Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Teoría de Identificación de Sistemas Aplicado al Diagnóstico y Tratamiento de Señales Bio-médicas

presentada por
MC. Moisés Bulmaro Ramos Martínez

como requisito para la obtención del grado de
Doctor en Ciencias en Ingeniería Electrónica

Director de tesis
Dr. Víctor Manuel Alvarado Martínez

Codirectora de tesis
Dra. Ma. Guadalupe López López

Cuernavaca, Morelos, México. Julio de 2020.

Cuernavaca, Mor., a 09 de julio de 2021.

Dr. Carlos Manuel Astorga Zaragoza
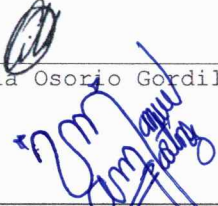Subdirector Académico
P r e s e n t e

At'n: **Dr. Carlos Daniel García Beltrán**
Presidente del Claustro Doctoral
del Departamento De Ing. Electrónica

Los abajo firmantes, miembros del Comité Tutorial del estudiante **Moisés Bulmaro Ramos Martínez**, manifiestan que después de haber revisado el documento de tesis titulado **"Teoría De Identificación De Sistemas Aplicada Al Diagnóstico Médico Y Tratamiento De Señales Bio-Médicas"**, realizado bajo la dirección del **Dr. Víctor Manuel Alvarado Martínez**, el trabajo se ACEPTA para proceder a su impresión.
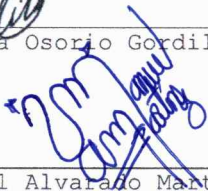
A T E N T A M E N T E

———————————————————
Dr. Gerardo Vicente Guerrero Ramírez
CENIDET

———————————————————
Dra. Gloria Lilia Osorio Gordillo
CENIDET

———————————————————
Dr. Víctor Manuel Alvarado Martínez
CENIDET

———————————————————
Dr. José Francisco Gómez Aguilar
CENIDET

———————————————————
Dr. Christophe Corbier
Université Jean Monnet/IUT de Roanne

———————————————————
Dra. Ma. Guadalupe López López
CENIDET

Reciba un cordial saludo.

c.c.p: M.E. Guadalupe Garrido Rivera / Jefa del Departamento de Servicios Escolares.
c.c.p: Dr. Hugo Estrada Esquivel / Jefe del Departamento de Ingeniería Electrónica.
c.c.p: Expediente.

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca, Mor., 12/julio/2021
No. de Oficio: SAC/95/2021
Asunto: Autorización de impresión de tesis

MOISÉS BULMARO RAMOS MARTÍNEZ
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS
EN INGENIERÍA ELECTRÓNICA
P R E S E N T E

Por este conducto tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado "**Teoría De Identificación De Sistemas Aplicada Al Diagnóstico Médico Y Tratamiento De Señales Bio-Médicas**", ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
"Excelencia en Educación Tecnológica
"Educación Tecnológica al Servicio de México"

EDUCACIÓN

CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
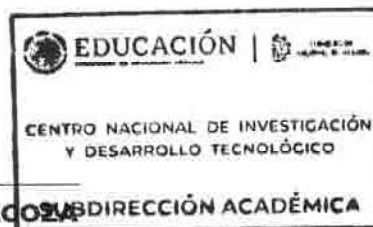
C.c.p. M.E. Guadalupe Garrido Rivera. Jefa del Departamento de Servicios Escolares
Expediente
CMAZ/CHG

Dedicatoria

A mi madre Cecilia y a mi padre Bulmaro.

Agradecimientos

A mis padres, que sin su apoyo incondicional nunca hubiera salido de mi ciudad natal y explorar el mundo como lo he hecho al día de hoy.

A mis directores de tesis, el Dr. Victor y la Dra Guadalupe, por todos los años que han sido pacientes, que se han tomado su tiempo para guiarme y producir temas interesantes de investigación.

Special Thanks to Christophe Corbier for being patient, guiding me, and working with me in a different culture, city, and environment.

A mis revisores de tesis, el Dr. Guerrero, la Dra. Gloria y el Dr. Francisco, por sus lecciones y comentarios en cada final de semestre.

Al Centro Nacional de Investigación y Desarrollo Tecnológico y al Departamento de Ingeniería Electrónica.

Al Consejo Nacional de Ciencia y Tecnologia por el apoyo otorgado a través de la beca de doctorado. Ya que sin éste, no hubiese sido posible llevar acabo tal trabajo de investigación.

Resumen

Este documento presenta el desarrollo y diseño de una nueva metodología basada en la característica promedio de Euler-Poincaré aplicada a las señales biomédicas.

Esta investigación analiza una de las señales biomédicas del cuerpo humano, específicamente las señales eléctricas del corazón. Un electrocardiógrafo adquiere esta señal, y el electrocardiograma (EKG) es su representación. El objetivo es crear un modelo de señales EKG utilizando métodos de identificación de sistemas. Sin embargo, los métodos de identificación clásicos no son factibles para este propósito, por lo que desarrollamos y diseñamos un nuevo enfoque.

La metodología propuesta para el modelado de EKG se divide en seis pasos; el primero es realizar un pretratamiento de la señal utilizando polinomios para filtrar y suavizar el EKG. El segundo es la construcción de un campo aleatorio no gaussiano utilizando la señal del EKG filtrada. El tercero trata de obtener sus propiedades geométricas centradas en sus conjuntos de excursión $A_u(phi, T)$. Posteriormente, el cuarto paso es transformar el conjunto de excursiones en una imagen 2D. Luego, el quinto paso es extraer la característica invariante topológica llamada característica promedio de Euler-Poincaré. Y finalmente, la señal procesada se trata matemáticamente para obtener un nuevo modelo probabilístico explícito, denominado característica promedio de Euler-Poincaré descompuesta (DMEPC).

El método propuesto produce un modelo reducido con una interpretación viable para diferentes afecciones cardíacas investigadas para los datos emitidos a partir de registros Holter. En este número se estudiaron los eventos de infarto de miocardio, desmayo y derrame cerebral.

El segundo caso es la extensión de la aplicación de la metodología antes mencionada a la tarea de clasificación. Con este objetivo en mente, se modificó el paso

final. En su lugar, utilizamos un modelo basado en conjuntos KNN para clasificar la apnea del sueño a partir de las grabaciones de Holter EKG.

Abstract

This document presents the development and design of a new methodology based on the mean Euler-Poincaré characteristic applied to the biomedical signals.

This research analyzes one of the human body's biomedical signals, specifically the electric heart signals. An electrocardiograph module acquires this signal, and the electrocardiogram (EKG) is its representation. The objective is to create a model of EKG signals using systems identification methods. However, classical identification methods are not feasible for this purpose, thus, we developed and design a new approach.

The proposed methodology for EKG modeling is divided into six steps; the first is to lead a pre-treatment of the signal using polynomials to filter and smooth the EKG. The second is building a non-Gaussian random field using the filtered EKG signal. The third deals with getting their geometric properties centered on their excursion set $A_u(\phi, T)$. Afterward, the fourth step is to transform the excursion set to a 2D image. Then, the fifth step is to extract the topological invariant feature called mean Euler-Poincaré characteristic. And finally, the processed signal is mathematically treated to obtain a new explicit probabilistic model, referred to as the decomposed mean Euler-Poincaré characteristic (DMEPC).

The proposed method produces a reduced model with a viable interpretation for different heart conditions investigated for data issued from Holter EKG recordings. In this issue, the myocardial infarction, stroke, and syncopal events were studied.

The second issue is the extend the application of the before stated methodology to classification task. With this goal in mind, the final step was modified. Instead, we used a model based on ensemble KNN to classify sleep apnea from Holter EKG recordings.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A system is a set of elements characterized by many variables that interact and produce observable signals. These signals are called *outputs* and are essential to describe the behavior of the system. External signals influence this behavior. Those that the *observer* can manipulate are called *inputs* and are known signals. Others are called disturbances, which are unknown signals, but their influence on the *outputs* can be observed. A system is described in Figure 1.1.
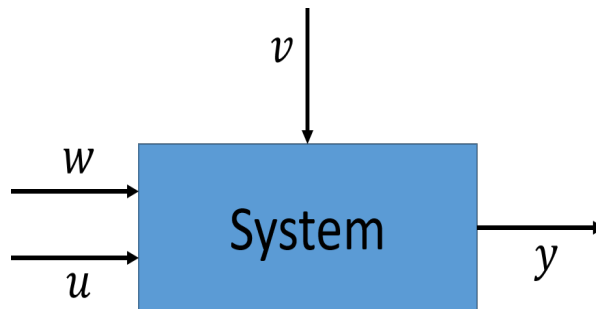
Figure 1.1: A system with an output y, an input u, measured disturbances w and unmeasured disturbances v.

System identification theory is widely used to obtain a mathematical representation of the dynamical behavior of many systems like electrical, mechanical, industrial processes, etc. System models obtained by identification techniques can be formulated from measured data, based on three entities, described by Ljung [1] as:

1. *Data recording.* Imply the input-output data record of the system behavior.

2. Collection of estimated candidate models.

3. Rules to determinate the *best model* in the set, using the data.

System identification theory keeps building a model using input-output data from a process. The procedure has three essential elements: the data set, the collection of models with a defined structure, and the rules for evaluating the model.

Ljung [1] proposed a methodology of identification that starts by considering the model's final goal, followed by the experimental design, ending with the model formulation. Figure 1.2 represents this methodology and is described below:

1. Get Record input-output data set: this step has to do with direct experimentation in the process, it consists of applying a signal in the input, to generate an output, then record the input-output evolution with a certain time frequency.

2. Conduct data pre-processing: most of the time, noise or undesired signal accompany the acquired data. Then, in order to get more manageable data, filtering is the main action.

3. Choose the model structure: in this step the aim is to determine the mathematical identification structure and the adequate order of the process to identify. A priori knowledge and engineering insight make the selection easier.

4. Obtain the "best" model: this step has to do with the identification method. Here, the parameters are estimated attempting to match the model's response to the input-output data acquired from the experiments.

5. Model validation: the last step is applying several tests to the model to determine if it satisfies the chosen criterion. However, the model can have many deficiencies for many causes:

   (a) The numerical procedure failed to find the best model according to our criterion.

   (b) The criterion was not well chosen.

   (c) The model set was not appropriate because it did not contain any "good enough" description of the system.

(d) The data set was not informative enough to provide guidance in selecting good models.
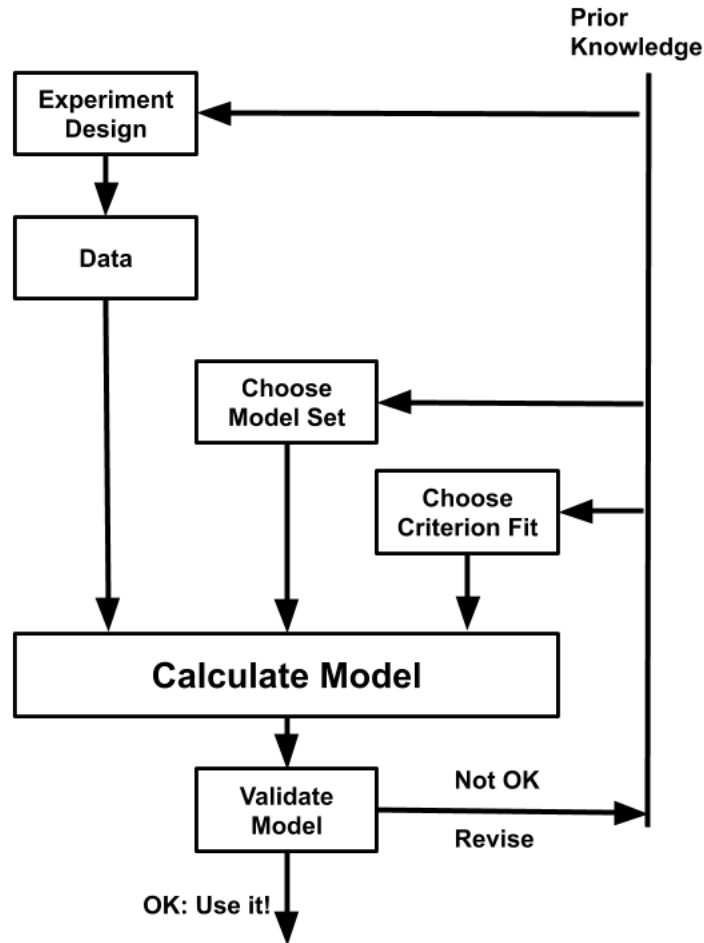
Figure 1.2: The system identification loop. From Ljung [1].

A well suited book on the topic, by Ljung [1], described parametric and nonparametric methods, parameter estimation methods in the prediction error framework, reported frequency domain data and interpretations, various ways to compute estimates, recursive estimation techniques, model validation,and case studies. Also, when input or output signals have static nonlinearities, we can use structures where we have a static non-linear block and a dynamic linear block like NARX, Hammerstein, Wiener, or a combination Hammerstein-Wiener.

Deterministic mathematical models, expressed in terms of an algebraic, differential, partial differential, or integral set of equations, as well as the standard identification procedures outlined earlier, are not always fitting to model some systems.

For biomedical signals, the modeling task needs the best understanding of the biological phenomenon. In the human body, many variables can be observed and measured; in this work, we focus on the heart's electrical signals.

An electrocardiograph can detect the heart's electrical behavior, and the data recorded by this device is called an electrocardiogram (ECG, or EKG). The test detects cardiovascular problems and monitors the state of the heart.

According to the identification methodology, the first step is to record the input and output data set. In the EKG signals, we can not apply a signal in the input to generate an output. So that, our approach considered an EKG as a unique signal.From this point of view, we only have an output.

This thesis introduces a new methodology to model the heart's electrical behavior and a new method to classify cardiac diseases. This approach combines geometry and statistics.

The next sections aim to describe the various studies published about the EKG signals modeling. Then, the objectives of the thesis are given, and the study case is presented.

1.1 Background

The compilation of works taken as a reference for this work was divided into two sections, a section that gathers research related to the treatment of electrocardiographic signals. The other covers discussions about geometry, random fields, and mean Euler-Poincaré characteristic, which are concepts highly related to to the layed out methodology.

1.1.1 Treatment of EKG signals

To construct a mathematical model for representing a dynamic system by equations that involve many parameters is a challenging task. For example, McSharry et al. [2] proposed a dynamical model to generate synthetic EKG signals using ordinary

differential equations. An alternative method to obtain a model is handling input-output data of the system.

The methodology that uses input-output data is named system identification theory. Ljung [1] compiled the principles of this method.

Models for several engineering systems, such as electrical, mechanical, industrial processes, and the automotive industry have widely been conceived by the system identification method. However, for use in biomedicine and medicine, this method is less practiced. A few examples are describing below, Jalali et al. [3] obtained a nonlinear autoregressive exogenous model (NARX) that estimate the heart rate (HR) baroreflex mechanism. The model relates and the HR as output and the arterial blood pressure(ABP) as input. Another example is explained in [4]; the authors obtained a MIMO model for predicting the HR in a Lokomat walking. This model used two outputs, the power of the human and the treadmill speed, and one input, the (HR) variation.

In this thesis we worked with biomedical signals, especially with electrical signals from the heart, recorded by EKG. To handle this kind of signals with the aim to obtain a model via classical methods described by Ljung [1] is a cumbersome task, so we looked for another approach that would enable the development of a model for this type of application.

There exist a lot of biomedical signals, so we focus only on EKG signals. In the literature, we found several methods to analyze them. We summarize them below.

Pan and Tompkins [5] developed an algorithm for detecting the QRS complex of the EKG cycles. This method treated the signal using a bandpass filter and adapted a threshold technique for searching the beats. It was evaluated using a 24 h MIT/BIH arrhythmia database and reached 99.3 percent of the QRS complexes search. Also, we used this algorithm to detect the R-peaks in the methodology proposed.

Heart rate variability (HRV) analysis is a tool to observe the behavior of the heart rate (HR) from unpredictable stimuli. The methods for analysis HRV can be dividing into several domains and features, for example, the time and frequency domain, non-linear and geometrical features. A review of these characteristics is in [6]. HRV analysis is an important tool in cardiology, and various methods use it.

Guo et al. [7] worked with HRV features analysis using a support vector machine

(SVM) and principal component analysis(PDA) for classifying emotions.

In the literature we find several works about the classification of heart diseases, because it is the principal cause of death around the world. Some research articles concern coronary artery disease (CAD), myocardial infarction (MI), fibrillation auricular (FIB), or others.

The principal heart disease in recent years in Mexico is the CAD [8], this affects the supply of blood into the heart muscle because of narrowed heart arteries. When the blood flow is blocked, the heart cells die, thus a heart attack or MI is produced. Some methods to classify the heart diseases are discussing below.

Acharya et al. [9] presented a methodology to classify different types of myocardial infarction using the wavelet transform as a base to extract various no lineal features and analyze them with a KNN classifier. This method is shown in Figure 1.3.

Figure 1.3: Block diagram of the Acharya's method for classifying MI, from Acharya et al. [9].

The Acharya's block diagram method can be divided into three main phases that can be adjusted by applying different approaches. The first phase is the preprocessing for filtering signals, the second is the processing for extract features, and the third is the classification.

The filtering step is primordial to remove the baseline wander and eliminate the noise. In the literature, we can find a decomposition using wavelet functions as Dauchebies6 (db6) in [9],[10],[11], and [12]. Banerjee and Mitra [13] used a discrete wavelet method that decomposed the EKG and made reconstruction to eliminate the noise and baseline wander, and analyze patterns in the EKG for classifying normal and abnormal data. Banerjee's method for filtering is also appropriate (see [14]).

In [15] it was also used a wavelet function to detect the noise and eliminate it from the raw signal, then the features were extracted, and finally, by means of a *k-nearest* neighborhood(KNN) classifier it was possible to detect MI. T wave amplitude, Q wave, and ST level deviation got from 12 lead EKG, constructed the feature vector

to be analyzed by the classifier.

Dolatabadi et al. [16] implemented a bandpass filter to eliminate the noise and detect the R-peaks using the Pan-Tompkins algorithm. Then they extracted HRV features and used PCA and SVM to classify coronary artery disease (CAD).

Sun et al. [17] used a discrete cosine transform over bandpass filter from [18]. Then, they used the ST segment to create a sixth-order polynomial, and took into account 12 leads of the EKG. Later, they created a vector of 72 coefficients. Finally, they used a latent topic multiple instance learning (LTMIL) for automated ECG classification.

Sleep apnea is another disease typically studied with the aid of EKG signals; da Silva Pinho et al. [19] managed the noise with a median filter, then extracted HRV and ECG-derived respiration (EDR) signals features. These are the inputs for an artificial neuronal network with a hyperbolic tangent sigmoid transfer function and linear transfer function in the layers to classify these inputs as sleep apnea or nop-apnea patient.

1.1.2 Geometric approach

This work deals with a method to transform an EKG signal into a geometrical structure. The main idea is to gather random polynomial and topological indicators from NGRF. So the background for this idea is the following.

Random algebraic polynomials (RAP) are relevant to engineering, physics, or economics goals. Bharucha-Reid and Sambandham [20] wrote an excellent work that describes the properties of random orthogonal and trigonometric polynomials. A big deal for using RAP is to identify the number of real zeros. In this interest, Kac [21] gave the formula to get the expected number of real zeros of a RAP. Other works that implement the Kac's formula are, for example, [22], [23], and [24]. In our method, RAP is used as a filter.

Random fields (RF) give a statistical description of complex random patterns of change and relationships from physical data sets [25, 26, 27]. The geometry and properties of RF have been discussed in the literature. Adler and Taylor [25] wrote many estimations for the geometric structure of the excursion set for a Gaussian RF using differential topology and integral geometry. They also wrote about the mean

Euler-Poincaré characteristic in these Gaussian RF.

For the RF, we found that the distribution can be Gaussian or non-Gaussian. Related to the Gaussian cases, we found several works with applications in mathematics, sciences, and engineering [28]. GRF are considered a tool. A noticeable issue to increase their applications has been accessing information from biomedical images. For this interest, Shaikh et al. [29] used Markov irregular fields (MRF) to address biomedical image analysis by image segmentation, object labeling, and 3D vision. Later, Zhang et al. [30] developed a Bayesian framework that introduces an adjustable parameter on a Generalized Adaptive Gaussian Markov random field (GAGMRF) model to adjust the image quality for X-ray Luminescence tomographies. Researches that engage in RF concern also medical analyses. As an example, we refer to the work in [31], where statistical methods assisted morphometric analyses of specific subregions of the brain. In this work, they used GRF to differentiate the shape of the amygdala and the hippocampus of normal subjects facing patients with attention-deficit/hyperactivity disorder (ADHD). Another type of RF usage points to disease mapping. For example, Ferreira and De Oliveira [32] proposed a Bayesian analysis of a Gaussian Markov random field (GMRF) to determine the spatial variability of lip cancer cases in Scotland in five years.

The theory of Adler and Taylor has been extended to Gaussian-related RF (namely, χ^2, $\mathcal{F}$, and t fields) [33]. Later, Adler et al. [34] further study the excursion sets of non-Gaussian random fields (NGRF) holding high levels. High peaks in an RF are due to noise, high frequencies, states, or other behaviors. Adler and Taylor [25] presented a special description of NGRF of the form $f(t) = F(y(t)) = F(y_1(t), \ldots, y_k(t))$ where the $y_j(t)$ are a collection of independent, identically distributed (i.i.d.) GRF. Because some cases are defined in this way, and it is hard to get a function F to encode the behavior of NGRF, we propose to focus on the MEPC.

1.2 Research problem

The biomedical signals extract the behavior of many phenomena from the human body that can be measured. So researchers have been studying these kinds of signals

due to their relevance. Generally, they use images to create models, classify symptoms, or analyze shapes or conditions. So, in the case of the electrical heart behavior, what happens? These signals are time series with regular and random components.

The electrocardiograms are processed using many methods or algorithms based for example on Wavelet, Fourier, or other transformations. A set of characteristics can also be obtained from measures in the time-domain or frequency, stated as geometrical, or no-lineal, among others. With these features, the researchers can found a behavior to many cardiovascular diseases.

Thus, this research attempts to find new methods for modelling this kind of biomedical signals, with a focus on EKG signals. Besides, it is kept in mind the objective to achieve a straightforward form to visualize the information and to detect cardiovascular diseases. As study cases, the EKG of patients with myocardial infarction and sleep apnea will be studied.

1.3 Main Objective

To apply system identification theory tools for the numerical processing of biomedical signals in order to provide support in the diagnosis and clinical treatment

1.3.1 Specific objectives

- To develop a mathematical model for the clinical diagnosis of myocardial infarction in Holter EKG.

- To develop a methodology that allows to obtain a heart disease model.

- To apply a geometric approach based on random field theory for modeling biomedical signals.

1.4 Justification

Cardiovascular diseases (CVDs) are the first reason of global deaths. CVDs are a group of disorders of the heart and blood vessels, and more than 17 million people die for this reason each year [35].

Classical strategies to detect, diagnose or classify the CVDs implement a transformation in the frequency domain, namely Wavelet, Hilbert, and Fourier. This transformation can hide or mislead the researchers data that can be useful to find the CVDs.

In this research, we alternatively used a geometrical and statistical approach to studying this type of phenomena' behavior for modeling the EKG signal and classify myocardial infarction. A new model based on non-Gaussian RF was developed rather than applying classical Gaussian RF theory.

1.5 Thesis outline

This work is divided into six chapters. Chapter 1 contains a brief background and the definition of the objectives. Chapter 2 is a review of the techniques and methodologies encountered in the literature regarding the various aspects of this study. Chapter 3 is dedicated to explain the proposed method from a geometrical and statistical point of view. In Chapter 4, the first research case with application to myocardial infarction disease is presented. A probabilistic model produced by the technique outlined in this work is given as a result of the processing of Holter EKG recordings. In Chapter 5, the second case study is documented. In this challenge, we worked with a KNN classifier for sleep apnea detection. Finally, the conclusions and future works are described in chapter 6.

Chapter 2

Theoretical framework

2.1 Introduction

This section shows the theoretical bases for the new methodology showing in this thesis. It will explain the theory of random fields and the extraction of mean Euler-Poincaré characteristic. Also, the basic concepts for the selected biomedical signal and the KNN classifier.

2.2 Concepts of electrocardiograms

In this section, we talk about the electrocardiogram (EKG) signal. EKG records the electrical activity produced by the heart. The electrical activity in the internal cells can be divided into two actions: the relaxation phase, the cells in the heart muscle are charged, or repolarization occurs, and when they are stimulated electrically, it contracts . The contraction causes a depolarization [36].

An EKG presents these electrical activity effects in the cardiac cells and identifies rhythm, disturbances, and conduction abnormalities. Figure 2.1 shows the principal components in an EKG signal.

P-wave is the first element in a normal EKG; it represents the atrial depolarization. So that, it means the contraction in the two atrial chambers. Then, a pause of 1/10 s is produced to fill the ventricular chambers.

The second element in a normal EKG is the complex QRS. This phase repre-

sents the onset of ventricular contraction. The physical phenomenon of ventricular contraction lasts much longer than the QRS complex. Here, we considered the depolarization of the ventricular chambers and produced the ventricular contraction. This contraction ejects blood from the ventricles and pumps it through the arteries, creating a pulse.

T-wave is the third element in a normal EKG. It represents repolarization or ventricular recovery.
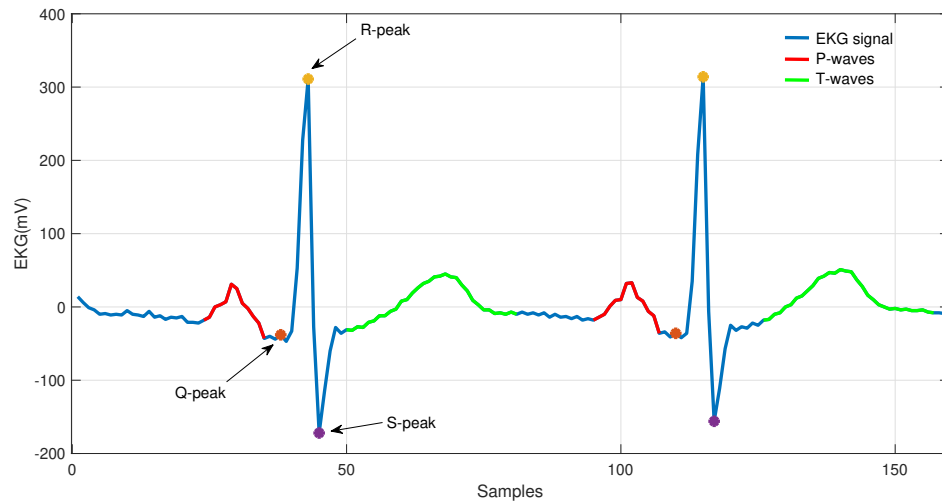
Figure 2.1: Representation of the electrocardiogram or an EKG.

2.3   Random Fields Theory

In this thesis, the random field theory is applying to transform the signal into a geometric structure with a parameter space of dimensionality 2. The principles are addressing below to understand this idea.

Adler and Taylor [25], along with extensions of Worsley [33, 37, 38], combined the geometry of topology with probability and statistics. The resulting theory has become a standard framework for analyzing random fields.

A random field (RF) refers to a stochastic process, habitually taking values in a Euclidean space, and a determined parameter space of dimensionality at least 1.

The concept field refers to the geometric structure with a parameter space. From Adler and Taylor [25] was taken a formal definition,

*Definition **2.1** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and T a topological space.[1] Then a measurable mapping $f : \Omega \to \mathbb{R}^T$ (the space of all real-valued functions on T) is called a real-valued random field. Measurable mappings from Ω to $(\mathbb{R}^T)^d$, $d > 1$, are called vector-valued random fields. If $T \subset \mathbb{R}^N$, we call f an (N, d) random field, and if $d=1$, simply an N-dimensional random field.*

All random fields are separable. This property involves the topological space and the measurable mapping. An example is showed in Figure 2.2. This property involves conditions on both T and f. In particular, an $\mathbb{R}^d$-valued random field f, on a topological space T, is called separable if there exists a countable dense subset $D \subset T$ and a fixed event N with $\mathbb{P}\{N\} = 0$ such that, for any closed $B \subset \mathbb{R}^d$ and open $I \subset T$,

$$\{\omega : f(t, \omega) \in B \forall t \in I\} \Delta \{\omega : f(t, \omega) \in B \forall t \in I \cap D\} \subset N. \tag{2.1}$$

Here, Δ denotes the usual symmetric difference operator, so that

$$A \Delta B = (A \cap B^c) \cup (A^c \cap B) \tag{2.2}$$

where A^c is the complement of A.

