Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Quantitative analysis of smartphone PPG data for heart monitoring

Relatore:
Prof. Daniel Remondini

Correlatore:
Prof. Hae-Won Uh
Yosef Safi Harb

Presentata da:
Francesco Bussola

Anno Accademico 2017/2018

Sommario

Il monitoraggio dell'attività cardiaca tramite PPG e app è promettente, ma classificare il ritmo cardiaco in normale o fibrillazione atriale (FA) è difficile in caso di misure rumorose.

In questo lavoro caratterizziamo un dataset di 1572 soggetti che hanno fornito i loro segnali raccolti tramite un'app e la videocamera dello smartphone. Studiamo le distribuzioni di tre proprietà dei segnali: l'area, l'ampiezza del picco e l'intervallo temporale tra i picchi successivi. Valutiamo se qualche fattore influisca sulle distribuzioni, scoprendo che gli effetti principali si hanno per l'età e il BMI. Valutiamo l'accordo fra i risultati sui canali R G B, trovandolo buono per i primi due.

Dopo aver identificato indici di qualità dalla letteratura, ne usiamo alcuni per una classificazione, combinandoli con una grandezza data dal dynamic time warping, una tecnica che ottimizza l'accordo fra due segnali, uno di riferimento e un soggetto. Otteniamo un'accuratezza dell'89% sul test, per una classificazione binaria.

Sulle serie temporali assunte caotiche, valutiamo l'aspetto dei diversi ritmi nei Poincaré plots e quantifichiamo i risultati tramite una misura di dispersione 3D, su un dataset di 20 soggetti, 10 sani e 10 con FA, che risultano significativamente differenti secondo la loro morfologia 3D. Estendiamo l'analisi al dataset maggiore, ottenendo ulteriori risultati significativi.

Abstract

The field of app-based PPG monitoring of cardiac activity is promising, yet classification of heart rhythms in normal sinus rhythm (NSR) or atrial fibrillation (Afib) is difficult in the case of noisy measurements.

In this work, we aim at characterizing a dataset of 1572 subjects, whose signals have been crowdsourced by collecting measurements via a dedicated smartphone app, using the embedded camera. We evaluate the distributions of three features of our signals: the peak area, amplitude and the time interval between two successive pulses. We evaluate if some factors affected the distributions, discovering that the strongest effects are for age and BMI groupings. We evaluate the results agreement between the R G B channels of acquisition, finding good agreement between the first two.

After finding signal quality indexes in literature, we use a subset of them in a classification task, combined with dynamic time warping distance, a technique that matches a signal to a template. We achieve an accuracy of 89% on the test set, for binary quality classification.

On the chaotic temporal series we evaluate the appearance of different types of rhythms on Poincaré plots and we quantify the results by a measure of their 3D spread. We perform this on a set of 20 subjects, 10 NSR and 10 Afib, finding significant differences between their 3D morphologies. We extend our analysis to the larger dataset, obtaining some significant results.

Contents

1 Introduction

1.1 Atrial Fibrillation

Atrial fibrillation (Afib) is a medical condition that occurs when the heart atria contract rapidly, irregularly and incompletely due to irregular electrical signals that interfere with the normal physiological pace making of the heart and also with ventricular contraction, which becomes irregular too [1]. The normal sequence of the systolic phase, where blood is pumped from the atria to the ventricles and subsequently to the body and lungs, is thus altered.

Atrial fibrillation is one of the most frequent pathologies affecting cardiac health, as it is the most common arrhythmia in the world [2, 3], with a conservatively-estimated population of more than 30 million of patients worldwide [4], leading to hundreds of thousands of deaths.

The financial burden placed by atrial fibrillation on health systems is also heavy: worldwide are being spent tens billions of dollars each year for treatment and management of atrial fibrillation and of its consequences (more than 6 billion of dollars in the US alone) [5].

Several factors have been linked to an increased probability of developing atrial fibrillation [6], among which are: older age, being overweight, smoking, binge drinking (including acute episodes), other cardiac conditions (such as atrial flutter).

Current treatment options can control the rhythm and revert it to a normal sinus rhythm (cardioversion) [7].

Diagnosing atrial fibrillation is not always straightforward, since it may arise in several different clinical manifestation, or even be asymptomatic, therefore accounting for the conservative epidemiological estimates previously reported.

A summary of these different manifestations could be as follows [2]:

- paroxysmal atrial fibrillation, with occasional episodes, lasting from some seconds up to two days and then disappearing without treatment. Paroxysmal atrial fibrillation clearly represents one of the least detectable manifestations of atrial fibrillation, as continuous monitoring is required and might still not prove successful;

- persistent atrial fibrillation consists of single episodes that last longer than seven days or less;

- long-standing persistent atrial fibrillations occurs for episodes of a duration longer than a year;

- permanent atrial fibrillation is always present.

Symptomatic patients will experience symptoms such as palpitations, shortness of breath, chest discomfort, being tired, but atrial fibrillation episodes can be asymptomatic [3].

The diagnosis is usually performed by evaluating an electrocardiogram performed at a medical facility, or by using a continuous-monitoring device such as the Holter, worn by the patient over the course of a day [8]. Hence, especially in high-risk populations, it might prove useful to access more cost-effective and potentially more widespread diagnostic methods, in order to achieve greater coverage and more frequent measurements.

1.2 About the project

Heart for Heart is a crowd-sourced initiative promoted by the Arrhythmia Alliance, the Atrial Fibrillation Association, Happitech and other partners, which aims at gathering a million of cardiac measurements [9]. It also aims at increasing awareness of atrial fibrillation and accelerating the pace of progress on atrial fibrillation diagnostic technology. Happitech has conducted clinical trials of its technology in the past in collaboration with hospitals in Amsterdam (The Netherlands) and is now (as of time of writing, in 2018 and 2019) conducting more trials involving also UMC Utrecht (Utrecht, The Netherlands).

1.3 Goals and structure of the thesis

We present data collected by using a technique called *photoplethysmography* (PPG). This thesis aims at:

1. characterizing the data available;
2. identifying possible techniques suitable to classify two different rhythms in data, normal sinus rhythm (NSR) and atrial fibrillation (Afib);
3. analysing the issue of signal quality.

In order to achieve these objectives, we will proceed in the following fashion: after introducing useful terminology and relevant context for the analysis in Chapter 1, in Chapter 2 we will present the findings currently available in specialized literature regarding signal quality indexes for PPG signals.

We will focus on *dynamic time warping*, a technique for aiding in assessing signal quality in Section 3.2 and subsequently, in Section 3.1 we will explore whether some factors play a role in the values of the different features of the PPG signal.

Building on the exploratory evidence from our previous analysis, in Section 3.3 we will study our data from a physical perspective: considering the temporal data series as chaotic, we use recurrence techniques, namely Poincaré plots and spread measures, to assess ulterior properties of the two rhythms.

1.4 Photoplethysmography

Photoplethysmography (PPG) is a technique first used in 1938 [10] to assess physiological changes due to the variable blood perfusion in skin and tissues. Its etymology [11] comes from the Greek words for light, increase and record, meaning recording increases (of volume) by exploiting light.

1.5 The signal

The principles

The genesis of the technique is rooted in two principles, one physiological and one physical in nature.

The first reason lies in the fact that arteries and arterioles are more heavily affected by the flow of blood pumped into them than veins are. When blood is pumped into them, in the systolic phase, they enlarge and dilate, determining a characteristic increase in volume which is not present in surrounding tissue or veins [12].

The second reason is that different tissues interact differently with incident light: oxygen-rich blood especially has a higher absorption of light than venous blood or surrounding tissues [12].

By measuring changes in light absorption in bodily tissues it is thus possible to detect changes related to the blood flow and to cardiac activity.

Building the signal

In order to build a meaningful signal upon these principles, several steps must be undertaken.

At first, light penetration in human skin must be evaluated [13]. For visible light, there is a window in the red, which extends to the near infrared region of the electromagnetic spectrum, namely from circa 600 nm of wavelength up to circa 2500 nm.

Further infrared light and ultraviolet are absorbed by water in tissue, whereas blue light is the most strongly absorbed region of the visible spectrum [12]. Also, yellow and green light are strongly absorbed, but both deoxy- and oxy- haemoglobin have

high absorptivity for green light, thus resulting in a generally higher signal-to-noise ratio (SNR) for green light than for other, more intense, wavelengths [12].

Measurement devices can exist either in transmission mode, where light travels from one end of the anatomical part to the opposite one, such as from the front to the rear of the ear lobe or of the fingertip, or in reflective mode, where both the light source and the detector are on the same side of the part under analysis. We will now assume, for clarity, that we are operating with reflected light, but the following description holds true also for transmitted light.

So, if we shine either visible light or near infrared LASER onto the skin and detect the reflected light component, we obtain a signal that has two components [12]:

1. a direct component (DC), which is due to absorption by tissues and venous blood. This is mostly stationary, with a low-frequency oscillation due to the respiratory rate (that, in fact, it can be estimated by PPG signals [14, 15, 16, 17]);

2. an alternate component (AC), superimposed to the direct component, due to the different amount of light absorption in arterial blood, depending on cardiac activity.

The latter component is pulsatile: a decrease in intensity of reflected light is due to an increased absorption, due to the (delayed) systolic phase that has increased the amount of blood in the vessels, and conversely, an increase in the intensity of reflected light is associated to the diastolic phase. The delay is related to the fact that we are observing cardiac activity when it has reached the region of the body under analysis, therefore later than the actual cardiac cycle phases.

Light sources are usually LEDs, but also ambient light can be used. Dedicated sensors and commercial consumer cameras, both professional or embedded in smartphones are commonly used as detectors [18].

To effectively perform the measurement, suitable regions of the body must be found. Most commonly used are ear lobes, fingertips, wrists, forehead [12] which all have high perfusion and not much surrounding tissue.

In a typical acquisition session, the device is placed in contact with the chosen body region, such as the fingertip, a light is shone into the skin and recording starts. In our

case, the users would place their fingertips on the camera of their smartphones; the app would start recording a video and lighting the tissue and blood with the embedded flash light, continuously on. For RGB cameras, for channel of the camera sensor, from the video frame several regions of interests (ROIs) are extracted. Within each region, the average intensity of the signal is computed, and the process is repeated over every sample, typically corresponding to a frame, thus producing the signal intensity value at each time sample. The typical visualization of the PPG signal mirrors the actual intensity: signal is shown to be higher when intensity is lower, that is when the amount of blood in arteries and arterioles is increasing.

1.6 Applications

Photoplethysmography finds application both in clinical and non-clinical contexts, and its potential of assessing physiological processes is of growing interests in diverse fields, including security and emotional response detection.

Clinical PPG sensors

Clinical PPG sensors and devices are available and in clinical use, for continuous monitoring. Often, they are operating in transmission mode, applied to the fingertip of the patient, who is generally lying in bed. One of the usages of these sensor is that of triggering arrhythmia alarms [19]. Since this alarm usually results in a scramble of the assisting staff to the patient in order to face the arrhythmia, there is a growing need of reduction of the false alarm rate. Most pulse oximeters [20] to detect oxygen saturation in blood are indeed PPG devices, and also blood pressure can be recovered from the PPG signal [21].

Non-clinical measurements

Most usage of the PPG technology outside of the medical facilities is now happening on wearable or mobile devices, which are notoriously becoming ubiquitous. All that is needed is a camera and a flash light (truly, an LED), but several wearables incorporate dedicated sensors and infrared laser light sources. In some cases, webcams or surveillance cameras are used.

The increased availability of PPG-based monitoring calls for both improved reliability of results and assessments, such as for sporty people, and for exploiting new opportunities in reliably evaluating the health status of high-risk populations and enhanced screening.

Noise and challenges

Noise and artefacts play a pivotal role in the feasibility of using PPG technology reliably. Two main sources of disturbance can dramatically influence the quality of the signal:

1. ambient light, or changes in intensity of the light source, which result in changes of the baseline level. This effect can also be due to automatic exposure settings on smartphone video recording applications, which change the ISO sensitivity of the sensor;

2. motion artefacts, due to a number of reasons: breathing, talking, vibrations, variations in the pressure between the body and the device. These effects are much harder to identify.

Finally, when applicable, the power source frequency might influence the signal trends [21].

1.7 Terminology

As the growth of the PPG-based field is relatively recent compared to the electrocardiogram (ECG) analysis, there is often some misconception or inaccuracy in the terminology employed [22], especially since most of the quantities of value in PPG-based analysis are derived from those exploited in ECG.

A typical ECG lead presents five important points around each beat, marked and commonly named as P, Q, R, S and T. The time interval occurring between two successive beats is often measured as the difference between two successive R points, and it is therefore referred to as RR. Since not every such interval will be properly detected, the set of intervals will be deprived of those deemed of unworthy quality. The remaining high-quality intervals are now referred to as NN intervals.

In the PPG there is not yet a consensus on nomenclature: we will refer only to each peak as a pulse, to its top-point as P, thus marking the time interval between two successive pulses as PP time interval (see, for instance, [23]). Each heart beat is called pulse when it is detected via its distorted form in peripheral circulation. Since we do not perform any removal of low-quality data, we will only use PP intervals and we will refer to them, sometimes, as pulse or beat time interval or duration, meaning the time interval between two successive pulses. We will refer to the amplitude of each pulse, from trough to peak, as peak or pulse amplitude, and we will refer to the area under it as peak or pulse area (area of the peak).

2 Literature review

In this Chapter we will present our findings on methods and features used in literature to assess signal quality for PPG signals.

2.1 Introduction

As part of our research, we aimed at assessing the quality of the PPG signal, since this is crucial to distinguishing atrial fibrillation from normal sinus rhythm, especially in the case of measurements that are collected in uncontrolled environments and conditions, where noise and artefacts could be significantly present. Therefore, we searched over the literature to identify some metrics that might prove to be useful in assessing signal quality. We refer to these metrics as *signal quality indexes* (SQIs).

2.2 Relevant findings

Hereby we present the most interesting approaches to signal quality assessment that we encountered, followed by a note on how to best label signal quality itself.

The bibliographical search was performed by looking for the key

(Photoplethysm* AND (signal quality)) OR (Photoplethysm* AND (SQI)) OR (PPG AND (signal quality)) OR (PPG AND (signal quality))

on PUBMED and by evaluating also interesting entries found in the references of papers we evaluated. We obtained 58 articles after evaluating titles and abstracts.

Our aim was to identify a set of SQIs that were used commonly, to exploit them too, later in our analysis. We also included some features used to assess the presence of atrial fibrillation, since we are interested in that rhythm too. Signal quality for

PPG is not clearly defined, *i.e.* there is not one single commonly-accepted definition, but generally a good quality signal enables us to identify morphological features and, more importantly, reliably estimating heart rate and its variability, whereas a bad signal would fail in doing so.

After reading the papers, we assessed that 22 of them were meeting our needs for approaches to the signal quality problem, that is, were proposing clearly-defined indexes and pipelines to assess the signal quality, and that tested their techniques on datasets of at least tens of subjects (the minimum sample size in the papers we accepted was 36 subjects). Of these 22 papers, we will now present 15, having discarded seven: three focused on filtering and preprocessing without providing actual signal quality indexes, two were reviews of the field, one focused exclusively on spectral approaches, whereas we were interested mostly on time-domain features, and the last one explored the possibilities of different colour spaces but did not provide actual signal quality indexes.

Signal quality indexes

Several papers focused on *ad hoc* approaches to the problem of assessing signal quality, developing novel and elaborated frameworks and pipelines, which we didn't identify as a right fit to our search.

One of the most straightforward approaches to signal quality is to investigate the causes of signal degradation itself, namely motion artefacts. Therefore, if available, analysing accelerometer magnitude could recover important information. This approach has been developed by Nemati *et al.* in [24], where they included the magnitude as a feature to a neural network to identify atrial fibrillation. They achieved very promising results (accuracy of approximately 97%) on a small-sized sample of 36 patients.

A second straightforward approach would be that of exploiting some *a priori* knowledge of the physiology of the processes involved in order to discard unrealistic data values. Orphanidou *et al.* [25] only accept segments whose heart rates are within the range of 40 bpm to 180 bpm. Moreover, they place an upper limit to the duration of a beat, either by ECG or PPG measurement, that has to be shorter than 3 seconds and they expect the ratio between the duration of the maximum time interval

between two beats and the minimum duration to be not higher than 2.2, within a 10 seconds long windows. They acknowledge the fact that their approach might not be suitable in the case of highly irregular beats, but it might still be sufficient or effective depending on the applications envisioned.

Firoozabadi *et al.* in [23] devise a strategy that aims at reliably assessing parameters from which is defined heart-rate variability (HRV), that is the fact that the time interval between two successive beats is not regular, in a healthy heart (the degree of HRV exhibited by a subject may indicate the presence of some pathologies). They refer to heart rate variability as pulse rate variability (PRV), in the context of PPG. Such parameters are the same used for ECG analysis, such as the percentage of beats whose duration is longer of 50 ms (pNN50), the standard deviation of the set of said beats (SDNN), its mean, the standard deviation of the subsequent differences of the duration of a beat and of the following one (SDSD), the root mean square of the subsequent differences (RMSSD). In order to achieve these measures, which they compare to the corresponding values obtained from simultaneous ECG recording leads, they estimate the inter-beat interval (IBI) as the median value of the beat duration estimated by three different points: from the peak of one pulse to the peak of the subsequent peak (PP), from trough to trough (TT) and from upslope to upslope (UU), assessed as the point where the slope of the rising signal is maximum.

Silva *et al.* [26] approached the issue with multi-channel filtering and found that their *ad hoc* signal quality index was related to the magnitude of the signal noise ratio (SNR) and proved to perform similar to the human-assessed labels of binary quality, either bad or good.

Sološenko *et al.* [27] built templates of the PPG waveform by linear combination of a log-normal distribution and two gaussians, to use in pair with public annotated databases, to extract pulse duration from which they could evaluate heart rate and type of rhythm.

