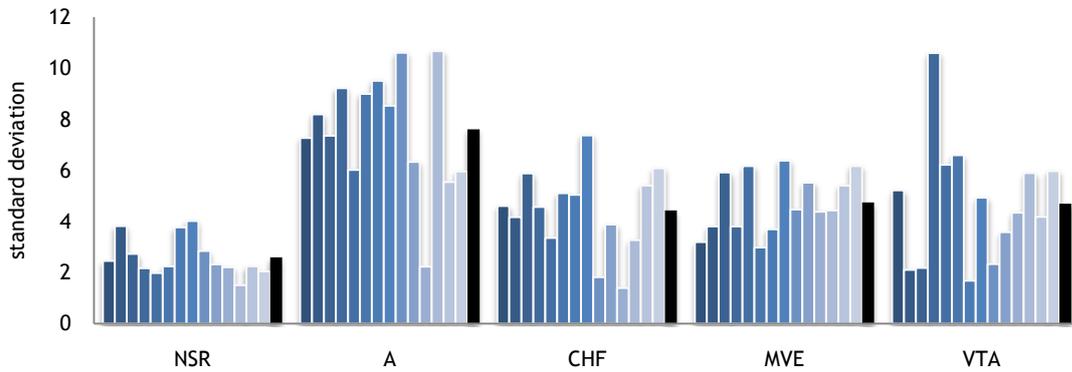
Figure 39. Estimation of the standard deviation value of entropy for 14 subjects of each type of classification (the black bar represents the standard deviation mean value for each set of values).

Figure 40 shows a significant difference in amplitude between signals of healthy subjects and subjects who have been diagnosed with pathology. The entropy estimate for a healthy subject has, in general, lower amplitude than for an unhealthy subject and are quite consistent from subject to subject.

The plot bars for entropy amplitude for subjects diagnosed with Arrhythmia show high and variable values. The entropy estimation for some subjects shows a low variability and, considering only this parameter, may be false-classified as healthy, as subject 11. The same happens with subjects 9 and 10 in Congestive Heart Failure. Due to its low amplitude, they may be false-classified as healthy subjects.

The plot bars for Ventricular Tachyarrhythmia shows high variability. There are subjects who, according to the reduced amplitude of the entropy estimate, could be considered healthy, although the mean amplitude for all 14 subjects is high (Table 8).
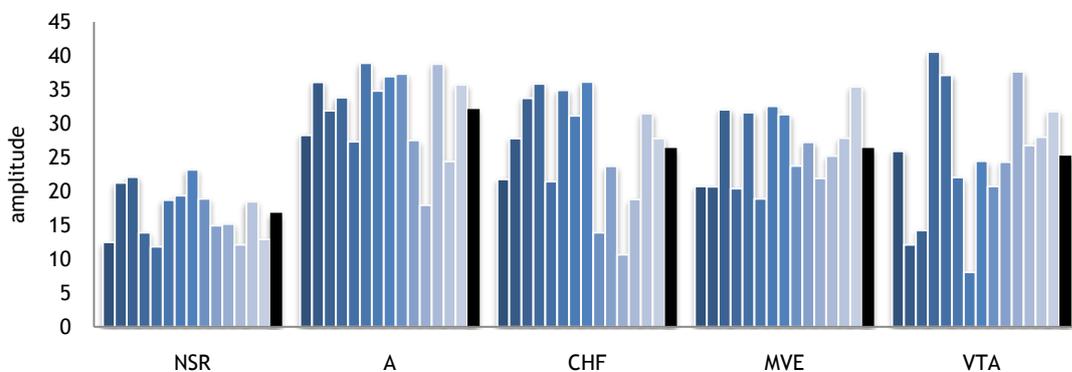

Figure 40. Estimation of the amplitude value of entropy for 14 subjects of each type of classification (the black bar represents the amplitude mean value for each set of values).

Thus, according to M. Costa [30], disease systems, because they are associated with the emergence of more regular behavior, show reduced multiscale entropy values when compared to healthy systems. In our research, and regarding Shannon's entropy estimated in a windowed manner, we obtained high entropy estimates both for healthy and non-healthy systems; nevertheless, in general, a non-healthy system has higher variability than a healthy one.

5.3. The relation between the Hurst parameter and entropy

Analyzing the results of the Hurst parameter (*H*) and entropy estimation, it is possible to confirm a relation between these two parameters.

When compared to signals from subjects who have been diagnosed with pathology, signals from healthy subjects show higher values for the Hurst parameter estimation; yet values from entropy estimate for all sets of ECG data are also high but quite similar.