---

[1]When we move to manifolds as parameter spaces, we shall replace T by M. The points in M will still be denoted by t.
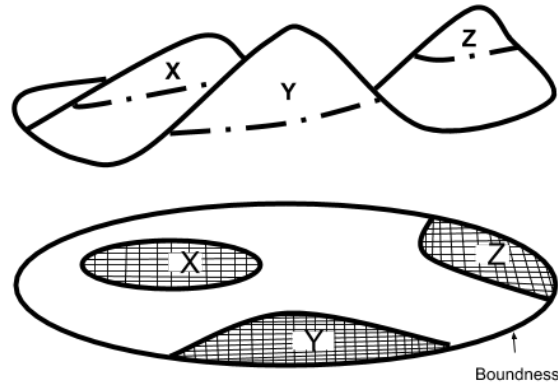
Figure 2.2: It shows a random field at the top, and at the bottom, it presents the concepts of boundedness and field when we used a threshold u to cut., because one property of these fields, it is that can be separated.

Adler and Taylor [25] estimated the geometrical structure of the excursion sets from a Gaussian random field using the integral geometry (IG) and the differential topology (DT). Santaló [39] described quite comprehensive studies of IG, where the geometry, group theory, and probability interact with each other in several problems. On the other hand, Lee [40] discussed the manifolds in the context of differential topology.

A manifold describes objects such as curves and surfaces in $\mathbb{R}^N$. A manifold maintains a dimension, i.e., several "parameters" declare a point. To write a definition for the manifold, first, a description for "locally" in a Euclidean space is necessary.

A topological space T is said to be *locally Euclidean* of dimension n if every point $t \in T$ has a neighborhood that is homeomorphic to an open subset $I \in \mathbb{R}^N$. This neighborhood is called a Euclidean neighborhood of t. The next lemma said that an open subset could replace it with an open ball in $\mathbb{R}^N$ [40].

Lemma **2.1** *A topological space T is locally Euclidean of dimension n if and only if either of the following proporties holds:*

(a) Every point of T has a neighborhood homeomorphic to an open ball in $\mathbb{R}^N$.

(b) Every point of T has a neighborhood homeomorphic to $\mathbb{R}^N$.

They call M a topological N-manifold if M is locally compact Hausdorff space[2] such that for every $t \in M$, there exists an open $U \subset M$ containing t, an open $\tilde{U} \subset \mathbb{R}^N$, and a homeomorphism $\varphi : U \to \tilde{U}$.

With these concepts of DT and IG, we can analyze some types of random fields, especially the Gaussian and Non-Gaussian. Some properties, examples, and excursion probabilities are in the next sections.

2.3.1 Gaussian Random Field

In this section, the theory of Gaussian random fields explains about the excursion probabilities, some examples of the fields like the Brownian sheet, and some properties of IG and DT.

An essential case of random fields is the Gaussian fields. Gaussian processes have rich, detailed, and well-understood theory. That's one reason to choose this type of field first, and the second reason is for the application, it is possible to derive formulas to predict or compare with experiment.

Definitions

A real-valued random variable X is said to be Gaussian (or normally distributed) if it has the density function

$$\varphi(x) \triangleq \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2}, x \in \mathbb{R},$$

for some $m \in \mathbb{R}$ and $\sigma \geq 0$. Also, the mean of X is m and the variance σ^2, and the characteristic function is given by

$$\phi(\theta) = \mathbb{E}\{e^{i\theta X}\} = e^{i\theta m - \sigma^2\theta^2/2}.$$

Also, we can abbreviate as $X \sim N(m, \sigma^2)$. The case $m = 0, \sigma^2 = 1$ is rather particular, and in this situation, we say that X has a standard normal distribution. Generally, if a random variable or process has zero mean, we call it centered.

---

[2]A topological space T is said to be a Hausdorff space if given any pair of distinct points $q_1, q_2 \in X$, there exist neighboorhoods U_1 of q_1 and U_2 of q_2 with $U_1 \cup U_2 = \varnothing$ [40].

An $\mathbb{R}^d$- valued random variable X is said to be multivariate Gaussian if, for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$, the real-valued variable $\langle \alpha, X' \rangle = \sum_{i=1}^{d} \alpha_i X_i$ is Gaussian. In this case there exists a mean vector $m \in \mathbb{R}^d$ with $m_j = \mathbb{E}\{X_j\}$ and a nonnegative definite $d \times d$ covariance matrix C, with elements $c_{ij} = \mathbb{E}\{(X_i - m_i)(X_j - m_j)\}$, such that the probability density of X is given by

$$\varphi(x) = \frac{1}{(2\pi)^{d/2} |C|^{1/2}} e^{-\frac{1}{2}(x-m)C^{-1}(x-m)'} \tag{2.3}$$

where $|C| = \det C$ is the determinant of C. We can write as $X \sim N(m, C)$, or $X \sim N_d(m, C)$ if it is necessary to emphasize the dimension.

Boundedness and continuity

In this section, the conditions for a centered Gaussian field on a parameter space T to be almost indeed bounded and/or continuous are determined, i.e., determining requirements for which

$$\mathbb{P}\left\{\sup_{t \in T} |f(t)| < \infty\right\} = 1 \; or \; \mathbb{P}\left\{\lim_{s \to t} |f(t) - f(s)| = 0, \forall t \in T\right\} = 1.$$

When we talk about continuity, i.e., for the notation $s \to t$ above, T must have some topology, and we assume that (T, τ) is a metric space. Also, that continuity is in terms of the τ-topology.

The first step is to define a new metric d on T by

$$d(s,t) \triangleq \{\mathbb{E}[(f(s) - f(t))^2]\}^{1/2} \tag{2.4}$$

where d is the canonical metric for T and/or f. Also, d is only a pseudometric, because $d(s,t) = 0$ does not necessarily imply that $s = t$.

We are interested in the f continuity so, we suppose that $\sup_T \mathbb{E}\{f_t^2\} < \infty$ and that f is a.s. continuous. Then

$$\lim_{s \to t} d^2(s,t) = \lim_{s \to t} \mathbb{E}\{(f(s) - f(t))^2\} = \mathbb{E}\left\{\lim_{s \to t}(f(s) - f(t))^2\right\} = 0,$$

this exchange of limit and expectation said that since f is Gaussian, boundedness in L^2 implies boundness in all L_p or a.s continuity of f implies the continuity of d.

Here is the lemma establishing the irrelevance of τ to the continuity question [25].

*Lemma **2.2** Let f be a centered Gaussian process on a compact metric space (T, τ). Then f is a.s. continuous with respct to the τ-topology if and only if it is a.s. continuous with respect to the d (psedo)topology. More precisely, with probability one, for all $t \in T$,*

$$\lim_{s:\tau(s,t)\to 0} |f(s) - f(t)| = 0 \Leftrightarrow \lim_{s:d(s,t)\to 0} |f(s) - f(t)| = 0.$$

In the Lemma above the parameter space T was assumed to be compact. From now on, this assumption will always be taken as the case to be analyzed.

The next definition talks about when we need to measure the size of T using a metric entropy.

*Definition **2.2** Let f be a centered Gaussian field on T, and d the canonical metric 2.4. Assume that T is d-compact, and write*

$$B_d(t,\varepsilon) \triangleq \{s \in T : d(s,t) \leq \varepsilon\} \tag{2.5}$$

for the d ball centered on $t \in T$ and of radius ε. Let $N(T, d, \varepsilon) \equiv N(\varepsilon)$ denote the smallest number of such balls that cover T, and set

$$H(T, d, \varepsilon) \equiv H(\varepsilon) = \ln(N(\varepsilon)) \tag{2.6}$$

Then N and H are called the (metric) entropy and log-entropy functions for T (or f). We shall refer to any condition or result based on N or H as an entropy condition/result.

Since we are assuming that T is d-compact, it follows that $H(\varepsilon) < \infty$ for all $\varepsilon > 0$. Furthermore, we define

$$diam(T) = \sup_{s,t \in T} d(s,t) \tag{2.7}$$

then $N(\varepsilon) = 1$ and so $H(\varepsilon) = 0$ for $\varepsilon \geq diam(T)$. Next, we have the main result about boundedness and continuity from Adler and Taylor [25].

Theorem **2.1** *Let f be a centered Gaussian field on a d-compact T, d the canonical metric, and H the corresponding entropy. Then there exists a universal constant K such that*

$$\mathbb{E}\left\{\sup_{t\in T} f_t\right\} \leq K \int_0^{diam(T)/2} H^{1/2}(\varepsilon)d\varepsilon \tag{2.8}$$

With this result, we define the modulus of continuity ω_F of a real valued function F on a metric space (T,τ) as

$$\omega_F(\delta) \equiv \omega_{F,\tau}(\delta) \triangleq \sup_{\tau(s,t)\leq\delta} |F(t) - F(s)|, \ \delta > 0 \tag{2.9}$$

The modulus of continuity of f can be thought of as the supremum of the random field $f_{s,t} = f_t - f_s$ over a certain neighborhood of T x T, in that

$$\omega_{f,\tau}(\delta) = \sup_{\substack{(s,t)\in T \ x \ T \\ \tau(s,t)\leq\delta}} f(s,t). \tag{2.10}$$

And finally, two trivial observations but essential are discussed below.

Observation 1.- If f is a separable process on T then $sup_{t\in T} f_t$ is a well-defined random variable.

Observation 2.- If f is a separable process on T and X a centered random variable, then

$$\mathbb{E}\left\{\sup_{t\in T}(f_t + X)\right\} = \mathbb{E}\left\{\sup_{t\in T} f_t\right\}.$$

Examples

This section shows examples of the Gaussian random field. To start, let f_t be a centered Gaussian process on a compact $T \subset \mathbb{R}^N$ and define

$$p^2(u) = \sup_{|s-t|\leq u} \mathbb{E}\left\{|f_s - f_t|^2\right\} \tag{2.11}$$

where $|\cdot|$ is the usual Euclidean metric. If f is stationary, then

$$p^2(u) = 2 \sup_{|t| \leq u} [C(0) - C(t)]. \tag{2.12}$$

Theorem 2.2 is for the family of random fields in $\mathbb{R}^N$ [25].

Theorem **2.2** *If, for some $\delta > 0$, either*

$$\int_0^\delta (-\ln u)^{\frac{1}{2}} dp(u) < \infty \ \text{or} \ \int_\delta^\infty p\left(e^{-u^2}\right) du < \infty \tag{2.13}$$

then f is continuous and bounded on T with probability one. A sufficient condition for either integral in Eq. (2.13) to be finite is that for some $0 < C < \infty$ and $\alpha, \eta > 0$,

$$\mathbb{E}|f_s - f_t|^2 \leq \frac{C}{|\log|s - t||^{1+\alpha}}, \tag{2.14}$$

for all s,t with $|s-t| < \eta$. Furthermore, there exists a constan K, dependent only on the dimension N, and a random $\delta_0 > 0$ such that for all $\delta < \delta_0$,

$$\omega_f(\delta) \leq K \int_0^{p(\delta)} (-\ln u)^{1/2} dp(u), \tag{2.15}$$

where the modulus of continuity ω_f is taken with respect to the Euclidean metric.

Next, we show an example of the Brownian family of processes. This kind of field needs a Gaussian noise based on a finite ν measure. This noise is a random field $W : \mathcal{T}_\nu \to \mathbb{R}$ such that for all $A, B \in \mathcal{T}_\nu$,

$$W(A) \sim N(0, \nu(A)), \tag{2.16}$$

$$A \cap B = \emptyset \Rightarrow W(A \cup B) = W(A) + W(B) \ a.s., \tag{2.17}$$

$$A \cap B = \emptyset \Rightarrow W(a) \ and \ W(B) \ are \ independent. \tag{2.18}$$

In the property of Eq. (2.17) is mentioned that W is not generally finite, and in Eq. (2.18) said that W has independent increments. With these properties, a new Theorem emerged.

Theorem **2.3** *If $(T, \mathcal{T}, \nu)$ is a measure space, then there exists a real-valued Gaussian noise, defined for all $A \in \mathcal{T}_\nu$, satisfying Eqs. (2.16)-(2.18).*

When a space is restricted in rectangles of the form $[0, t] \subset \mathbb{R}_+^N$, where $t \in \mathbb{R}_+^N = \{(t_1, \ldots, t_N) : t_i \geq 0\}$, then a random field is defined on $\mathbb{R}_+^N$ and denoted

$$W(t) = W([0, t]) \tag{2.19}$$

W_t is called the Brownian sheet on $\mathbb{R}_+^N$, or multiparameter Brownian motion. This field has the next covariance.

$$\mathbb{E}\{W_s W_t\} = (s_1 \wedge t_1) \ x \ \cdots \ x \ (s_N \wedge t_N), \tag{2.20}$$

where $s \wedge t \triangleq \min(s, t)$. A simulation of a Brownian sheet with its contour lines is in the Figure 2.3.
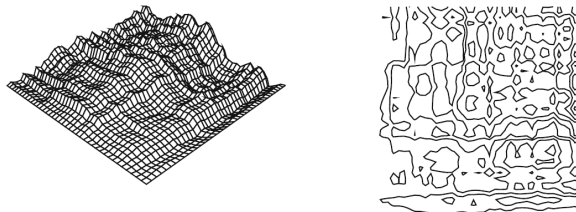
Figure 2.3: A simulated Brownian sheet on $[0, 1]^2$, along with its contour lines at the zero level.

Above in the text, it was mentioned that one of the difficulties in stochastic processes is the excursion probabilities. The next section talks about it.

Excursion sets

A principal concept in the random field theory is the excursion set. A definition from Adler and Taylor [25] is mentioned below,

Definition **2.3** *Let f be a measurable real-valued function on some measurable space, and let T be a measurable subset of that space. Then, for each $u \in \mathbb{R}$,*

$$A_u \equiv A_u(f, T) \triangleq \{t \in T : f(t) \geq u\} \tag{2.21}$$

The excursion set is just the set of points where the field exceeds a fixed threshold value u [33]. Also, it helps to determinate the probability of the RF, that is a difficult problem [25].

$$\mathbb{P}\left\{\sup_{t\in T} f(t) \geq u\right\} \tag{2.22}$$

where f is the random field at a compact space T. Another way of looking at this problem is with the eq. (2.21) of the random field f on the set A over the level u, since

$$\mathbb{P}\left\{\sup_{t\in T} f(t) \geq u\right\} \equiv \mathbb{P}\left\{A_u \neq \varnothing\right\} \tag{2.23}$$

Eq. (2.23) represents an approximation of the probability using the excursion sets but is not the same as the direct method using Eq. (2.21) [41]. Indeed the simple case $[0, T]$ is arduous to compute the excursion probabilities. In this thesis, the case of study is a non-Gaussian random field. The next section talk about the principal concepts for a Non-Gaussian Random Field.

2.3.2 Non-Gaussian Random Field

The term "non-Gaussian" is not well marked and covers a wide range of generalizations. In this section, we talk about the random fields of the form

$$f(t) = F(y(t)) = F(y_1(t), \ldots, y_k(t)), \tag{2.24}$$

where the $y_i(t)$ are a collection of independent, identically distributed Gaussian random fields over a parameter space M, and $F : \mathbb{R}^k \to \mathbb{R}$ is a smooth function.

If we choose $k = 1$ and $F(x) = x$, then we back to the Gaussian case, but other choices are interesting random fields. For example, if we considered that the y_j are centered and of unit variance and consider the following three options for F, where in the third we set $k = n + m$:

$$\sum_{i=1}^{k} x_i^2, \qquad \frac{x_1\sqrt{k-1}}{\left(\sum_{i=2}^{k} x_i^2\right)^{1/2}}, \qquad \frac{m\sum_{i=1}^{n} x_i^2}{n\sum_{i=n+1}^{n+m} x_i^2} \tag{2.25}$$

These random fields are known as χ^2 fields with k degrees of freedom, Student's t field with $k-1$ degrees of freedom, and the F field with n and m degrees of freedom.

For this kind of field, the excursion set of a real-valued non-Gaussian $f = F \circ y$ above a level u, is equivalent to the excursion set for a vector-valued Gaussian y in $F^{-1}[u, \infty]$ under appropriate assumptions on F, is a manifold with piecewise smooth boundary.

$$A_u(f, M) = A_u(F(y), M) = \{t \in M : (F \circ y)(t) \geq u\}$$
$$= \{t \in M : y(t) \in F^{-1}[u, \infty)\} = M \cap y^{-1}(F^{-1}[u, +\infty)). \tag{2.26}$$

An important matter is the study of the mean Lipschitz-Killing curvatures,

$$\mathbb{E}\{\mathcal{L}_i(A_u(f, M))\} = \mathbb{E}\left\{\mathcal{L}_i(M \cap y^{-1}\left(F^{-1}([u, +\infty)))\right)\right\}. \tag{2.27}$$

This function is general, so is the set $F^{-1}([u, +\infty))$, which will generally be a stratified manifold in $\mathbb{R}^k$. But,when we take a suitable subset $D \subset \mathbb{R}^k$, and we have suitable stratified manifolds M, we have

$$\mathbb{E}\left\{\mathcal{L}_i(M \cap y^{-1}(D))\right\} = \sum_{j=0}^{\dim M - i} \begin{bmatrix} i + j \\ j \end{bmatrix} (2\pi))^{-j/2} \mathcal{L}_{i+j}(M)\mathcal{M}_j^\gamma(D), \tag{2.28}$$

where $\begin{bmatrix} n \\ m \end{bmatrix}$ are the combinatorial flag coefficients and the $\mathcal{M}_j^\gamma = \mathcal{M}_j^{\gamma_{\mathbb{R}^k}}$ are the generalized (Gaussian) Minkowski functionals. In Eqs. 2.27) and (2.28) are computed with respect to the Riemannian metric induced on M by the individual Gaussian fields y_j.

The parameter spaces for the random fields in this section are denoted by M and not T. It means that we work with manifolds,and the structure of $F^{-1}([u, \infty))$ is also a manifold.

The first step to compute a general form for expectation in Eq. (2.28), shows that it must be of the form

$$\mathbb{E}\{\mathcal{L}_i(M \cap y^{-1}(D))\} \sum_{j=0}^{\dim M - i} \mathcal{L}_{i+j}(M)\tilde{p}(i, j, D), \qquad (2.29)$$

where $\tilde{p}$ depends on all the parameters given, but not on the distribution of the underlying Gaussian fields y_j.

With Eq. (2.29), we need to establish the form of the function $\tilde{p}$. In order to find it, we can freely choose the manifold M and the Gaussian fields y_j, since the result can not depend on these choices. The choice selected was a specific rotationally invariant field restricted to subsets of a sphere, and a theorem is describing it in [25].

Theorem **2.4** *Let M be an N-dimensional, regular stratified manifold embedded in $\tilde{M}$, also of dimension N, and let D be a regular, stratified manifold in $\mathbb{R}^k, k \leq 1$.*
Let $y = (y_1, \ldots, y_k) : M \to \mathbb{R}^k$ be a vector-valued Gaussian field, the components y_i of which are independent, identically distributed, zero-mean, unit-variance Gaussian fields satisfaying the conditions of Corollary 11.3.5.
Let $\mathcal{L}_j, j = 0, \ldots, N$, be the Lipschitz-Killing measures on M with respect to the metric induced by the y_i as defined in (10.7.1). Then there exist functions $\tilde{p}(i, j, D)$ dependent on all the parameters displayed, but not on the distribution of the underlying Gaussian fields y_j, such that

$$\mathbb{E}\{\mathcal{L}_i(M \cap y^{-1}(D))\} = \sum_{j=0}^{N-i} \mathcal{L}_{i+j}(M)\tilde{p}(i, j, D). \qquad (2.30)$$

2.4   Euler-Poincaré Characteristic

The Euler-Poincaré characteristic is a topological invariant, is a property in a topological space that is invariant under homeomorphisms, describing the topological space's shape or structure regardless of the way it is bent.

The Euler characteristic made by Leonard Euler (1707 - 1783) observed that the number of vertices minus the number the edges plus the number of faces always equals 2 in the ordinary polyhedra [42].

$\chi = V - E + F$ represents the topological invariant called Euler characteristic (EC), and in Fig. 2.4 shows some examples. As you note, the polyhedra observed in Figure 2.4, all have the value of EC equal 2.

Tetrahedron, $\chi = 4 - 6 + 4 = 2$ Cube, $\chi = 8 - 12 + 6 = 2$

Octahedron, $\chi = 6 - 12 + 8 = 2$ Dodecahedron, $\chi = 20 - 30 + 12 = 2$
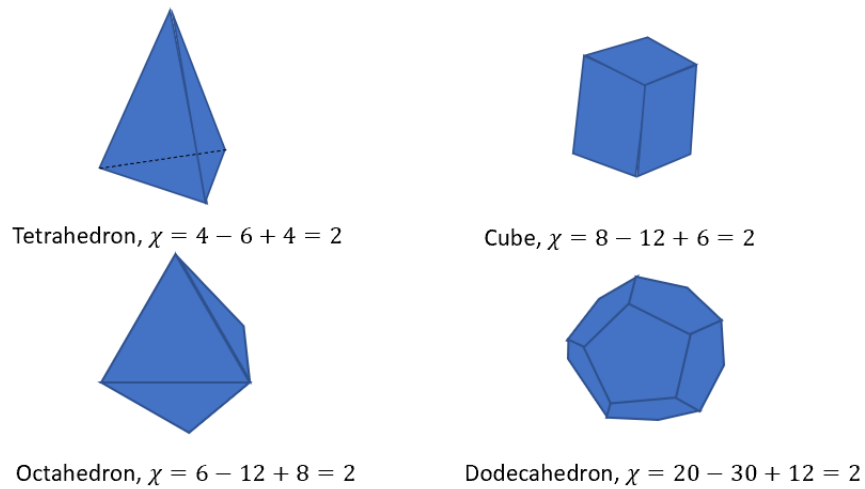
Figure 2.4: Some examples of the values from Euler characteristic $\chi = V - E + F$ in some polyhedra show that all the EC values are equal to 2. Remember that $V =$ vertices, $E =$ edges, and $F =$ faces.

Some applications with images, in particular in the fields of astronomic and treatment of brain images, they used the EC to analyze the expansion of the groups of bodies [cite] and the main signals in the brain [cite], and others used the EC as inputs of a classifier [cite].

In this research, we studied the random fields, and used a centered, stationary Gaussian process on a rectangle $T \in \mathbb{R}^N$ and satisfying the following Corollary 2.1 [25].

Corolary **2.1** *Let f be a centered Gaussian field over a finite rectangle T. If for each $t \in T$, the joint distributions of $(f_i(t), f_{ij}(t))_{i,j=1,\ldots,N}$ are nondegerate, and if for some finite K and all $s, t \in T$,*

$$\max_{i,j} |C_{f_{ij}}(t,t) + C_{f_{ij}}(s,s) - 2C_{f_{ij}}(s,t)| \leq K |\ln |t-s||^{-(1+\alpha)}, \tag{2.31}$$

then the sample functions of f are, with probability one, Morse functions [3] *over*

---

[3]The following characteristics can define Morse functions:

1. f is C^2 on an open neighborhood of T.

2. The critical points of $f_{|\partial_k T}$ are nondegenerate for all $k = 0, \ldots, N$.

T.

As usual, C denotes the covariance function, and ν the spectral measure. Then f has variance $\sigma^2 = C(0) = \nu(\mathbb{R}^N)$. The second-order spectral moments are in this section as

$$\lambda_{ij} = \int_{\mathbb{R}^N} \lambda_i \lambda_j \nu(d\lambda). \tag{2.32}$$

Λ will denote a N x N matrix of these moments—also, the differentiation via subscripts, so that $f_i = \partial f/\partial t_i$, $f_{ij} = \partial^2 f/\partial t_i \partial t_j$, etc., we have

$$\mathbb{E}\{f_i(t)f_j(t)\} = \lambda_{ij} = -C_{ij}(0). \tag{2.33}$$

Therefore Λ is also the variance-covariance matrix of ∇f. The random field and its second-order derivatives are, correlated, and

$$\mathbb{E}\{f(t)f_{ij}(t)\} = -\lambda_{ij} \tag{2.34}$$

Finally, the N x N Hessian matrix (f_{ij}) by $\nabla^2 f$ was denoted.

Lemma **2.3** *Let f y T be describe above, and set*

$$\mu_k = \#\{t \in T : f(t) \geq u, \nabla f(t) = 0, index(\nabla^2 f) = k\} \tag{2.35}$$

then for all $N \geq 1$,

$$\mathbb{E}\left\{\sum_{k=0}^{N}(-1)^k \mu_k\right\} = \frac{(-1)^N |T| |\Lambda|^{1/2}}{(2\pi)^{(N+1)/2}\sigma^N} H_{N-1}\left(\frac{u}{\sigma}\right) e^{-u^2/2\sigma^2} \tag{2.36}$$

where H_{N-1} is the Hermite polynomial [4]. This result depends on the covariance

---

3. $f_{|\partial_k T}$ has no critical points on $\bigcup_{j=0}^{k-1} \partial_j T \ \forall \ k = 1, \ldots, N$.

[4]k-th hermite polynomial is the function

$$H_n(x) = n! \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{(-1)^j x^{n-2j}}{j!(n-2j)!2^j}, \ n \geq 0, x \in \mathbb{R}, \tag{2.37}$$

where $\lfloor a \rfloor$ is the integer greater or equal than a.

f and the second moment spectral.

For calculate the Euler-Poincaré characteristic, we need

$$T = \prod_{i=1}^{N} [0, T_i], \tag{2.38}$$

a rectangle on $\mathbb{R}^N$.

*Theorem **2.5** Let f be a centered. stationary Gaussian field on a rectangle $T \in \mathbb{R}^N$. For real u, let $A_u = A_u(f, T) = \{t \in T : f(t) \geq u\}$ be an excursion set, and let φ be the Euler characteristic. Then*

$$\mathbb{E}\{\varphi(A_u)\} = e^{-u^2/2\sigma^2} \sum_{k=1}^{N} \sum_{J \in \mathcal{O}_k} \frac{|J||\Lambda_J|^{1/2}}{(2\pi)^{(k+1)/2}\sigma^k} H_{k-1}\left(\frac{u}{\sigma}\right) + \Psi\left(\frac{u}{\sigma}\right) \tag{2.39}$$

Because of the *manifolds* and our proposes. We assume that f is isotropic. And with this assumption we have the next corollary.

*Corolary **2.2** In addition to the conditions of Theorem 2.5, lef f be isotropic and T the cube $[0, T]^N$. If λ_2 denotes the variance of f_i, then*

$$\mathbb{E}\{\varphi(A_u)\} = e^{-u^2/2\sigma^2} \sum_{k=1}^{N} \frac{\binom{N}{k} T^k \lambda_2^{k/2}}{(2\pi)^{(k+1)/2}\sigma^k} H_{k-1}\left(\frac{u}{\sigma}\right) + \Psi\left(\frac{u}{\sigma}\right) \tag{2.40}$$

2.4.1   Examples of the Mean Euler-Poincaré characteristic

Now, with Eq. (2.40), we can observe the case when N=1, so that T is simply an interval [0, T]. Then, using the definition of the Hermite polynomials given by Eq. (2.37), the mean Euler-Poincaré characteristic for this case is

$$\mathbb{E}\{\varphi(A_u(f, [0, T]))\} = e^{-u^2/2\sigma^2} \frac{T\lambda_2^{1/2}}{2\pi\sigma} + \Psi(u/\sigma) \tag{2.41}$$

The Eq. (2.41) show the Rice formula[25]. Figure 2.5 gives two examples, with $\sigma^2 = 1$, $\lambda_2 = 200$, and $\lambda_2 = 1000$.
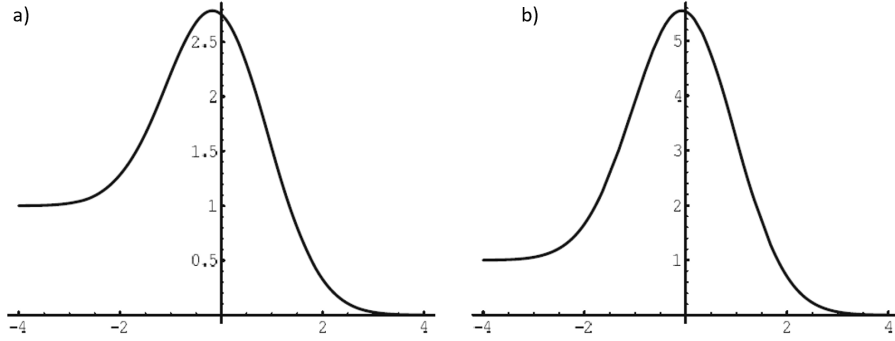
Figure 2.5: $\mathbb{E}\{\varphi(A_u)\} : N = 1$. a) $\lambda_2 = 200$, b) $\lambda_2 = 1000$ with $\sigma^2 = 1$. From [25].