Yu *et al.* [28] use two features that are fed into an SVM classifier: the fraction of aligned waves and the pulse rate variability.

Many papers focused on features of the frequency domain, where the signal is transformed to its counterpart in the frequency domain by means of the Fourier

Transform, which can be expressed [29] for a finite discrete signal $f(t_k)_{k=0,\dots,N-1}$ as

$$\tilde{F}_n = \sum_{k=0}^{N-1} f_k \exp\left(-\frac{2\pi}{N} nik\right)$$

$n = 0, \dots, N-1$, where we obviously renamed $f(t_k)$ with f_k.

Krishnan *et al.* [30] identify kurtosis of the spectrum as an SQI, whereas Yu *et al.* [28] had also evaluated high-frequency power, low-frequency power and their ratio.

Spectral and log-accelerometer powers were instead included by Wander et Morris [31].

Integrated approaches or template matching were used in several instances, especially on the single pulse level, such as [16, 32, 20], either by cross-correlation to the previous beat or by comparison to a golden standard of reference.

Of particular interest is the work by Li and Clifford [19, 33], who also apply a technique called dynamic time warping to enhance the match between signal and a reference template. They combine those techniques with autocorrelation and cross correlation measures, to create fused SQIs.

Elgendi [34] and Liang *et al.* [14] make use of different quantities, such as perfusion index, entropy, SNR, zero crossing, spectral power, but most importantly kurtosis and especially skewness emerge as significant SQIs, as also in [30].

Furthermore, and lastly, Elgendi finds [34] that a signal quality classification in three categories is generally more effective than the binary classification in good and bad quality. Signal quality is thus assessed as good, if both morphological features such as the dicrotic notch and the systolic and diastolic waves are visible, acceptable if the waveforms are visible but there is a lack of dicrotic notch, and bad quality, when morphology is severely compromised.

3 Analyses and results

3.1 Exploratory statistics

In this Section we will report on our research into the general properties of the dataset, especially when divided into subgroups of different classes pertaining to the subjects' properties, such as age, lifestyle etc. We investigate significant differences in the data distribution for the different groups of each class and we assess the underlying distribution models.

Description of the dataset

Data The complete dataset is crowdsourced via a campaign and data collection is performed via a dedicated application. It is comprised of 1572 subjects. For each one of them, we have the PPG signal for the R, G, B channels of the camera sensor, each frame's time-stamp and the values of the accelerometer magnitude along three orthogonal axes $\hat{X}$, $\hat{Y}$, $\hat{Z}$. All of these values are in arbitrary units, apart from the time, which is measured in seconds (we will use milliseconds very frequently, as it is common practice in the field). Frame rate per second is 30 fps and each acquisition lasts about 90 seconds. The user was guided by the app in placing his or her fingertip upon the camera of the smartphone. The LED flash would then illuminate the finger for the entirety of the data acquisition process, which must be performed at rest, as much as possible.

Metadata Each user has also voluntarily provided additional information: their age, sex, weight, height, location (city), any known heart condition and further details on it, their lifestyle, whether they are smokers or not. The app is also capable of logging the model of the device which has been used to perform the measurement. A unique,

anonymous, ID is also associated to each subject. None of the additional information is required, therefore it is not infrequent to encounter missing or null values.

Reliability The reliability of the metadata could not be independently verified, and in addition to that, it must be noted that some classes are not uniquely-defined: *i.e.* there is no indication as of which answer an ex-smoker should have provided, or how the three possible categories of lifestyle (active, moderate, sedentary) are defined. Even the labels regarding any known cardiac condition are not to be taken for granted: it is not unlikely that those who have received a diagnosis are also under active treatment for the condition, which may have reverted any physiological symptom back to the normal state, thus removing any difference with respect to a healthy subject's signal altogether.

Therefore, we will treat any analysis as unsupervised, with exceptions for the more reliable labels such as Sex and Device, which are respectively unlikely to be misreported or automatically collected by the application and, hence, reliable.

Review A review of the red channel of the signal is available for most of the subjects. This review is performed by people trained in labelling PPG signals and it assesses the following:

1. **class** it determines the type of rhythm displayed in the data stream. It may be any of the following categories: normal sinus rhythm, atrial fibrillation, flutter, arrhythmia, undetermined, strange, damaged, premature atrial or ventricular contractions;

2. **quality** any segment of data could be labelled as high, medium or low quality. Common criteria in literature [34], [35], [25] to define signal quality are related to the possibility of identifying morphological features of the pulse and assessing heart-rate variability (HRV);

3. **confidence** a score regarding the degree of confidence of the quality assessment performed. It can be high, medium or low and it is assigned by the reviewer.

Matching As a general rule, each data stream is divided in three segments of 30 s of duration each, starting from a so-called *guard point* located after the initial transient

has damped. For some acquisitions, less than three segments are available. The review file also contains several duplicates, which have been removed.

Matching each review to the proper data file is not a straightforward task, as no reference to the ID is available in the review. Anyway, it is present an index that matches the progressive data index of each subject in the dataset, therefore we performed an association based on said number.

The analysis

Preliminary selection

We decided to focus on a subset of the metadata available, specifically limiting our analysis to the classes of sex, device, age, lifestyle, height and weight (combined in the body-mass index). Each category has been grouped according to these criteria:

1. sex in male and female, presented in Table 3.1, in fairly balanced groups;

2. device in 13 groups - iPhone models 4S, SE, 5, 5C, 5S, 6, 6S, 6 Plus, 6S Plus, 7, 7 Plus and iPod 7.1 and iPad Pro (9.7in). The corresponding number of subjects per group is reported in Table 3.2;

3. age in 8 groups, corresponding to the decades from 10 to 90 years old, as in 18-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89. We present the relevant number of subjects in Table 3.3;

4. BMI as defined by [36] in the groups underweight ($\mathrm{BMI} < 18.5$), normal ($18.5 \leqslant \mathrm{BMI} < 15$), overweight ($25 \leqslant \mathrm{BMI} < 30$), obese ($\mathrm{BMI} \geqslant 30$). The imbalance between the classes can be noticed by reading Table 3.4;

5. lifestyle as provided by the users, it can be active, moderate or sedentary, as shown in Table 3.5. With the exception of the sex class, the other groups present a degree of class imbalance for the number of subjects per class.

Data preprocessing

Raw data presents several components in the signal that need to be accounted for in order to enable a more accurate analysis. In Figure 3.1 we display an example of such

Sex	
Male	664
Female	679
n.a.	229

Table 3.1: Number of subjects for grouped by sex. Missing entries have not been analysed. The classes are balanced.

Age			
iPhone 4S	34	iPhone 6 Plus	63
iPhone SE	168	iPhone 6S Plus	52
iPhone 5	42	iPhone 7	165
iPhone 5C	35	iPhone 7 Plus	113
iPhone 5S	303	iPod 7.1	4
iPhone 6	325	iPad Pro (9.7in)	5
iPhone 6S	263	n.a.	0

Table 3.2: Number of subjects for grouped by device. As this label is automatically collected by the device, it is reliable and there are no missing entries. The group size is unbalanced, from a few entries to a fifth of the number of total subjects.

Age	
18-19	26
20-29	130
30-39	171
40-49	283
50-59	388
60-69	261
70-79	78
80-89	7
n.a.	228

Table 3.3: Number of subjects for grouped by age groups. Missing entries have not been analysed. Class imbalance is strong for the youngest and eldest age groups and for the most abundant, of people in their fifties.

BMI	
Underweight	13
Normal	578
Overweight	525
Obese	193
n.a.	263

Table 3.4: Number of subjects for grouped by BMI. Missing or null entries of height or weight have not been analysed. There is an important class imbalance as far as the obese is concerned, and even more so for the underweight class.

Lifestyle	
Sedentary	251
Moderate	772
Active	314
n.a.	235

Table 3.5: Number of subjects for grouped by lifestyle. Missing entries have not been analysed. The moderate group has more than twice the number of subjects of the other groups.

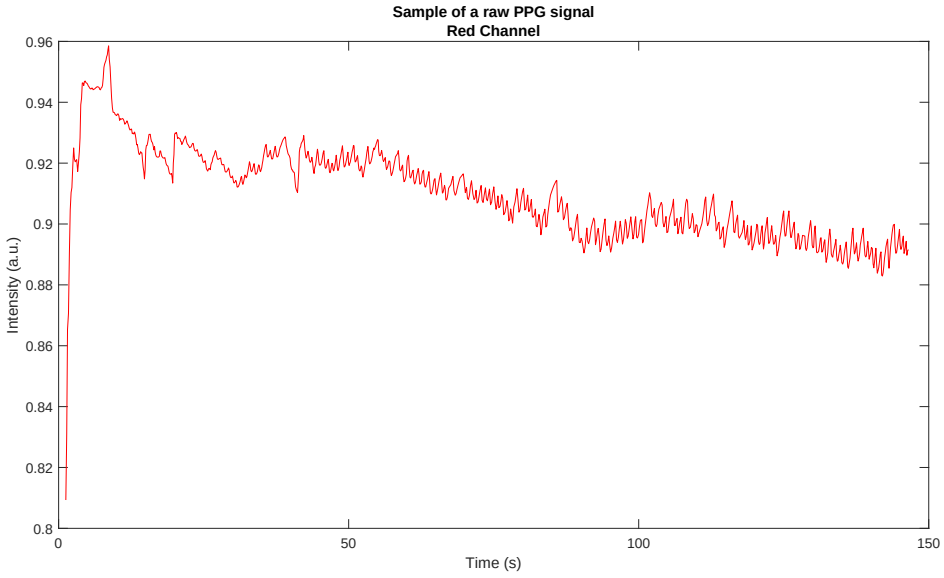


Figure 3.1: Sample of raw signal for the red channel.

a signal, which is given by the sum of three major components: the PPG signal itself, a variable baseline which can be described as a DC component, associated [12] with venous perfusion modulated by the respiratory rate, noise and artefacts of unknown frequency. We also consider that shifts in baseline might be caused by the adaptive exposure level of the camera sensor of the recording devices.

Each signal is filtered with a proprietary algorithm. An example of a filtered and detrended signal is shown in Figure 3.2.

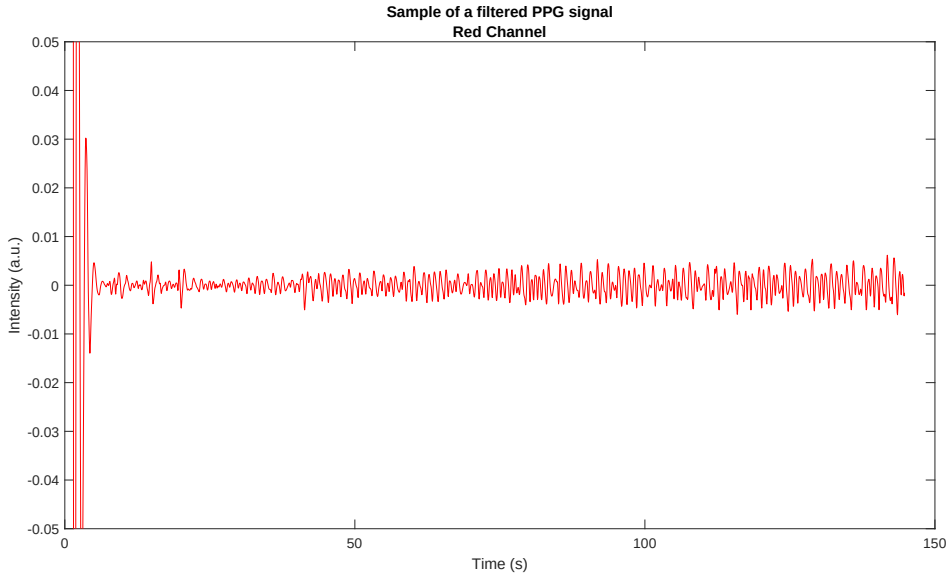


Figure 3.2: Sample of the detrended signal for the red channel. It is a close-up that crops out the complete amplitude of the initial transient, for the sake of visualization. The sample is the same of Figure 3.1.

Methods

From each patient's data, for the three channels separately, we extracted the following three features: PP time interval (the time interval between two successive pulses, measured from peak to peak), amplitude and area of the pulses (which we will also call peaks).

The following processes must be intended as performed on each repetition possible over the classes and their further groupings, for each possible feature analysed: data

of a feature (i.e. peak amplitude) is grouped according to a class (*e.g.* lifestyle) into the groups of that class (*e.g.* active, moderate, sedentary).

Empirical test of subjects' data

At first, we evaluate the distribution of the feature values for a random subset of 100 subjects, to get an intuition of what the underlying distributions might be. Their distribution is not normal or log-normal, and often it is very sharply peaked. We performed a fit with the Cauchy distribution and a fit with the normal distribution and checked whether any clustering was displayed regarding the groups of each class in the parameters plane, $\mu - \sigma$ for the normal and $\tau - \gamma$ for the Cauchy, where τ is its location parameter and γ the scale parameter. The general expression [37] for the Cauchy distribution is

$$P(x) = \frac{1}{\gamma\pi\left(1 + \left(\frac{x-\tau}{\gamma}\right)^2\right)}$$

for a random variable $x \in \mathbb{R}$.

No clustering appeared to be present, in any channel. For instance, the reader may refer to Figures 3.3, 3.4, 3.5, that show some typical distributions of the values of the features for 20 of those 100 subjects of the random subset for the three different features extracted. We only display 20 subjects for the sake of clarity and readability.

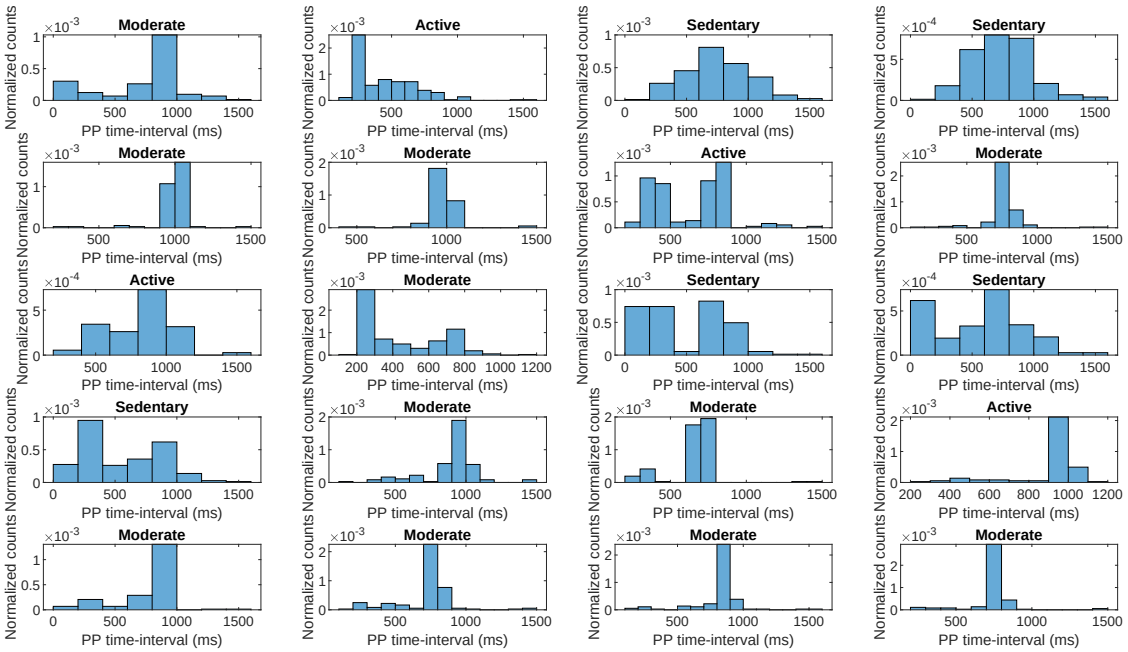


Figure 3.3: Distribution of the values of the PP time interval feature for twenty random subjects out of the 100-subject subset that we used to assess which distributions were likely to match the data distributions. Original data is from the red channel. PP time interval is measured as the time interval between the peaks of two successive pulses.

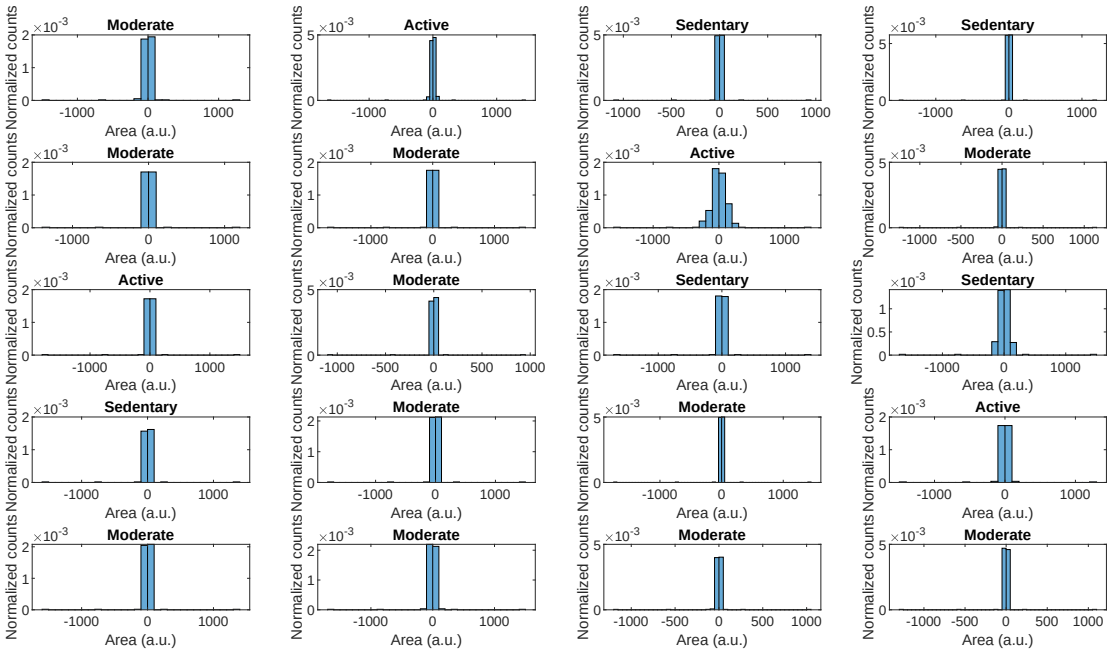


Figure 3.4: Distribution of the values of the pulse area feature for twenty random subjects out of the 100-subject subset that we used to assess which distributions were likely to match the data distributions. Original data is from the red channel. Pulse area is measured as the area under each pulse.