To allow comparison, we need to evaluate the variability of the entropy with the variability of the *H* parameter, which have both shown to be significant. The variability values are presented in Figure 41. Here, we can see a relationship between the Hurst parameter and entropy estimates: signals from healthy subjects show less variable *H* and entropy values than non-healthy subjects, meaning the dispersion from the mean values is lower and all signals from healthy subjects present quite similar mean values. Besides, from a system to the other, when the variability of the Hurst parameter increases, the entropy variability also increases, and vice versa.

	NSR	A	CHF	MVE	VTA
mEBP	0,039	0,086	0,078	0,102	0,044
Entropy	2,598	7,605	4,428	4,741	4,702
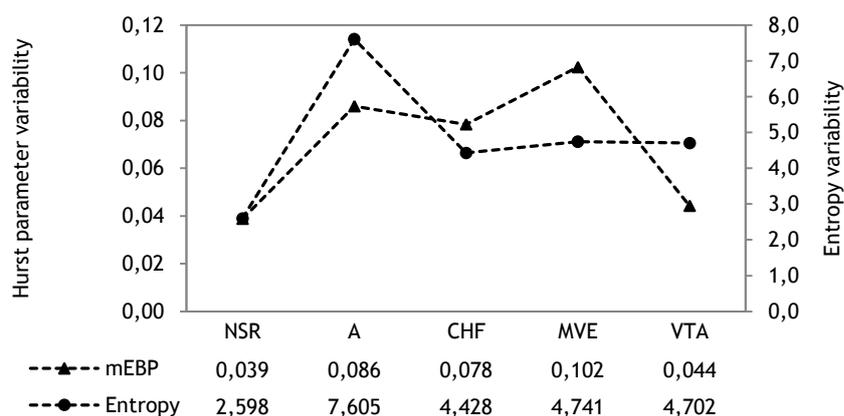
Figure 41. Relation between the mEBP Hurst parameter estimation and entropy estimation variability (mean values from 14 recordings).

Considering the methods used to estimate the Hurst parameter, the modified Embedded Branching Process shows the lower standard deviation, *i.e.*, the values are less variable. So in conclusion, the mEBP method and the entropy estimate behave similarly for the studied data.

Chapter 6. Conclusions and future work

The study of self-similarity and entropy of ECG signals is not new. In fact, several authors [3, 7, 35, 36] have published results showing that younger and healthier subjects can produce signals that have a higher degree of self-similarity and a higher level of entropy. Yet, these measurements have all been based on the study of the ECG time series, either using inter-hear-beat intervals or other distinctive features.

The Hurst parameter (*H*) and entropy estimates are methods that allow the distinction of a set of healthy ECG signals from non-healthy signals. To estimate *H*, we assessed the efficiency of several Hurst parameter estimators and, finally, used the modified Embedded Branching Process (mEBP); to estimate the entropy, we studied several previously published results and decided on Shannon's entropy. Both measurements were performed in a windowed manner as to allow the recognition of the variation of finer details and to prevent the influence of the signal history in the overall measurement.

Having a new approach on the study of ECG series, it was decided to assess such metrics on the spatial domain of the signal rather than on the time domain, being this topic already well known in the scientific community.

Results show that ECG signals from healthy subjects are very different from the majority of the signals of subjects who have been diagnosed with some form of pathology. Following the analysis of the charts, we further evaluated several statistics such as mean, standard deviation and amplitude of the signals obtained and we conclude that it is possible to differentiate healthy from non-healthy subjects. Healthy subjects have higher Hurst parameter estimate and lower standard deviation and amplitude than non-healthy subjects, showing that these signals have a higher degree of self-similarity even when studied for sample sizes that are no bigger than a hundred or a thousand values.

These results allow us to confirm what had been previously presented by [13] and [33], who state that signals from healthy (or young) subjects are more self-similar than signals from diseased (or elder) subjects, although the research in [13] and [33] was focused in the time distribution of the signals considering their inter-beat time. Moreover, our results show that a signal, whose self-similarity is low, or very variable, belongs to a subject who is not healthy. However, we can not state the opposite, because a non-healthy subject is not necessarily less self-similar. This is the case of some signals for some subjects who have been diagnosed with Congestive Heart Failure Thus, the Hurst parameter alone does not guarantee the correct classification of a subject as healthy or non-healthy.

Compared to other methods for estimating the Hurst parameter, the mEBP presents the highest values for Normal Sinus Rhythm signals, indicating their highest self-similarity. Moreover, the obtained *H* values from mEBP are less variable from subject to subject.

The results for Shannon's entropy estimation of the ECG signals show that it is also possible to distinguish healthy subjects' signals from non-healthy ones by the entropy estimated values. We evaluated the median, standard deviation and amplitude statistics. Considering only the average value of entropy can be misleading, because the median values for entropy are very variable between subjects and don't allow making a distinction between healthy and non-healthy systems. This study was based on standard deviation values which are more significant. In general, values of entropy estimate for healthy subjects are less variable than values for non-healthy subjects. However, some healthy subjects may be false-classified as non-healthy, because they present high variability; and the same happens to non-healthy subjects when they present low variability.