In this case, when $u \to -\infty$, we have $\mathbb{E}\{\varphi(A_u)\} \to 1$. The excursion set geometry behind this is simple. Once $u < inf_T f(t)$ we have $A_u \equiv T$, and so $\varphi(A_u) = \varphi(T)$, which in the current case is 1. This is a general phenomenon, independent of dimension or the topology of T.

Figure 2.6 shows the case when $N = 2$, with the following conditions $\sigma^2 = 1, \lambda_2 = 200$, and $\lambda_2 = 1000$.

$$\mathbb{E}\{\varphi(A_u)\} = e^{-u^2/2\sigma^2} \left[\frac{T^2 \lambda_2}{(2\pi)^{3/2}} u + \frac{2T\lambda_2^{1/2}}{2\pi} \right] + \Psi(u) \qquad (2.42)$$
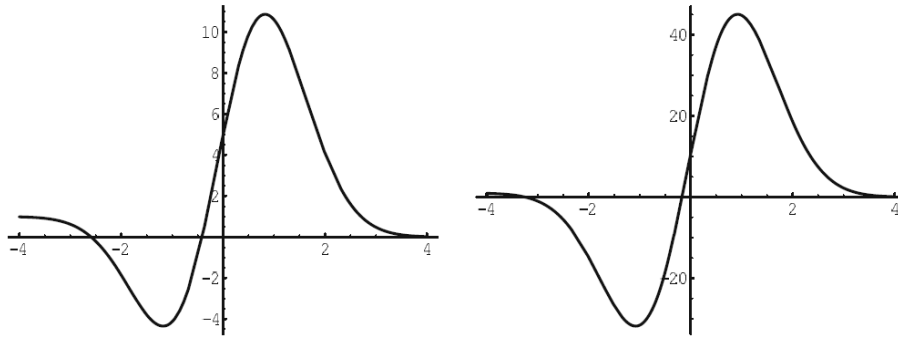
Figure 2.6: $\mathbb{E}\{\varphi(A_u)\} : N = 2$. a) $\lambda_2 = 200$, b) $\lambda_2 = 1000$. From [25].

In cases with one and two dimensions, some analogies are similar in both cases, and they are:

1. Some different power series (one in T, one in u, and one in $\sqrt{\lambda_2}$) represents the expression before the exponential term.

2. The geometric meaning of the negative values from Eq. (2.42) are worth understanding. They are due to the excursion sets, in the mean, have more holes than connected components for negative values of u.

3. The impact of the spectral moments is not clear in higher dimensions as it in one. With the Brownian sheet, we can observe the behavior of the spectral moments. When the spectral moments are significant, it generates large numbers of small islands (or lakes, depending on the level of the excursion set), leading to more significant variation in the values of the $\mathbb{E}\{\varphi(A_u)\}$.

In three dimensions, there are a number of different power series appearing in the Eq. (2.39). There are positive peaks because A_u is primarily composed of several only connected components for large u and primarily of single holes for negative u.

2.5 Machine learning classifiers

Machine learning algorithms don't work for each problem. Sometimes identifying the correct algorithm is a heuristic task. Knowing the principal features of each type of algorithms can help decide which ones to try first and understand your making tradeoffs. Table 2.1 shows a list of features of some popular classifiers.

In these classifiers, the method consists of predicting a class from a set of data points. We divided this set into two: the training data and the validation data. So for the training data, we have targets, and the work of the classifier is to learn the classes from this set. Then, the classifier predicts the categories for the validation data.

Below, we presented the main characteristic of each type of classifier mentioned in table 2.1.

2.5.1 Support Vector Machine(SVM):

Vapnik developed this method. This algorithm is a supervised learning model with associated learning algorithms that analyze data for classification and regression

analysis.

A set of points X, a set X in a significant space $\mathcal{T}$, each $xinX$ has one of two possible categories, the algorithm based in SVM builds a model that predicts a new point x is in one or other class.

The SVM searches a hyperplane that separates the best possible way the points from one category to the other, then this new set is projected in a superior dimensionality space.

The dimension of a dataset can be transformed by combining or modifying any of its dimensions. For example, you can change a two-dimensional space into one of three by applying the following function:

$$f(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{2.43}$$

Eq. (2.43) is just one of the infinite possible transformations, how to know which is the right one? We have tools called kernels come into play. A kernel (K) is a function that returns the result of the dot product between two vectors carried out in a new dimensional space different from the original space in which the vectors are found. Although the mathematical formulas used to solve the optimization problem have not gone into detail, it contains a dot product.

If a kernel replaces this dot product, the support vectors (and the hyperplane) in the dimension corresponding to the kernel are obtained directly. It is often known as the kernel trick because, with only a slight modification of the original problem, the result can be obtained for any dimension thanks to the kernels. There are many different kernels. Some of the most used are:

$$\textbf{Lineal Kernel:} \quad K(x, x') = x \cdot x' \tag{2.44}$$

$$\textbf{Polynomial Kernel:} \quad K(x, x') = (x \cdot x' + c)^d \tag{2.45}$$

when $d = 1$ and $c = 0$, the polynomial kernel's result is the same as the linear kernel. If $d > 1$, limits of no lineal decision appears, generating no lineality when d grows. It is not recommendable to use $d > 5$ because of overfitting.

$$\textbf{Gaussian Kernel:} \quad K(x, x') = \exp(-\gamma ||x - x'||^2) \quad (2.46)$$

Where γ controls the kernel's behavior, when it is small, the final model is similar to the linear kernel; when it is large, the model's flexibility grows.

2.5.2 K-th nearest neighbor (KNN):

The k-nearest neighbors algorithm (KNN) is a non-parametric classification method developed by Fix and Hodges in 1951. This method is applied to classify data and regression cases. In both ideas, the input consists of the k nearest training examples in the data set. The output depends on the application of the KNN. It could be for classifying or regression, but we focus on the classification method in this research.

The KNN classifier categorizes unlabeled observations by assigning them to the class of the most similar labeled examples. The features of the unlabeled marks are collected from the training and test set. To simplify the visualization, we plot two characteristics at the same time on a two-dimension plot. If there are many predictors, we can extend the labeled examples to incorporate any number of features.

Zhang [43] mentions a KNN example to classify the sweet potato in the middle of three groups: fruit, grain, and vegetables. In his example, he choose the four nearest kinds of food, they are apple, green bean, lettuce, and corn. Because the vegetable wins the most votes, sweet potato is assigned to the class of vegetable. You can see that the critical concept of KNN is easy to understand.

We need to know two main concepts for solving the above example; The first is how to calculate the distance between sweet potato and other kinds of food. The second concept is the parameter k which decides how many neighbors will be chosen for the KNN algorithm.

How we calculate the distance in the KNN algorithm? We need to use a Euclidean distance which can be calculated with the following equation:

$$D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_N - q_N)^2} \quad (2.47)$$

where p and q are subjects to be compared with n features. There are also other methods to calculate distance such as Manhattan distance [44].

The k is another central concept. A large k reduces the impact of variance caused by random error but runs the risk of ignoring small but essential patterns. The key to choose an appropriate k value is to strike a balance between overfitting and underfitting [45].

Performance of the KNN model

The KNN algorithm assigns a category to observations in the test dataset by comparing them to the training dataset observations. Because we know the actual class of observations in the test dataset, we can evaluate the KNN model's performance with this information. One of the most commonly used parameter is the average accuracy that is defined by the following equation:

$$Accuracy = \sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i}/l \tag{2.48}$$

where TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative.

Sensitivity and specificity

Sensitivity is a measure of the proportion of positives that correctly identify positive points. Specificity is a measure of the balance of genuinely negative negatives. These measures are commonly used to measure the performance in the test data. If we have a perfect fit model that can predict 100% accuracy, these measures' sensitivity and specificity are 100%. These measures are calculated separately for each class. The equations are the follows.

$$Sen_i = \frac{TP_i}{(TP_i + FN_i)} \tag{2.49}$$

$$Sp_i = \frac{TN_i}{(TN_i + FP_i)} \tag{2.50}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative. The index i indicates category.

Table 2.1: Table of different classification algorithms and some characteristics [46]

Algorithm	Prediction Speed	Training Speed	Memory Usage	Required Tuning	General Assessment
Logistic Regression (Linear SVM)	Fast	Fast	Small	Minimal	Good for small problems with linear decision boundaries
Decision Trees	Fast	Fast	Small	Some	Good generalist, but prone to overfitting
(Nonlinear)SVM	Slow	Slow	Medium	Some	Good for many binary problems, and handles high- dimensional data well
Nearest Neighbor	Moderate	Minimal	Medium	Minimal	Lower accuracy, but easy to use and interpret
Naïve Bayes	Fast	Fast	Medium	Some	Widely used for text, including spam filtering
Ensambles	Moderate	Slow	Varies	Some	High accuracy and good performance for small to medium sized datasets
Neural Network	Moderate	Slow	Large	Lots	Popular for classification, compression, recognition, and forecasting

Chapter 3

Decomposed mean Euler-Poincaré characteristic methodology

This chapter is published as:

M. Ramos-Martinez, C. Corbier, V. M. Alvarado and G. L. Lopez, "Decomposed Mean Euler-Poincaré Characteristic Model for a Non-Gaussian Physiological Random Field," in *IEEE Access*, vol. 9, pp. 21180-21191, 2021.[47].

In this chapter, we focus on the algorithms to extract a model from the EKG signals. Figure 3.1 shows a general view of the method; the first point is to generate the polynomials and then create a non-Gaussian random field using them after extracting the Euler-Poincaré characteristic, and finally describe it mathematically.
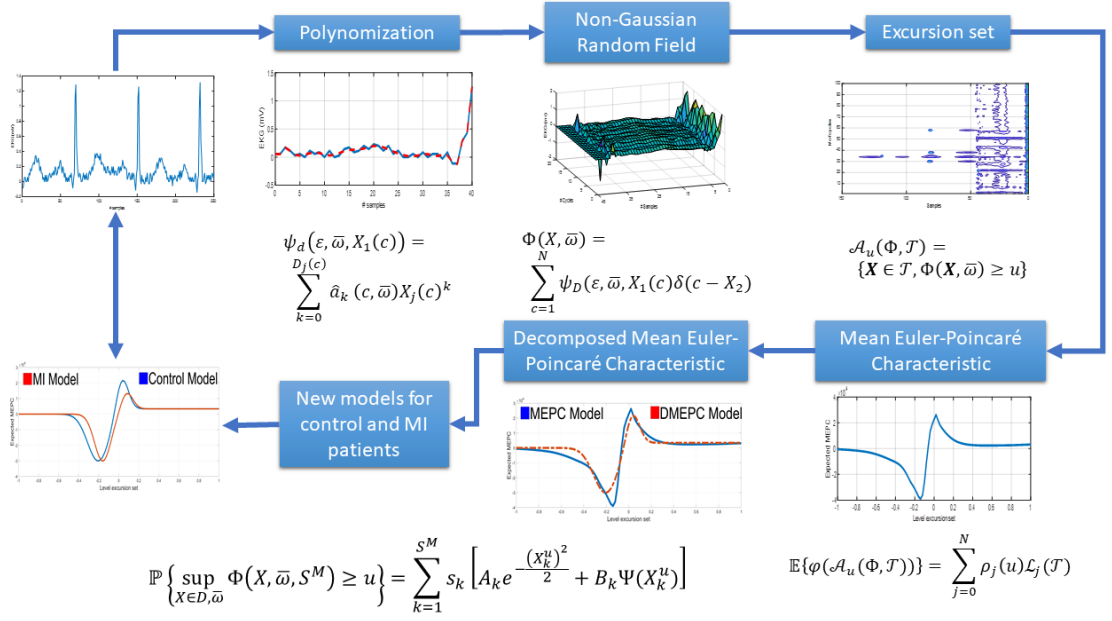
Figure 3.1: The resume of the methodology.

Essentially, the methodology can be divided into three main sections:

- The first section, the pre-processing part, applies a filter to reduce noise and smooth the signal.

- The second section, the processing part, transforms the filtered signal into a geometrical structure. In the literature, we can found that the time-series signal is converted to a frequency signal using a Fourier transformation or any related transform in this section.

- In the third section, the post-processing part, after the transformation, we get a feature that identifies hide hints of the time-series into the geometrical structure or the literature cases into the power spectral signal.

The methodology developed in this research have five steps, enumerated below:

1. Polynomization: the first step, using the random algebraic polynomial (RAP) as a filter. RAP is the filtered signal of the EKG raw signal. Here is the pre-processing part of the methodology.

2. Building the non-Gaussian random field: using the filtered signal, we stacked all the filtered cardiac cycles and made this random field. In other words, we constructed the geometrical structure using the filtered signals.

3. Excursion sets: a principal concept in the random field theory, is simply the set of points where the RF exceeds a fixed threshold value $u \in \mathbb{R}$. These belong to the RF.

4. Getting the mean Euler-Poincaré characteristic: using the excursion sets level, we extract this topological feature that indicates the number of connected components in the excursion set minus the number of holes.

5. Decomposed mean Euler-Poincaré characteristic: Once we have the MEPC of each EKG signal that belongs to a patient, we have an MEPC from an NGRF that doesn't have an explicit formula that describes this behavior. Adler and Taylor [25] presented a description of the form $f(t) = F(y(t)) = F(y_1(t), y_2, \cdots, y_k(t))$ where the $y(t)$ are a collection of identically, independent, and distributed (i.i.d.) GRF. It isn't easy to get a function F to encode the behavior of NGRF, so that we introduce an approach focused on the MEPC. We propose a new characteristic through a decomposition.

3.1   Polynomization

Polynomization is a modeling based on RAP denoted $\psi(\bullet, X_1)$ of a EKG cycle c $(c = 1...N)$ where, for a fixed threshold, each of them can be decomposed over two intervals $I_1(c) = \begin{bmatrix} 1 & X^R(c) \end{bmatrix}$ and $I_2(c) = \begin{bmatrix} X^R(c) & X^E(c) \end{bmatrix}$ with $X^R(c)$ index-time of the R wave. Let $\{a_k(c, \overline{\omega})\}_{k=1}^{D(c)}$ be a sequence of i.i.d. random variables and $\psi_B(\mathcal{E}, \overline{\omega}, X_1(c))$ be a basic RAP of order $D(c)$ defined by

$$\psi_B(\mathcal{E}, \overline{\omega}, X_1(c)) = \sum_{k=0}^{D(c)} a_k(c, \overline{\omega}) X_1(c)^k, \qquad (3.1)$$

with $\mathcal{E} = (\overline{\Omega}, \overline{\Sigma}, P)$ a complete probability space where $\overline{\Omega}$ denotes the sample space, $\overline{\Sigma}$ a σ-algebra on $\overline{\Omega}$ and P a probability measure on $\overline{\Sigma}$. For each interval $I_m(c)$ $(m = 1, 2)$ define a RAP as $\psi_D^m(\mathcal{E}, \overline{\omega}, X_1(c)) = \psi_B(\mathcal{E}, \overline{\omega}, X_1(c)) \mathbf{1}_{I_m(c)}$ where $\mathbf{1}_X$

is the unit function over X. The polynomization over one cycle c can then be written as

$$\psi_D(\mathcal{E}, \overline{\omega}, X_1(c)) = \sum_{k=0}^{D_1(c)} a_k^{I_1}(c, \overline{\omega})(X_1^{I_1}(c))^k + \sum_{k=0}^{D_2(c)} a_k^{I_2}(c, \overline{\omega})(X_1^{I_2}(c))^k.$$

3.2 Modeling Non-Gaussian Random field

Let $\Phi(\boldsymbol{X}, \overline{\omega})$ be a NGRF at $\boldsymbol{X}$ with $\boldsymbol{X} = [X_1(c) \ \ X_2]^T$ where $X_1(c) \in I_1(c) \cup I_2(c)$. Let $\mathcal{E} = (\overline{\Omega}, \overline{\Sigma}, P)$ be a complete probability space and $\mathcal{T}$ a topological space, where $\overline{\Omega}$ denotes the sample space, $\overline{\Sigma}$ a σ-algebra on $\overline{\Omega}$ and P a probability measure on $\overline{\Sigma}$. Then a measurable mapping $\Phi : \overline{\Omega} \to \mathbb{R}^{\mathcal{T}}$ is called a real-valued random field. Here $\boldsymbol{X} \in \mathcal{T} \subset \mathbb{R}^2$ and Φ is called an 2-dimensional NGRF over N cycles defined as

$$\Phi(X_1, X_2, \overline{\omega}) = \Phi(\boldsymbol{X}, \overline{\omega}) = \sum_{c=1}^{N} \psi_D(\mathcal{E}, \overline{\omega}, X_1(c))\delta(c - X_2), \qquad (3.2)$$

where δ is the Dirac's distribution. Eq.(3.2) means that each EKG signal for a fixed threshold is decomposed into basic RAP $\psi_D(\mathcal{E}, \overline{\omega}, X_1(c))$ to form a NGRF according to the coordinate X_2.

3.2.1 Non-Gaussian random field

The most important concept in the random field theory is named *excursion set*. Let $\Phi(\boldsymbol{X}, \overline{\omega})$ be a NGRF, $\boldsymbol{X} \in \mathcal{T} \subset \mathbb{R}^2$, defined inside a set $\mathcal{T}$. The excursion set is a geometrical object defined as

$$\mathcal{A}_u(\Phi, \mathcal{T}) = \{\boldsymbol{X} \in \mathcal{T}, \Phi(\boldsymbol{X}, \overline{\omega}) \geq u\}. \qquad (3.3)$$

Thus $\mathcal{A}_u$ of $\Phi(\boldsymbol{X}, \overline{\omega})$ above a threshold u is the set of points in $\mathcal{T} \subset \mathbb{R}^2$ where $\Phi(\boldsymbol{X}, \overline{\omega})$ exceeds u. Remember that stratified manifolds $\mathcal{T}$ in $\mathbb{R}^2$ are basically sets that can be partitioned into the disjoint union of manifolds as $\mathcal{T} = \cup_{j=0}^{dim\mathcal{T}} \partial_j \mathcal{T}$ where each stratum $\partial_j \mathcal{T}$ is itself a disjoint union of a number of j-dimensional manifolds. Here non-Gaussian properties of Φ involve to consider a class of generalization of random fields of the form

$$\Phi(\boldsymbol{X}, \overline{\omega}) = F\left(\phi_*(\boldsymbol{X}, \overline{\omega})\right) = F\left(\phi_*^1(\boldsymbol{X}, \overline{\omega}), ..., \phi_*^m(\boldsymbol{X}, \overline{\omega})\right), \tag{3.4}$$

are defined, where $\phi_*^j(\boldsymbol{X}, \overline{\omega})$ are a collection of i.i.d. Gaussian random fields (GRF), all defined over a topological space $\mathcal{T}$ and $F : \mathbb{R}^m \to \mathbb{R}$ is a smooth function to be picewise C^2, along with appropriate side conditions. The excursion set of a real-valued non-Gaussian $\Phi = F \circ \phi_*$ above a level u is equivalent to the excursion set for a vector-valued Gaussian ϕ_* in $F^{-1}[u, \infty)$ and given by

$$\mathcal{A}_u(\Phi, \mathcal{T}) = \mathcal{A}_u(F \circ \phi_*, \mathcal{T}) = \{\boldsymbol{X} \in \mathcal{T}, (F \circ \phi_*)(\boldsymbol{X}, \overline{\omega}) \geq u\}. \tag{3.5}$$

In our framework let $\Sigma_s = (s_1, ...s_{S^M})$ be the set of *signatures* with s_k the associated signed unit weight (SUW) for each sub GRF $\phi^{s_k}(\boldsymbol{X}, \overline{\omega})$ and $S^M = \sum_{k=1}^M (-1)^k s_k$. Let us define the NGRF given by Eq.(3.2) as

$$c\Phi(\boldsymbol{X}, \overline{\omega}, S^M) = (s_1 \phi^{s_1}(\boldsymbol{X}, \overline{\omega}), ..., s_{S^M} \phi^{s_{S^M}}(\boldsymbol{X}, \overline{\omega})), \tag{3.6}$$

where $s_m \phi^{s_m}(\boldsymbol{X}, \overline{\omega})$ $(m = 1...S^M)$ is a collection of i.i.d. *signed* GRF with $\boldsymbol{X} \in \mathbb{R}^2, \overline{\omega} \in \overline{\Omega}$ and $s_{2n} = +1, s_{2n+1} = -1$.

3.3 Mean Euler-Poincaré characteristic

Now denote the EPC of a set $\mathcal{A}_u(\Phi, \mathcal{T})$ by $\varphi(\mathcal{A}_u(\Phi, \mathcal{T}))$. For numerous trials, consider that $\mathbb{E}\left\{\varphi(\mathcal{A}_u(\Phi, \mathcal{T}))\right\}$ is computable. Therefore the random fields theory impose some regularity conditions on $\Phi(\boldsymbol{X}, \overline{\omega})$ to ensure both, that its realisations are smooth and the boundary $\partial\mathcal{T}$ is smooth. Then $\mathcal{T}$ is a regular C^2 domain in a compact t subset of $\mathbb{R}^2$ bounded by a regular 1-dimensional manifold $\partial\mathcal{T}$ of classe C^2. Consider $\Phi(\boldsymbol{X}, \overline{\omega})$, $\boldsymbol{X} = [X_1, X_2] \in \mathbb{R}^2$ a stationary non isotropic random field and $\dot{\Phi}_j(\boldsymbol{X}, \overline{\omega}) = \dfrac{\partial\Phi(\boldsymbol{X}, \overline{\omega})}{\partial X_j}$, $\ddot{\Phi}_{jk}(\boldsymbol{X}, \overline{\omega}) = \dfrac{\partial^2\Phi(\boldsymbol{X}, \overline{\omega})}{\partial X_j \partial X_k}$, $j, k = 1, 2$. The moduli of continuity of $\dot{\Phi}_j(\boldsymbol{X}, \overline{\omega})$ and $\ddot{\Phi}_{jk}(\boldsymbol{X}, \overline{\omega})$ inside $\mathcal{T}$ are given by $\omega_j(h) = \sup\limits_{\|\boldsymbol{X}-\boldsymbol{Y}\|<h, \overline{\omega}} |\dot{\Phi}_j(\boldsymbol{X}, \overline{\omega}) - \dot{\Phi}_j(\boldsymbol{Y}, \overline{\omega})|$ and $\omega_{jk}(h) = \sup\limits_{\|\boldsymbol{X}-\boldsymbol{Y}\|<h, \overline{\omega}} |\ddot{\Phi}_{jk}(\boldsymbol{X}, \overline{\omega}) - \ddot{\Phi}_{jk}(\boldsymbol{Y}, \overline{\omega})|$, respectively. To ensure that realisations of $\Phi(\boldsymbol{X}, \overline{\omega})$ are sufficiently smooth, consider the following conditions

- C1: $\mathbb{P}\left(\max_{j,k}\{\omega_j(h), \omega_{jk}(h)\} > \epsilon\right) = o(h^N)$ as $h \downarrow 0$,

- C2: Hessian matrix $\ddot{\boldsymbol{\Phi}}$ of $\ddot{\Phi}_{jk}$ has finite variance conditional on $\left(\Phi, \dot{\boldsymbol{\Phi}}\right)$ with $\dot{\boldsymbol{\Phi}}$ grandient of $\dot{\Phi}_j$,

- C3: the density of $\left(\Phi, \dot{\boldsymbol{\Phi}}\right)$ is bounded above, uniformly for all $\boldsymbol{X} \in \mathcal{T}$.

At a point $\boldsymbol{X} \in \partial\mathcal{T}$, let $\dot{\Phi}_\perp$ be the gradient of Φ in the direction of the inside normal to $\partial\mathcal{T}$, let $\dot{\boldsymbol{\Phi}}_{\boldsymbol{T}}$ be the gradient 1-vector in the tangent plane to $\partial\mathcal{T}$, let $\ddot{\boldsymbol{\Phi}}_{\boldsymbol{T}}$ be the 1×1-Hessian matrix in the tangent plane to $\partial\mathcal{T}$ and let $\boldsymbol{r}$ be the 1×1 inside curvature matrix of $\partial\mathcal{T}$. Let *sign* be the sign function. Consider the notation of Knuth [48] where a logical expression in parentheses takes the value one if true and zero otherwise. Under conditions C1-C3, then the EPC is

$$
\begin{aligned}
\varphi(\mathcal{A}_u\left(\Phi, \mathcal{T}\right)) = &\sum_{\boldsymbol{X} \in \mathcal{T}} (\Phi \geq u)\left(\dot{\boldsymbol{\Phi}} = \boldsymbol{0}\right)\left[\det(-\ddot{\boldsymbol{\Phi}})\right] \\
&+ \sum_{\boldsymbol{X} \in \partial\mathcal{T}} (\Phi \geq u)\left(\dot{\boldsymbol{\Phi}}_{\boldsymbol{T}} = \boldsymbol{0}\right)\left(\dot{\Phi}_\perp <, 0\right) sign\left[\det(\ddot{\boldsymbol{\Phi}}_{\boldsymbol{T}} - \dot{\Phi}_\perp \boldsymbol{r})\right],
\end{aligned}
\tag{3.7}
$$

with probability one. In $\mathbb{R}^2$ integral geometry defines $\varphi(\mathcal{A}_u\left(\Phi, \mathcal{T}\right))$ as (the number of connected components)-(the number of holes) in $\mathcal{A}_u\left(\Phi, \mathcal{T}\right)$. Moreover remember that the expectation of $\varphi(\mathcal{A}_u\left(\Phi, \mathcal{T}\right))$ for multiple realisations is

$$
\begin{aligned}
\mathbb{E}\left\{\varphi(\mathcal{A}_u\left(\Phi, \mathcal{T}\right))\right\} = &\int_{\mathcal{T}} \mathbb{E}\left\{(\Phi \geq u)\det(-\ddot{\boldsymbol{\Phi}})\dot{\boldsymbol{\Phi}} = \boldsymbol{0}\right\}\theta(\boldsymbol{0})d\boldsymbol{X} \\
&+ \int_{\partial\mathcal{T}} \mathbb{E}\left\{(\Phi \geq u)(\dot{\Phi}_\perp < 0)\det(-\ddot{\boldsymbol{\Phi}}_{\boldsymbol{T}} - \dot{\Phi}_\perp \boldsymbol{r}\dot{\boldsymbol{\Phi}}_{\boldsymbol{T}} = \boldsymbol{0}\right\}\theta_T(\boldsymbol{0})d\boldsymbol{X},
\end{aligned}
\tag{3.8}
$$

where $\theta(.)$ is the density of $\dot{\boldsymbol{\Phi}}$ and $\theta_T(.)$ is the density of $\dot{\boldsymbol{\Phi}}_T$. Worsley in [37] showed that under the slightly more general condition where the boundary of $\mathcal{T}$ is composed of a finite number of piecewise smooth components, then the expectation of the excursion sets of a random field with zero mean and unit variance is

$$
\mathbb{E}\left\{\varphi(\mathcal{A}_u\left(\Phi, \mathcal{T}\right))\right\} = \sum_{j=0}^{N} \rho_j(u)\mathcal{L}_j(\mathcal{T})
\tag{3.9}
$$

where $\rho_j(u) = (2\pi)^{-(j+1)/2} H_{j-1}(u) \exp(-u^2/2)$ is the *intensity* of the EPC per unit volume, H_j the *jth* Hermite polynomial and $\mathcal{L}_j(\mathcal{T})$ the Lipschitz-Killing curvatures.