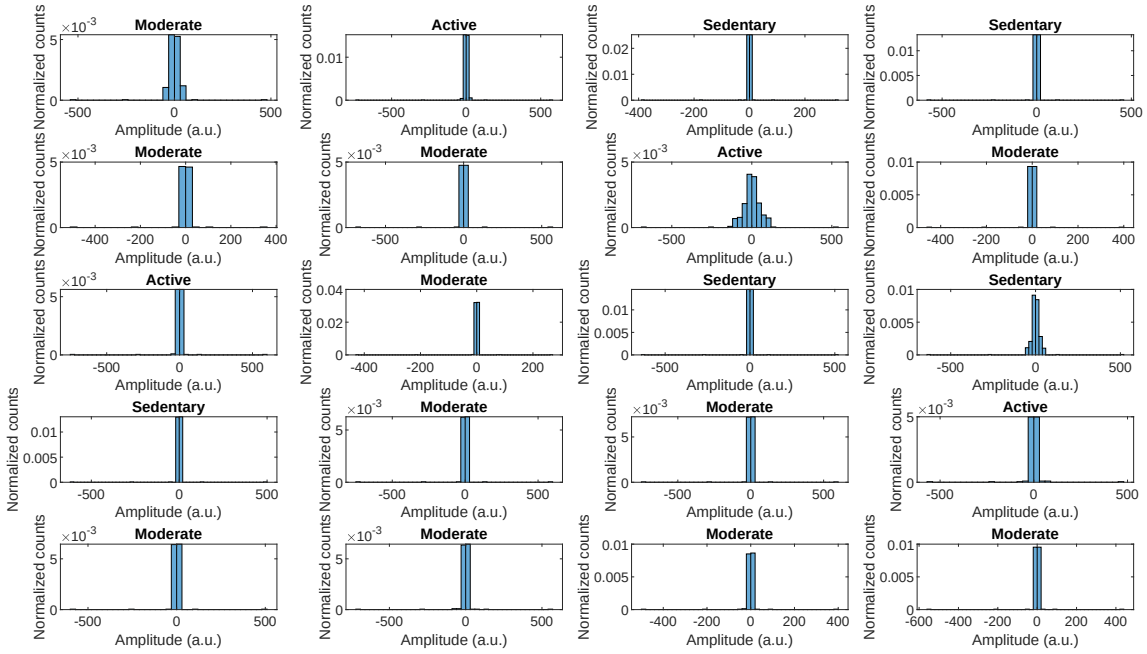


Figure 3.5: Distribution of the values of the peak amplitude feature for twenty random subjects out of the 100-subject subset that we used to assess which distributions were likely to match the data distributions. Original data is from the red channel. Peak amplitude feature is measured as the average value of the amplitude of the signal, from the peak to both the preceding and successive trough.

Distribution of the means

Methods We proceeded to get the distribution of the mean values of the feature for the patients and test it for several possible probability density distributions: normal, log-normal, Weibull, gamma, exponential.

We briefly recap here their salient properties. The normal distribution is very well known, and its probability density function is of the form

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $x \in \mathbb{R}$ is a random variable, μ is the mean and σ^2 the variance. In presence of a strictly positive quantity whose logarithm $\ln(x)$ is normally distributed, we have the log-normal distribution [38, 39]:

$$P(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$$

with $\sigma > 0$ acting as a shape parameter; μ is the mean of the log of the random variable x, and σ^2 is its variance.

The 2-parameter Weibull distribution [40, 41] for a random variable $x \geqslant 0$ can be expressed as

$$P(x) = \frac{B}{A}\left(\frac{x}{A}\right)^{B-1} \exp\left(-\left(\frac{x}{A}\right)^{B}\right)$$

where B is the shape parameter and A the scale parameter. The other distributions that we considered are the Gamma distribution, defined [42, 43], as

$$P(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} \exp\left(-\frac{x}{b}\right)$$

for the random variable $x \geqslant 0$ with a shape parameter and b scale parameter. It is used commonly in when modelling lifetimes. It is a general version of the exponential distribution, which is obtained when $a = 1$, thus resulting in [44, 45]

$$P(x) = \frac{1}{b} e^{-\frac{x}{b}}$$

$\Gamma(x)$ denotes the gamma function, which for a positive real number $x > 0$ is defined as $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$.

We made use of the Anderson-Darling test to a significance level of $\alpha = 5\%$ to check whether the null hypothesis H_0 of the data coming from the distribution under examination was rejected or not.

The Anderson-Darling test [46, 47] can assess the hypothesis that a given set of data comes from a certain distribution. For $x_{i\,i=1,\dots,N}$ monotonically increasing ordered data points, it computes the statistic

$$A^2 = -N - \sum_{i=1}^{N} \frac{2i-1}{N} \left(\ln F(x_i) - \ln(1 - F(x_{N+1-i}))\right)$$

which is compared to the relevant critical value, and where $F(x)$ is the cumulative distribution function of the distribution under evaluation. The test requires the data samples to be monotonically increasing, but it can estimate distribution parameters and is useful when data are not normally distributed.

In order to address the issue of multiple testing, we used the Bonferroni correction [48], which divides the chosen significance level α by the number m of tests being performed.

The issue stems from the following fact: when performing multiple tests, there is a chance of obtaining significant results when there is no actual difference. This chance is in fact not small, as we are about to show.

Let's consider a value of $\alpha = 0.05$, as we – and others – commonly assume, and let's perform $m = 30$ tests. Then the probability [49] of obtaining at least a significant result is

$$P(1^+ \text{ s.r.}) = 1 - P(0 \text{ s.r.}) = 1 - (1 - \alpha)^m$$

which for us evaluates to

$$1 - (1 - 0.05)^{30} \approx 0.785 \approx 79\%$$

not really a neglectable effect!

Results In most situations, log-normal and Weibull were found to be significantly better performing than the other distributions, and comparably so among each other. For the blue channel, the lognormal distribution had an edge, whereas the Weibull outperformed the log-normal on the green channel data. The red channel showed a less crisp difference in performance, with the Weibull better performing on the PP time interval data and the log-normal on the other features (area and amplitude of the peaks). The scores and best-performing choices of these two distributions for the channels and features examined are presented in Table 3.6 and Table 3.7 respectively.

		Red channel					Green channel					Blue channel				
		Sex ·/2	Lifestyle ·/3	Device ·/13	BMI ·/4	Age ·/8	Sex ·/2	Lifestyle ·/3	Device ·/13	BMI ·/4	Age ·/8	Sex ·/2	Lifestyle ·/3	Device ·/13	BMI ·/4	Age ·/8
Peak amplitude	Lognormal	0	1	10	2	5	0	0	6	1	3	0	1	7	2	6
	Weibull	0	0	6	0	4	0	2	9	2	6	0	0	10	1	3
Peak area	Lognormal	0	2	10	3	6	0	0	6	1	3	0	1	8	1	5
	Weibull	0	0	3	0	2	0	2	9	3	6	0	0	9	1	3
PP time interval	Lognormal	0	0	7	1	3	0	0	8	1	2	0	0	7	1	3
	Weibull	0	1	10	2	4	0	2	8	2	5	0	0	3	0	1

Table 3.6: Number of times that the test did not reject the null hypothesis at 5% significance level. Aggregating over the five classes and taking the most frequently accepted distribution as a model generates Table 3.7. We used Bonferroni correction to address the issue of multiple comparisons.

	PP time interval	Peak area	Peak amplitude
R	Weibull	Lognormal	Lognormal
G	Weibull	Weibull	Weibull
B	Lognormal	Lognormal	Lognormal

Table 3.7: Overall best performing probability density function fit for the distributions of the means of the features, based on the number of times of acceptance of the null hypothesis across the different classes at 5% significance level.

The data did not display meaningful distributions on a log-log plot either, thus ruling out, for example, the possibility of a power-law distribution.

Discussion It is important to note that in the case of the blue channel, where the lognormal was found to be the best fit consistently, there was a very low number – five – of instances of in which the null hypothesis was not rejected at the 5% significance level, therefore we will not pursue further analysis of its data, since it is seldom properly modelled by our distributions.

Moreover, it must be noted that the Sex class has consistently rejected to be modelled by any distribution, apart from some sporadic occasions in which the exponential has been successful.

Scatter plots of the distribution parameters

Distribution of the means We plotted the values of the parameters A and B of the Weibull and μ and σ of the lognormal distribution (when more appropriate) and grouped them within each class: points are too scarce to show any clustering, but they never overlap or come close to it, as is shown in Figure 3.6. In Figure 3.7 we present

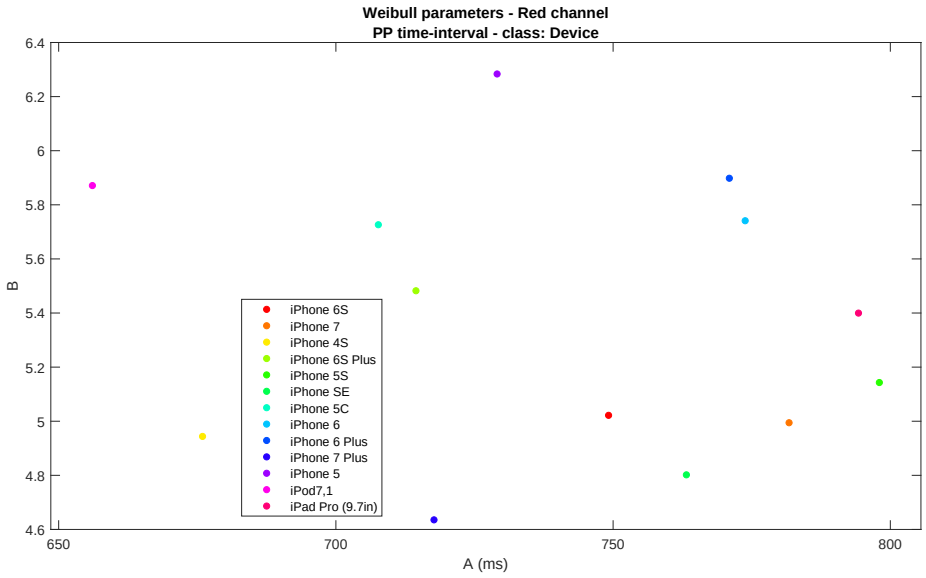


Figure 3.6: Example of the scatter plot of the parameters of the Weibull distribution for means of PP time interval grouped by device type. A refers to the scale parameter and B to the shape parameter.

the results of fitting the Weibull distribution to the means of the PP time interval to the data grouped per the BMI class.

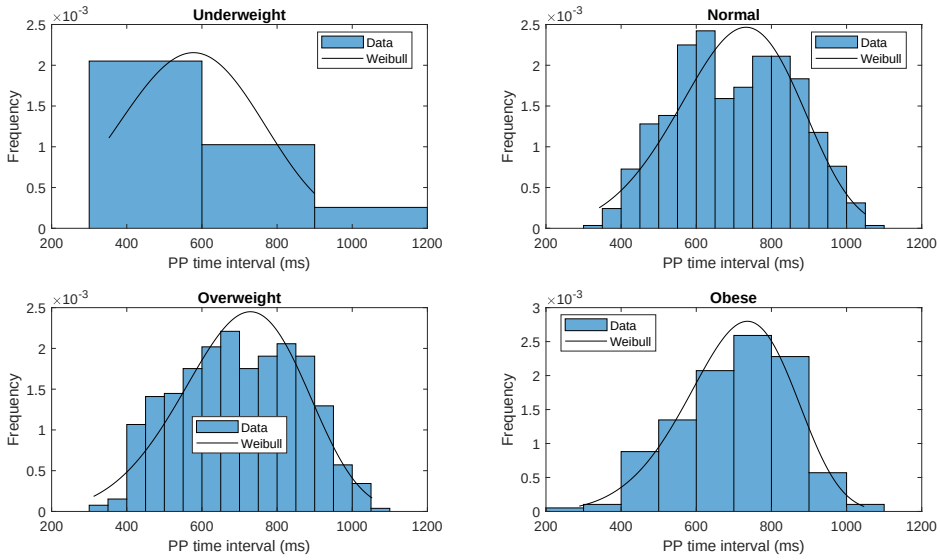
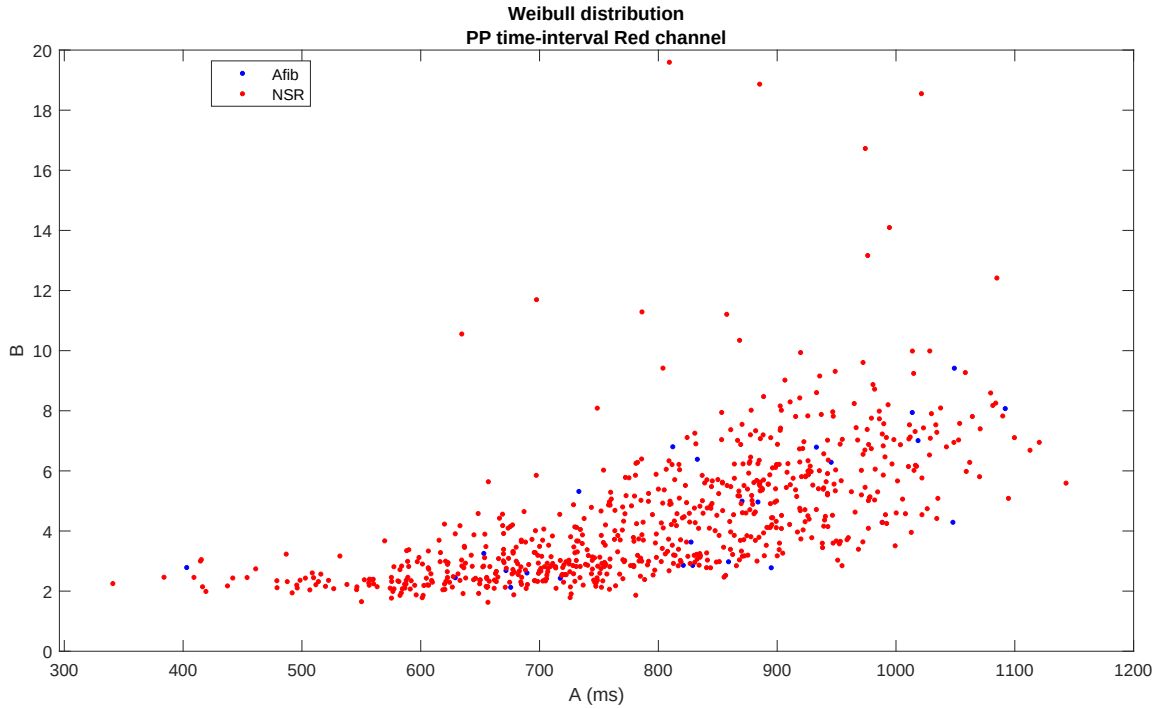
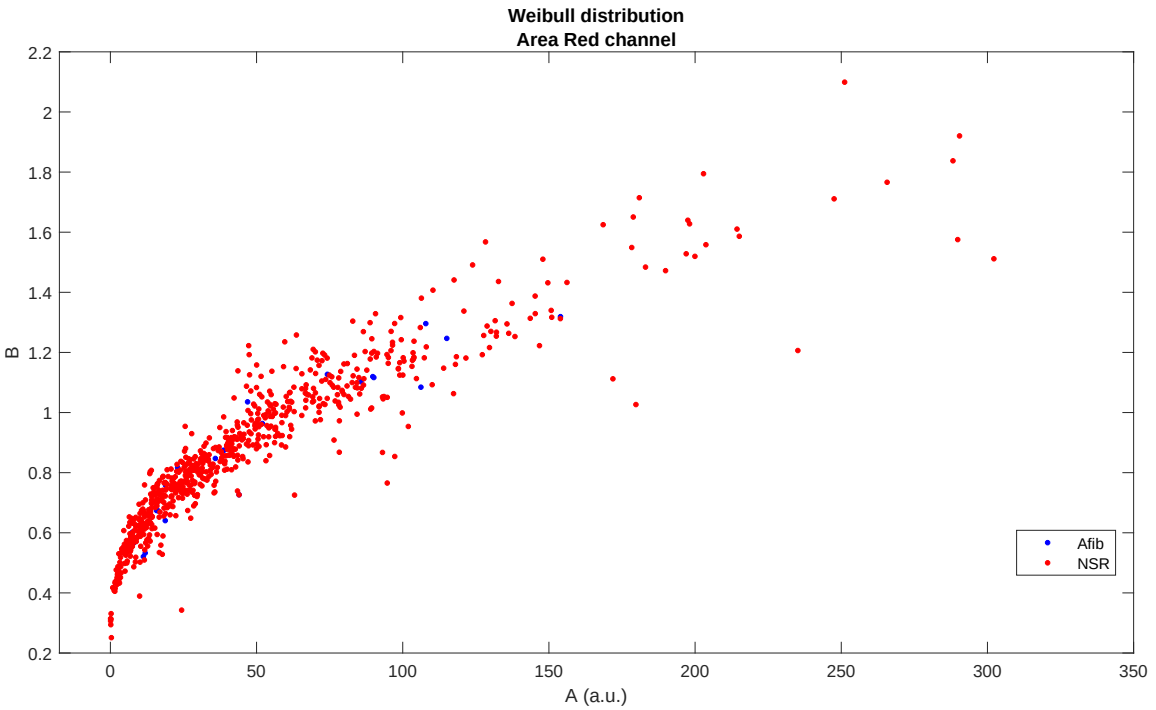


Figure 3.7: Weibull fit of the distribution of the mean PP time interval for the subjects, grouped according to their BMI. Only two out of the four groups did not reject the null hypothesis: the underweight and obese groups. The different appearances of the underweight and obese groups are due to their smaller sample sizes, respectively of 13 and 193 subjects, whereas normal has 578 samples and overweight has 525 subjects.