Although our results are not consistent with previous results based on multi-scale entropy applied to physiological signals, including M. Costa's [33, 34], our results do allow a distinction between healthy and non-healthy systems. In addiction, the results were obtained by Shannon's entropy on the spatial distribution of ECG signals, differently of M. Costa's Multiscale entropy applied on the temporal distribution of the signals.

Another conclusion allows us to say that we can relate the Hurst parameter to the entropy estimate. We evaluated the variability of the estimated values for all ECG data sets. Signals from healthy subjects show low variability both for Hurst parameter and entropy values, then all healthy subjects have quite similar Hurst parameter and entropy estimates. Of all the methods we used to estimate the Hurst parameter, the modified Embedded Branching Process has the lower standard deviation, *i.e.*, the values are less variable. So we can say that mEBP is the estimator which shows better coherence with the entropy estimate.

Regarding a more conclusive classification algorithm, and following the conclusions in [13], further research will take into consideration other variables, such as gender and age. In addition, it is worthwhile to use bigger samples and to elaborate this study with signals from other and variable pathologies. Moreover, the initial algorithm may be supplemented with new algorithms which determine other parameters, in order to substantiate the results obtained so far.

All these improvements will require more elaborate algorithms having as goal the creation of a full classification method between healthy and non-healthy subjects.

References

[1]   H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste*, Paris: Gauthiers-Villars, 1892-1899.

[2]   R. C. Eberhart, "Chaos theory for the biomedical engineer," *IEEE Eng Med Biol Mag,* vol. 8, no. 3, pp. 41-5, 1989.

[3]   A. L. Goldberger, and B. J. West, "Fractals in physiology and medicine," *Yale J Biol Med,* vol. 60, no. 5, pp. 421-35, Sep-Oct, 1987.

[4]   C. Musso, J. Bezic, S. Christiansen *et al.*, "Histological Glomerular Patterns as Chaotic Attractors," *Rev Electron Biomed / Electron J Biomed* no. 1, pp. 21-27, 2007.

[5]   V. Valle, "Chaos, Complexity and Deterrence," National War College, 2000.

[6]   A. L. Goldberger, "Is the normal heartbeat chaotic or homeostatic?," *News Physiol Sci,* vol. 6, pp. 87-91, Apr, 1991.

[7]   M. D. Costa, C. K. Peng, and A. L. Goldberger, "Multiscale analysis of heart rate dynamics: entropy and time irreversibility measures," *Cardiovasc Eng,* vol. 8, no. 2, pp. 88-93, Jun, 2008.

[8]   M. Costa, and J. A. Healey, "Multiscale Entropy Analysis of Complex Heart Rate Dynamics: Discrimination of Age and Heart Failure Effects," *Computers in Cardiology*, pp. 705-708, 2003.

[9]   B. B. Mandelbrot, "Stochastic models for the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands," *Proc Natl Acad Sci U S A*, vol. 72, no. 10, pp. 3825-3828, Oct, 1975.

[10]   K. Park, and W. Willinger, *Self-similar Network Traffic and Performance Evaluation*, New York: John Wiley & Sons, Inc., 2000.

[11]   A. L. Goldberger, C. K. Peng, and L. A. Lipsitz, "What is physiologic complexity and how does it change with aging and disease?," *Neurobiol Aging,* vol. 23, no. 1, pp. 23-6, Jan-Feb, 2002.

[12]   C. K. Peng, J. M. Hausdorff, and A. L. Goldberger, "Fractal mechanisms in neural control: Human heartbeat and gait dynamics in health and disease," *Self-Organized Biological Dynamics and Nonlinear Control*, Cambridge: Cambridge University Press, 2000.

[13]   A. L. Goldberger, L. A. Amaral, J. M. Hausdorff *et al.*, "Fractal dynamics in physiology: alterations with disease and aging," *Proc Natl Acad Sci U S A,* vol. 99 Suppl 1, pp. 2466-72, Feb 19, 2002.

[14]   H. E. Hurst, "Long-Term Storage Capacity of Reservoirs," *Transactions of American Society of Civil Engineers,* vol. 116, pp. 770-799, 1951.

[15]   I. Kaplan. "Estimating the Hurst exponent," October 29, 2009, 2009; http://www.bearcave.com/misl/misl_tech/wavelets/hurst/.

[16]   P. R. M. Inácio, "Study of the Impact of Intensive Attacks on the Self-Similarity Degree of the Network Traffic in Intra-Domain Aggregation Points," PhD Thesis, University of Beira Interior, Covilhã, Portugal, 2009.