3.4 Volume of tubes

Volume of tubes approach for a class of Gaussian processes is linked to a finite Karhunen-Loéve expansion. These are processes defined on a manifold $\mathcal{T}$ that can be expressed as

$$\phi(\boldsymbol{X}, \overline{\omega}) = \langle \alpha(\boldsymbol{X}), \xi(\overline{\omega}) \rangle_{\mathbb{R}^l} = \sum_{j=1}^{l} \alpha_j(\boldsymbol{X}) \xi_j(\overline{\omega}) \tag{3.10}$$

for some smooth mapping $\alpha : \mathcal{T} \longrightarrow S(\mathbb{R}^l)$ where ξ_j are independent standard Gaussians and $S(\mathbb{R}^l)$ the sphere in $\mathbb{R}^l$, $l \geq 1$. Consider a metric space $(\mathcal{T}, \tau)$ as $S(\mathbb{R}^l)$ with a geodesic metric $\tau(x, y) = \cos^{-1}(\langle x, y \rangle)$. Then a tube of radius ρ around a closed set A is $Tube(A, \rho) = \left\{ x \in S(\mathbb{R}^l) : \sup_{y \in A} \langle x, y \rangle \geq \cos(\rho) \right\}$. In the case where X is a random vector uniformly distributed on $S(\mathbb{R}^l)$ with distribution η_l then

$$\mathbb{P}\left\{ \sup_{y \in A} \langle X, y \rangle \geq \cos(\rho) \right\} = \eta_l\left(Tube(A, \rho)\right).$$

Consider $\mathcal{H}_N$ the N-dimensional Hausdorff measure associated with the geodesic metric τ. Remember the Weyl's tube formula on $S(\mathbb{R}^l)$. Assume $\mathcal{T}$ is a C^2, locally convex, Whitney stratified submanifold of $S_\lambda(\mathbb{R}^l)$, the sphere of radius λ. For $\rho < \rho_c(\mathcal{T}, S_\lambda(\mathbb{R}^l))$

$$\mathcal{H}_{l-1}\left(Tube(\mathcal{T}, \rho)\right) = \sum_{j=0}^{N} \left(\sum_{n=0}^{\lfloor \frac{j}{2} \rfloor} \frac{(-4\pi)^{-n} \lambda^{l-1+j} j! G_j - 2n, l - 1 + 2n - j(\frac{\rho}{\lambda})}{n!(j-2n)!} \right) \mathcal{L}_j(\mathcal{T}),$$

$$\tag{3.11}$$

where $G_{a,b}(\rho) = \dfrac{\pi^{b/2}}{b\Gamma(b/2+1)} \int_0^\rho \cos^a(r) \sin^{b-1}(r) dr$, $\lfloor n \rfloor$ is the integer less than n

and $\mathcal{L}_j(\mathcal{T})$ the Lipschitz-Killing curvatures (LKC) given for a measurable subsets U by

$$\mathcal{L}_j(\mathcal{T}) = \frac{(-2\pi)^{-(N-j)/2}}{((N-j)/2)!} \int_U Tr^{\mathcal{T}}(R^{(N-j)/2}) Vol_g$$

if $N - j$ is even and 0 otherwise. Tr is the trace, R is the Riemannian curvature tensor and Vol_g is the Riemannian volume form on the metric tensor g. Remark that LKC is the EPC for $j = 0$, the perimeter of $\partial\mathcal{T}$ for $j = 1$ and the area of $\mathcal{T}$ for $j = 2$.

3.5 Decomposed Mean Euler-Poincaré Characteristic

Now we present our approach related to a new modeling of MEPC based decomposition. In our context of sub signed GRF $\phi^{s_k}(\boldsymbol{X}, \overline{\omega})$ to form the NGRF $\Phi(\boldsymbol{X}, \overline{\omega}, S^M)$ the volume of tubes is required to prove our following Lemma.

Lemma **3.1** *Let $\mathcal{T}$ be a C^2 manifold in $\mathbb{R}^N$ ($N \geq 1$), locally convex, Whitney stratified submanifold of $S_1(\mathbb{R}^l)$. Let $s_k\alpha^{s_k}(\boldsymbol{X}), k = 1, ..., S^M$ be deterministic coefficients for $s_k \in \Sigma_s$ and $\xi(\overline{\omega})$ be independent standard Gaussians. Let $s_k\phi^{s_k}(\boldsymbol{X}, \overline{\omega})$ be mappings and sub GRF such that from a finite Karhunen-Loéve expansion $s_k\phi^{s_k}(\boldsymbol{X}, \overline{\omega}) = \langle s_k\alpha^{s_k}(\boldsymbol{X}), \xi(\overline{\omega}) \rangle_{\mathbb{R}^l}$. Then*

$$\mathbb{P}\left\{\sup_{\boldsymbol{X}, \overline{\omega}} (s_k\phi^{s_k}(\boldsymbol{X}, \overline{\omega})) \geq s_k u\right\} \approx \mathbb{E}\left\{\varphi\left(\mathcal{A}_{s_k u}(s_k\phi^{s_k}, \mathcal{T})\right)\right\} = s_k\mathbb{E}\left\{\varphi\left(\mathcal{A}_u(\phi^{s_k}, \mathcal{T})\right)\right\}$$

$$= \frac{\Gamma(l/2)}{2\pi^{l/2}}\sum_{j=0}^N \mathbb{E}\left\{\tilde{G}_{j,l}\left(\cos^{-1}(u/|\xi|)\right)\mathbf{1}_{|\xi|\geq u}\right\}\mathcal{L}_j(s_k\alpha^{s_k}),$$

(3.12)

where $\tilde{G}_{j,l}(\rho) = \sum_{n=0}^{\lfloor j/2 \rfloor} \frac{(-4\pi)^{-n} j!}{n!(j-2n)!} G_{j-2n,l-1+2n-j}(\rho)$.

See proof in Appendix.

Let $(\mathcal{T}, g)$ be an 2-dimension Riemannian manifold where g is the Riemannian metric tensor. For each $\boldsymbol{X} \in \mathcal{T}$, there is an inner product $g_{\boldsymbol{X}} : T_{\boldsymbol{X}}\mathcal{T} \times T_{\boldsymbol{X}}\mathcal{T} \to \mathbb{R}$

such that $(\xi_{\boldsymbol{X}}, \sigma_{\boldsymbol{X}}) \mapsto g_{\boldsymbol{X}}(\xi_{\boldsymbol{X}}, \sigma_{\boldsymbol{X}})$ where $T_{\boldsymbol{X}}\mathcal{T}$ is the tangent space of $\mathcal{T}$ at $\boldsymbol{X}$. Let $\langle \xi_{\boldsymbol{X}}, \sigma_{\boldsymbol{X}} \rangle$ the inner product $g_{\boldsymbol{X}}(\xi_{\boldsymbol{X}}, \sigma_{\boldsymbol{X}})$. For a coordinate system with basis $\left\{ \dfrac{\partial}{\partial X_i} |_{\boldsymbol{x}} \right\}_{i=1,2}$ we have for $i, j = 1, 2$

$$g_{ij}\left(\boldsymbol{X}\right) := g_{\boldsymbol{X}} \left(\frac{\partial}{\partial X_i} |_{\boldsymbol{x}}, \frac{\partial}{\partial X_j} |_{\boldsymbol{x}} \right) = \langle \frac{\partial}{\partial X_i} |_{\boldsymbol{x}}, \frac{\partial}{\partial X_j} |_{\boldsymbol{x}} \rangle \tag{3.13}$$

which are called the components of g at $\boldsymbol{X}$ under this coordinate system. Let $G(\boldsymbol{X}) = (g_{ij}\left(\boldsymbol{X}\right))$ for $i, j = 1, 2$ be a symmetric and positive definite matrix from the definition of g. For a real-valued function ϕ on $\mathcal{T}$, let $\phi_i = \dfrac{\partial}{\partial X_i}\phi$ and $\phi_{ij} = \dfrac{\partial^2}{\partial X_i \partial X_j}\phi$. Define the length of the differentiable curve $\gamma : [a, b] \to \mathcal{T}$ by $\mathcal{L}(\gamma) = \int_0^1 \sqrt{\langle \gamma'(t)\gamma(t) \rangle}dt$. Then the distance on $\mathcal{T}$ by the Riemannian metric g is $d_{\mathcal{T}}(\boldsymbol{X}, \boldsymbol{Y}) = \inf\limits_{\gamma \in D^1([0,1];\mathcal{T})_{\boldsymbol{X},\boldsymbol{Y}}} \mathcal{L}(\gamma)$. Riemannian metric g is closely related to the covariance function $C(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left\{\phi_p \phi_q\right\}$. In particular for $\boldsymbol{X_0} \in \mathcal{T}$, it follows that

$$g_{\boldsymbol{X_0}} = \xi_{\boldsymbol{X_0}} \sigma_{\boldsymbol{X_0}} C(\boldsymbol{X}, \boldsymbol{Y})|_{\boldsymbol{X}=\boldsymbol{Y}=\boldsymbol{X_0}}. \tag{3.14}$$

Lemma **3.2** *Let $\phi_* = \{\phi_*(\boldsymbol{X}, \overline{\omega}), \boldsymbol{X} \in \mathcal{T}, \overline{\omega} \in \overline{\Omega}\}$ be a GRF satisfying $C(\boldsymbol{X}, \boldsymbol{Y}) = 1 - cd_{\mathcal{T}}^2(\boldsymbol{X}, \boldsymbol{Y})(1 + o(1))$, $c > 0$, where $(\mathcal{T}, g)$ is an 2-dimension Riemannian manifold. Let $D \subset \mathcal{T}$ be an 2-dimension compact submanifold on $\mathcal{T}$. Then,*

$$\mathbb{P}\{\sup_{\boldsymbol{X} \in D, \overline{\omega}} \phi_*(\boldsymbol{X}, \overline{\omega}) \geq u\} = \sum_{j=1}^{m} \mathbb{P}\{\sup_{\boldsymbol{X} \in D_j, \overline{\omega}} \phi_*^j(\boldsymbol{X}, \overline{\omega}) \geq u\}, \tag{3.15}$$

where $D = \cup_{j=1}^m D_j$, $\phi_*^j(\boldsymbol{X}, \overline{\omega}) \in D_j$ and $\phi_*(\boldsymbol{X}, \overline{\omega}) = (\phi_*^1(\boldsymbol{X}, \overline{\omega}), ..., \phi_*^m(\boldsymbol{X}, \overline{\omega}))$.

See proof in Appendix.

Now we present the main Theorem of our method related to the decomposition of EPC. Let $s_j \phi^j(\boldsymbol{X}, \overline{\omega})$ be a collection of i.i.d. signed sub GRF. Let $D = \cup_{j=1}^m D_j$ be a composition such that $\phi^j(\boldsymbol{X}, \overline{\omega}) \in D_j$. Let $\Phi(\boldsymbol{X}, \overline{\omega}, S^M) = (s_1 \phi^{s_1}(\boldsymbol{X}, \overline{\omega}), ..., s_{SM} \phi^{s_{SM}}(\boldsymbol{X}, \overline{\omega}))$ be a NGRF where $\Sigma_s = (s_1, ...s_{SM})$ with s_j the SUW of each $\phi^{s_j}(\boldsymbol{X}, \overline{\omega})$ and $S^M = \sum_{j=1}^M (-1)^j s_j$. Then

$$\mathbb{P}\{\sup_{\boldsymbol{X}\in D,\overline{\omega}} \Phi(\boldsymbol{X},\overline{\omega}, S^M) \geq u\} = \sum_{k=1}^{S^M} s_k \left[A_k \exp(-(X_k^u)^2/2) + B_k \Psi(X_k^u)\right], \qquad (3.16)$$

with $X_k^u = (u - \mu_k)/\sigma_k$, $\mu_k, \sigma_k, A_k, B_k \in \mathbb{R}$, Ψ the tail probability function, and the mother-wave defined as

$$W(X_k^u) = A_k \exp(-(X_k^u)^2/2) + B_k \Psi(X_k^u). \qquad (3.17)$$

See proof in Appendix. Eq.(3.16) means that all signed decomposed NGRF has a decomposed mean Euler-Poincaré Characteristic. This formula allows to treat complex cases of signal containing outliers such that R waves as shown in experimental data.

Chapter 4

Modelling myocardial infarction

Myocardial infarction (MI), most famously known as a heart attack, represents the first cause of death worldwide. Mexico is not the exception; 800 000 deaths are estimated in the USA [35], and 105 000 in our country because of the acute myocardial infarction [49].

MI is the irreversible death of heart muscle secondary to prolonged lack of oxygen supply. The well-known symptoms of heart attack are chest pain and briefness of breath, but there exist others like:

- pressure or tightness in the chest

- pain in the chest, back, jaw, and other areas of the upper body that persists a few minutes or comes and goes

- shortness of breath

- sweating

- nausea

- vomiting

- anxiety

- a cough

- dizziness

- a fast heart rate

The heart and the blood vessels integrate the cardiovascular system. The arteries take oxygen-rich blood to your body and all of your organs. When the arteries are blocked due to an accumulation of plaque, the blood flow decreases or stops, so a heart attack happens.

The unhealthy habits can lead to a condition where MI occurs. Bad cholesterol, saturated fats, and trans fat can blockage the arteries.

Unhealthy habits can lead to a condition where MI occurs. Bad cholesterol, saturated fats, and trans fat can blockage the arteries. The most risk people have high blood pressure, high cholesterol levels, high triglyceride levels, diabetes and high blood sugar levels, obesity, smoking, age, or family history.

We mentioned a brief description of MI and the causes, but what about the diagnoses? A doctor will listen to the heart and check for irregularities in the heartbeat, check the blood pressure, and do an EKG.

The EKG can show a lot of visible information, but hidden information exists, and we can analyze it with the appropriate method. We can interpret the signal using transformation as Wavelet, Hilbert, or Fourier as mentioned in the works cited in this research's background.

So this chapter threatens this disease and shows the results using our geometrical approach to get a new model to represent the EKG signal with MI.

4.1   Case of study

EKG data is a collection of the *Smart Health for Assessing the Risk of Events via EKG databases (SHAREE)* [50] database. This material includes 139 hypertensive patients; we have downloaded a total of 22 patients. The signals were digitized at 128 Hz, have a duration of 24 hours or around 94000 cardiac cycles. The patients were monitored 12 months after the recordings to record a cardiovascular event. In total, 17 patients presented a type of event; 11 had a myocardial infarction, three strokes, and three syncopal events. Therefore, we analyzed 11 patients without events that we will call control and 11 patients with myocardial infarction.

Figure 4.1 represents the two types of EKG signals from the SHAREE database. The top of the figure represents a control patient and the bottom a myocardial

infarction patient.
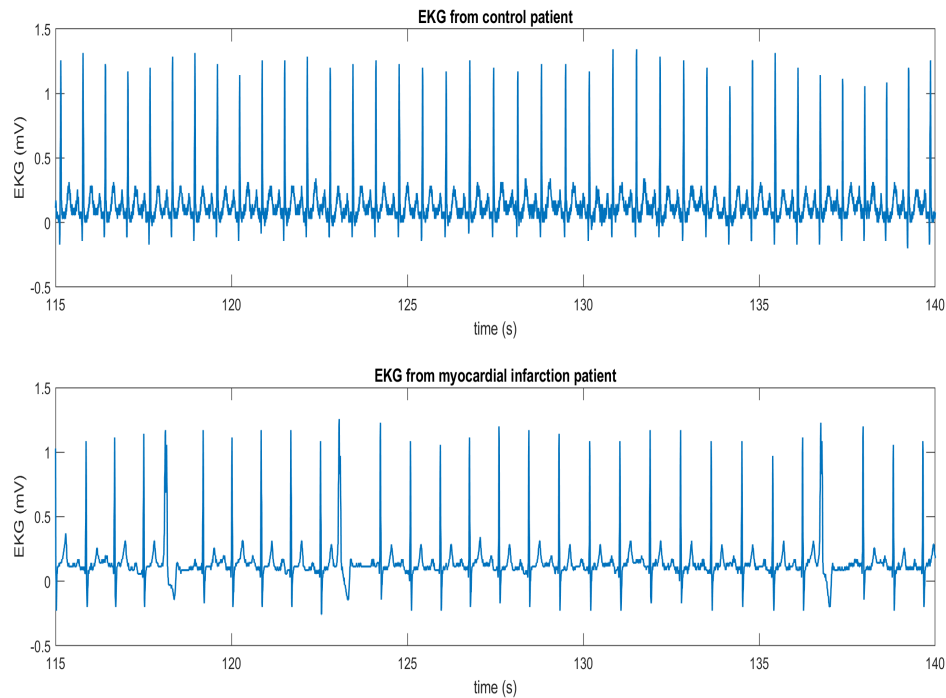
Figure 4.1: EKG signals extracted from the SHAREE database.

EKG is a standard medical test to diagnose cardiovascular conditions. Among these, ventricular tachycardia (VT), atrial fibrillation (AFIB), coronary artery disease (CAD), and myocardial infarction (MI) [51]. The SHAREE database [52] collects Holter EKG that spans EKG and respiration signals of patients with a cardiac event prior to the Holder EKG record. It is well-known that it is indeed feasible to recognize precise cardiac events through the analysis of abnormalities in primary waves (PQRST) of a cardiac cycle [53]. Hence, we propose to study the behavior pattern of control and MI patients by assessing the MEPC from the NGRF formed with a RAP model of the P-waves from Holter EKG records.

4.2 Polynomization

In this section, we get the first model representing the EKG signal using random algebraic polynomials (RAP). In this step, we separate the PQRST waves of the EKG in two parts, the first represents from the P-wave to the R-peak, and the second represents the R-peak to the T-wave.

The final equation is represented in 3.5 in section 3. In each cycle, we get two series of coefficients $a_k(c, \bar\omega)$, which the filtered signals for the P-wave and T-wave section.

Before, we need to calculate the order of the polynomials, and then the method is described.

4.2.1 Order of polynomials

In the polynomization step, it is essential to know the number of order of the polynomials to represent each EKG cycle for maintaining homogeneity. We selected 100 000 samples of each patient (of total 22).

The following procedure was carried out to obtain the order:

1. Find the R-peaks in each cycle using the Pan-Tompkins algorithm.

2. Find the middle points between the interval R-R.

3. For each interval, we get the results with the order from 20 to 25.

4. The measure of the fit for each order and the raw EKG is getting using the Eq. (4.1).

5. Choose the best fit in each cycle.

6. A frequency histogram plots the relationship between the fit of the polynomials and the orders between 20 and 25. Figures 4.2 and 4.3 shows these results.

$$FIT = \max \left\{ 100 \left(1 - \frac{\left\| P_j(c) - \sum_{k=0}^{D_j(c)} \hat{a}_k(c, \bar\omega) X_j(c)^k \right\|}{\left\| P_j(c) - \bar{P}_j(c) \right\|} \right) \right\}_{j=1,2.} \tag{4.1}$$
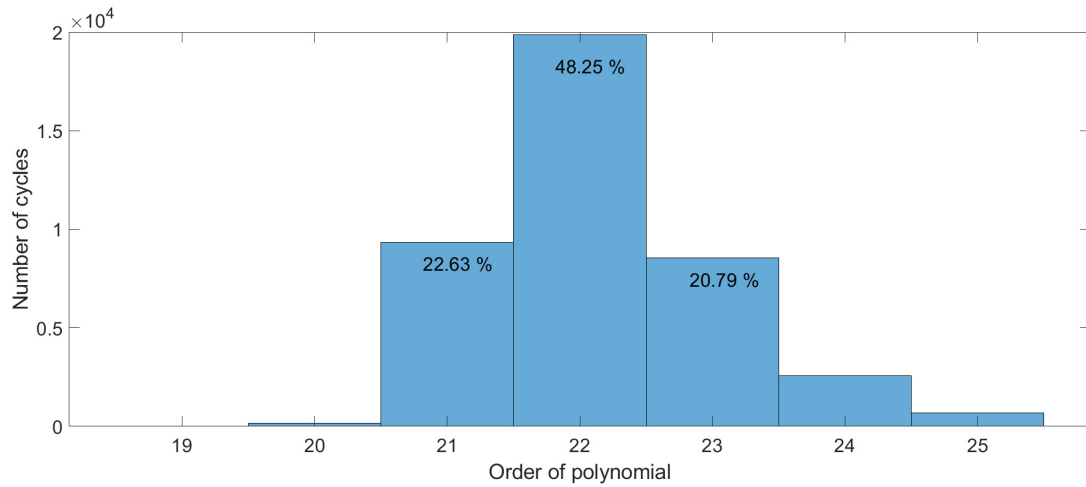
Figure 4.2: Histogram to select the best order of the polynomials P_1 (P-waves) for each cardiac cycle.
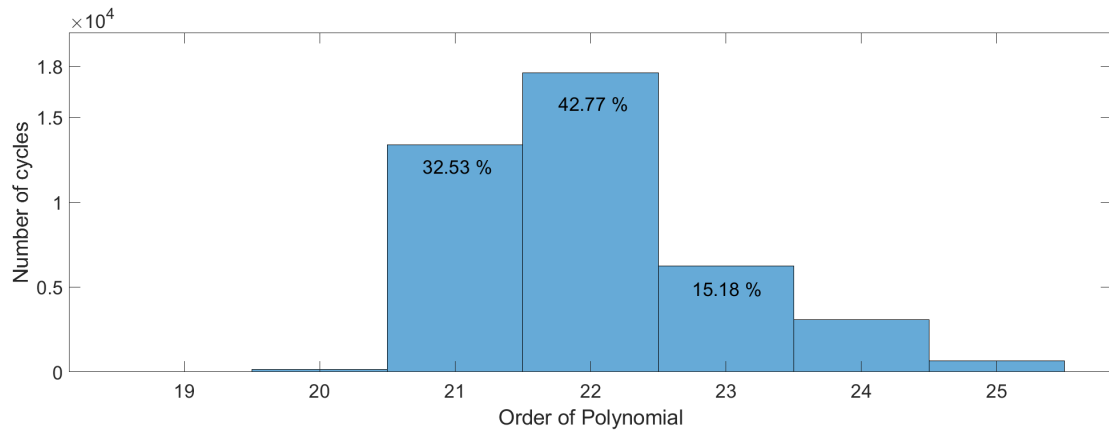
Figure 4.3: Histogram to select the best order of the polynomials P_2 (T-waves) for each cardiac cycle.

From now on, the order of the polynomials is set at $D_1(c) = 22$ and $D_2(c) = 22$. In this way, the polynomials $I_1(c)$ and $I_2(c)$ are obtained in terms of the coefficients P_j. Algorithm 1 dictates the entire procedure. As a result, it returns the structure arrays P_1 and P_2.

So that, each cycle has a *down cycle*, and an *up cycle* that formed the full cycle, as we can observe in Figure 4.4.
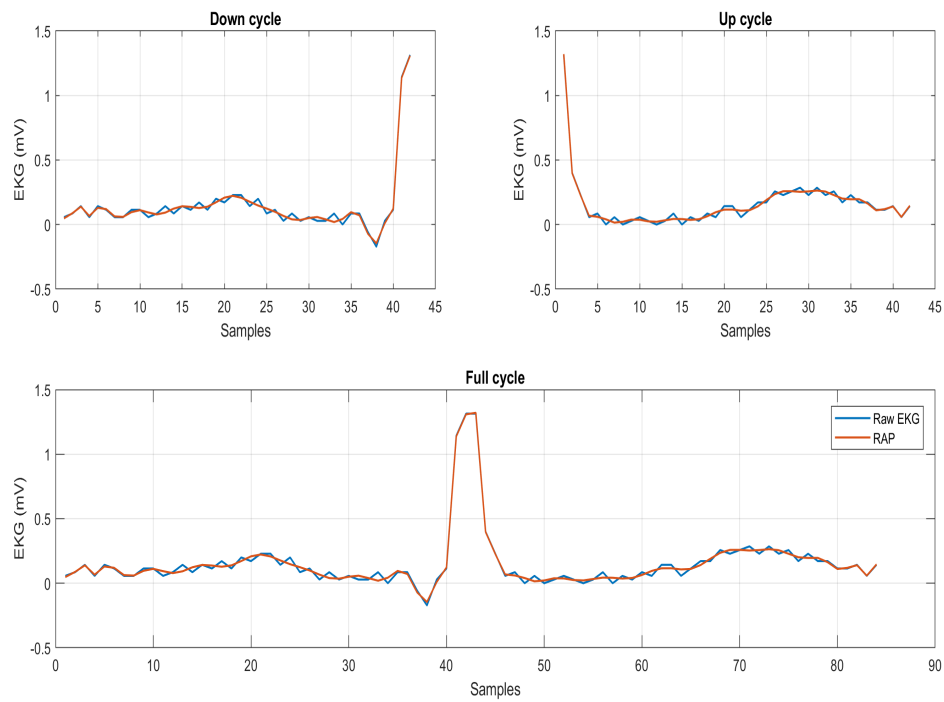
Figure 4.4: The graph *Down cycle* shows the polynomization for the P-wave; in the *Up cycle* shows the polinomization for T-wave. The *Full cycle* shows the raw EKG and the $P_1(c)$ and the $P_2(c)$.

4.2.2   Algorithm for polynomization

Algorithm 1: Algorithm for calculating random algebraic polynomials

   Data: Holter EKG

   Result: Polynomials $P_1(c)$ and $P_2(c)$

1 initialization;

2 Detect the peaks position R_k, $k = 1, \ldots, N_c$ using the Pan-Tompkin algorithm [5];

3 Detect the middle points $m_k = \frac{R_{k+1} - R_k}{2}$, $k = 1, \ldots, N_c - 1$;

4 Determine the order of polynomials using the FIT (eq. 4.1)

5 Calculate the polynomials $I_1(c)$, $c \in [m_k, R_{k+1}]$ and $I_2(c)$, $c \in [R_{k+1}, m_{k+1}]$ using the Vandermonde matrix;

6 Evaluate polynomials and add results at structure array $P_1(c)$ and $P_2(c)$;

4.3   Non-Gaussian Random Field

In this step, the polynomization results allow to create the Non-Gaussian random field (NGRF). This field needs to be homogeneous, so that to maintain the homogeneity, we add zeros to the polynomials in each cardiac cycle.

Why do we need to add zeros? The number of samples in each cardiac cycle changes because of the cardiac frequency for different activities as rest, exercise, or others. If each cycle is different and we need homogeneity, the solution was to add zeros to keep it.

Let $\Phi(\boldsymbol{X}, \bar{\omega})$ be a NGRF at $\boldsymbol{X}$ with $\boldsymbol{X} = [X_1(c) \quad X_2]^T$ where $X_1(c) \in I_1(c) \cup I_2(c)$. Let $\mathcal{E} = (\bar{\Omega}, \bar{\Sigma}, P)$ be a complete probability space and $\mathcal{T}$ a topological space, where $\bar{\Omega}$ denotes the sample space, $\bar{\Sigma}$ a σ-algebra on $\bar{\Omega}$ and P a probability measure on $\bar{\Sigma}$. Then a measurable mapping $\Phi : \bar{\Omega} \to \mathbb{R}^{\mathcal{T}}$ is called a real-valued random field. Here $\boldsymbol{X} \in \mathcal{T} \subset \mathbb{R}^2$ and Φ is called an 2-dimensional NGRF over N cycles defined as

$$\Phi(X_1, X_2, \overline{\omega}) = \Phi(\boldsymbol{X}, \overline{\omega}) = \sum_{c=1}^{N} \psi_D(\mathcal{E}, \overline{\omega}, X_1(c)) \delta(c - X_2), \qquad (4.2)$$

where δ is the Dirac's distribution. Eq.(4.2) means that each EKG signal for a fixed threshold is decomposed into basic RAP $\psi_D(\mathcal{E}, \overline{\omega}, X_1(c))$ to form a NGRF according

to the coordinate X_2.