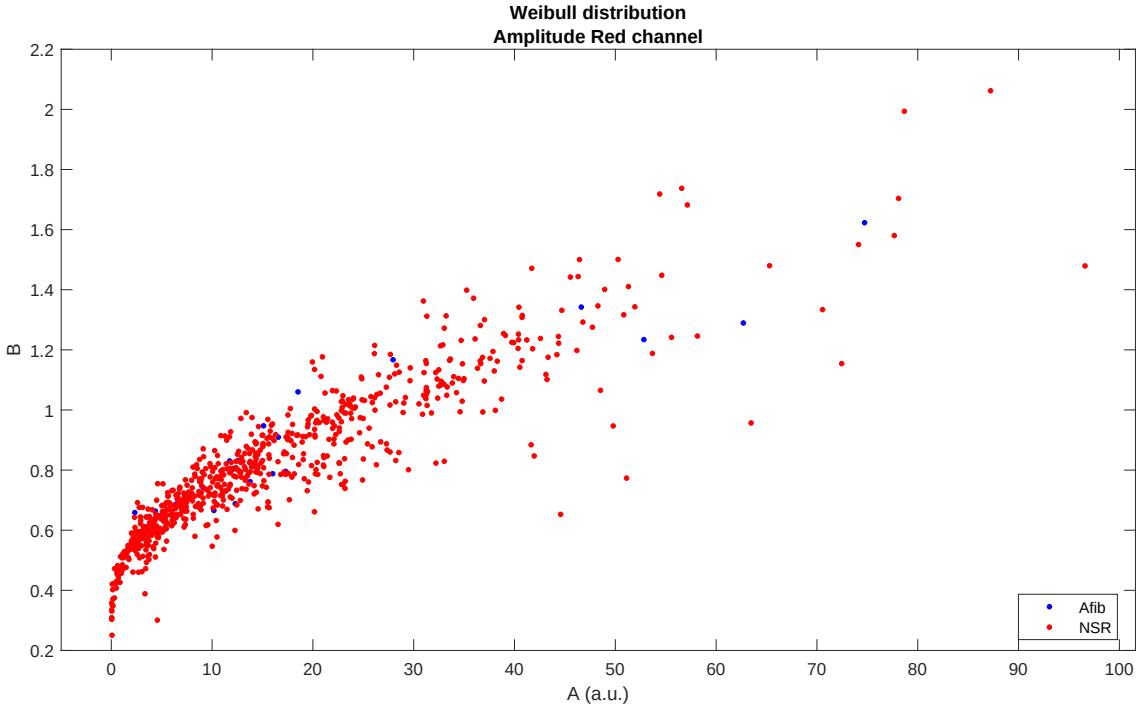
Fitting to each subject's data We wondered if NSR or Afib rhythms would result in clusters on the parameter space. Hence, we fitted each subject's features with the Weibull, gamma, normal, lognormal distributions and grouped the subjects' points according to their rhythm class, as obtained by the reviews of the dataset. To improve our chances of noticing any effect, if present, we restricted our analysis only to high-quality and high-confidence NSR and Afib rhythms, for the entire duration of the acquisition. We obtained respectively subjects 905 and 30 subjects for the NSR and Afib classes. We only limited our analysis to the three usual features extracted from the red channel data, since we do not have any label regarding confidence and quality for neither the blue nor the green channel. As displayed in Figure 3.8, which is a typical result obtained, no clusters were observed.



(a) Feature: PP time interval for the red channel only. A refers to the scale parameter and B to the shape parameter.



(b) Feature: peak area for the red channel only. A refers to the scale parameter and B to the shape parameter.



(c) Feature: peak amplitude for the red channel only. A refers to the scale parameter and B to the shape parameter.

Figure 3.8: Representation of the best-fit parameters for a Weibull distribution on each subjects' data, for the three features.

Significant effects within the classes

In order to establish whether any significant difference was present among the values of the different groups of each class (intra-class significant effects), we visually compared by means of boxplots the distributions of the mean values, as shown in Figure 3.9.

Subsequently we performed a Kruskal-Wallis test and compared the combinations, at a significance level of $\alpha = 5\%$ and applied the Bonferroni correction to deal with the multiple comparisons problem (it is likely to randomly obtain statistically significant differences when performing multiple tests between different classes). The Bonferroni correction [48] is a conservative countermeasure that divides α by the number m of tests that are being performed, thus considering an effective significance level $\alpha_E = \alpha/m$.

The Kruskal-Wallis test is a non-parametric test that compares population medians by ranking the samples according to their magnitude (smallest value is 1, and then increasing up to the number N of data samples), summing the ranks R_i for each group of data and computes the statistic

$$H = \frac{12}{N(N+1)} \sum_{i+1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where k is the number of groups and n_i the number of observations for the $i-th$ group [50, 51]. The statistic is approximately distributed as a χ^2 statistic with $k-1$ degrees of freedom, as long as $n_i > 4$, $\forall\, i = 1, \ldots, k$. We hereby summarize which combination of groups significantly differed from each other, for each channel, class and feature.

Device When considering results grouped by device type, we obtain a significant difference between the distributions of the means of the values for:

1. **iPhone 5S and iPhone 7 Plus**, as found consistently across the R G B channels for the PP time interval feature;

2. **iPhone 5S and iPhone 7**, as found consistently across the R G B channels for the peak amplitude;

Sex When considering results grouped by sex, we obtain a significant difference between males and females, as found consistently across the R G B channels for the PP time interval feature.

BMI When considering results grouped by BMI, we obtain a significant difference between:

1. **underweight and normal**, as found consistently across the R G B channels for the peak area feature;
2. **underweight and obese**, as found consistently for the red channel both from the peak area feature and the PP time interval.

Age If we consider people grouped by their decades of age, we obtain a significant difference between:

1. people **under thirty** and those in their **forties and sixties**, as found consistently across the R G B channels for the PP time feature;
2. people aged **30-39 and 40-49**, as found consistently across the three channels for the PP time interval.

Lifestyle Lifestyle does not yield any significant difference consistently in any feature or channel.

We must take with caution the potential value of these results for practical applications, such as a correction on the values that accounts for intrinsic biases. Generally speaking, for personalised monitoring, a provider would be interested in consistent deviations from a single baseline model. These results do not allow the provider to predict the group of an individual, except in extreme cases, as the overwhelming majority of the subjects would end up in a region of great overlap of values, as was shown by the whiskers of the box plots, for instance in Figure 3.9. The region within the whiskers represent where more than 99% of the values of that distribution are, and by looking at the plots along the groups it is clear that most of these regions overlap for the different groups. We only obtained paired differences, that is a group against another single group, not a group standing out from all the others (apart from the sex class, which is binary). The results however could be the tip of more findings

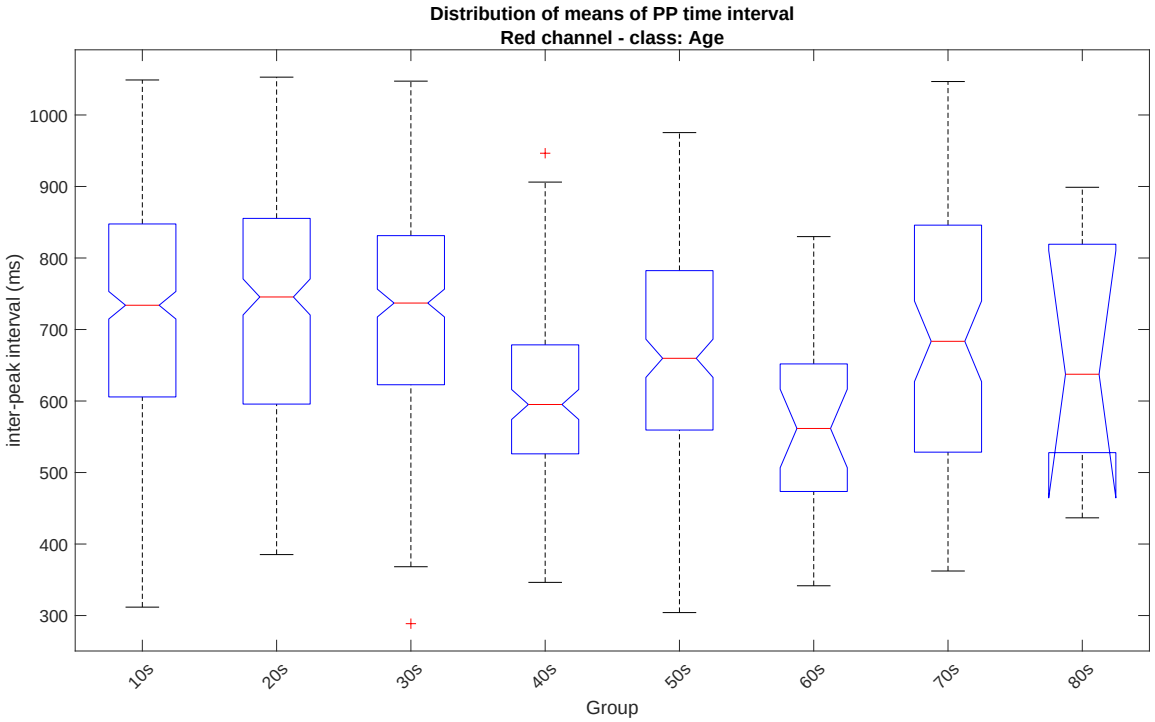
similar to these ones, which could be hidden at the moment by confounding factors. A multivariate analysis could discern the presence of such effects, and it would be best performed on large, reliable datasets.

Discussion

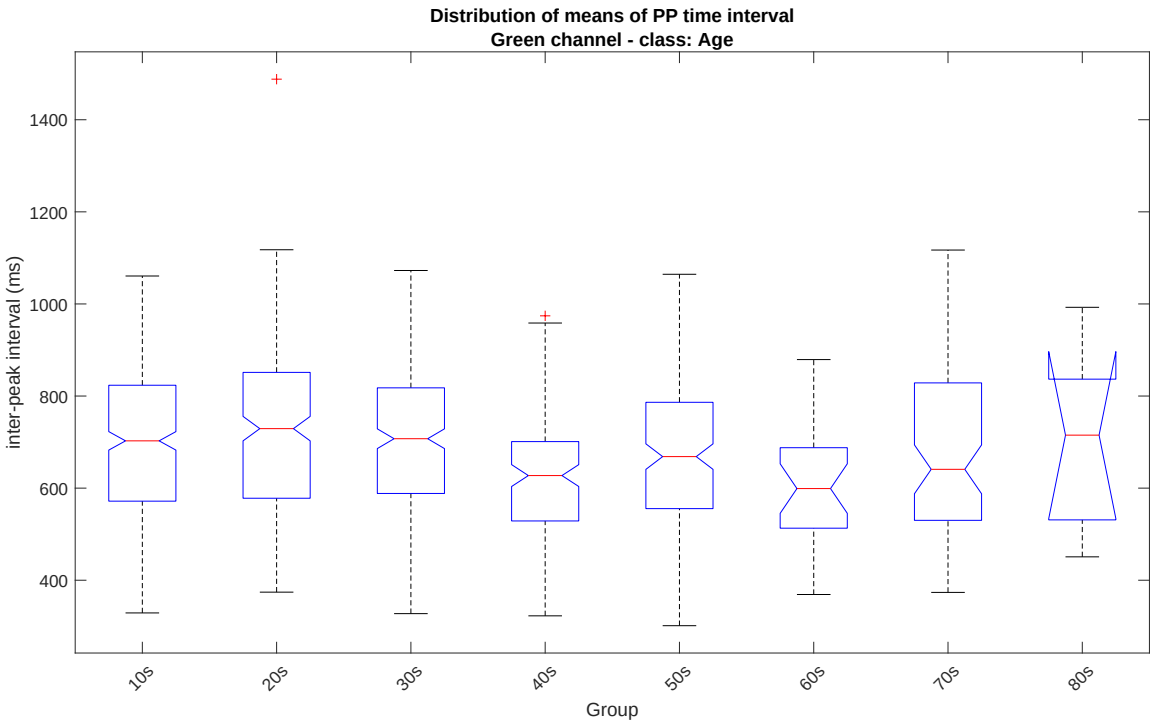
The statistically significant difference between groups of a class might not be of practical importance, even though is a part of the characterization of the dataset. This is likely to be the case for the Device class, where the intra-class differences do not necessarily equate directly to practical applications, and even more, in a field where the technology is changing rapidly.

It is important to stress once more that most of these features are self-reported, ambiguously defined and often subjective: a diagnosed patient who has successfully achieved cardioversion by means of the treatment, would likely display an NSR rhythm but might still self-report Afib as rhythm class, because of the previous diagnosis.

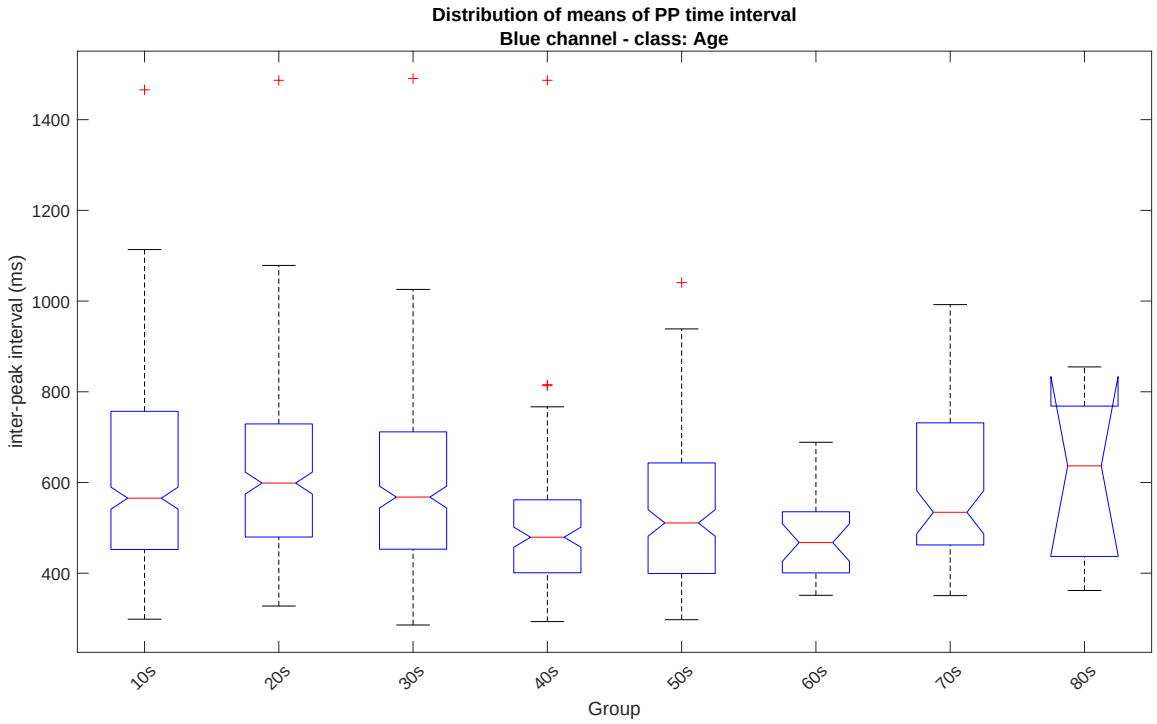
Another limitation of our analysis is that we did not perform a multivariate analysis, which could have discerned more classes.



(a) Red channel.



(b) Green channel.



(c) Blue channel.

Figure 3.9: Typical boxplots of the means values of a given feature (here: PP time interval) grouped within a class.

Comparison of results across the R G B channels

We already showed that the blue channel leads to poor performances, both in finding a good distribution to model the values and in having values systematically different from the other channels, as in Figure 3.9c. The box plots and comparisons within the classes that we presented also highlight the poor performance of the blue channel, especially over the area and amplitude features, whose results often do not agree where there is a match between those of the green and red channels.

Red and green channel often yield similar results when the PP time interval is used, whereas over the other features there is seldom a match.

3.2 Signal quality estimation

By building on our findings presented in Chapter 2, in this Section we present our process to build a set of features that we exploited to classify signal quality of our data.

Introduction

In this Section we present a technique called dynamic time warping, which is a candidate to be a part of a processing pipeline that aims at effectively classifying the quality of the signals, without relying on reviewed input, such as peak labels provided by an expert.

We will briefly introduce the concept of dynamic time warping, then report on our findings on its application to our signals and eventually discuss its potential.

We finally combine it with other signal quality indexes and classify the quality of the signals in our dataset.

Dynamic Time Warping

Given two signals, which we will denote by A and B, each comprised of p points $A_i = (t_i; a_i)$ and $B_j = (t_j; b_j)$ with both $i, j = 1, \ldots, p$, where t_i and t_j are the time points of collection of the two signals of respective amplitudes a_i and b_j, in whatever units are appropriate. The algorithm generally works also for two signals of different sizes, as long as their dimensionality is the same, but we only deal with signals of the same duration and sampling rate, therefore size and we will restrict our presentation to this particular case.

Dynamic time warping (DTW) [52] is a technique that maps A and B into two so-called warped signals A_W and B_W which are obtained via the values of the signals A and B takes at two sets of indexes iA and iB, both of length $q \geqslant p$, obtained from the original indexes $1, \ldots, p$ (repetition of values is accepted) according to the following criterion: *the warped paths minimize their distance*, as it is commonly stated.