[17]   P. Ulanovs, and E.Petersons, "Modeling Methods of Self-similar Traffic forNetwork Performance Evaluation," *Scientific Proceedings of RTU. Series 7. Telecommunications and Electronics,* vol. 2, 2002.

[18]   H.-D. J. Jeong, D. McNickle, and K. Pawlikowski, "Comparisson of Various Estimators of Hurst Parameter in Simulated FGN," *TR-COSC 02/06 Technical Report*, 2006.

[19]   O. D. Jones, and Y. Shen, "Estimating the Hurst Index of a Self-Similar Process via the Crossing Tree," *IEEE Signal Processing Letters,* vol. 11, no. 4, pp. 416-419, 2004.

[20]   O. D. Jones, "Fast efficient on-line simulation of self-similar processes," *World Scientific*, pp. 1-12, 2003.10.08, 2003.

[21]   M. Hajduczenia, N. M. Garcia, P. Monteiro *et al.*, *Monitoring method and apparatus of processing of a data stream with high rate/flow* European Patent Filed in the European Patent Office Patent number 05023580.3-2416,to Siemens SA, 2005.

[22]   O. D. Jones, and Y. Shen, "Estimating the Hurst Index of a Self-Similar Process via the Crossing Tree," *IEEE Signal Processing Letters,* vol. 11, no. 4, pp. 419-419, 2004.04, 2004.

[23]   M. Hajduczenia, N. Garcia, H. Silva *et al.*, "Real Time Evaluation of Self-Similarity Degree through Hurst Parameter Estimation using modified Embedded Branching Process (mEBP) " Siemens S.A., ed., 2005.

[24]   S. Carnot, *Reflections on the Motive Power of Fire – and other Papers on the Second Law of Thermodynamics.*, Gloucester: Peter Smith, 1977.

[25]   R. Clausius, *The Mechanical Theory of Heat – with its Applications to the Steam Engine and to Physical Properties of Bodies*, London: John van Voorst, 1 Paternoster Row, MDCCCLXVII.

[26]   L. Boltzmann, "On the mechanical meaning of the second law of the theory of heat," *Wiener Berichte*, no. 53, 1866.

[27]   J. W. Gibbs, *Elementary Principles in Statistical Mechanics*, New York: Charles Scribner's Sons, 1902.

[28]   I. Prigogine, and I. Stengers, *Order Out of Chaos, Man's Dialogue with Nature*, New York: Bantam Books, 1984.

[29]   C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal,* vol. 27, pp. 379-423, 623-656, 1948.

[30]   M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys Rev Lett,* vol. 89, no. 6, pp. 068102, Aug 5, 2002.

[31]   S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc Natl Acad Sci U S A,* vol. 88, no. 6, pp. 2297-301, Mar 15, 1991.

[32]   D. E. Lake, J. S. Richman, M. P. Griffin *et al.*, "Sample entropy analysis of neonatal heart rate variability," *Am J Physiol Regul Integr Comp Physiol,* vol. 283, no. 3, pp. R789-97, Sep, 2002.

[33]   M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy to distinguish physiologic and synthetic RR time series," *Comput Cardiol,* vol. 29, pp. 137-40, 2002.

[34]   M. Costa, C. K. Peng, A. L. Goldberger *et al.*, "Multiscale entropy analysis of human gait dynamics," *Physica A,* vol. 330, pp. 53 – 60, 2003.

[35]   J. S. Richman, and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am J Physiol Heart Circ Physiol,* vol. 278, no. 6, pp. H2039-49, Jun, 2000.

[36]   M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Phys Rev E Stat Nonlin Soft Matter Phys,* vol. 71, no. 2 Pt 1, Feb, 2005.

[37]   G. E. Miller, *Artificial Organs*, First ed.: Morgan & Claypool Publishers, 2006.

[38]   R. E. Klabunde, *Cardiovascular Physiology Concepts*: Lippincott Williams & Wilkins, 2005.

[39]   V. Schulte-Frohlinde, Y. Ashkenazy, A. L. Goldberger *et al.*, "Complex patterns of abnormal heartbeats," *Physical Review E*, no. 66, 2002.

[40]   "Instructor's Resource CD-ROM to Accompany Care Nursing: A Holistic Approach," Lippincott Williams & Wilkins, 2005.

[41]   A. L. Goldberger, L. A. Amaral, L. Glass *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation,* vol. 101, no. 23, pp. E215-20, Jun 13, 2000.

[42]   N. M. Garcia. "mEBP Java estimator," July 20, 2010, 2010; http://allab.di.ubi.pt.

[43]   T. Karagiannis, and M. Faloutsos, "SELFIS: A Tool For SelfSimilarity and LongRange Dependence Analysis."

[44]   O. Rose, *Estimation of the Hurst Parameter of Long-Range Dependent Time Series*, Institute of Computer Science, University of Wurzburg, 1996.