Eq. (4.2) means that we take all the waves of the EKG cycle. However, we mentioned that we only took the section of P-waves to analyze the MEPC. Hence, $X_1(c) \in I_1(c)$. Figure 4.5 shows an example of NGRF of P-waves using the polynomials.
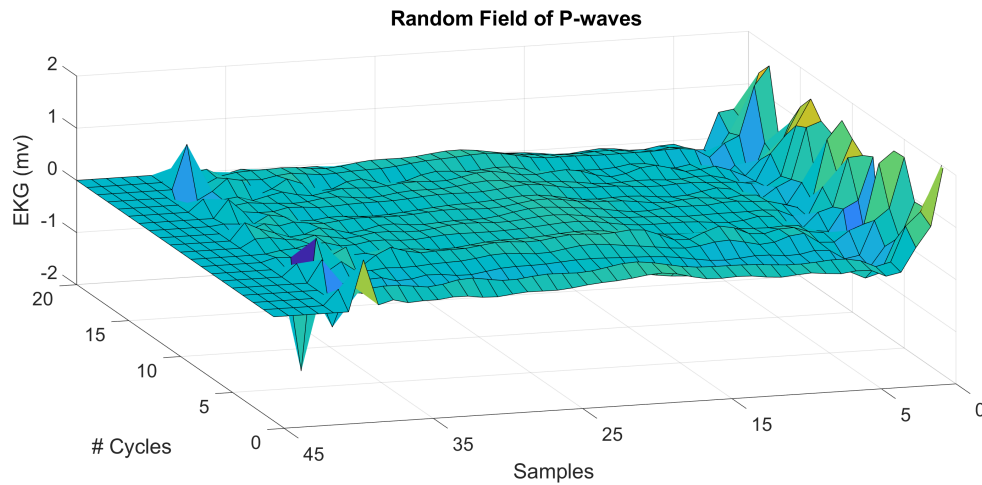
Figure 4.5: An example of NGRF of P-waves.

4.3.1 Excursion sets

Excursion sets are widely studied; in this case, we used them for formed new geometric structures from the NGRF and extracted the MEPC in each level u. The definition for Gaussian random fields is written in Eq (2.21).

If the case is a non-Gaussian random field, the excursion set is a geometrical object defined by

$$A_u(\Phi, \mathcal{T}) = \{\boldsymbol{X} \in \mathcal{T} : \Phi(\boldsymbol{X}, \bar{\omega}) \geq u\} \tag{4.3}$$

where $\Phi(\boldsymbol{X}, \bar{\omega})$ is the NGRF. Thus $\mathcal{A}_u$ of $\Phi(\boldsymbol{X}, \bar{\omega})$ above a threshold u is the set of points in $\mathcal{T} \subset \mathbb{R}^2$ where $\Phi(\boldsymbol{X}, \bar{\omega})$ exceeds u. The level u is well-chosen between the interval $[-1, 1]$. Figure 4.6 shows the level u as a plane cutting the random field, the points from the level u to the maximum is the excursion set.
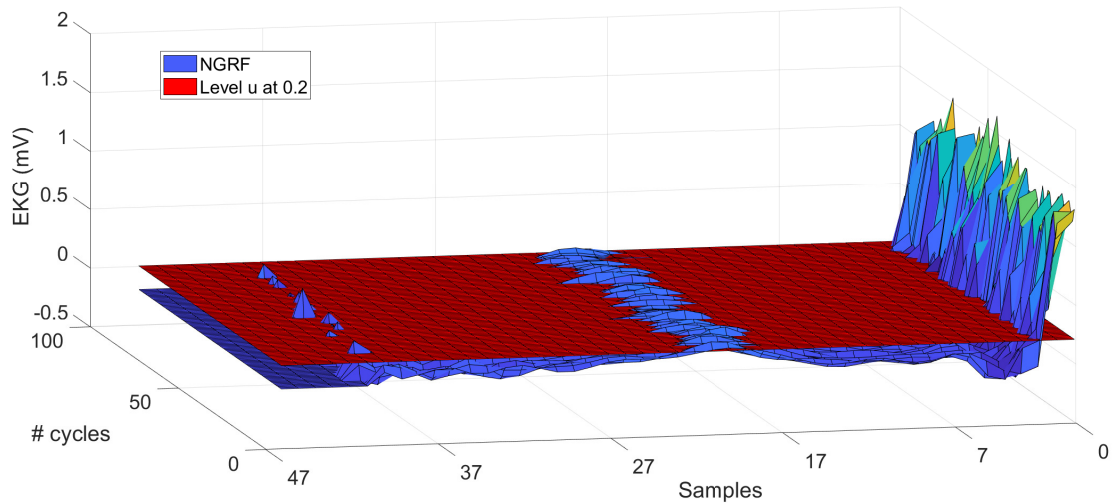
Figure 4.6: Non-Gaussian random field cut by a level $u = 0.2$.

The region upper the level u is the excursion set, and we transform it from a geometrical structure to a binary image that allows us to extract the Euler-Poincaré characteristic. Figure 4.7 shows the excursion set from a control patient, and Figure 4.8 shows the same for a myocardial infarction patient, both with a level $u = 0.1$.
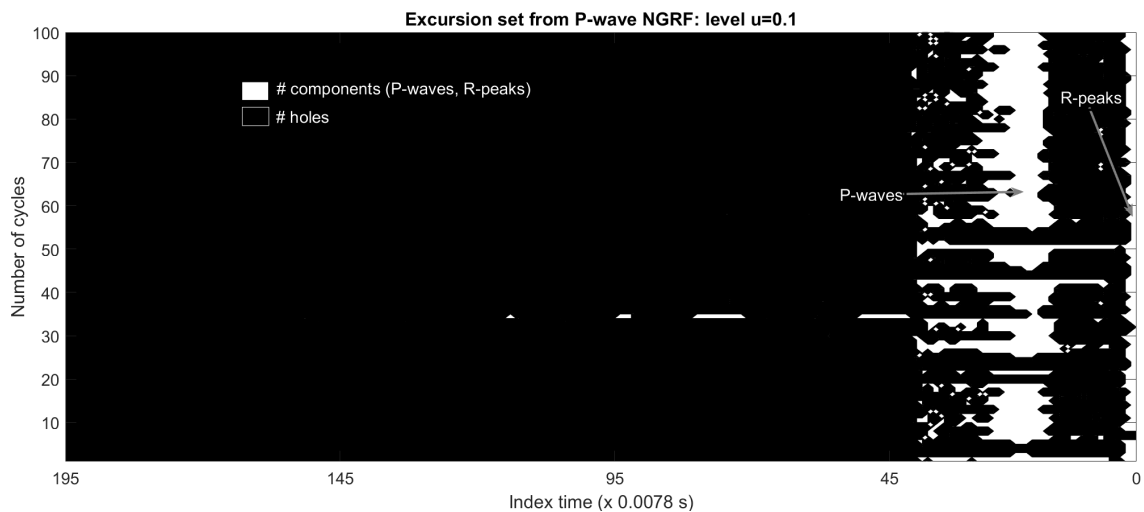
Figure 4.7: Excursion set of the Non-Gaussian Random Field at $u = 0.1$ for a control patient's P-waves.

Excursion set from P-waves NGRF MI patient: level u=0.1
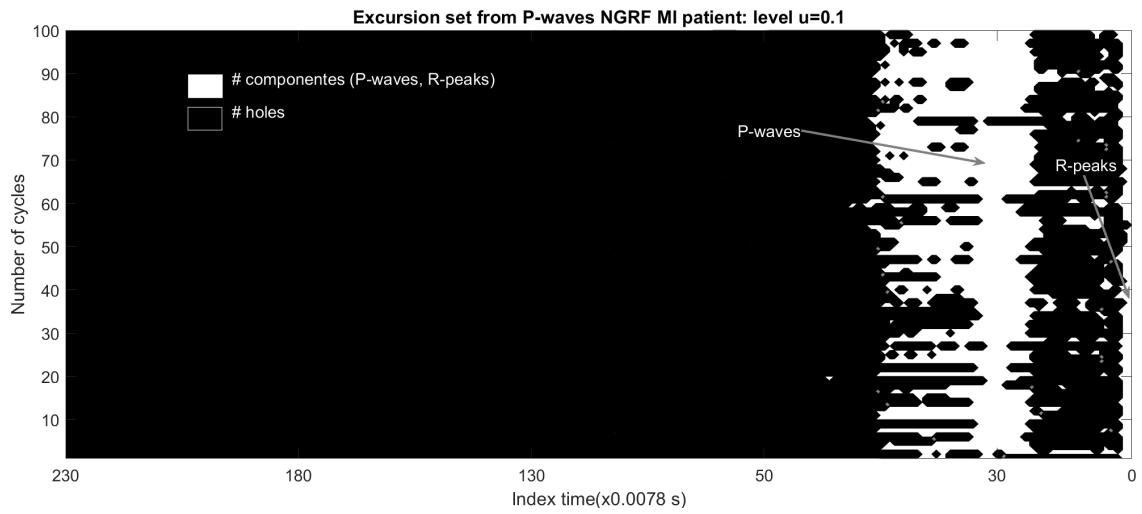
Figure 4.8: Excursion set of the Non-Gaussian Random Field at $u = 0.1$ for a MI patient's P-waves.

4.3.2 Algorithm to create the Non-Gaussian Random Field

Algorithm 2: Algorithm to build the non-Gaussian random fields

Data: Polynomials $P_1(c)$ and $P_2(c)$

Result: *Non-Gaussian Random Field* $\Phi(\boldsymbol{X}, \bar{\omega}, S^M)$

1   initialization;

2   **for** $k = 1 : N_c$ **do**

3     $A(k) = length(P_l(c)_k), l = 1, 2$

4   **end**

5   $M = \max\{A\} | M \in \mathbb{R}$; **for** *k=1:Nc* **do**

6   **if** $A_k < M$ **then**

7       $\Phi_1([k, 1 : M], \bar{\omega}, S^M) = [zeros, P_1(c)_k]$ and
       $\Phi_2([k, 1 : M], \bar{\omega}, S^M) = [P_2(c)_k, zeros]$

8   **else**

9       $\Phi_1([k, 1 : M], \bar{\omega}, S^M) = P_1(c)_k$ and $\Phi_2([k, 1 : M], \bar{\omega}, S^M) = P_2(c)_k$;

10   **end**

11   **end**

12   Calculate the level u of excursion sets: $u = [u_1 : \frac{u_n - u_1}{N_u} : u_n]$, where $u_1 = \min\{\Phi(\mathbf{X}, \bar{\omega}, S^M)\}, u_n = \max\{\Phi(\mathbf{X}, \bar{\omega}, S^M)\}$, and N_u=number of levels;

4.4 Mean Euler-Poincaré Characteristic

In this step, a new transform to our original EKG signal shows a new concept of a model in this type of signal. In this methodology, we transform the raw EKG signal to a polynomial form. Then we create a geometrical structure from the excursion sets of the non-Gaussian random field. Now, we transform the geometrical structure to a binary image in order to get the Euler-Poincaré characteristic.

When the value of $t \in \Phi(\mathbf{X}, \bar{\omega})$ is higher than the level u, so the value for the binary image is 1, otherwise it is 0. Figures 4.7 and 4.8 represent an example of a binary image from an excursion set.

The EPC is a scalar value representing the number of connections minus the number of holes in those elements [42]. To get these values, we used Gray's algorithm

[54] to the binary image of each excursion set A_u. The levels u selected were in the interval $[-1, 1]$ divided into 100 points.

Remember that we analyze 22 signals/patients, so each signal acquired 100 values that formed the EPC model due to the levels. Figures 4.9 and 4.10 show the model EPC from P-waves for each patient analyzed.
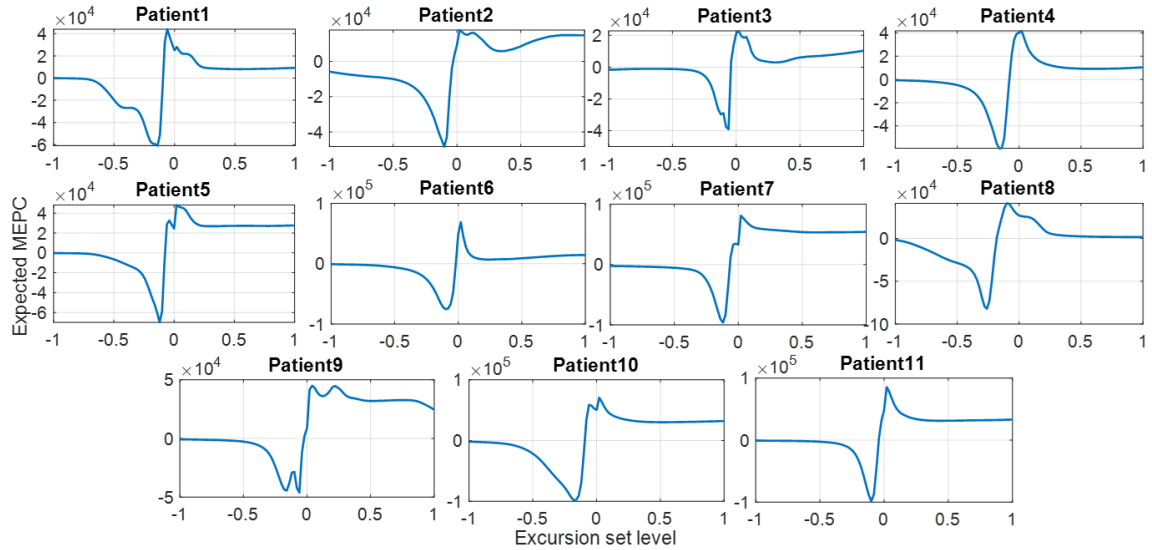
Figure 4.9: Expected EPC from the 11 control patients.
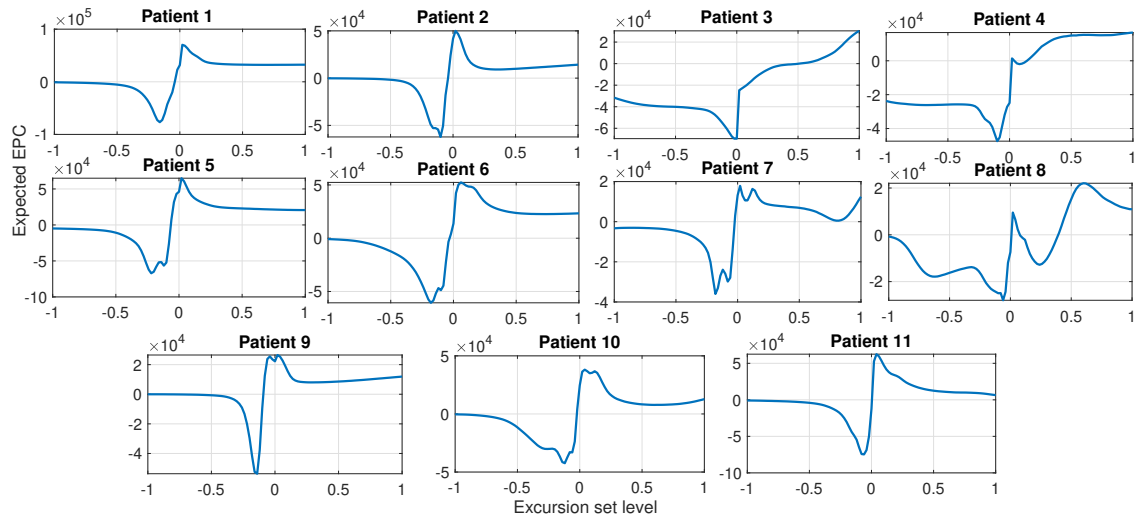
Figure 4.10: Expected EPC from the 11 myocardial infarction patients.

The mean Euler-Poincaré characteristic model is the final result using the mean of all the values of EPC of the eleven patients, as shown in Figure 4.11.
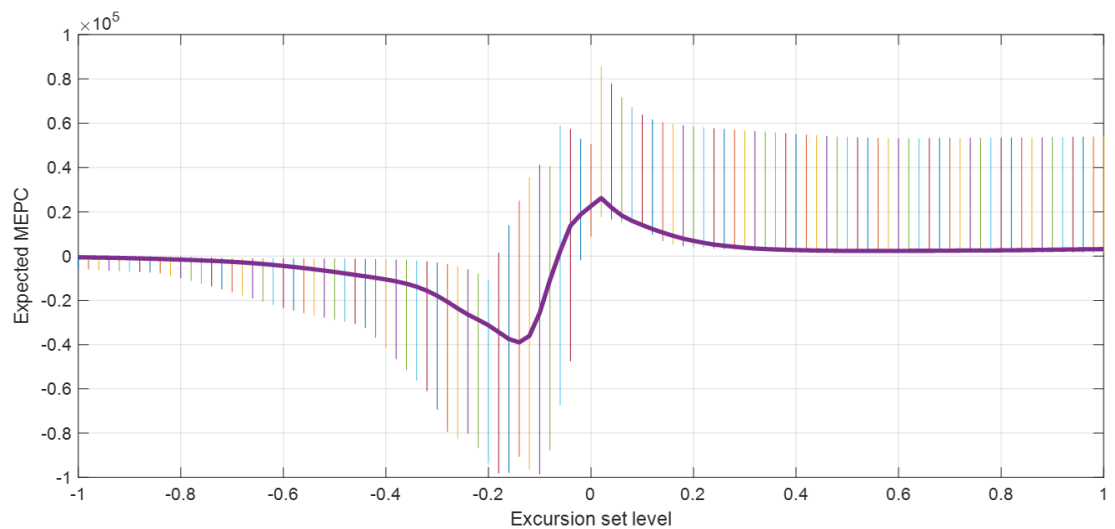
Figure 4.11: Mean Euler-Poincaré characteristic model from the EPC values of the P-waves of the control patients.

The mean Euler-Poincaré characteristic of control and MI patients is the mean of all EPC values in each case. Figure 4.12 shows the MEPC from control and MI patients.
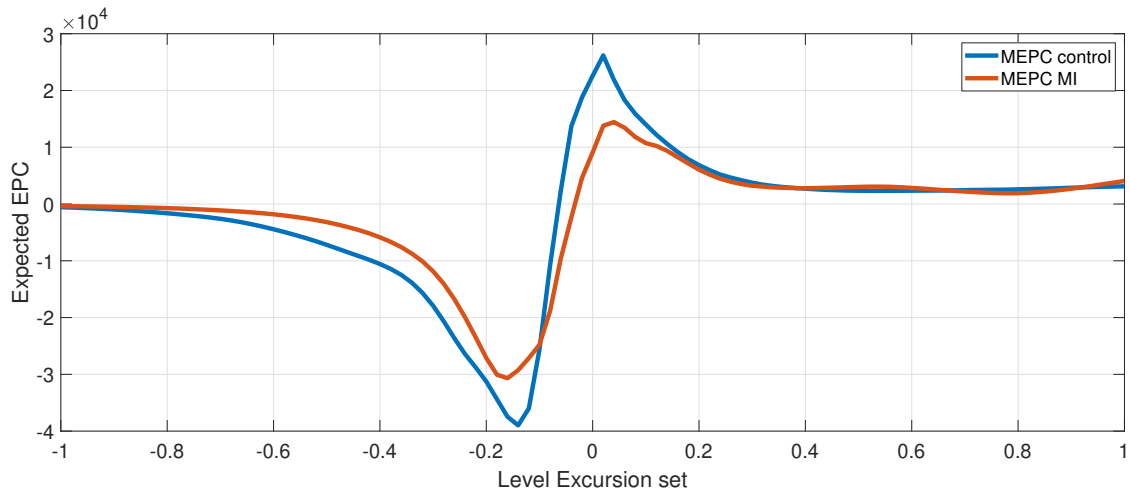
Figure 4.12: Mean Euler-Poincaré characteristic from P-waves of the control and MI patients.

4.4.1 Algorithm to get the mean Euler-Poincaré characteristic

---

Algorithm 3: Algorithm to get the values of mean Euler-Poincaré characteristic from *excursion set*

Data: *Non-Gaussian Random Field* $\Phi(\boldsymbol{X}, \bar{\omega}, S^M)$, *u (level)*

Result: *Euler-Poincaré Characteristic* $\varphi(\mathcal{A}_u)$

1 initialization;

2 for $k = 1 : Nc$ **do**

3   **for** $j = 1 : M$ **do**

4   **if** $\Phi([k,j], \bar{\omega}, S^M) >= \mu$ **then**

5 $\mathcal{A}_u(k,j) = 1$

6   **else**

7 $\mathcal{A}_u(k,j) = 0$;

8   **end**

9   **end**

10 end

11 $\varphi(\mathcal{A}_u) = \frac{1}{4}[n\{Q_1\} - n\{Q_3\} + 2n\{Q_D\}]$;

4.5 Decomposed Mean Euler-Poincaré Characteristic

The main proposal of this approach is the relationship between the non-Gaussian random field and the model based on the mean Euler-Poincaré characteristic. Both can be decomposed in manifolds of the topological space T.

With the literature boarded, we know that the shape of these MEPC models is similar to the Gaussian MEPC with dimension $N = 2$. However, the Adler-Taylor model (Eq. (2.42)) cannot represent them, so we used the DMEPC model based on the mother-wave defined in Eq. (3.17) and with $S^M = 2$. Where S^M are the segments into which the MEPC model is divided.

We estimated the parameters A, B, mu, and sigma of the eq 15 using the Particle Swarm Optimization (PSO) method. First, we scaled the MEPC models using a ratio 10000:1. Then, we divided the MEPC model into two sections, the negative values and the positive values of the expected MEPC.

The particles ρ are the parameters to estimate, and the cost function CJ is set as the maximum fit that the estimate of the parameters attains. The constraints for the parameters are the following:

1. A_k should be less than ten but greater than -10.

2. B_k should be higher than -10 but less than 10.

3. σ_k is an interval between 0 to 1.

4. μ_k can take the values between -1 to 1.

Fig. 4.13 shows the DMEPC model from the control patients, and this model fits 67% of the measured signal. Table 4.1 shows the parameters obtained for this model.
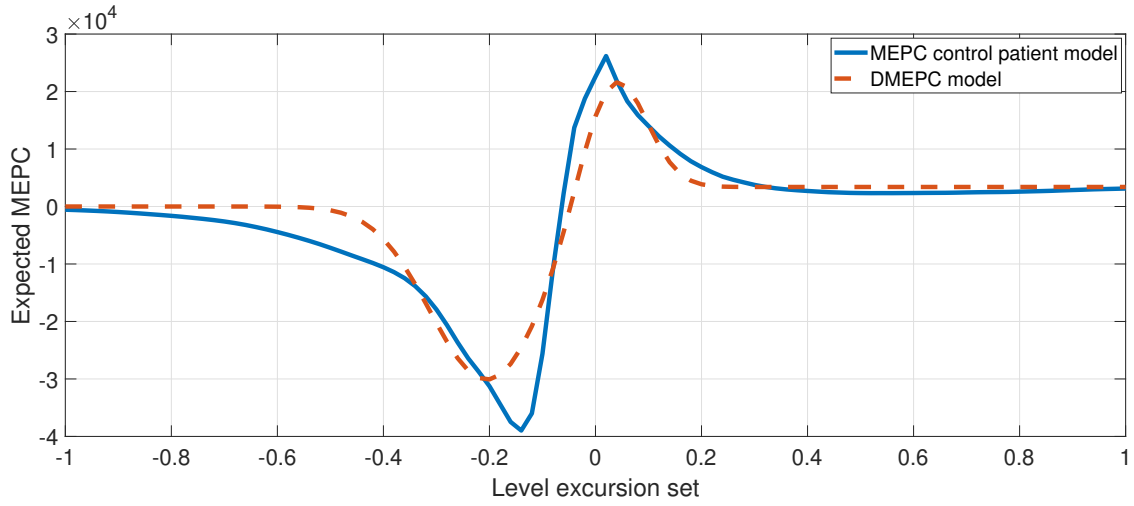
Figure 4.13: DMEPC model from control patients with respect to the threshold u.

Table 4.1: Parameters DMEPC control subject model.

k	A_k	B_k	μ_k	σ_k
1	0.023	-3.024	0.107	-0.205
2	0.317	2.203	0.061	0.031

Figure 4.14 shows the DMEPC model from the MI patients, and this model fits 72% of the measured signal. Table 4.2 shows the parameters obtained for this model.
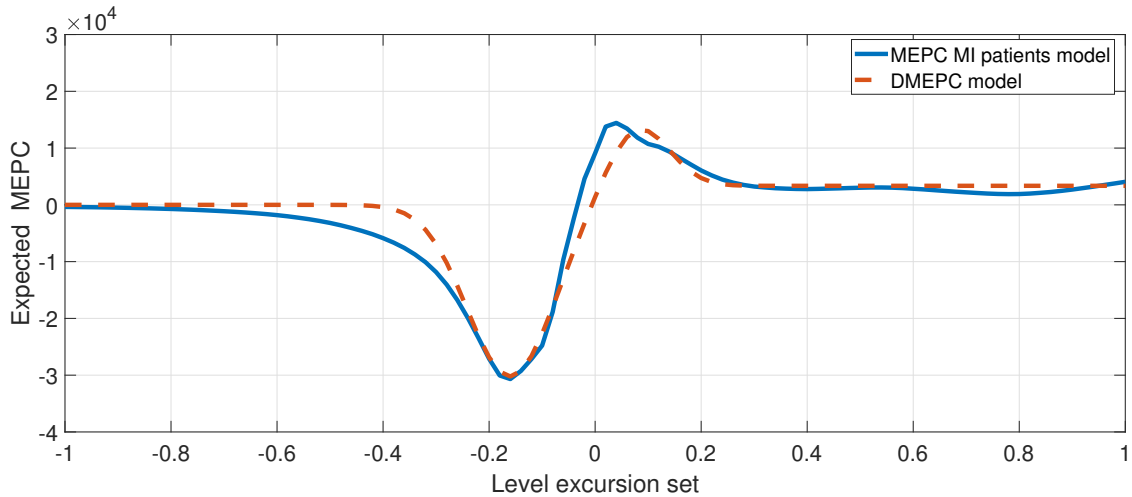
Figure 4.14: DMEPC model from MI patients with respect to the threshold u.

Table 4.2: Parameters DMEPC MI suject model.

k	A_k	B_k	μ_k	σ_k
1	0.045	-3.046	0.081	-0.160
2	0.290	1.165	0.060	0.077

4.5.1 Comparing DMEPC models

In this section, we compared the MEPC models vs. the DMEPC models for each type of patient. We have four kinds of models: control, myocardial infarction, syncope, and stroke.

In the first place, we compared the DMEPC model from control and MI vs. the myocardial infarction patients. We can observe that the minimum and maximum values in each model are different so that the gradient line formed by these values and their position are different for each type of patient.