We will follow suit in using this expression, but not without having highlighted the fact that not every possible combination is explored by the algorithm, but only those within a certain region [53, 54], close to the identical match (a signal to itself). Not

only this makes the problem tractable, it also reflects the expectation that a match between two similar signals should not somehow present extreme variations from the identity match.

Thus, *distance* is the minimum value of the sum of the Euclidean distances between the template and sample points, after the warping (other metrics could be employed).

Let's consider the $p \times p$ matrix D of the distances between each point of one signal to the other points in the second signal, *i.e.* the $(h, k) - th$ element of D, $d_{h,k}$, is the distance between the $h - th$ point A_h of A and the $k - th$ sample B_k of B, according to the metric chosen (we use the Euclidean distance).

Starting from position $d_{1,1}$ (the distance between the two first points A_1 and B_1 of the signals), a path is built that reaches the end point $d_{p,p}$ (the distance of the final points A_p and B_p). From each position, only horizontal ($(h, k) \mapsto (h + 1, k)$), vertical ($(h, k) \mapsto (h, k + 1)$) or diagonal ($(h, k) \mapsto (h + 1, k + 1)$) increments of a single step are accepted, as if we were moving a king across a chessboard.

This is best explained visually, with an example, as Figure 3.12 illustrates: the path starting in the bottom-left corner and reaching the upper-right corner of the distance matrix D is the one that minimizes the sum of the distances and doesn't violate any of the additional constraints: it must start from $d_{1,1}$ and end in $d_{p,p}$, without any interruption, loops or turning back. How is it obtained, though? Let's consider two signals, which in this example are two NSR good quality segments, shown in Figure 3.10.

The algorithm starts at the initial point of each signal, which is fixed, as a boundary condition. It computes the distance between the two points according to the metric chosen, in our case, Euclidean. If we indexed the signal samples as a sequence of the likes of $1, 2, 3, \ldots$, we have started from the two initial points of the two signals of indexes $h = 1$ and $k = 1$. Now the algorithm progresses to a new couple of points. These can be chosen in one of three possible ways, corresponding to the previously mentioned chess-king moves. It can proceed horizontally to the right, or vertically up, or diagonally. The corresponding couples of indexes would be, respectively: $(h = 2, k = 1)$, $(h = 1, k = 2)$, $(h = 2, k = 2)$ which are the indexes of the previous points either increased by one or not at all (as the king does on the chessboard). The distances between the corresponding signal points are then computed. For example, for the pair $(h = 2, k = 1)$ the distance would be computed between the second point

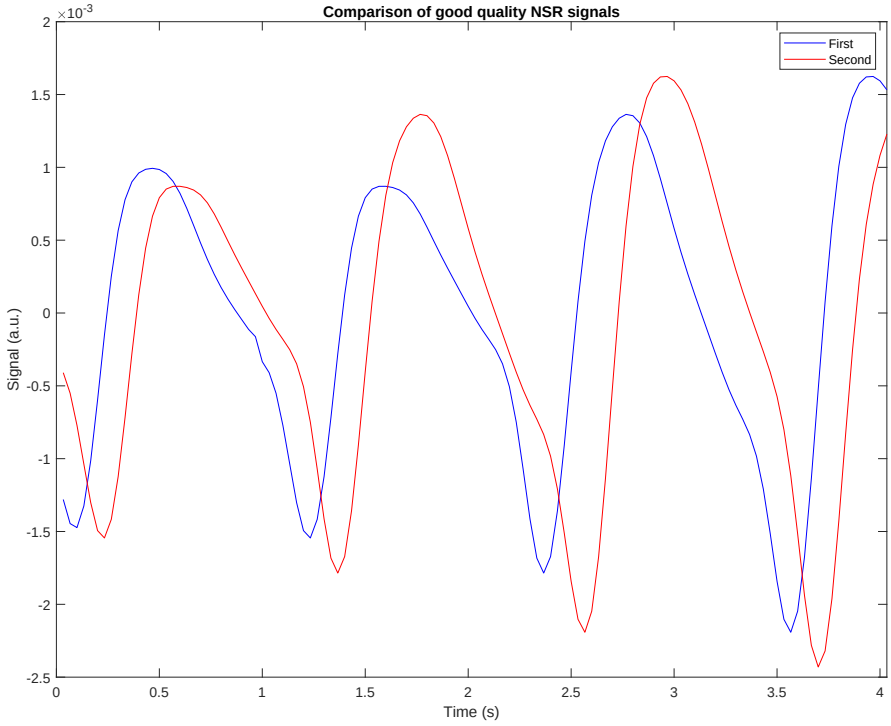


Figure 3.10: Two high quality NSR signals, before being processed by the DTW algorithm. There are 121 sample points in each signal.

($h = 2$) of the first signal and the first point ($k = 1$) of the second signal, which is the same that was used in the previous step. The algorithm progresses to the index combination that has the minimum distance of those computed and builds the new indexes arrays of minimum distance. For example, if of the three possible moves the vertical one was found to be the one that minimizes the distance, the arrays would be: $iA = [1, 1]$ and $iB = [1, 2]$, where the first entries are from the first points and the second entries from the successive step.

When we will use these vectors to warp the signals, we will index the first signal with iA and thus build a signal whose first value would be as follows:

1. the first point is always constrained to be the first point of the original signal;

2. for the *second* sample point of the warped signal, we look into the *second* position of the index array (iA or iB). We take the index in that position, let's say $iA(2) = \mathbf{1}$, and then assign the value of the **first** sample point of the original signal to the warped signal. Thus, if $A(\mathbf{1}) = 4$ (a.u.), we will have $A_W(2) = A(iA(2)) = A(\mathbf{1}) = 4$ (a.u.);

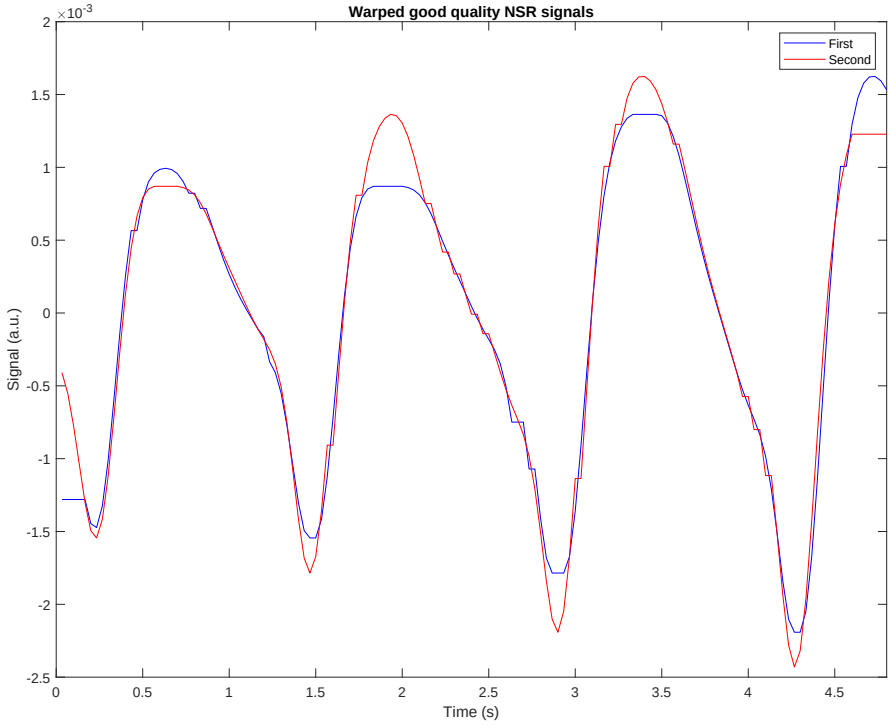


Figure 3.11: The two signals of Figure 3.10 after being warped, that is, re-indexed according to the best path, which is the new sequence of indexes that minimise the sum of the Euclidean distances between the points of the signals. The reader may notice that the time duration of the signal has increased, due to the stretch intrinsic to the warping procedures, as in the length q of the arrays of indexes being longer than the original sample size p of each signal.

3. the procedure is repeated until both signals reach the last point.

The arrays would expand to a length $q \geqslant p$, until both signals are indexed to their end points. Then, the algorithm repeats the procedure until it reaches the end point, which is fixed too, corresponding to the two final points of both signals. The path of minimum distance across the distance matrix for our example is displayed in Figure 3.12. The warped signals are obtained by indexing the original signals with the new sequences of indexes, $\mathfrak{i}A$ and $\mathfrak{i}B$, and we present the result in Figure 3.11.

We were inspired by the approach of [33], who used dynamic time warping in their estimation of signal quality. They fused different features into a signal quality index (SQI), which they submitted as a feature in a support vector machine classifier, which had to assess the belonging of each segment of a signal to one out of three quality classes.

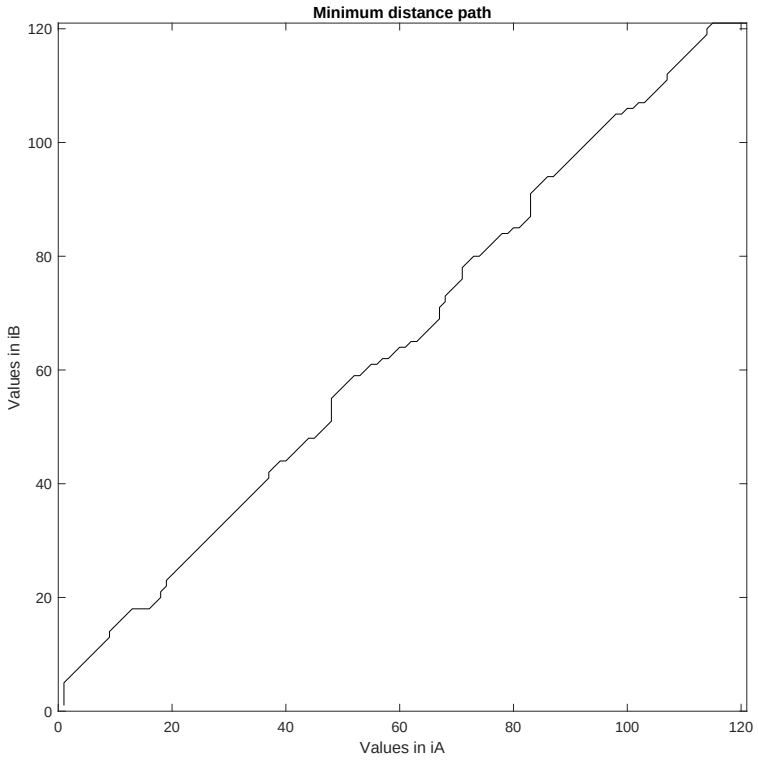


Figure 3.12: Matrix D of the distances with $p = 121$ and optimal path, the one that minimizes the cumulative distance of the two signals. The axis are the values within the new indexes arrays, iA, iB.

They worked on a single-beat basis, obtaining features based on the degree of correlation between a template beat and a signal to evaluate. The template beat for each signal was created iteratively by averaging together the beats of each signal segment, provided that they were similar enough to the running template beat, a decision made on the basis of their correlation.

Three features were based on the direct match of the signal, a linear stretch (or compression) of it to the same duration of the template beat, and a DTW-processed match. The fourth SQI that was fused was the percentage of the non-clipped signal, meaning non-saturated around the peak or the minimum.

However, we wanted to explore the possibilities of DTW not to a single beat but applied to a whole segment of the signal, which contained several beats.

Methods

We proceeded to apply dynamic time warping to a segment which includes some peaks. The first step was to label the filtered data to evaluate their quality. Data had

been previously labelled by reviewers according to the type of rhythm (we were only concerned with NSR and Afib rhythms), and only the red channel is employed, since it's the only one that has been labelled. The labels regarded both the rhythm class and the type of each peak. Each data stream was segmented by a proprietary algorithm, according to a user-provided time duration. The algorithm would then label each segment as either *good* or *bad quality*, based on the frequency of the types of peaks present in each segment.

We proceeded to define a proper template for each possible combination of labels, totalling four classes: Afib good or bad quality, NSR good or bad quality. Template definition could either be by averaging over all the elements of a given class or by hand-picking some segments which could be set as a reference. We evaluate both the signal itself or its successive difference, a measure proportional to its first derivative, that reflects, for instance, how fast the signal is rising or diving, when considered along its duration. In practice, the *rhythm class* label (Afib or NSR) was provided by the human reviewer, whereas the binary *quality class* label was determined by the algorithm, taking into account the *peak class* labels provided by the human reviewers.

Optimization was performed by comparing the distance distributions for the classes at different combinations of parameters. We tried to maximize the separation of the distributions for the different classes: *e.g.* trying to warp an NSR good quality signal to an Afib bad quality template should result in a higher minimum distance (worse match) than the distance of it to a NSR good quality template. We did not have a clear expectation on the outcomes of a possible hierarchy in the case that one of the two categories being the same: should the distance be higher when matching an NSR good quality to an Afib good quality or to an NSR bad quality? Our main goal was to achieve good discrimination between the good and bad quality classes, therefore our optimization process has prioritized said outcome, which would prove useful in providing real-time feedback to the user of the monitoring application.

Results

As far as the optimization process is concerned, we soon discovered that a situation was problematic: generating the templates by averaging would distort heavily the expected features of each class, resulting in a template that would not be easily

classifiable. This is shown in Figures 3.13 and 3.14, for the segments of 4 and 7 seconds of duration. It is clear even to the untrained eye that the templates even appear reversed: the Afib look like NSR and vice versa - they have not been inverted. One often-proposed solution is that of aligning the starting point of each sample. This has been performed and did not lead to any improvement: since the pulse is not periodic, the averaging process still occurs between points that are unaligned even if the initial part of each stream has been aligned, thus cancelling out the signature appearances of the different signals. Moreover, dynamic time warping itself is essentially a technique to align the signals as much as possible, so, even if it were somehow possible, it would have been redundant to align the signals first.

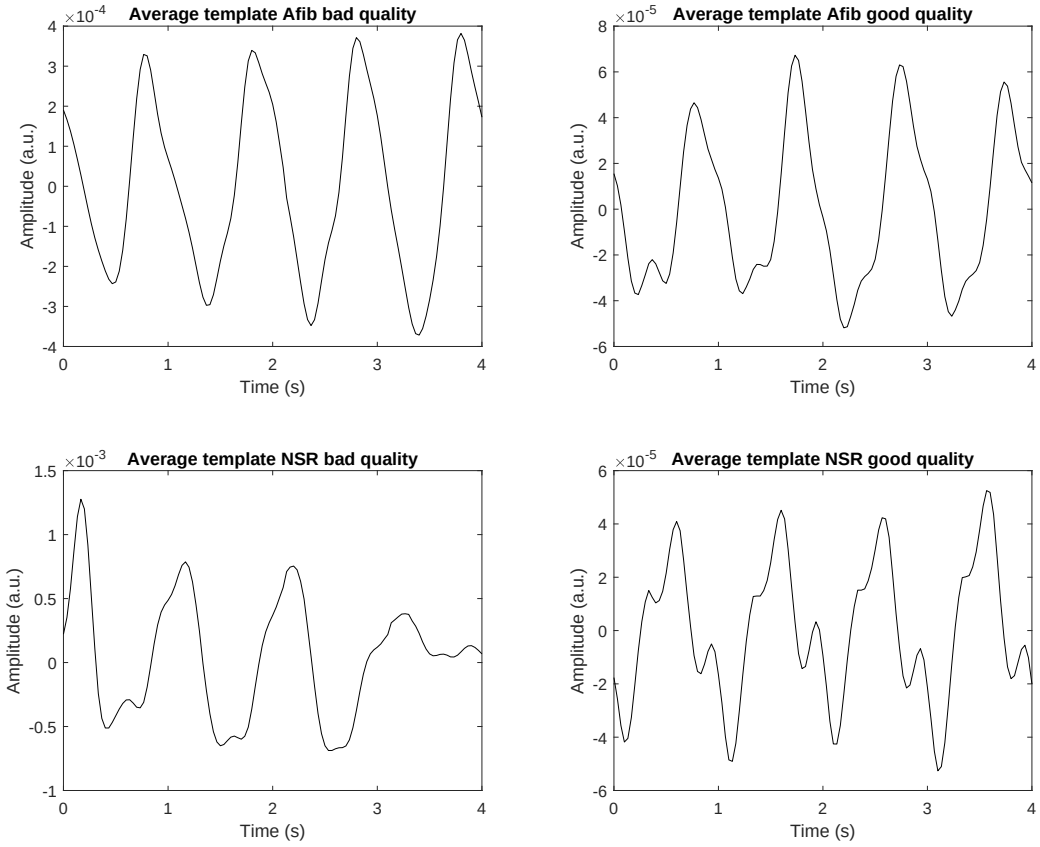


Figure 3.13: Average templates for the 4 seconds long segments. The averaging took place over the whole dataset. At first glance, it would seem that we mistook NSR good quality for Afib bad quality and vice versa! This is not the case. The unpredictable appearance of the templates, far from the expectations, is due to the averaging process. It's clear why we proceed to hand-pick the reference templates, displayed in Figure 3.15.