Figures 4.15 and 4.16 represent the first comparison between the DMEPC models from control and MI patients vs. the Adler-Taylor models. In this case, it is not a conventional comparison because the Adler-Taylor model presented in Eq. (2.42) is for GRF. Therefore, the geometric structures are from an NGRF, so the behavior is from an MEPC for an NGRF.
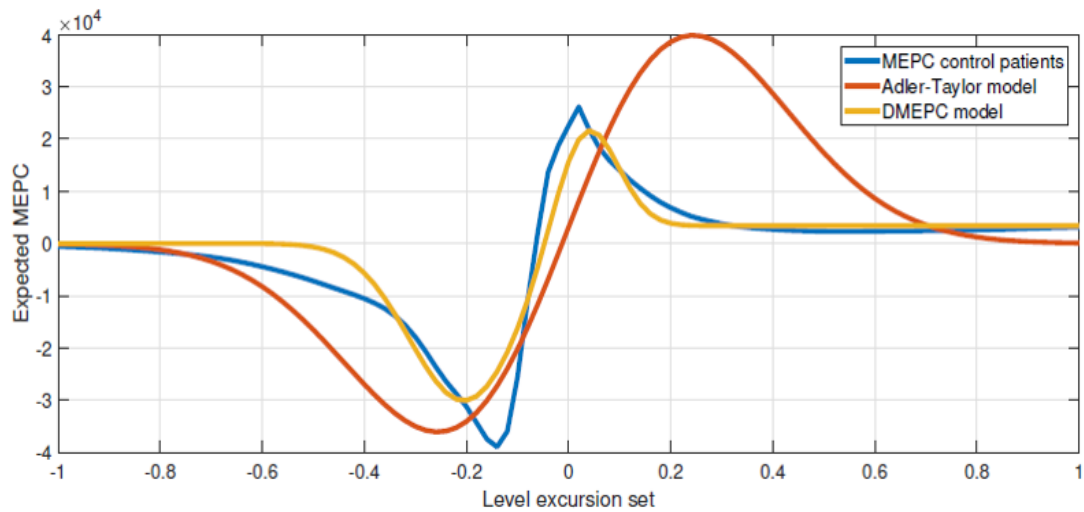
Figure 4.15: DMEPC model vs Adler-Taylor model for control patients.
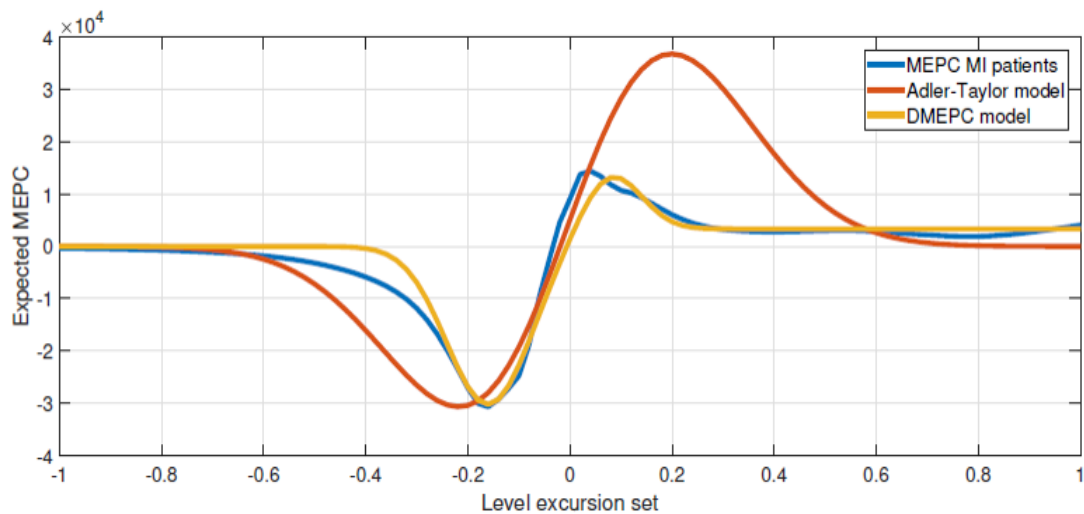
Figure 4.16: DMEPC model vs Adler-Taylor model for MI patients.

Eq. (2.42) can not give us the liberty to find values that match the behavior of the MEPC models because these behaviors are not from GRF. Table 4.3 showed the found parameters' values for Eq. (2.42) using the PSO algorithm.

Table 4.3: Parameters for the Adler-Taylor models usign PSO algorithm.

Model	T	λ^2	σ
Control	100.226	98.165	1
MI	150.345	80.265	1.2

Then, we compared the Control DMEPC model vs. the MI models using our approach. We used six patients because we do not want to cram the graph with information. Figure 4.17 showed the results; observing this graph, we can assume that our DMEPC models can be a way to find abnormalities between two cases. But, we need to analyze with detail this information to make sure this assumption. Table 4.4 present the parameters' values for the DMEPC models of the MI patients. The fitness in these models is between 65-77 %.
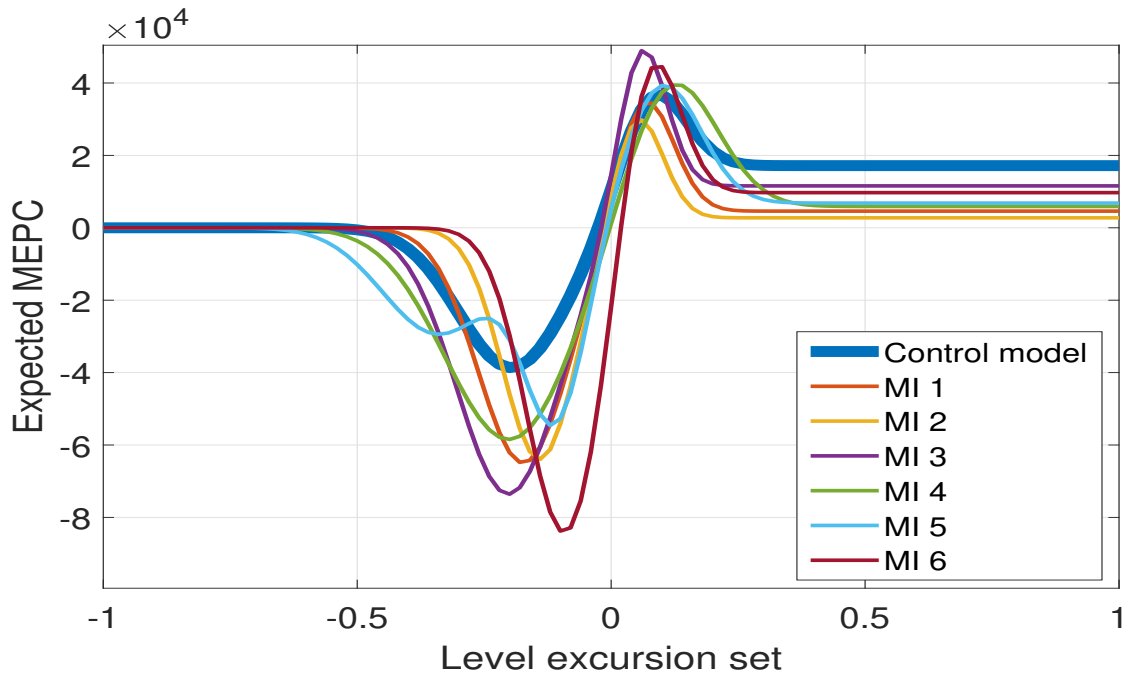
Figure 4.17: A comparison beetween the DMEPC of control patients vs. the DMEPC model for each MI patient

Table 4.4: Parameters for the MI models.

Model	FIT	k	A	B	σ	μ
MI 1	77%	1	-65211.07	476.1843	0.089	-0.173
		2	33858.71	4130.40	0.056	0.065
MI 2	73%	1	-64230.47	513.4826	0.073	-0.141
		2	29872.87	2246.857	0.049	0.051
MI 3	72%	1	-74014.29	809.90	0.101	-0.202
		2	45085.14	10765.96	0.048	0.057
MI 4	72%	1	-58996.29	877.56	0.127	-0.200
		2	38491.64	5014.94	0.090	0.115
MI 5	78%	1	35731.51	6213.03	0.075	0.099
		2	-52232.17	246.88	0.062	-0.110
		3	-29548.24	363.37	0.110	-0.340
MI 6	65%	1	-85156.30	867.54	0.076	-0.091
		2	45424.63	8843.88	0.056	0.076

Then, if you remember, in the study case appeared two more events, syncope and stroke. We treated these events with our methodology and got the DMEPC model for each syncope and stroke patient. Figures 4.18 and 4.19 showed the resulting models and were compared with the control model. And Tables 4.5 and 4.6 present the parameters' values for the DMEPC models of the syncope and stroke patients. The fitness in these models is between 60-79 %.

Table 4.5: Parameters for the syncope models.

Model	FIT	k	A	B	σ	μ
Syncope 1	76%	1	-32686.33	50.88	0.057	0.016
		2	23019.93	5427.64	0.04	0.189
Syncope 2	74%	1	-18907.67	281.21	0.099	-0.02
		2	24857.1	144.1	0.042	0.179
		3	12523.71	3517.81	0.116	0.506
Syncope 3	81%	1	-39960.47	142.21	0.046	0.036
		2	31070.26	1384.91	0.045	0.194
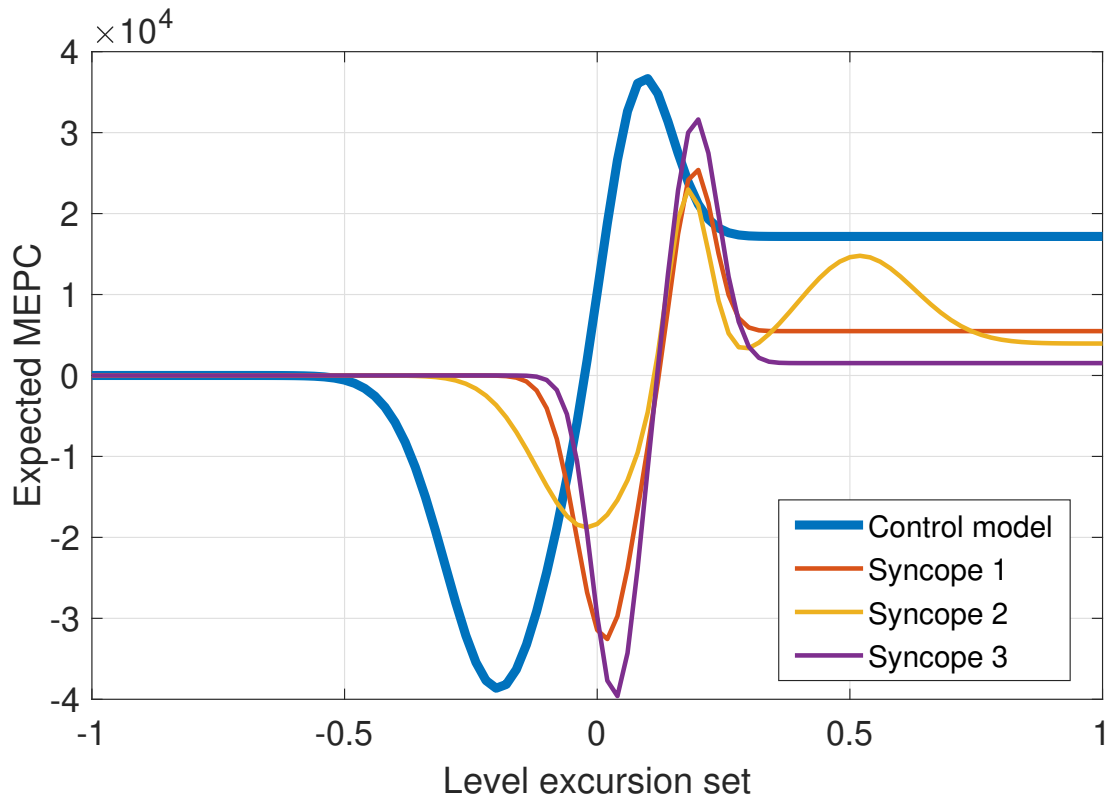
Figure 4.18: A comparison beetween the DMEPC of control patients vs. the DMEPC model for each syncope patient

Table 4.6: Parameters for the stroke models.

Model	FIT	k	A	B	σ	μ
		1	-48441.39	183.59	0.041	0.067
Stroke 1	79%	2	35719.09	1025.19	0.043	0.201
		3	6138.78	1786.5	0.072	0.829
Stroke 2	74%	1	-46420.87	458.39	0.077	0.02
		2	37937.06	3746.21	0.067	0.239
Stroke 3	60%	1	-17973.33	-228.9	0.091	-0.053
		2	26010.7	876.06	0.069	0.117
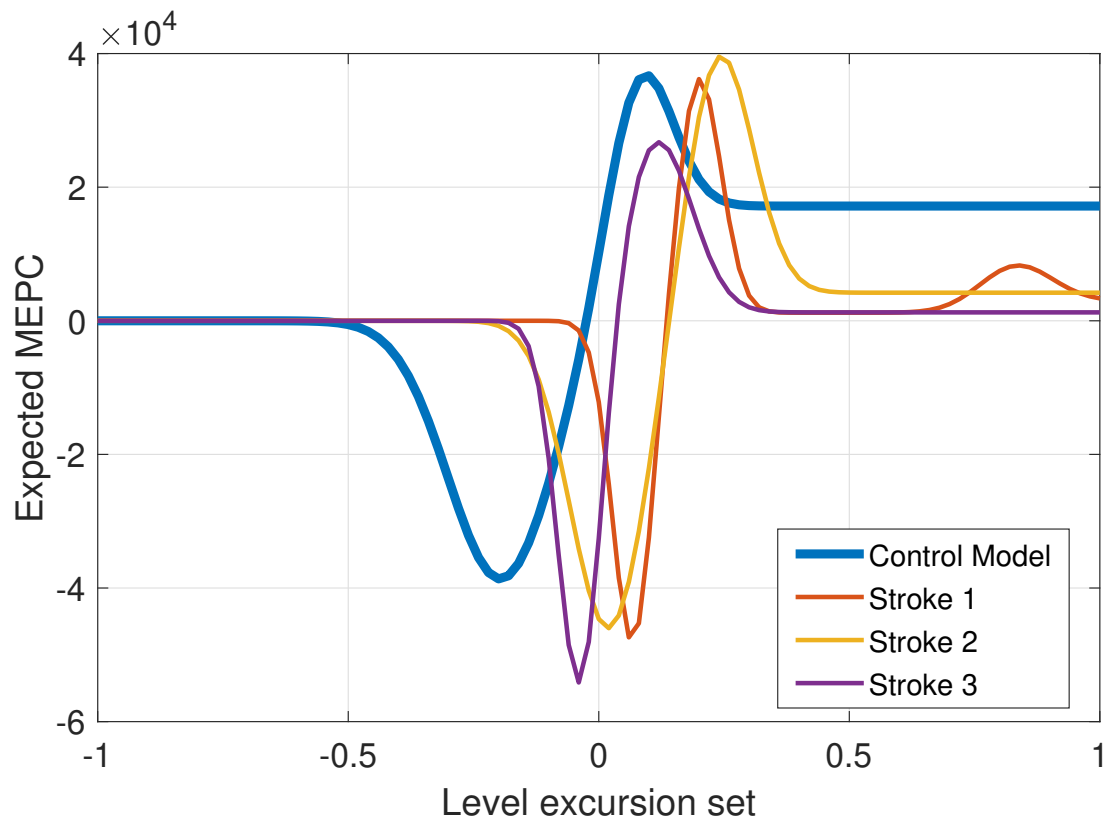
Figure 4.19: A comparison beetween the DMEPC of control patients vs. the DMEPC model for each stroke patient

Observing the DMEPC models, the minimum and maximum values change. In the syncope cases, the minimums are after the cero excursion set level, and the maximums are close to the 0.19 excursion set level. The minimums are relative to a cero in stroke, and maximums are in a 0.2 excursion set level.

Finally, we showed in Figure 4.20 the four different DMEPC models for control, MI, syncope, and stroke cases. Table 4.7 present the parameters' values for the DMEPC models.
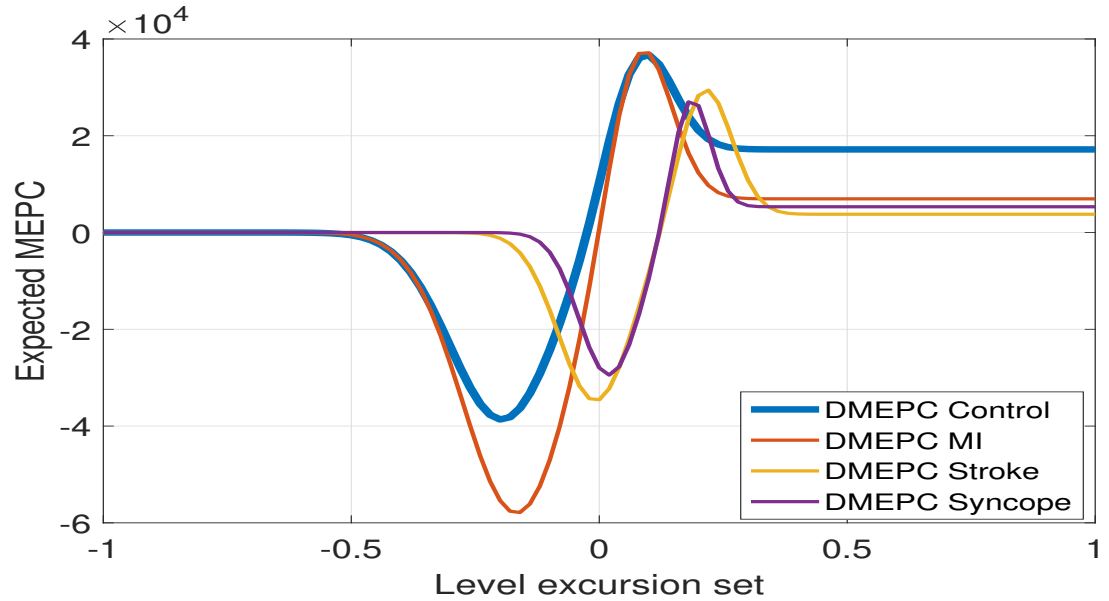
Figure 4.20: A representation of the MEPC based on DMEPC for the fourcases (control MI, syncope, and stroke).

Table 4.7: Parameters for the mean DMEPC models .

Model DMEPC	FIT	k	A	B	σ	μ
Control	77%	1	-38940.19	560.41	0.104	-0.195
		2	28017.40	16613.80	0.065	0.075
MI	71%	1	-58392.27	755.79	0.108	-0.166
		2	37425.67	6204.29	0.062	0.078
Stroke	78%	1	-34871.91	175.72	0.073	-0.008
		2	27881.90	3599.18	0.053	0.211
Syncope	77%	1	-29566.22	202.83	0.059	0.019
		2	25297.32	5121.67	0.038	0.183

4.5.2 Algorithm to find the DMEPC parameters

This section defines the algorithm to get the parameters of the Eq. (3.17) for each segment of the MEPC model.

Algorithm 4: Algorithm for get the parameters of the Decomposed Mean Euler-Poincaré Characteristic Model using PSO

Data: *Mean Euler-Poincaré Characteristic* $\mathbb{E}\{\varphi(Au)\}$

Result: $A_k, B_k, C_k, \sigma_k, \mu_k$ parameters

1 Initialize velocity w, c_1 and c_2 constants, and define the cost function CJ;

2 Select the number of particles N_p and create random particles $\rho(4, N_p)$ using the constraints for each parameter and random velocities $V(4, N_p)$;

3 Evaluate the particles in the eq. 3.17.

4 Calculate the fitness of each particle (F_p);

5 If $F_p > pBest$ in the iteration so the set of values of F_p is the new $pBest$

6 Choose the particle with the best $CJ = 100 \left(\frac{\|y - \hat{y}\|}{\|y - \bar{y}\|} \right)$ of all the particle as the $gBest$;

7 Calculate particle velocity using

$$V = \omega V + c_1(pBest - \rho) + c_2(gBest - \rho), \qquad (4.4)$$

where $c_1, c_2 \in \mathbb{R}$, V = velocity of the particle and ρ = particle ;

8 Update particle position with $\rho = \rho + V$

9 Update particle using the next constraints:

$$-10 <= A_k <= 10,$$
$$-10 <= B_k <= 10,$$
$$-1 <= \mu <= 1,$$
$$0 <= \sigma <= 1$$

10 When the number of maximum iterations or maximum fit is accomplished, the approximate solution is the particle with best fitness and STOP; contrarily add iteration and turn to 3.

4.6   Conclusion

This chapter presents the methodology applied to the myocardial infarction cases from the EKG Holter.

- We suggest using the Normalized Root Mean Square Error (NRMSE) and histograms to know the order of the polynomials utilizing a portion of each signal. The polynomials orders are 22nd for the P and T wave sections.

- We adequate the filtered signals into a Non-Gaussian random field, collapsing each EKG cycle filtered signal and add some zeros to homogenized.

- We search the excursion set at each level U. This part transforms the signal and the NGRF into a geometrical structure.

- For extracting a topological invariant feature of the excursion set, we transform the excursion set into a 2D image and extract the values of the mean Euler-Poincaré characteristic using Gray's algorithm.

- Then, we decomposed the MEPC model to get a mathematical model to express the behavior of the MEPC from an NGRF. DMEPC parameters were obtained using the PSO algorithm, and for the "separations/divisions," we use an empirical method.

- The DMEPC models fit the MEPC models with the following percentages 77% for Control, 71% for MI, 78% for stroke, and 77% for syncope.

- The developed method can be a helpful tool for a doctor since it extracts a single image of the long-term recording (Holter EKG).

Chapter 5

Classify Sleep Apnea

Sleep apnea is a disorder that affects many people in the world. In the USA, 22 million people suffer from sleep apnea, and almost 80 % of the cases sustained with severe obstructive sleep apnea undiagnosed. Obstructive sleep apnea (OSA) can complicate everything if left untreated, leading to heart problems like atrial fibrillation, heart failure, high blood pressure, stroke, and others. It is also associated with type 2 diabetes and depression and is a risk factor because of sleepy drivers that don't know about this condition.

The public and the health community are generally aware of the population's increasing obesity, a phenomenon related to the increase in sleep apnea. However, few people from the sleep medicine community are aware that too little good sleep appears to be a factor in obesity as too much food and too little exercise.

OSA do not care about the people's age, it includes infants and children, but it is most frequently seen in men over 40, significantly overweight or obese men.

During sleep, cessation of breathing can last seconds or minutes when sleep apnea is presented. It effect on the cardiovascular system, and these effects are associated with physiological functions such as systemic hypertension and it increased sympathetic activity that compromises the heart (ref).

da Silva Pinho et al. [19] analyzed the EKG signal to detect sleep apnea using a median filter to obtain the HVR and the EKG-derived respiration (EDR), then extracts features from these analyses, and finally trained an artificial neural network.

Burgos et al. [55] developed an alternating decision trees classifier using the per-

centage of oxygen saturation signals taken from pulse oximetry systems (SpO2). Also, Lee et al. [56] used these signals to classify sleep apnea but, they implemented a Wavelet transform and measured the time when the SpO2 decreased.

McNames and Fraser [57] use the EKG to detect sleep apnea using the RR variability values, S-wave and T-wave amplitude, and EKG pulse energy into thresholds about 5 min. Other research used spectral features using the EDR and RR signals minute by minute [58] or using a linear and quadratic discriminant classifier [59, 60].

In our approach, we get the MEPC model from the EKG signal, and the OSA can be detected using the EKG Holter data. The EKG apnea database [61] was analyzed using our approach, but for classifying, we took another tool.From many models of classifiers, we selected an ensemble KNN classifier. In the next section, the methodology proposed for this study case is presented.

5.1   Methodology

In this section, we present an approach to classify the EKG signal and detect sleep apnea. Figure 5.1 shows a general view of the method; the first point is to generate the polynomials and then create a non-Gaussian random field using them after extracting the Euler-Poincaré characteristic, and finally describe it mathematically. But in this case, instead of acquiring the mathematical model, we add a classifier. Figure 5.1 shows the new method.

The methodology developed in this research have five steps, enumerated below:

1. Polynomization: the first step, using the random algebraic polynomial (RAP) as a filter. RAP is the filtered signal of the EKG raw signal. Here is the pre-processing part of the methodology.

2. Building the non-Gaussian random field: using the filtered signal, we stacked all the filtered cardiac cycles and made this random field. In other words, we constructed the geometrical structure using the filtered signals.

3. Excursion sets: a principal concept in the random field theory, is simply the set of points where the RF exceeds a fixed threshold value $u \in \mathbb{R}$. These belong to the RF.

4. Getting the mean Euler-Poincaré characteristic: using the excursion sets level, we extract this topological feature that indicates the number of connected components in the excursion set minus the number of holes.

5. An ensemble KNN classifier model is proposed to classify sleep apnea from Holter EKG. In this case, we select several EKG data for the training and validation set to have the final results.
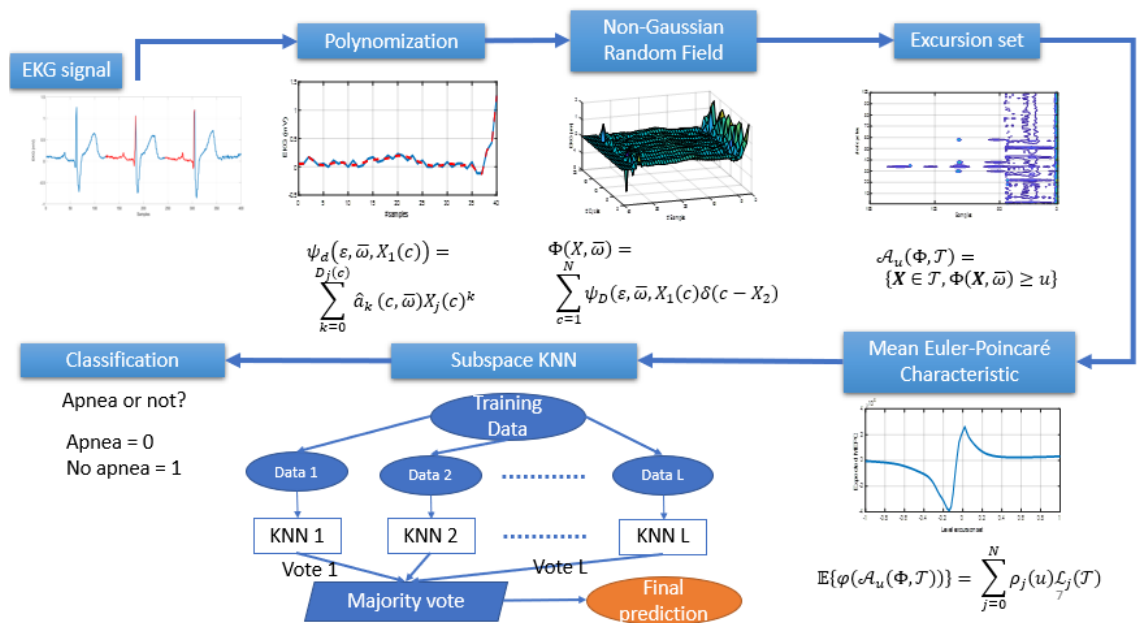
Figure 5.1: The resume of the new methodology.

5.2   Study case

The dataset has 70 records, divided into two. The first set is the learning set and have 35 records, where we have 20 severe cases, 5 started instances, and ten no apnea cases. The second set is the validation with 35 records. All of them downloaded from the Apnea-ECG database [61]. All of the EKG signals have a bunch of apnea and QRS annotations. The EKG signal was digitized at 100 Hz with a 12-bit resolution of approximately 8 h, corresponding to men and women between 27 and 60 years old.

5.3   Implementing the methodology

For this case, the first four steps are the same as the chapter 4 for myocardial infarction. This section has the changes in each step, the first step needs a new polynomial order, and we used a classifier for this patient.

5.3.1   Polynomization

Remember that we need to choose an order of polynomials for maintaining homogeneity; in this case, we selected 100 000 samples of each patient from the training set.

The following procedure was carried out to obtain the order:

1. Find the R-peaks in each cycle using the QRS annotations.

2. Find the middle points between the interval R-R.

3. For each interval, we get the results with the order from 15 to 25.

4. The measure of the fit for each order and the raw EKG is getting using the Eq. (4.1) and the ITAE performance index.

5. Choose the best fit in each cycle.

6. A frequency histogram plots the relationship between the fit and IAE of the orders polynomials from 15 and 25. Figures 5.3 and 4.3 shows these results.

On this occasion, we used both histogram results, so with the FIT criterion, the polynomial order is 15, and with the IAE criterion, the 21th polynomial order.

To build the NGRF, we selected the filtered signals from 15 and 21 orders and constructed the geometrical structure for these EKG signals.
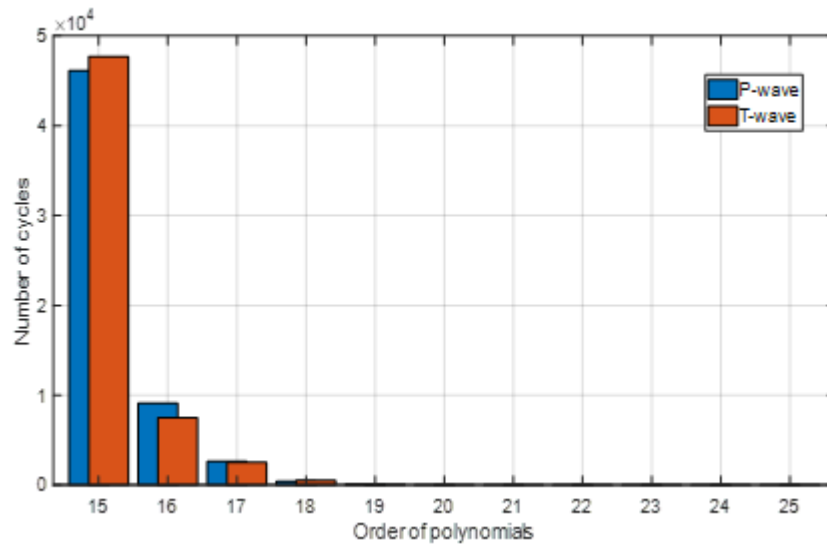
Figure 5.2: Histogram to select the best order of the polynomials (P-waves and T-waves) for each cycle using the FIT criterion
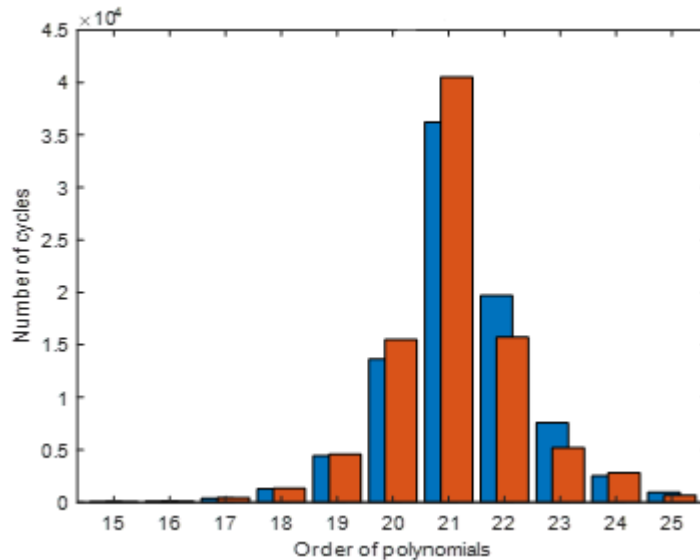
Figure 5.3: Histogram to select the best order of the polynomials (P-waves and T-waves) for each cycle using the IAE criterion

5.3.2   Mean Euler-Poincaré characteristic

In this step, we transform the geometrical structure to a binary image to get the Euler-Poincaré characteristic.

When the value of $t \in \Phi(\mathbf{X}, \bar{\omega})$ is higher than the level u, so the value for the binary image is 1, otherwise it is 0. Figures 4.7 and 4.8 represent an example of a binary image from an excursion set.

The EPC is a scalar value representing the number of connections minus the number of holes in those elements [42]. To get these values, we used Gray's algorithm [54] to the binary image of each excursion set A_u. The levels u selected were in the interval $[-1, 1]$ divided into 100 points.

Remember that we analyze the training set in the first place, so each signal acquired 100 values that formed the EPC model due to the levels. Figures 5.4 and 5.5 show the model EPC from P-waves for each patient analyzed.

By not having a similarity in the MEPC models' behaviors and not finding an adequate function representing these behaviors in their entirety, we have obtained to use another tool to classify these EKGs.
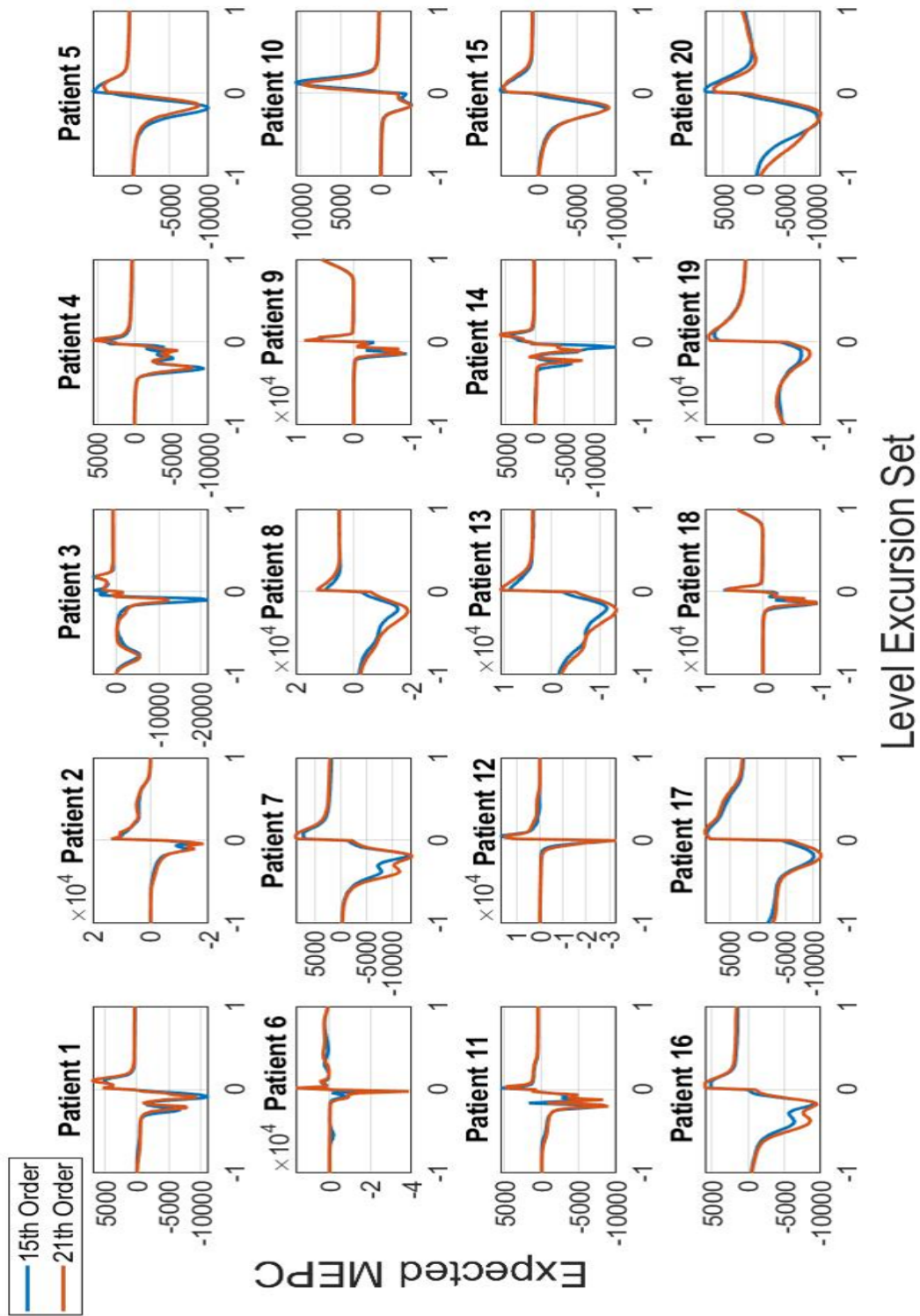
Figure 5.4: Expected EPC from the OSA patients from the training set.
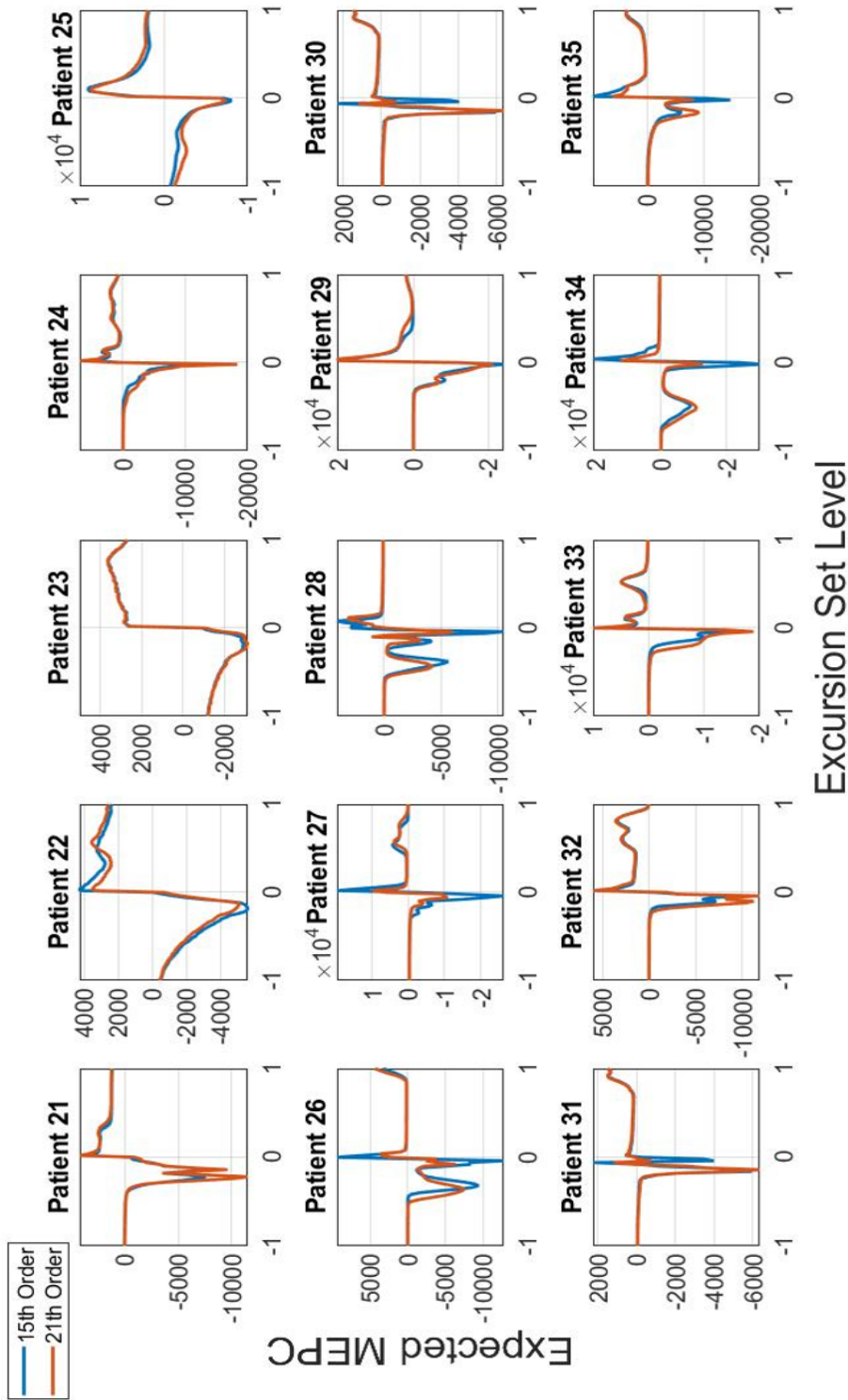
Figure 5.5: Expected EPC from bordeline apnea (21-25) and no apnea (26-35) patients from the training set.

5.3.3 Ensemble KNN classifier

The KNN classifier has only two parameters, the Euclidean distance and the number of neighbors k compared with the neural networks and decision tree classifiers with more parameters. The KNN represents a modern and straightforward way to classify.

The KNN classifier has a "good" performance in classifying different classes from a series of data. But, when we have a "bad" accuracy using the training and test data. One solution is to connect multiple classifiers, known as ensemble methods, to improve learning algorithms' prediction performance, especially in non-informative features in the data sets.

We don't have non-informative features; instead, we have an MEPC model that describes the behavior of the high peaks in different levels of the NGRF. In the case of MI, we notice that the action of the MEPC is similar in each patient. In this case, Figures 5.4 and 5.5 represent the MEPC model for each patient, and we know it is challenging to analyze them. But we know that this model revealed a lot of information.

To get the ensemble KNN classifier, we use the 101 values from each patient's EPC model. These values were the classifiers' predictors, and the classes are sleep apnea (representing the value 0) and no apnea (with 1) in the 102 columns in the X input of the classifier. The matrix below represents the matrix X of inputs.

$$\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 & t_1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 & t_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^m & x_2^m & \cdots & x_n^m & t_m \end{bmatrix}$$

where x_i^j are the predictors with $i = [1, 101]$, and t_i is the class of each patient.

The random subspace method relies on a stochastic process that randomly selects several components of the given feature vector in constructing each classifier [62]. In KNN, that means when a test sample is compared to a prototype, only the selected features have nonzero contributions to the distance. Geometrically this is equivalent to projecting all the points to the chosen subspace, and the k nearest neighbors are found using the projected distances. Each time a random subspace is selected, a new set of k nearest neighbors are computed. The k nearest neighbors in each

selected subspace are then assembled for a majority vote on the test sample's class membership. The same training sample may appear more than once in this ensemble if it happens to be among the k nearest neighbors in more than one selected subspace.

So that, we have a set of N points in an $n-$dimensional feature space

$$(x_1, x_2, \ldots, x_n)|x_i \text{is real } \forall 1 \leq i \leq n,$$

we consider the $m-$dimensional subspaces

$$(x_1, x_2, \ldots, x_n)|x_i = 1 \text{for } i \in I, x_i = 0 \text{ for } i \notin I,$$

where I is an $m-$element subset of $1, 2, \ldots n$, and $m < n$.In each pass, a subspace is chosen by randomly selecting an I from $C(n,m)-$many choices. All points are projected into the subspace chosen. For each testing point, k nearest neighbors $(1 \leq k \leq N)$ passes, the test point is assigned to the class with the most frequent occurrences in the list C. The method is a derivative of stochastic discrimination where many stochastically created weak classifiers are combined for a nearly monotonic increase inaccuracy [63, 64].

In this research, we implement an ensemble random KNN classifier using the tool of Matlab, and the results showed below are using an AMD Ryzen 5 3600, 16 GB of RAM, and the parallel pool option. Using this tool, we have the next diagram to explain the ensemble.
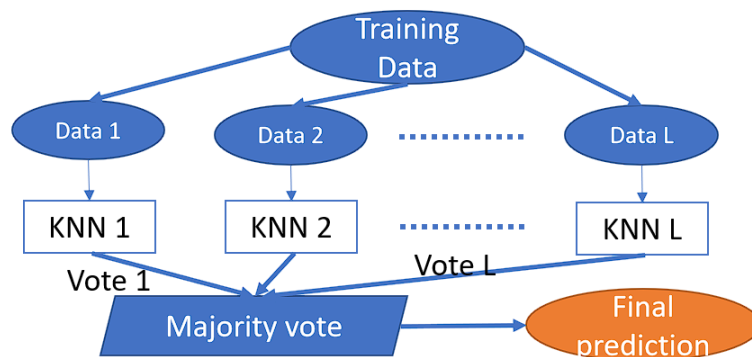
Figure 5.6: Diagram of the KNN ensemble classifier.

We choose 30 learners, which are 30 subspaces, with 65 random entries in each learner (predictors)—a total of 35 observations (35 MEPC model values) for the training phase and 35 ones for the test phase. The next section explains the confusion matrix and the results of this model.

5.4   Results and discussion

This chapter found a classifier model for sleep apnea and not apnea cases using an ensemble KNN classifier. The first result is with the training phase. The dataset comprises 20 sleep apnea with more than an hour of activity, 5 with less than an hour, and ten patients without apnea. The result is the next confusion matrix showed in Figure 5.7. In this phase, we get an accuracy of the 80%, a sensitivity of 82%, and a specificity of 71%.
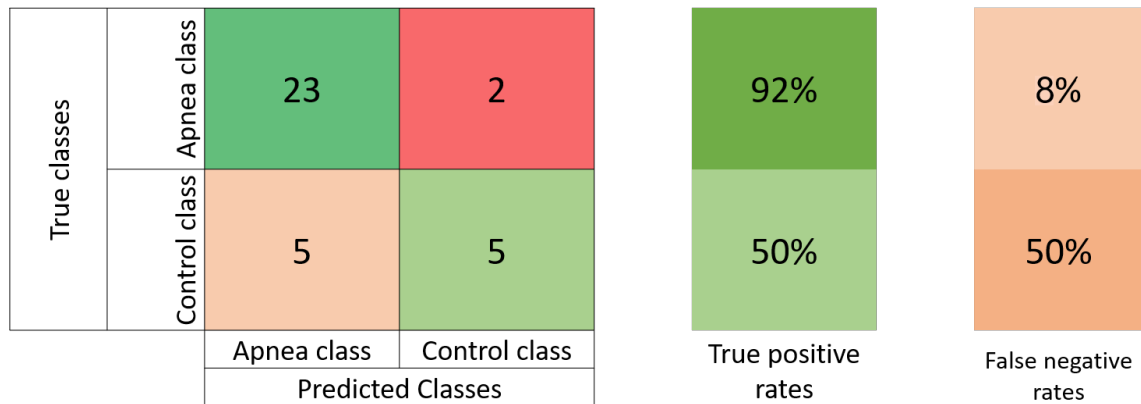
Figure 5.7: Confusion matrix for the training phase

We observed that we have 92% of accuracy in the apnea class and 50% in the control class. To get the values of the accuracy, sensitivity, and specificity, we used the following formulas.

$$Accu = \frac{TP + TN}{P + N} = \frac{23 + 5}{35} * 100\% = 80\%$$

$$Sen = \frac{TP}{TP + FN} = \frac{23}{23 + 5} * 100\% = 82\%$$

$$Spec = \frac{TN}{TN + FP} = \frac{5}{5 + 2} * 100\% = 71\%$$

where TP is true positive, TN is true negative, P are the positive cases, N the negative cases, FN is false positive, and FP is false negative.

Remember, we have two different polynomials to filter the signal in the pre-processing step, which are 15 and 21 orders. In both cases, the training phase was the same.

For the test phase, we have two different confusion matrices. The first result using 15 orders in the polynomials to filter the signal is the next confusion matrix shown in Figure 5.8. And the accuracy is 97%, sensitivity 96%, and specificity 100%.
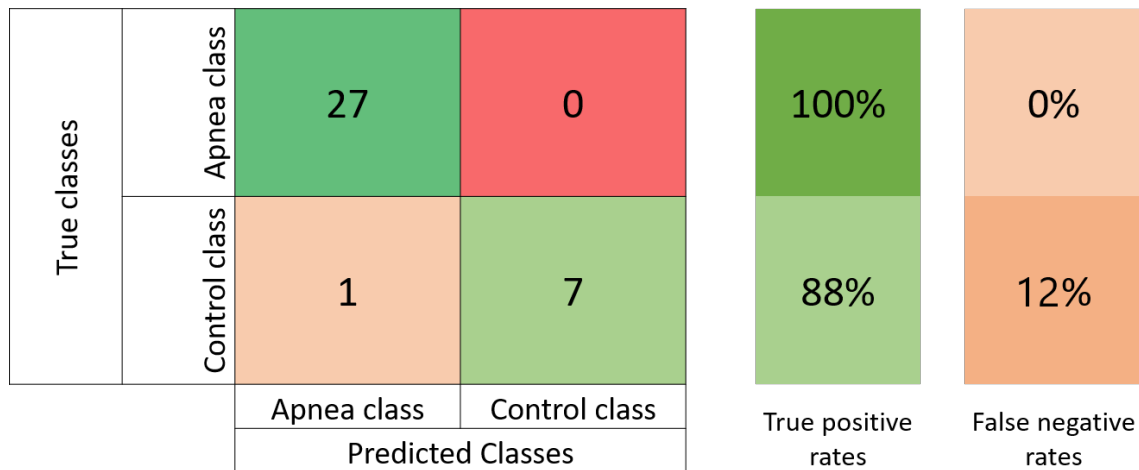
Figure 5.8: Confusion matrix for the test phase using 15 order polynomials.

We observed that we get 100% of accuracy in the apnea class and 88% in the control class. To get the values of the accuracy, sensitivity, and specificity, we used the following formulas.

$$Accu = \frac{TP + TN}{P + N} = \frac{27 + 7}{35} * 100\% = 97\%$$

$$Sen = \frac{TP}{TP + FN} = \frac{27}{27 + 1} * 100\% = 96\%$$

$$Spec = \frac{TN}{TN + FP} = \frac{7}{7 + 0} * 100\% = 100\%$$

And finally, Figure 5.9 shows the last confusion matrix with the 21 order of polynomials to filter the signal. We obtain an accuracy of 94%, a sensitivity of 93%, and a specificity of 100%.
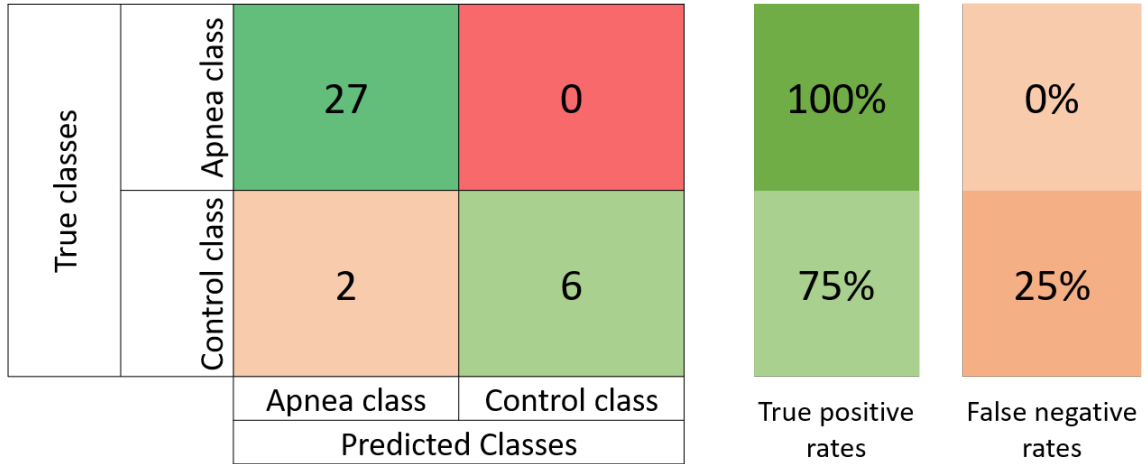
Figure 5.9: Confusion matrix for the test phase using 21 order polynomials

We observed that we have 100% of accuracy in the apnea class and 75% in the control class. To get the values of the accuracy, sensitivity, and specificity, we used the following formulas.

$$Accu = \frac{TP + TN}{P + N} = \frac{27 + 6}{35} * 100\% = 94\%$$

$$Sen = \frac{TP}{TP + FN} = \frac{27}{27 + 2} * 100\% = 93\%$$

$$Spec = \frac{TN}{TN + FP} = \frac{6}{6 + 0} * 100\% = 100\%$$

Specificity is the ability to correctly exclude individuals who do not have a given disease or disorder. For example, if we have a value of 90% specific. If 100 healthy individuals are tested with that method, only 90 of those 100 healthy people will be found to be a control patient. The other ten people also do not have the disease, but their test results indicate they do. For that, 10% of their "abnormal" findings are misleading false-positive results. In this case, we have a value of 100% of specific, which means our model could correctly classify all the control cases. So the model

did not produce false-positive results.

Sensitivity is the ability to identify people who have a given disease or disorder correctly. For example, if we have 90% sensitivity in a test. It means that if 100 people are known to have a particular disease, the test will correctly identify 90 of those 100 cases of the disease. The other ten people who were tested also have the disease, but the test will fail to detect it. For that 10%, the finding of a "normal" result is a misleading false-negative result. A test's sensitivity becomes particularly important when you are seeking to exclude a dangerous disease. Our model gets a 96% and a 93%, which means the model can detect correctly with these percentages and give 1 or 2 false negatives.

So, if we need to select one model, we choose the first one, using a 15th polynomial order to smooth the signal.

5.5   Comparative with the literature

In this section, we can say that we produced a classifier that can detect sleep apnea using a Holter EKG signal. Other works in the literature also can see the same disease and classified according to different perspectives.

Some works try to decide if a patient suffers apnea or not using an overall diagnostic over the whole biomedical signal. For example, they used the δ index, which is the differences in oxygen saturation in a 12s interval [65, 66], and non-linear methods as central tendency measure (CTM) and Lempel–Ziv (LZ) complexity [67].

Other works used fragments of the signal (minute by minute [57], or using windows of variable size [68]) to identify the apnea episodes and do not attempt to make a general diagnosis of SAHS, which is considered as a responsibility of the specialists.

Some works use their database created with data obtained from many patients [65, 66, 67]. Others used the same database that we use, which is the Apnea-ECG Database [68].

Different biomedical signals are used to classify and diagnoses the SAHS. For example, photopletismography signal [69], S_pO_2 signal [56, 67, 55], oronasal air flow signal [70], ECG signal [66, 57], etc.

Those works use different techniques or methods to analyze these kinds of sig-

nals. For example, RR variability [57, 59, 71], wavelet transform [56, 58], neuronal networks [70], etc.

Our approach is different as the previous works, an following these considerations:uses the whole recording of each patients, but only the P-waves sections,uses the Apnea-ECG Database, analyzes only the P-wave from the EKG signal after the filtering phase, and employs the mean Euler-Poincare characteristic and a neuronal network.

The first works, [57], [58], and [59], are results from the 2000 challenge organized by Physionet and Computers in Cardiology [61] and are the best results. The challenge consisted of displaying the effectiveness of EKG-based methods for sleep apnea detection using the EKG signal using the Apnea-ECG database.The middle works, i.e., [56], [71], [60], and [55], are to show the many techniques implemented and also their performance results using EKG and respiratory signals.

McNames and Fraser [57] got an accuracy of 92.62% by doing a spectral analysis of heart rate variability and using different ECG-derived parameters like ECG pulse energy and amplitude of the S-component. Raymond et al. [58] got an accuracy of 92.30% using the T-wave amplitude's spectral analysis using wavelet transform (WT). Also, they get some power spectral features from the EKG-derived respiratory signal (EDR). And the RR interval tachogram was computed.

de Chazal et al. [59] got an accuracy of 89.36% by spectral analysis of heart rate variability and R-wave amplitude. They created a set of characteristics based on RR and PR intervals, obtained from each minute of data and others related to series of alternating bradycardia and tachycardia.

Lee et al. [56] worked with the S_pO_2 signals from the eight recordings with the EKG and the respiratory signs from the Apnea-ECG database. They compared the performance obtained by WT to three conventional algorithms. With the wavelet algorithm, they got an accuracy of 82%. The other traditional algorithms, wavelet algorithm (WA), amplitude duration algorithm (ADA), drop duration algorithm (DDA), and nervus algorithm (NA), got accuracy above 95% but using tuned parameters, which do not reflect acceptable medical values. When they adjusted their algorithms with good results, the performance decreased to 88%.

Corthout et al. [60] compared three strategies to evaluate obstructive sleep apnea

(OSA). The three methods, one based on the Empirical Mode Decomposition (EMD) and the Hilbert transform (HT), one based on EMD and reassigned spectrogram (RAS), and one based on wavelet analysis (WA) and HT, were able to classify the minute-by-minute signal with an accuracy of around 90%.

Burgos et al. [55] made a classifier using a bagging method in a decision tree (ADTree) that identifies sleep apnea from the blood oxygen saturation taken from SpO2. Their strategy reached 93.03% of accuracy, with a sensibility and specificity of 92.35% and 93.52%, respectively. Also, they get a receiver operating characteristic-area under the curve (ROC-AUC) of 98.5% on SpO_2 signals.

The last three works, i.e., [72], [73], and [74], are more recent works that also offer performance results to detect sleep apnea; Ye et al. [72] got a 99% using a frequency extraction network (FENet), using the RR-interval during one minute as inputs. FENet generates an implicit representation of respiration with a group of multi-frequency filters and a frequency feature extractor. Then a convolutional layers and fully connected layers are adopted to produce three independent classification results.