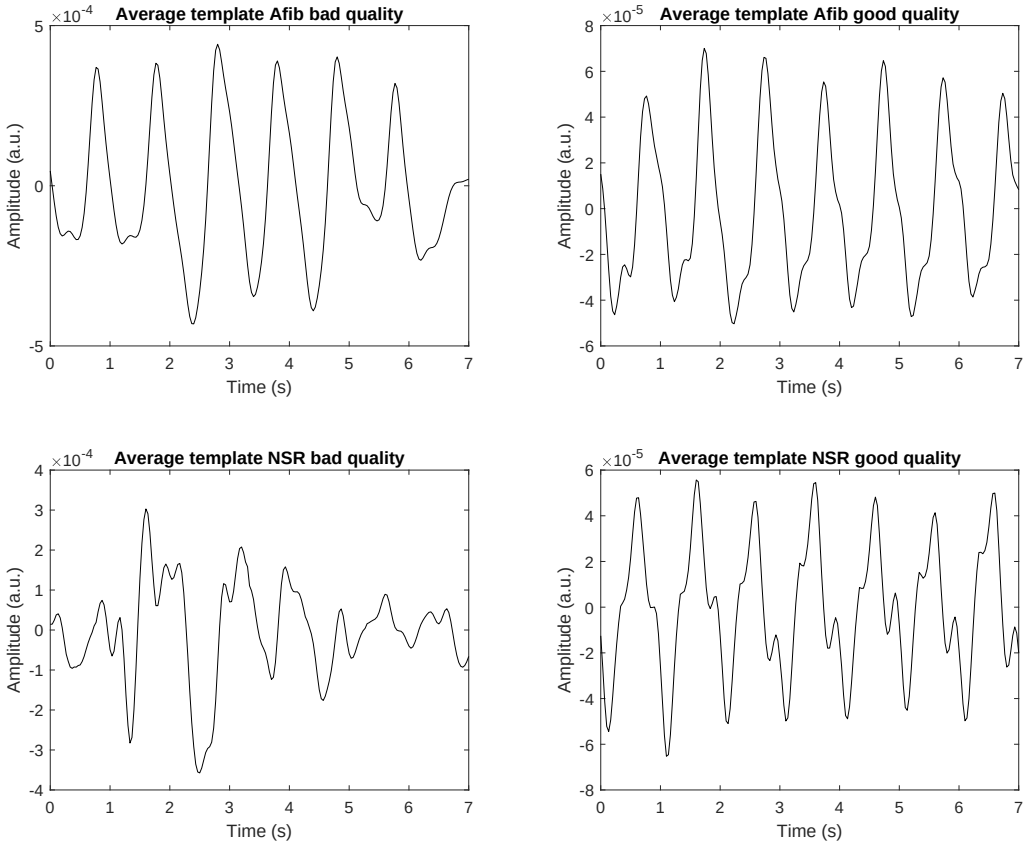


Figure 3.14: Average templates for the 7 seconds long segments. The averaging took place over the whole dataset. In this case too, as per Figure 3.13, the results are not acceptable as templates. Therefore, we selected manually the golden standards that we show in Figure 3.16.

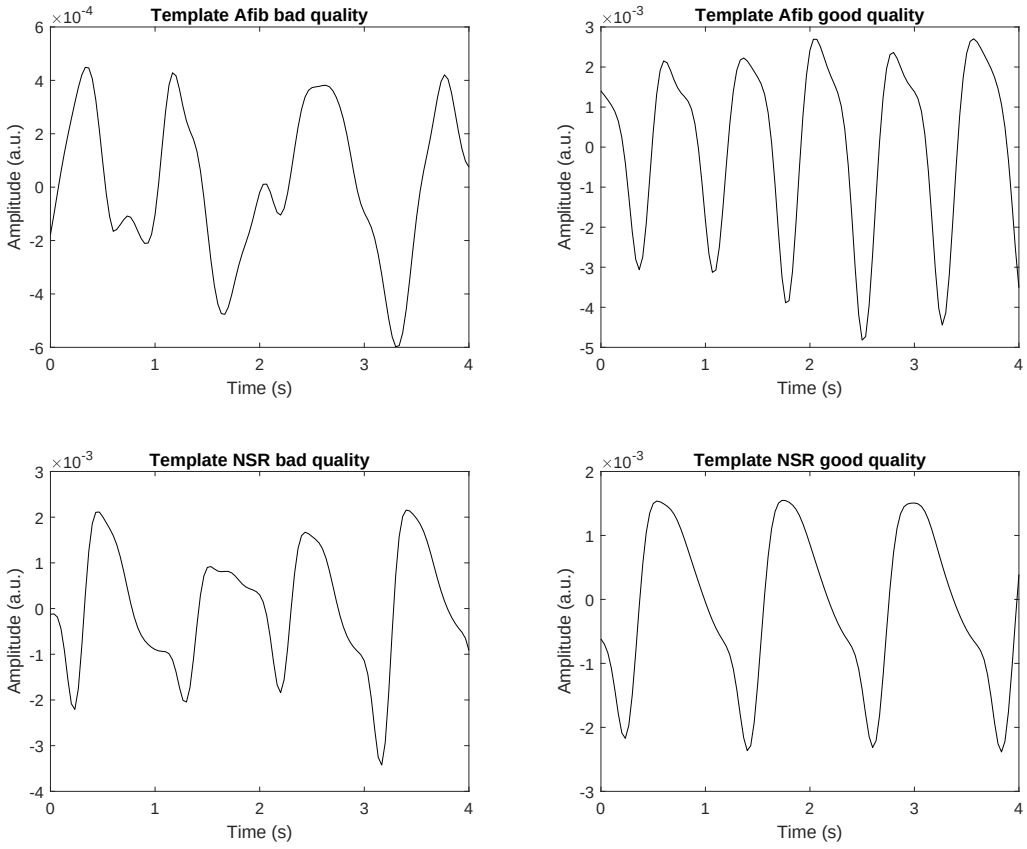


Figure 3.15: Templates for the 4 seconds long segments. We have chosen them manually. Now the appearance agrees with our expectations.

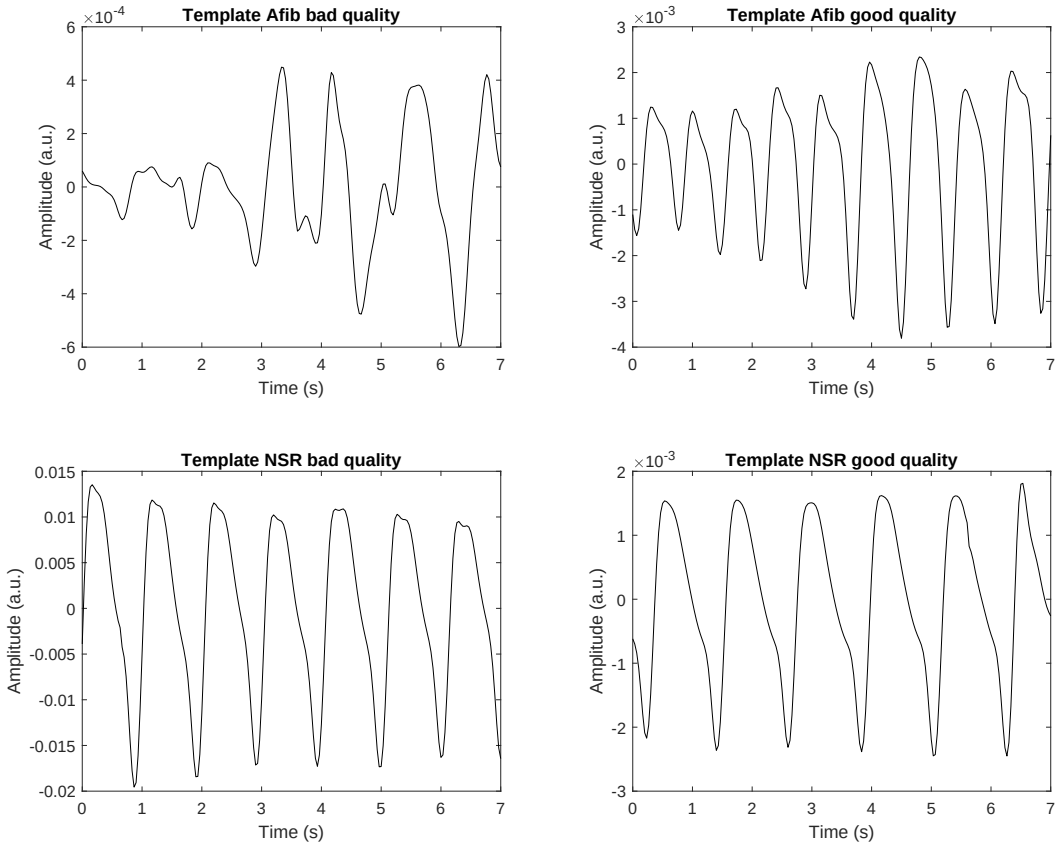


Figure 3.16: Templates for the 7 seconds long segments. These have been hand-picked so that their morphology corresponds to the notion of the different groups of rhythms and qualities.

Hence, we focused on 4 s and 7 s long segments, which we present in Figure 3.15 and 3.16, in which it is clearly visible the typical feature of an NSR pulse (bottom right). We performed a Kruskal-Wallis test between the distances between each template and the data grouped according to their class: *e.g.* we warped the four classes to a template, say NSR good quality, grouped the resulting distances according to the classes and tested them reciprocally. Also here we applied the Bonferroni correction for multiple tests. We obtained that matching the data to the Afib good quality templates results in all classes being significantly different ($\alpha = 5\%$ before correction) from each other and therefore we concluded that this is the template that discriminates the most between the different classes, as shown in Figure 3.17. A similar result is obtained using the 7 seconds long templates.

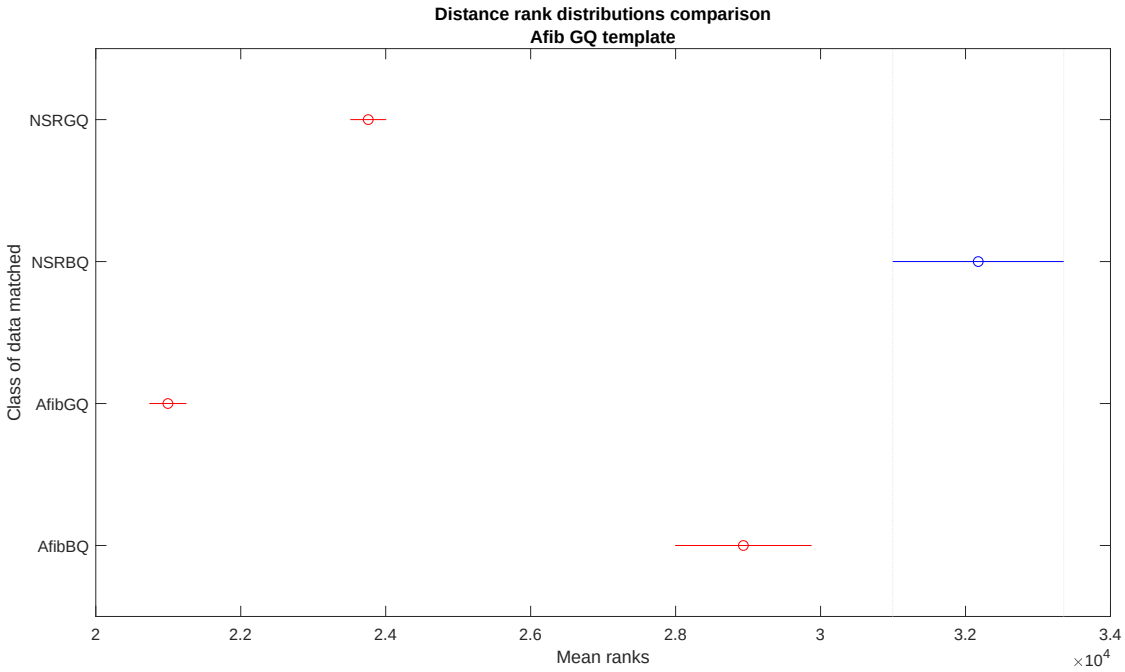


Figure 3.17: Box plot of the mean rank values for the classes when matched to the hand-picked 4-seconds-long Afib good quality template. In this case, the template used in the DTW is the one for which the mean ranks of the distances grouped by the four groups are significantly different. This means that the distribution of the distances for the four groups are different from each other. However, as displayed in Figure 3.18, the actual values of the distances still overlap.

However, lest we forget, that the typical distributions of the distances between the warped paths resemble that shown in Figure 3.18: the amount of overlap is significant.

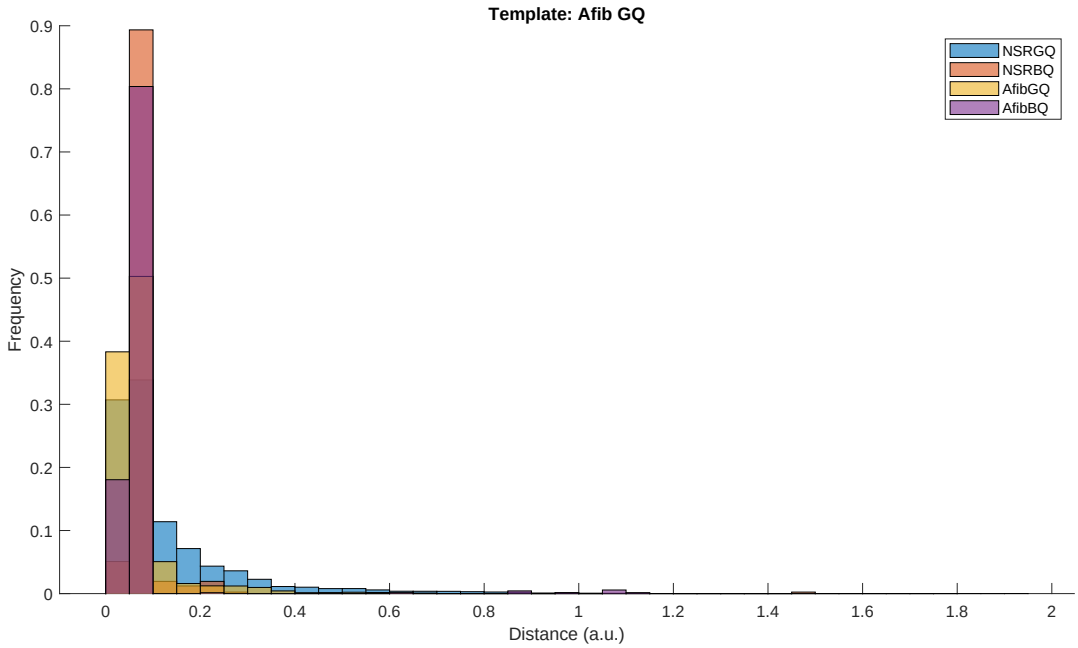


Figure 3.18: Typical appearance of the distribution of distances between the warped signals of each class and a template, in this case the Afib good quality.

Attempt at classification

Building upon our previously identified SQIs [34], we proceeded to attempt to classify the different classes, at least for the macro-classes of good and bad quality.

We employed a shallow neural network of five layers and trained it on the 4-seconds-long segments, aiming at classifying good and bad quality segments, regardless of rhythm, since our motivation is that of detecting bad quality data to warn the user.

As features we started with six SQIs: median of the signal, mean, standard deviation, skewness, kurtosis, DTW-distance from best-performing template (Afib good quality). Guided by intuition, we also introduced the number of local minima of a segment, since it is expected a higher number in bad quality and Afib, and after some tests we decided to drop the mean and standard deviation of the signal. Therefore, the network was trained upon 5 features.

We also chose to under-sample our data (per each class) to the number of elements of the least abundant class, so to avoid the issue of class imbalance, where the network achieves high accuracy by betting all on the most frequent label (*i.e.*, it almost always

predicts a sample to be of the most abundant class). We held out 10% of the samples of each class for testing, totalling 220 samples. Both under-sampling and hold-outs were performed randomly. On the remaining 90% of data (1652 samples) we performed a 5-fold cross validation which divides the data in 5 five groups: for five times, it takes out one group, trains on the other four and validates the performance on the group held out. In order to avoid data leakage, that is the transmission of information outside of each data group, which would render null the validation, we standardized the feature only within the 5-fold cross validation iterations and not before (the test set was standardized too).

We achieved a classification accuracy of almost 89% on the test set of the features extracted from the 4-seconds-long sets, as shown in Figure 3.19. This test was not seen before by the model.

If we take a look at the misclassified samples, we discover that:

- when the model wrongly predicts bad quality instead of good quality, 9 of the 16 samples are Afib and 7 are NSR;
- when the model wrongly predicts good quality instead of bad quality, 1 instance out of 9 is an atrial fibrillation episode, the remaining 8 are of normal sinus rhythm.

It is interesting the fact that, according to our judgement, for six of the seven NSR good quality signals misclassified, it is unclear why they have been labelled as good quality in the first place. This points us towards an iteration of our quality assessment algorithm, which can be further refined and finely tuned.

These results are to be considered preliminary, due to the small sample size and the limited tuning of the network, and reflect the fact that in the presence of fibrillation and real data, simple SQI metrics alone are often not sufficient to achieve the desired outcome; most of the studies we encountered focus either on fibrillation in a very controlled setting, to reduce noise and bad quality, or focus on controlled sources of disturbance to the signal and warn that the results and solutions devised could prove tricky in presence of fibrillation.

Probably, out of all the papers that we read, the one that is more directly comparable to our result is, once again, that by Li and Clifford [33]. Besides the type

of classifiers and the features employed, the main differences between our two approaches are in that their approach focuses on single beats, whereas we use time windows and that they use a three-class signal quality grouping, whereas ours is binary. Their accuracy of 95.2% is remarkable, and ours under-performs with that regard. However, we can reasonably hope that by fine tuning our model and possibly adding some more features we too can achieve more satisfying results.