Rajesh et al. [73] worked with the Apnea-ECG database, they used a random forest classifier (RFC) to get a 90% of accuracy. They extracted three sets of features for the RFC inputs, namely moments of power spectrum density (PSD), waveform complexity measures, and higher-order moments extracted from the 1-min segmented EKG subbands obtained from discrete wavelet transform (DWT). After, correlation-based feature selection with particle swarm optimization (PSO) search method is employed for getting an optimum feature vector.

Li et al. [74] evaluated the features from the EKG, SpO2, and the body mass index (BMI) using a multi-layer feed-forward neural network (FNN). They achieved with the multi-layer FNN an accuracy, sensitivity, and specificity of 97.8%, 98.6%, and 93.9%.

Table 5.1 summarizes each related work, including performance measures, the signal employed, the techniques used, and the decision method chosen to state whether there is an apnea or not on a specific signal interval. We can see that our approach offers suitable performance measures (sensitivity, specificity, and accuracy) from such results. Moreover, it is the only one that uses a geometrical approach.

Table 5.1: A comparative of our approach vs the literature approaches

Research work	Performance	Signal analyzed	Techniques employed	Decision method
McNames and Fraser [57]	Acc=92.6%	EKG	RR variability, EKG pulse Energy, S-wave amplitude, T-wave amplitude (WT)	Threshold (5 minute window)
Raymond et al. [58]	Acc=92.3%	EKG	EDR signal, RR signal	Shared mixture classifier (spectral features)
de Chazal et al. [59]	Acc=89.4%	EKG	RR variability, R-wave amplitude (PSD)	Classifier based on linear and quadratic discriminants (spectral and novel features)
Lee et al. [56]	Acc=88%	S_pO_2	WT, ADA,DDA,NA	Transform coefficients Threshold
Maier et al. [71]	Sens=81.3%, Spec=82.8%	EKG	RR series, MAV series	Threshold (120 ms. window)
Corthout et al. [60]	Acc=90%	EKG	EDM+HT, EDM+RAS, WA	Linear Discriminant classifier (feature set)
Burgos et al. [55]	Acc=93.03%, Sens=92.35%, Spec=93.52%	S_pO_2	Desaturation indixes, moving baseline	Bagging with ADTree
Ye et al. [72]	Acc=99.22%, Spec=99.02%	EKG	RR intervals, moving baseline	FENet
Rajesh et al. [73]	Acc=89%, ROC=96%	EKG	PSD, DWT, PSO	Random forest classifier
Li et al. [74]	Acc=97.8%, Sens=98.6%, Spec=93.9%	EKG, SpO_2, BMI	Feature extraction	Multi layer FNN
Our approach	Acc=89%, Sens=89%, Spec=86%	EKG	Random Field, Mean Euler-Poincaré Characteristic	Ensemble KNN classifier

5.6   Conclusion

This chapter presents a change in the methodology and applying a classifier to the sleep apnea cases from the EKG Holter.

- In the polynomization section, we selected the NRMSE vs. IAE criterion. The results are that with the NRMSE, the order of the polynomials was 15th, and with the IAE 21th order.

- We use the same method and exact algorithms for the NGRF, excursion set, and MEPC characteristic extraction.

- Then, we don't apply the DMEPC approach. We use an ensemble KNN classifier to detect apnea or no apnea cases.

- With this classifier, we get in the learning phase 80% of accuracy, 82 % of sensitivity, and 71% of specificity.

- And in the test phase, we got 97% accuracy, 96 % sensitivity, and 100% specificity.

- The developed method can be a helpful tool for a clinician since it can classify Holter EKG for sleep apnea disease.

Chapter 6

Conclusion

The procedure to get an EKG model using the random field theory from Adler and Taylor is new. It describes the steps to pass from an EKG signal to a geometrical structure called a random field, then transform from an RF to a 2D image, and finally to the MEPC model.

The Gaussian methods can not describe the mathematical expression for the MEPC model of our study cases, so the new approach based on decomposition was related to express these models. This method was called decomposition of the mean Euler-Poincaré characteristic.

The resume methodology is divided into polynomization, building NGRF, excursion sets, 2D image transforming, MEPC feature extraction, and getting a model.

In the case of the DMEPC that describes the behavior of the MEPC model mathematically, we get good approximations with 71-77 % of fit. Also, we observe that the place where the minimums and maximums are can be an indicator to differentiate one disease from another.

Also, in the MI case, where we use a Holter EKG, we construct a possible tool for the physicians to identify some diseases using the DMEPC models as templates and with some values as the difference from the minimum and maximum and the difference from level U where the values before are.

Using the idea below, we get four final templates for MI, stroke, syncope, and control cases. Remember, the sample space is a population of older people with some heart problems and heart diseases.

In the second case of the methodology, we applied an ensemble KNN. Because if we are against an NGRF where the mother wave proposed can not describe the MEPC model behavior, we can choose another method that helps to reach the goal.

An ensemble KNN can be chosen because it was easy to implement. If we took the learning and test phase results, we get these finals indicators: 89% sensitivity, which means how good the classifier is at detecting the sleep apnea cases; 85.5% specificity indicates how good the classifier is to see the normal status; and 88.5% accuracy, it measures the degree of veracity of the classifier to detect both cases (sleep apnea or no apnea).

In this case, we choose two criteria to select the order of polynomials, the NRMSE equation vs. IAE criterion, and when we compared all the graphs, we watch that the two are so closed. But we can conclude that the NRMSE criterion can satisfy in both study cases.

Also, the method used for sleep apnea based on the topological invariant MEPC and the KNN classifier can be considered a good result compared with the classical methods.

This research is an introduction to use a feature expected used in biomedical images, but not in time series biomedical signals. So this research has wide-open possibilities, as the name of this document, theory of identification systems applied to medical diagnoses and treatment of biomedical signals.

We used one of many biomedical signals, and we choose the Holter EKG as the first option. Our journey to discover more benefits for this approach will continue.

6.1   Future works

The methodology described before is not a static method. We can adapt many resources to create models of the Holter EKG, using different mother-waves and using classifiers after the DMEPC models. So that, we proposed the next works:

1. Get the MEPC model for other cardiac diseases as AFIB, coronary artery disease, MI in long terms (30 min), others.

2. The MEPC models will use as input of some classifiers (KNN, SVM,NN, others).

3. Get the MEPC model for breast cancer images and use a classifier.

4. Use the MEPC model with mechanical process, to analyze the vibrations.

5. DMEPC is based on the NGRF from the EKG time cycles. Another approach will consist of built the NGRF from the EKG frequency cycles using FFT and compare DMEPC in time and frequency.

Bibliography

[1] Lennart Ljung. *System Identification (2nd Ed.): Theory for the User*. Prentice Hall PTR, USA, 1999. ISBN 0136566952.

[2] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003.

[3] Ali Jalali, Ali Ghaffari, Parham Ghorbanian, and Chandrasekhar Nataraj. Identification of sympathetic and parasympathetic nerves function in cardiovascular regulation using anfis approximation. *Artificial Intelligence in Medicine*, 52(1): 27 – 32, 2011. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2011.01. 002.

[4] Alexander C. Koenig, Luca Somaini, Michael Pulfer, Thomas Holenstein, Ximena Omlin, Martin Wieser, and Robert Riener. Model-based heart rate prediction during Lokomat walking. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pages 1758–1761, 2009. ISSN 1557-170X. doi: 10.1109/IEMBS.2009.5333096.

[5] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[6] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.

[7] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 274–277. IEEE, 2016.

[8] Varios Autores. Diez principales causas de mortalidad de enero a marzo de 2019, 2019. URL https://www.cardiologia.org.mx/transparencia/transparencia_focalizada/estadisticas/.

[9] U Rajendra Acharya, Hamido Fujita, Vidya K Sudarshan, Shu Lih Oh, Muhammad Adam, Joel EW Koh, Jen Hong Tan, Dhanjoo N Ghista, Roshan Joy Martis, Chua K Chua, et al. Automated detection and localization of myocardial infarction using electrocardiogram: a comparative study of different leads. *Knowledge-Based Systems*, 99:146–156, 2016.

[10] U Rajendra Acharya, Hamido Fujita, Muhammad Adam, Oh Shu Lih, Vidya K Sudarshan, Tan Jen Hong, Joel EW Koh, Yuki Hagiwara, Chua K Chua, Chua Kok Poo, et al. Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ecg signals: A comparative study. *Information Sciences*, 377:17–29, 2017.

[11] U Rajendra Acharya, Vidya K Sudarshan, Joel EW Koh, Roshan Joy Martis, Jen Hong Tan, Shu Lih Oh, Adam Muhammad, Yuki Hagiwara, Muthu Rama Krishanan Mookiah, Kok Poo Chua, et al. Application of higher-order spectra for the characterization of coronary artery disease using electrocardiogram signals. *Biomedical Signal Processing and Control*, 31:31–43, 2017.

[12] Mohit Kumar, Ram Bilas Pachori, and U Rajendra Acharya. Characterization of coronary artery disease using flexible analytic wavelet transform applied on ecg signals. *Biomedical signal processing and control*, 31:301–308, 2017.

[13] Swati Banerjee and M Mitra. Cross wavelet transform based analysis of electrocardiogram signals. *International Journal of Electrical, Electronics and Computer Engineering*, 1(2):88–92, 2012.

[14] S. Banerjee, R. Gupta, and M. Mitra. Delineation of ecg characteristic features using multiresolution wavelet analysis method. *Measurement*, 45(3):474 – 487, 2012. ISSN 0263-2241. doi: https://doi.org/10.1016/j.measurement.2011.10.025.

[15] Muhammad Arif, Ijaz A Malagore, and Fayyaz A Afsar. Detection and localization of myocardial infarction using k-nearest neighbor classifier. *Journal of medical systems*, 36(1):279–289, 2012.

[16] Azam Davari Dolatabadi, Siamak Esmael Zadeh Khadem, and Babak Mohammadzadeh Asl. Automated diagnosis of coronary artery disease (cad) patients using optimized svm. *Computer methods and programs in biomedicine*, 138: 117–126, 2017.

[17] Li Sun, Yanping Lu, Kaitao Yang, and Shaozi Li. Ecg analysis using multiple instance learning for myocardial infarction detection. *IEEE transactions on biomedical engineering*, 59(12):3348–3356, 2012.

[18] Hang Sik Shin, Chungkeun Lee, and Myoungho Lee. Ideal filtering approach on dct domain for biomedical signals: index blocked dct filtering method (ib-dctfm). *Journal of medical systems*, 34(4):741–753, 2010.

[19] André Miguel da Silva Pinho, Nuno Pombo, and Nuno M Garcia. Sleep apnea detection using a feed-forward neural network on ecg signal. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, 2016.

[20] Albert T. Bharucha-Reid and M. Sambandham. *Random polynomials: Probability and mathematical statistics: a series of monographs and textbooks*. Academic Press, 1986. doi: https://doi.org/10.1016/B978-0-12-095710-1.50007-3.

[21] Mark Kac. On the average number of real roots of a random algebraic equation (II). *Proceedings of the London Mathematical Society*, s2-50(1):390–408, Jan. 1948. doi: 10.1112/plms/s2-50.5.390.

[22] Alan Edelman and Eric Kostlan. The road from Kac's matrix to Kac's random polynomials. In *In Proceedings of the 1994 SIAM Applied Linear Algebra Conference (Philadelphia)*, pages 503–507, 1994.

[23] Larry A Shepp and Robert J Vanderbei. The complex zeros of random polynomials. *Transactions of the American Mathematical Society*, 347(11):4365–4384, Nov. 1995. doi: 10.1090/S0002-9947-1995-1308023-8.

[24] Christopher P Hughes and Ashkan Nikeghbali. The zeros of random polynomials cluster uniformly near the unit circle. *Compositio Mathematica*, 144(3):734–746, May. 2008. doi: 10.1112/S0010437X07003302.

[25] Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer-Verlag New York, New York, 1 edition, 2007.

[26] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012. ISSN 1935-8237. doi: 10.1561/2200000013.

[27] E Vanmarcke. *Random Fields: Analysis and Synthesis*. MIT Press, Cambridge MA, 1983.

[28] Petter Abrahamsen. *A review of Gaussian Random Fields and Correlation Functions*. Norwegian Computing Center, Apr. 1997.

[29] R. A. Shaikh, J. Li, A. Khan, and I. Memon. Biomedical image processing and analysis using markov random fields. In *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 179–183. IEEE, Dec. 2015. doi: 10.1109/ICCWAMTIP. 2015.7493970.

[30] G. Zhang, S. Tzoumas, K. Cheng, F. Liu, J. Liu, J. Luo, J. Bai, and L. Xing. Generalized Adaptive Gaussian Markov Random Field for X-Ray Luminescence Computed Tomography. *IEEE Transactions on Biomedical Engineering*, 65(9): 2130–2133, Sep. 2018. doi: 10.1109/TBME.2017.2785364.

[31] Ravi Bansal, Lawrence H Staib, Dongrong Xu, Hongtu Zhu, and Bradley S Peterson. Statistical analyses of brain surfaces using gaussian random fields on 2-d manifolds. *IEEE Transactions on Medical Imaging*, 26(1):46–57, Jan. 2007. doi: 10.1109/TMI.2006.884187.

[32] Marco AR Ferreira and Victor De Oliveira. Bayesian reference analysis for gaussian markov random fields. *Journal of Multivariate Analysis*, 98(4):789–812, Apr. 2007. doi: 10.1016/j.jmva.2006.07.005.

[33] Keith J Worsley. Local maxima and the expected euler characteristic of excursion sets of χ 2, f and t fields. *Advances in Applied Probability*, 26(1):13–42, 1994.

[34] Robert J Adler, Gennady Samorodnitsky, Jonathan E Taylor, et al. High level excursion set geometry for non-gaussian infinitely divisible random fields. *The Annals of Probability*, 41(1):134–169, Jan. 2013. doi: 10.1214/11-AOP738.

[35] Salim S. Virani, Alvaro Alonso, Emelia J. Benjamin, Marcio S. Bittencourt, Clifton W. Callaway, April P. Carson, Alanna M. Chamberlain, Alexander R. Chang, Susan Cheng, Francesca N. Delling, Luc Djousse, Mitchell S.V. Elkind, Jane F. Ferguson, Myriam Fornage, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, Tak W. Kwan, Daniel T. Lackland, Tené T. Lewis, Judith H. Lichtman, Chris T. Longenecker, Matthew Shane Loop, Pamela L. Lutsey, Seth S. Martin, Kunihiro Matsushita, Andrew E. Moran, Michael E. Mussolino, Amanda Marma Perak, Wayne D. Rosamond, Gregory A. Roth, Uchechukwu K.A. Sampson, Gary M. Satou, Emily B. Schroeder, Svati H. Shah, Christina M. Shay, Nicole L. Spartano, Andrew Stokes, David L. Tirschwell, Lisa B. VanWagner, and Connie W. Tsao and. Heart disease and stroke statistics—2020 update: A report from the american heart association. *Circulation*, 141(9), March 2020. doi: 10.1161/cir.0000000000000757.

[36] Dale Davis. *Interpretación del ECG. Su dominio rápido y exacto.* Ed. Médica Panamericana, 2007.

[37] Keith J Worsley et al. Estimating the number of peaks in a random field

using the hadwiger characteristic of excursion sets, with applications to medical images. *The Annals of Statistics*, 23(2):640–669, 1995.

[38] K. J. Worsley. Boundary corrections for the expected euler characteristic of excursion sets of random fields, with an application to astrophysics. *Advances in Applied Probability*, 27(4):943–959, 1995. doi: 10.2307/1427930.

[39] L.A. Santaló. Integral geometry and geometric probability. *Encyclopedia of Mathematics and Its Applications*, 1(6), 1979. doi: 10.1002/zamm.19790590633.

[40] John M. Lee. *Introduction to Topological Manifolds*. Springer New York, 2011. doi: 10.1007/978-1-4419-7940-7. URL https://doi.org/10.1007/978-1-4419-7940-7.

[41] Robert J. Adler. On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability*, 10(1):1–74, 2000. ISSN 10505164. URL http://www.jstor.org/stable/2667187.

[42] Eitan Richardson and Michael Werman. Efficient classification using the euler characteristic. *Pattern Recognition Letters*, 49:99–106, 2014.

[43] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11), 2016. ISSN 2305-5847. URL https://atm.amegroups.com/article/view/10170.

[44] Scott Cost and Steven Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, 10(1):57–78, 1993.

[45] Zhongheng Zhang. Too much covariates in a multivariable model may cause the problem of overfitting. *Journal of thoracic disease*, 6(9):E196, 2014.

[46] MathWorks. *Mastering machine learning:A Step-by-Step Guide with MATLAB*. The MathWorks, 2019.

[47] M. Ramos-Martinez, C. Corbier, V. M. Alvarado, and G. L. Lopez. Decomposed mean euler-poincaré characteristic model for a non-gaussian physiological random field. *IEEE Access*, 9:21180–21191, 2021. doi: 10.1109/ACCESS.2020.3048712.

[48] Donald E Knuth. Two notes on notation. *The American Mathematical Monthly*, 99(5):403–422, May. 1992. doi: 10.1080/00029890.1992.11995869.

[49] Instituto Nacional de Estadística y Geografía. URL https://www.inegi.org. mx/programas/mortalidad/#Datos_abiertos. Last update: Jan. 27, 2021.

[50] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[51] Emelia J. Benjamin and et. al. Heart disease and stroke statistics—2019 update: A report from the american heart association. *Circulation*, 139(10), March 2019. doi: 10.1161/cir.0000000000000659.

[52] Paolo Melillo, Raffaele Izzo, Ada Orrico, Paolo Scala, Marcella Attanasio, Marco Mirra, Nicola De Luca, and Leandro Pecchia. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PLOS ONE*, 10(3):e0118504, March 2015. doi: 10.1371/journal.pone.0118504.

[53] Marcin Rosiak, Halina Bolinska, and Jan Ruta. P wave dispersion and p wave duration on SAECG in predicting atrial fibrillation in patients with acute myocardial infarction. *Annals of Noninvasive Electrocardiology*, 7(4):363–368, October 2002. doi: 10.1111/j.1542-474x.2002.tb00186.x.

[54] William K Pratt. Digital image processing. pages 589–595. John Wiley & Sons, Inc., 2001.

[55] A. Burgos, A. Goñi, A. Illarramendi, and J. Bermudez. Real-time detection of apneas on a pda. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):995–1002, 2010. doi: 10.1109/TITB.2009.2034975.

[56] Y. K. Lee, M. Bister, P. Blanchfield, and Y. M. Salleh. Automated detection of obstructive apnea and hypopnea events from oxygen saturation signal. In *The 26th Annual International Conference of the IEEE Engineering in Medicine*

and Biology Society, volume 1, pages 321–324, 2004. doi: 10.1109/IEMBS.2004. 1403157.

[57] J. N. McNames and A. M. Fraser. Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram. In *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, pages 749–752, 2000. doi: 10.1109/CIC.2000. 898633.

[58] B. Raymond, R. M. Cayton, R. A. Bates, and M. Chappell. Screening for obstructive sleep apnoea based on the electrocardiogram-the computers in cardiology challenge. In *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, pages 267–270, 2000. doi: 10.1109/CIC.2000.898508.

[59] P. de Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley. Automatic classification of sleep apnea epochs using the electrocardiogram. In *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, pages 745–748, 2000. doi: 10.1109/CIC.2000.898632.

[60] J. Corthout, S. Van Huffel, M. O. Mendez, A. M. Bianchi, T. Penzel, and S. Cerutti. Automatic screening of obstructive sleep apnea from the ecg based on empirical mode decomposition and wavelet analysis. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3608–3611, 2008. doi: 10.1109/IEMBS.2008.4649987.

[61] Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

[62] Tin Kam Ho. Nearest neighbors in random subspaces. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 640–648. Springer, 1998.

[63] EM Kleinberg. Stochastic discrimination. *Annals of Mathematics and Artificial intelligence*, 1(1-4):207–239, 1990.

[64] EM Kleinberg et al. An overtraining-resistant stochastic modeling method for pattern recognition. *Annals of statistics*, 24(6):2319–2349, 1996.

[65] Ulysses J. Magalang, Jacek Dmochowski, Sateesh Veeramachaneni, Azmi Draw, M. Jeffery Mador, Ali El-Solh, and Brydon J.B. Grant. Prediction of the apnea-hypopnea index from overnight pulse oximetry*. *Chest*, 124(5):1694–1701, 2003. ISSN 0012-3692. doi: https://doi.org/10.1378/chest.124.5.1694. URL `https://www.sciencedirect.com/science/article/pii/S0012369215333997`.

[66] Conor Heneghan, Chern-Pin Chua, John F. Garvey, Philip de Chazal, Redmond Shouldice, Patricia Boyle, and Walter T. McNicholas. A Portable Automated Assessment Tool for Sleep Apnea Using a Combined Holter-Oximeter. *Sleep*, 31 (10):1432–1439, 10 2008. ISSN 0161-8105. doi: 10.5665/sleep/31.10.1432. URL `https://doi.org/10.5665/sleep/31.10.1432`.

[67] D Álvarez, R Hornero, D Abásolo, F del Campo, and C Zamarrón. Nonlinear characteristics of blood oxygen saturation from nocturnal oximetry for obstructive sleep apnoea detection. *Physiological Measurement*, 27(4):399–412, mar 2006. doi: 10.1088/0967-3334/27/4/006. URL `https://doi.org/10.1088/0967-3334/27/4/006`.

[68] E. Gil, J. M. Vergara, and P. Laguna. Study of the relationship between pulse photopletismography amplitude decrease events and sleep apneas in children. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3887–3890, 2006. doi: 10.1109/IEMBS.2006.259997.

[69] E. Gil, V. Monasterio, P. Laguna, and J. M. Vergara. Pulse photopletismography amplitude decrease detector for sleep apnea evaluation in children. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 2743–2746, 2005. doi: 10.1109/IEMBS.2005.1617039.

[70] P. Varady, T. Micsik, S. Benedek, and Z. Benyo. A novel method for the detection of apnea and hypopnea events in respiration signals. *IEEE Transactions on Biomedical Engineering*, 49(9):936–942, 2002. doi: 10.1109/TBME.2002.802009.

[71] C. Maier, H. Dickhaus, and P. Laguna. Amplitude variability extraction from multi-lead electrocardiograms for improvement of sleep apnea recognition. In *Computers in Cardiology, 2005*, pages 355–358, 2005. doi: 10.1109/CIC.2005. 1588110.

[72] G. Ye, H. Yin, T. Chen, H. Chen, L. Cui, and X. Zhang. Fenet: A frequency extraction network for obstructive sleep apnea detection. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2021. doi: 10.1109/JBHI.2021.3050113.

[73] Kandala.N.V.P.S. Rajesh, Ravindra Dhuli, and T. Sunil Kumar. Obstructive sleep apnea detection using discrete wavelet transform-based statistical features. *Computers in Biology and Medicine*, 130:104199, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.104199. URL https://www.sciencedirect.com/science/article/pii/S0010482520305308.

[74] Zufei Li, Yanru Li, Guoqiang Zhao, Xiaoqing Zhang, Wen Xu, and Demin Han. A model for obstructive sleep apnea detection using a multi-layer feed-forward neural network based on electrocardiogram, pulse oxygen saturation, and body mass index. *Sleep and Breathing*, pages 1–8, 2021.

[75] Ikaro Silva and George B Moody. An open-source toolbox for analysing and processing physionet databases in matlab and octave. *Journal of open research software*, 2(1), 2014.

Appendix A

Publication produced

Decomposed Mean Euler-Poincaré Characteristic Model for a Non-Gaussian Physiological Random Field

MOISES RAMOS-MARTINEZ[1], CHRISTOPHE CORBIER[2], VICTOR M. ALVARADO[1],
AND GUADALUPE LOPEZ LOPEZ[1]

[1]TecNM/CENIDET, Int. Internado Palmira s/n, 62490 Cuernavaca, Mexico
[2]LASPI, IUT de Roanne, Université de Saint-Etienne, 42334 Saint-Etienne, France

Corresponding author: Victor M. Alvarado (victor@alvarado.fr)

ABSTRACT This paper introduces a new approach of the mean Euler-Poincaré characteristic for non-Gaussian random fields (NGRF), which is based on the decomposition by a basic function named *mother-wave*. The method is proved for long-term recorded, noisy physiological signals. A pretreatment allows the signal to become smooth as the original one is fitted through a Random Algebraic Polynomials (RAP)-based scheme. After that, the polynomized signals are merged by thresholding the RAP function at different levels u. In this way, it is formed a real-valued non-Gaussian physiological random field (NGPRF). Thereby, we deal with their geometric properties centered on their excursion sets $A_u(\Phi, \mathcal{T})$ and a topological invariant, such as the Euler Poincaré Characteristic (EPC) $\varphi(A_u(\Phi, \mathcal{T}))$. The highlight of this work is an explicit model, referred to as the decomposed mean Euler-Poincaré characteristic (DMEPC). The proposed method produces a reduced model with a viable interpretation for different heart conditions investigated for data issued from Holter EKG recordings.

INDEX TERMS DMEPC, Euler characteristic, random field.

I. INTRODUCTION

Random Fields (RF) give a statistical description of complex random patterns of change and relationships from physical data sets [1]–[3]. Consider a random variable $\Phi(X, \bullet)$ as a function of $X = [X_1, \ldots, X_n]^T$ with $\{X_k\}_{k=1}^n$ as the set of random continuous coordinates, then, a collection of $\Phi(X, \bullet)$ is named random field [4]. The geometry and regularity properties of RF have been largely discussed in the literature. These features have to do with continuity and differentiability notions [5] and still with the geometry generated by RF through their excursion sets over a level u [4]. Gaussian Random Fields (GRF) is a class of RF for which the finite-dimensional distributions are multivariable normal distributions that can be fully specified by expectations and covariances. GRFs lead to a large class of applications suited for use in mathematics, sciences, and engineering [6]. A noticeable issue has been accessing information from biomedical images. In this case, Gaussian Markov Random Fields (GMRF) provide a spatial-contextual knowledge that has often resulted in spatial statistics applications. The research by [7] used Markov irregular fields (MRF) to address biomedical image analysis by image segmentation, object labeling, and 3D vision. Later, [8] developed a Bayesian framework that introduces an adjustable parameter on a Generalized Adaptive Gaussian Markov random field (GAGMRF) model to adjust the image quality for X-ray Luminescence tomographies. Researches that engage in RF concern also medical analyses. As an example, we refer to the work in [9], where statistical methods assisted morphometric analyses of specific subregions of the brain. Here, using GRF differentiated the shape of the amygdala and the hippocampus of normal subjects face to patients with attention-deficit/hyperactivity disorder (ADHD). Another type of RF usage points to disease mapping. In this regard, we cite the work of [10] that propose a Bayesian analysis of a GMRF to determine the spatial variability of lip cancer cases in Scotland for a period of five years.

From the given literature review, we deduce that the theory of GRF supports quite well medical diagnosis practice by images. However, for some diseases, the diagnosis is made more by reading signals recorded over time. For example,

The associate editor coordinating the review of this manuscript and approving it for publication was Aniruddha Datta.

Appendix B

Libraries

B.1   WFDB Toolbox for MATLAB

The WFDB Toolbox for MATLAB and Octave is a collection of functions for reading, writing, and processing physiologic signals and time series in the formats used by PhysioBank databases (among others). The Toolbox is compatible with 64-bit MATLAB and GNU Octave on GNU/Linux, Mac OS X, and MS-Windows [75].

Prerequisites:

- MATLAB R2014a or later, with a working Java Virtual Machine (JVM) that supports Java 1.7 or later.

Installation:

- Open MATLAB.

- Go into the directory where you wish to install the Toolbox: *cd directoryname*

- Type (or copy and paste) the following commands into the MATLAB window:

```
[old_path]=which('rdsamp'); if(~isempty(old_path))
rmpath(old_path(1:end-8)); end
wfdb_url='https://physionet.org/physiotools/matlab/wfdb-app-matlab/wfdb-app-toolbox-0-10-0.zip';
[filestr,status]=urlwrite(wfdb_url,'wfdb-app-toolbox-0-10-0.zip');
unzip('wfdb-app-toolbox-0-10-0.zip');
cd mcode
addpath(pwd)
savepath
```

For more references, you can search at this site:

https://archive.physionet.org/physiotools/matlab/wfdb-app-matlab/