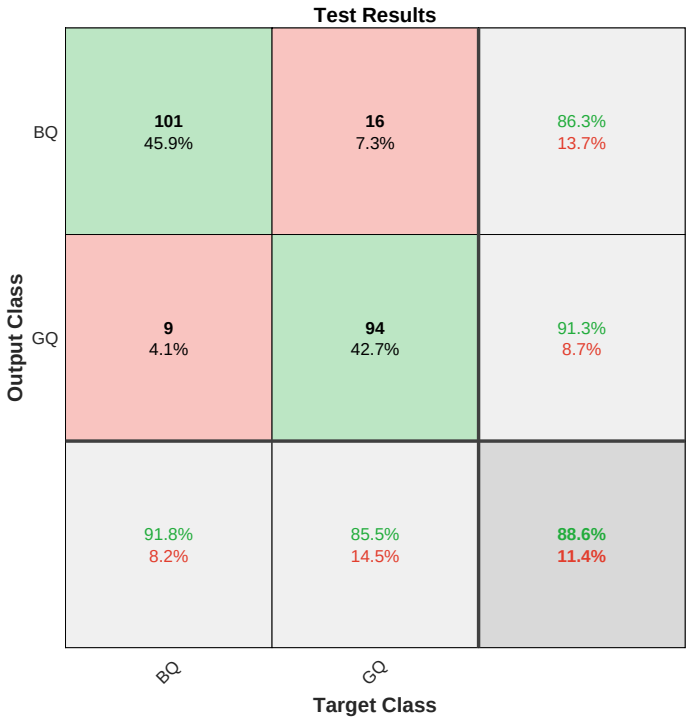


Figure 3.19: Confusion matrix of the performance of the classifier over the test set. The result displayed has been achieved over the test set, undisclosed to our model before of the final evaluation. The model was trained with a 5-fold cross validation.

3.3 Poincaré Plots

In this Section we explore a physical approach based on recurrence plots to evaluate the appearance and properties of the two different rhythms – NSR and Afib – on these plots.

Introduction

The study of heart beat and dynamics in terms of chaos theory has long been discussed and described [55], with more and more instances of applications of clinical relevance.

In particular, it is well known that the cardiac cycle is not a periodical phenomenon, as there is an amount of intrinsic variability that is actually an index of the health status of the heart itself.

This explains why many studies and techniques have focused on assessing heart-rate variability and not only on the beats-per-minute heartrate.

Poincaré plots are a kind of recurrence plot that have been used to assess heart-rate variability (HRV) and that are regarded as informative in the cardiology community [56, 57].

Given a set of n measurements of a certain variable $x(t)$ at different time points $t_1, \dots, t_n$, the Poincaré plot is constructed by plotting each value of $\{x_i\}_{i=1,\dots,n}$ against its subsequent one, where we imply $x_i = x(t_i)$. Hence, we are scattering the points of coordinates $\{(x_i, x_{i+1})\}_{i=1,\dots,n-1}$ on the bidimensional plane, or $\{(x_i, x_{i+1}, x_{i+2})\}_{i=1,\dots,n-2}$ in the three-dimensional space.

Methods

We performed our analysis on two independent sets of data. The first is a small sample of 10 Afib patients and 10 healthy control, validated clinically. We do not have any metadata associated to the data, apart from the rhythm labels. The second is a subset of our main dataset, from which we have extracted the high-quality, high-confidence, pure NSR and pure Afib rhythms, resulting in 905 and 30 subjects per class, respectively, as in Chapter 3.1. Our labels are only valid for the red channel for this database, therefore we did not analyze green and blue channel data.

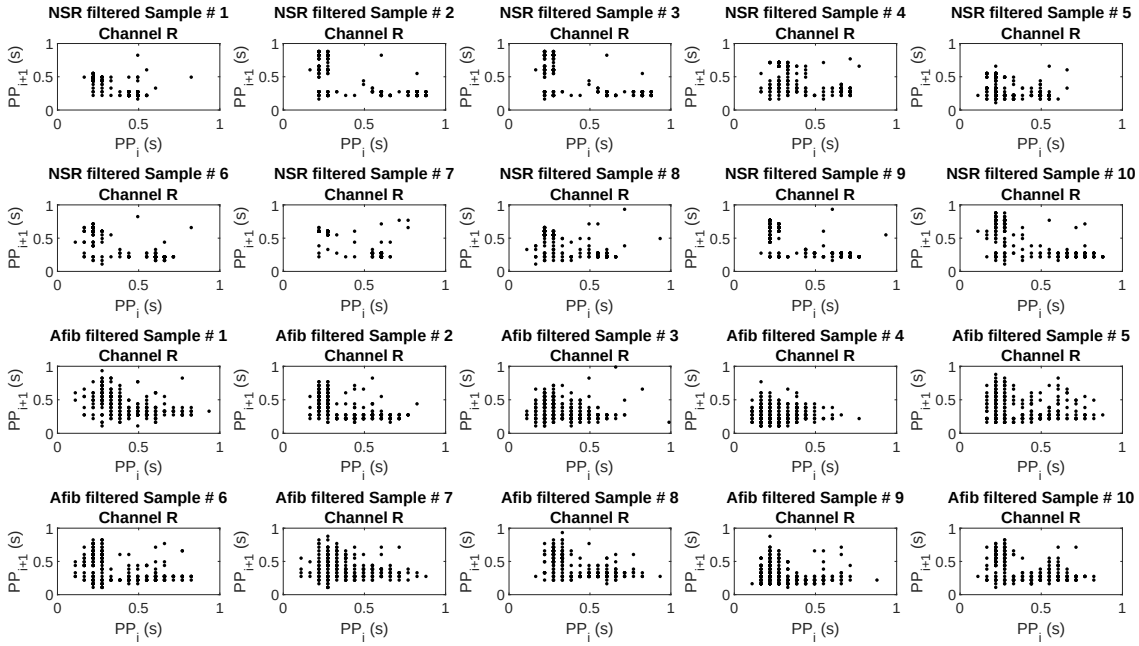


Figure 3.20: Poincaré plot of the PP time interval for the red channel data of the 20 subjects for which clinically-validated labels are available. Points appear more clustered for the NSR rhythm than for Afib.

We filtered the signals as per our standard procedure. We then proceeded to extract the PP time interval and amplitude of the peaks; we normalized the latter to its median value per each subject. After generating the three-dimensional Poincaré Plots, we computed the eigenvalues of the correlation matrix of the data, and we took the ratio between the two largest eigenvalues as a measure of the shape of the cluster in the Poincaré plot. Finally, we compared the values across the two classes, NSR and Afib. In Figures 3.20 and 3.21 we show the typical appearance of the Poincaré plots for the pilot dataset, respectively for the PP time interval and for the peak amplitude, both for the red channel, restricting to the bi-dimensional plots for greater clarity of visualization. In order to have a sense of the local density of the points, which may overlap, we produced heatmaps, which we present in Figure 3.22 for the PP time interval of the red channel and Figure 3.23 for the peak amplitude of the red channel.

Results

Comparing ratios The ratios $r = \lambda_i/\lambda_j$, where $(i, j) = \{(1, 2), (1, 3), (2, 3)\}$ of the three largest eigenvalues of the correlation matrix for the two classes are normally

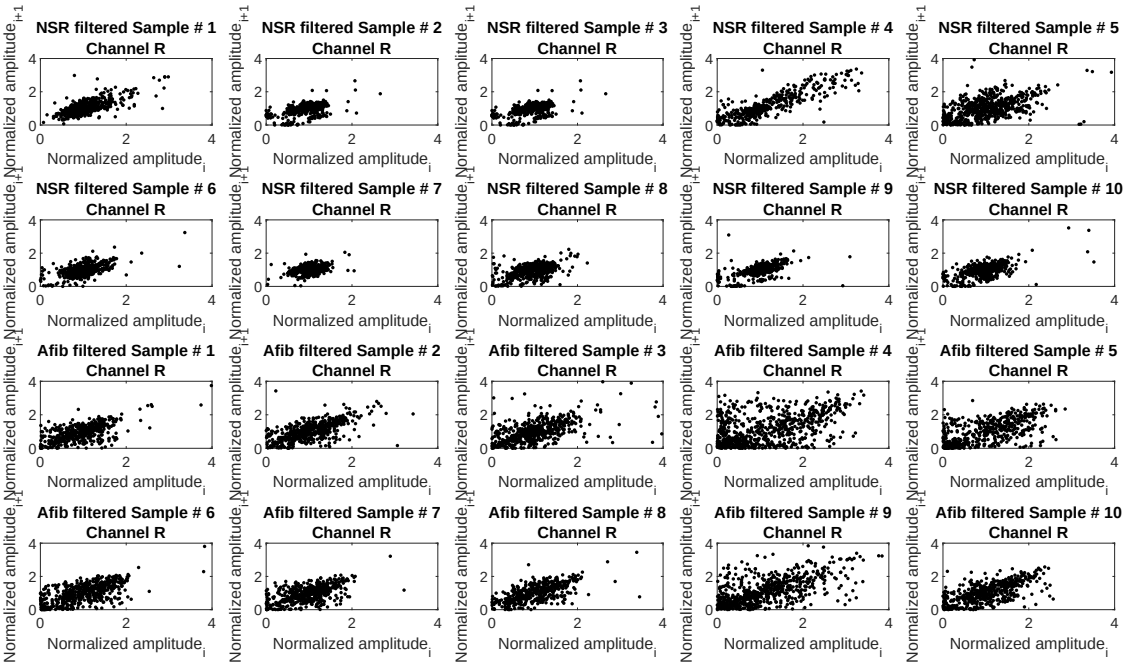


Figure 3.21: Poincaré plot of the normalized amplitude for the red channel data of the 20 subjects for which clinically-validated labels are available. Range has been restricted to the 0-4 interval for a clearer visualization. Data appears to be more clustered for NSR than for Afib, with the notable exceptions of samples NSR 4 and 5, which we suspect being corrupted by noise.

distributed, and the one-way ANOVA yields the results displayed in Table 3.8 for the normalized amplitude and in Table 3.9 for the PP time interval. The ANOVA assumes as null hypothesis that the mean is the same for the different classes and evaluates the F-statistic to compute the relevant p-value [58]. The F-statistic is defined as the ratio F between the variation from the mean between groups and the variation within each group from the mean estimate [59], normalized to their degrees of freedom ($k-1$ for the inter-group variation, and $N-1$ for the intra-group variation, where k, N are respectively the number of classes and the number of observations).

We tested whether the Afib and NSR results for the ratio r are consistent across the pilot and larger datasets: results are reported in Table 3.10. The larger dataset had 905 samples, whereas the pilot was made of 10 samples for the NSR class. We performed a second analysis in which we compared the large dataset with a balanced synthetic dataset generated from the pilot dataset, normally distributed and with the same mean and variance. The means were found to be significantly different ($p \approx 10^{-6}$), but this result will be further discussed later on.

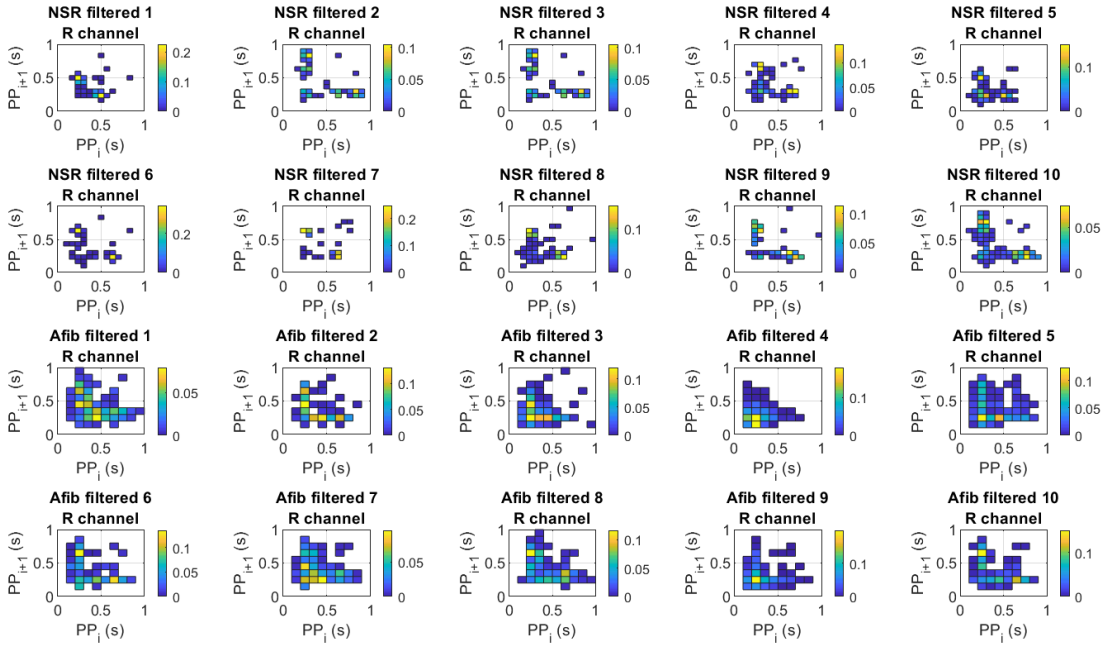


Figure 3.22: Tile-view of an intensity-normalized 3-D histogram of data of Figure 3.20. Areas in white indicate absence of data. NSR data presents often a bimodal appeareance and the heatmaps clearly display the greater degree of dispersion of the Afib data.

	λ_1/λ_2	λ_1/λ_3	λ_2/λ_3
Pilot	$p_R = 0.003$ $p_G = 0.4$ $p_B = 0.1$	$p_R = 0.02$ $p_G = 0.5$ $p_B = 0.2$	$p_R = 0.004$ $p_G = 0.8$ $p_B = 0.08$
Large	$p_R = 0.4$ $p_G = n.a.$ $p_B = n.a.$	$p_R = 0.3$ $p_G = n.a.$ $p_B = n.a.$	$p_R = 0.3$ $p_G = n.a.$ $p_B = n.a.$

Table 3.8: p-values from the ANOVA test between the classes NSR and Afib of each dataset. For the large dataset, only the labelled channel has been evaluated, that is the red channel. Feature analysed is the normalized amplitude.

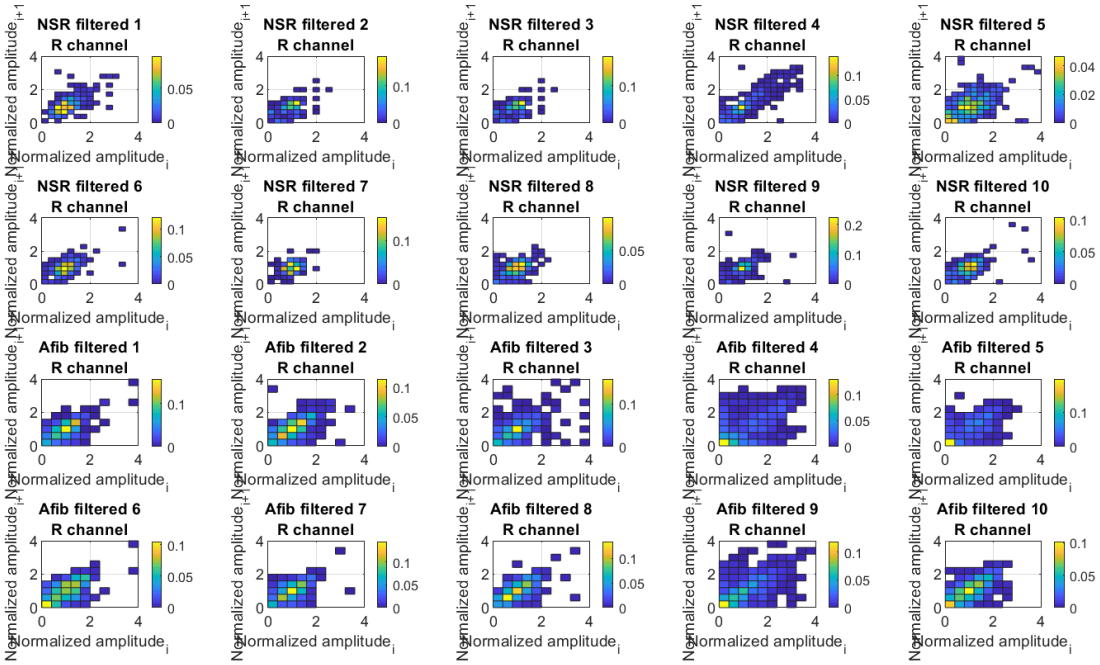


Figure 3.23: Tile-view of an intensity-normalized 3-D histogram of data of Figure 3.21. Areas in white indicate absence of data. Once more, the appearance of the data is more dispersed for Afib, with the exception of the two possibly-noisy samples of NSR. The restricted axis range of the plots hides the very high values of normalized amplitude that are present in the Afib population, which often exceeded 4.

	λ_1/λ_2	λ_1/λ_3	λ_2/λ_3
Pilot	$p_R = 0.1$ $p_G = 0.1$ $p_B = 0.7$	$p_R = 0.03$ $p_G = 0.02$ $p_B = 0.4$	$p_R = 0.3$ $p_G = 0.007$ $p_B = 0.5$
Large	$p_R = 3 \cdot 10^{-6}$ $p_G = \text{n.a.}$ $p_B = \text{n.a.}$	$p_R = 1 \cdot 10^{-6}$ $p_G = \text{n.a.}$ $p_B = \text{n.a.}$	$p_R = 0.6$ $p_G = \text{n.a.}$ $p_B = \text{n.a.}$

Table 3.9: p-values from the ANOVA test between the classes NSR and Afib of each dataset. For the large dataset, only the labelled channel has been evaluated, that is the red channel. Feature analysed is the PP time interval.

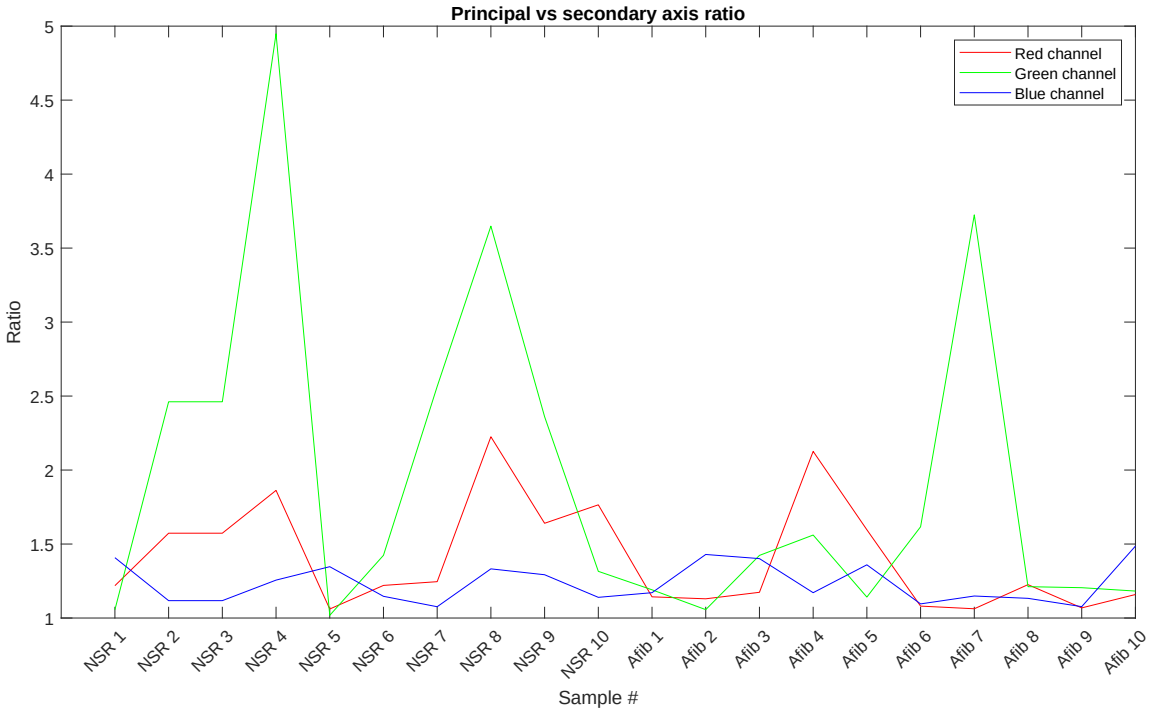


Figure 3.24: Ratio of the two largest eigenvalues of the correlation matrix for the pilot dataset, starting from the PP time interval values. The blue channel appears to be the most regular, whereas spikes are often present for both the red channel, to an extent, and for the green channel, more intensely.

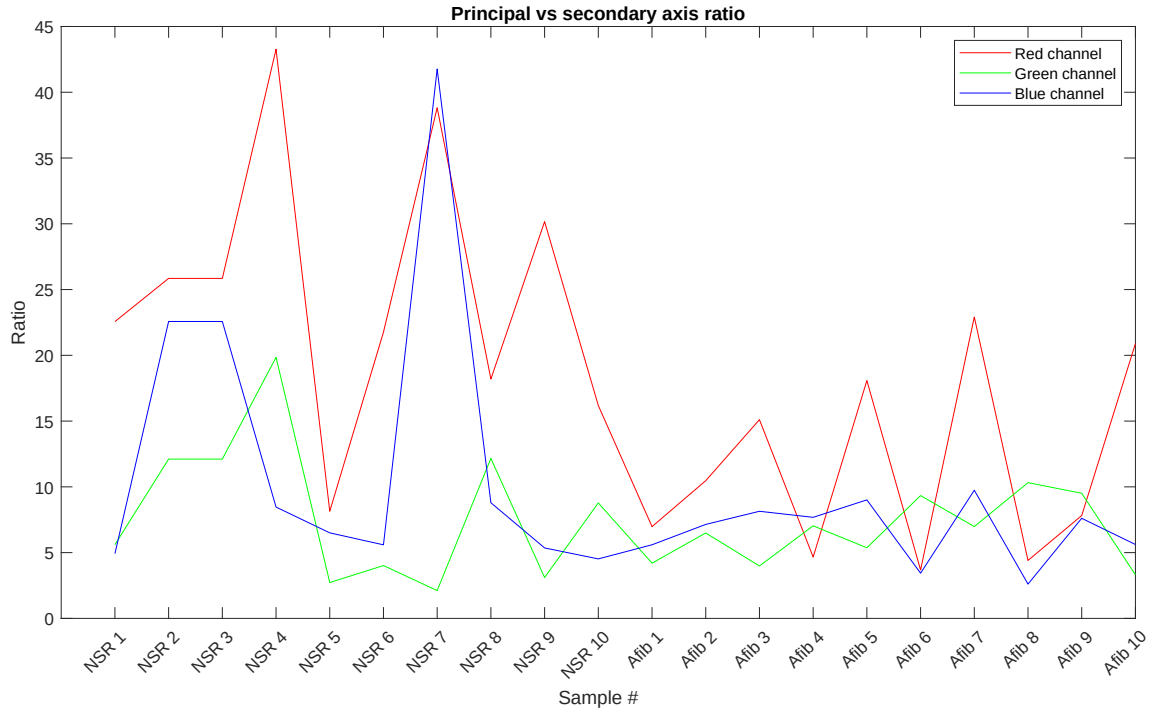


Figure 3.25: Ratio of the two largest eigenvalues of the correlation matrix for the pilot dataset, starting from the normalized amplitude values. For the red channel especially, values appear to be higher in the NSR population than in the Afib population.

Red channel	NSR	Afib
PP time interval	$p_{\lambda_1/\lambda_2} = 0.6$ $p_{\lambda_1/\lambda_3} = 0.5$ $p_{\lambda_2/\lambda_3} = 0.3$	$p_{\lambda_1/\lambda_2} = 0.3$ $p_{\lambda_1/\lambda_3} = 0.3$ $p_{\lambda_2/\lambda_3} = 0.5$
Normalized amplitude	$p_{\lambda_1/\lambda_2} = 0.07$ $p_{\lambda_1/\lambda_3} = 0.5$ $p_{\lambda_2/\lambda_3} = 0.5$	$p_{\lambda_1/\lambda_2} = 0.007$ $p_{\lambda_1/\lambda_3} = 0.04$ $p_{\lambda_2/\lambda_3} = 0.05$

Table 3.10: Comparing the consistency of result within classes across the two different datasets. Here we report the results of the three ratio comparisons for the red channel.

Discussion

As far as the 20-subjects dataset is concerned, Figures 3.22 and 3.23 present two interesting differences between the NSR and Afib classes: for PP time interval data, NSR tends to be bimodal and less spread than Afib, which presents only one peak and a larger area covered; amplitude data shows a larger spread for the Afib series, with some exceptions that may be due to noisy acquisitions both in the NSR and Afib groups, as shown by large transients in the unrestricted-range Poincaré plot.

Testing generally differentiates between the eigenvalue ratios of the two groups, NSR and Afib, only for the red channel of the clinically-validated dataset, for the PP time feature. Conversely, ratios are almost always comparable across the two datasets, with the only exception of the amplitude for Afib subjects for the λ_1/λ_2 ratio. Summing up, only the red channel seems to have some degree of useful information embedded. We must stress the fact that each dataset has its own drawbacks: the larger is not clinically validated, the smaller has very few samples per each group. This could be a factor influencing our synthetic dataset, which aims to reproduce the mean and variance of such a small set: the implicit hypothesis that 10-20 samples are representative of what a larger collection would yield is very strong and must be taken into consideration when assessing the strength of our previous result. Nevertheless, the results do look promising, with the potential for assessing different morphological features of the distributions of data on the Poincaré Plots, according to their rhythms. The robustness of further findings would benefit from a large dataset of accurately labelled signals and metadata.

4 Discussion and conclusion

4.1 Discussion

We now wish to retrace our main results and provide more context and perspective to these findings.

In **Section 3.1** we resolved to characterize the dataset, by assessing whether any factor, such as age, sex, BMI, device, lifestyle, would display signature properties of the dataset.

Our findings indeed recovered some differences, between single groups of a class, such as between the distributions of time intervals in iPhone 5S and iPhone 7 Plus and amplitude for iPhone 5S and iPhone 7. While this result in itself might be of limited practical interest, both because of the fast life-cycle of devices and because of the limited applicability to an analytical pipeline, we might speculate that it reflects some difference in the camera of the two devices.

BMI shows a degree of difference between the distribution values too, especially when considering peak area, between the underweight class and the overweight and normal class. While this result might prove true, it must be noted once again that the underweight class had just some tens of subjects, whereas the others were more abundant.

Age, we could summarize, shows a light divide at the age of forty, with the PP time interval feature different between the younger group and those above forty years old. We warn that this finding however is not bulletproof, since some classes do not result in significant differences, such as those older than 70 when compared to people under 30.

We found no difference within the lifestyle class, in contrast to our expectations, which would refer to the commonplace notion that sedentary lifestyles do worsen cardiac health. The noise level of the dataset could have contributed to this. Moreover,

we must remember that labels are self-reported and loosely defined and that our non-parametric tests are constrained by the Bonferroni correction which can lead to a high number of false negatives. Moreover, a lot of other co-factors might hide any underlying difference.

In fact, this might very well be the most important limitation of our exploratory analysis, that it wasn't multivariate and, as such, perhaps, it still has not delved deep enough into the layers of overlap.

Other findings denote that the Weibull distribution often (and, less often, the log-normal) might describe the data distributions, even though on several instances no distribution was found.

It is interesting to note that, in agreement with our expectations (from our readings in literature [12]), we noticed that red and green channel produce fairly similar results, whereas the blue channel does not.

In **Section 3.2** we followed through with our analysis of the issue of signal quality, that we started in Chapter 2, the correct classification of which is of crucial importance to the feasibility of accurate real-time monitoring of the cardiac rhythm type, and possibly diagnostics. It is important to note that several works either focused on small sized sets of controlled acquisitions to address signal quality alone or focused on classifying the rhythm class without a control on signal quality, apart from the pre-processing pipeline.

We chose to explore the possibilities of dynamic time warping, a technique to achieve a better match of two signals, on a multiple-peak segment-wise approach, and we combined it with some very straightforward signal quality indexes widely used in literature.

The optimization of the match between the signals proved tricky, but the classes were indeed distinguished. The solution of our optimization class-wise does not avert the fact that many distance features of the different classes overlap for the vast majority of their values, though. When combined with the other SQIs, the DTW distance feature produces a good result in a classification test by means of a neural network, achieving almost 89% accuracy on the test set, shown to the model only after the 5-fold cross validation was completed. This result is a good starting point, given the simplicity of the features provided and considering the small size of our dataset, which had to be under-sampled to achieve class balance and has thus to consider class

size as the limiting factor. However, it performs far worse than other techniques have achieved, which is even above 95% [33].

Building upon the notion of chaotic systems, as the heart beat is regarded to be, in **Section 3.3** we aimed at exploiting Poincaré plots, a kind of recurrence plots widely used in the assessment of heart rate variability, first to a small, clinically-labelled dataset and subsequently to our larger dataset, in order to check for different forms of the Poincaré plots, between the Afib and NSR classes.

We used the ratios of the magnitudes of the eigenvalues of the covariance matrix to assess the relative spread of the data in the Poincaré plots. We found significant differences on the distributions for the PP time values, when the ratio involved the largest eigenvalue. This dataset only had labels for the red channel.

The smaller pilot dataset shows instead significant values when the eigenvalues are divided by the third largest eigenvalue. Limitation of this analysis involve several factors: firstly, the clinically-reliable set is very small, but also there is no label referring to the quality of the data in it. We suspect that some of its data have a high degree of noise, however we were not able to unambiguously support or disprove this suspicion. It is worth looking deeper into these results: more robust conclusions could be drawn by performing the analysis on a large dataset, labelled reliably.

Conversely, the larger dataset has been labelled, even though not clinically or selectively for segments of the set, which would be preferable to the global quality and confidence scores that we had, as segments of partial lower quality might influence the results.

The comparison of results between the two datasets suffers from the class imbalance. Even the synthetic generation of samples does not fully address this problem, as the original dataset for generation is indeed very small and might not be sufficient to accurately represent a wider population. A large-size, reliably-labelled dataset would prove crucial for obtaining more robust results.

4.2 Conclusions

Concluding, we might state the promising and vast field of PPG-based cardiac health monitoring faces great challenges if it is to provide real-time feedback and diagnostics to end users or health care providers.

The open issue of dealing with noisy measurements when addressing different type of rhythms for clinical purposes is such, open. Our piece of research points to the direction of the strong need for accurately labelled measurements and multivariate analysis, to conclusively answer the questions of underlying effects in different groups.

A heart rhythm classification that takes into account also signal quality seems to be within reach, especially by exploiting the power of machine learning data analysis techniques.

The physical approach of recurrence of the chaotic time series yields promising findings as far as discerning between NSR and Afib rhythms is concerned. At this stage, classification is potentially possible via analysing the 3D spread of the Poincaré plots of the time intervals of a signal.

4.3 Possible developments

As the field is thriving with research, several possibilities can be further pursued.

Specifically, several times during our analysis, we felt the need of using a growing number of reliably labelled signals and clinical reports, so to assess the correspondence of user-reported data to the signal features or, more simply, to be able to rely on the labels provided. On that, a multivariate analysis of the data could discriminate different properties of the groups.

Some publicly accessible PPG databases are available, some are annotated, or present also simultaneous labelled ECG leads, which can be used to reliably assess the heart rhythm.

Signal quality classification might benefit from additional features and exploring the performances of other classifiers.

The potential of several quantities is still left to explore within the chaos theory approach; however, the crucial goal to achieve in future is to expand the size of reliably-labelled datasets, in order to obtain robust results.

References

[1] *What Is Atrial Fibrillation (Afib)?* URL: https://www.afibmatters.org/en_GB/About-atrial-fibrillation (visited on 02/25/2019).

[2] *Atrial Fibrillation*. 20 Oct 2017, 12:22 p.m. URL: https://www.nhs.uk/conditions/atrial-fibrillation/ (visited on 02/25/2019).

[3] *What Is Atrial Fibrillation?* URL: https://www.heartfoundation.org.nz/your-heart/heart-conditions/atrial-fibrillation (visited on 02/25/2019).

[4] Nileshkumar J Patel et al. "Global Rising Trends of Atrial Fibrillation: A Major Public Health Concern". In: *Heart* 104.24 (Dec. 2018), pp. 1989–1990. ISSN: 1355-6037, 1468-201X. DOI: 10.1136/heartjnl-2018-313350.

[5] *Atrial Fibrillation Fact Sheet|Data & Statistics|DHDSP|CDC*. URL: https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_atrial_fibrillation.htm (visited on 02/25/2019).

[6] Lau Dennis H. et al. "Modifiable Risk Factors and Atrial Fibrillation". In: *Circulation* 136.6 (Aug. 8, 2017), pp. 583–596. DOI: 10.1161/CIRCULATIONAHA.116.023163.

[7] Heart Rhythm Society. *Treatment*. URL: https://www.hrsonline.org/Patient-Resources/Treatment (visited on 02/25/2019).

[8] *Tests and Investigations*. URL: https://www.afibmatters.org/en_GB/Tests-and-investigations (visited on 02/25/2019).

[9] Sudler & Hennessey. *Heart For Heart*. URL: http://www.heartrateapp.com/ (visited on 02/25/2019).

[10] Alrick B. Hertzman. “THE BLOOD SUPPLY OF VARIOUS SKIN AREAS AS ESTIMATED BY THE PHOTOELECTRIC PLETHYSMOGRAPH”. In: *American Journal of Physiology-Legacy Content* 124.2 (Oct. 31, 1938), pp. 328–340. ISSN: 0002-9513. DOI: 10.1152/ajplegacy.1938.124.2.328.

[11] *Etymology of the Word Plethysmograph*. URL: http://clas.mq.edu.au/speech/physiology/respiration/etymology/index.html (visited on 02/25/2019).

[12] Toshiyo Tamura et al. “Wearable Photoplethysmographic Sensors—Past and Present”. In: *Electronics* 3.2 (Apr. 23, 2014), pp. 282–302. ISSN: 2079-9292. DOI: 10.3390/electronics3020282.

[13] R. Rox Anderson and John A. Parrish. “The Optics of Human Skin”. In: *Journal of Investigative Dermatology* 77.1 (July 1981), pp. 13–19. ISSN: 0022202X. DOI: 10.1111/1523-1747.ep12479191.

[14] Yongbo Liang et al. “A New, Short-Recorded Photoplethysmogram Dataset for Blood Pressure Monitoring in China”. In: *Scientific Data* 5 (Feb. 27, 2018), p. 180020. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.20.

[15] Ainara Garde et al. “Estimating Respiratory and Heart Rates from the Correntropy Spectral Density of the Photoplethysmogram”. In: *PLoS ONE* 9.1 (Jan. 22, 2014). Ed. by Derek Abbott, e86427. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0086427.

[16] Syed Ahmar Shah et al. “Respiratory Rate Estimation during Triage of Children in Hospitals”. In: *Journal of Medical Engineering & Technology* 39.8 (Nov. 17, 2015), pp. 514–524. ISSN: 0309-1902. DOI: 10.3109/03091902.2015.1105316. pmid: 26548638.

[17] Xiaorong Zhang and Quan Ding. “Respiratory Rate Monitoring from the Photoplethysmogram via Sparse Signal Reconstruction”. In: *Physiological Measurement* 37.7 (July 1, 2016). ISSN: 0967-3334, 1361-6579. DOI: 10.1088/0967-3334/37/7/1105.

[18] Sebastian Zaunseder et al. “Cardiovascular Assessment by Imaging Photoplethysmography – a Review”. In: *Biomedical Engineering / Biomedizinische*

Technik 63.5 (2018), pp. 617–634. ISSN: 0013-5585. DOI: 10.1515/bmt-2017-0119.

[19] Qiao Li and Gari D. Clifford. “Signal Quality and Data Fusion for False Alarm Reduction in the Intensive Care Unit”. In: *Journal of Electrocardiology* 45.6 (Nov. 1, 2012), pp. 596–603. ISSN: 0022-0736. DOI: 10.1016/j.jelectrocard.2012.07.015.

[20] J. Abdul Sukor, S. J. Redmond, and N. H. Lovell. “Signal Quality Measures for Pulse Oximetry through Waveform Morphology Analysis”. In: *Physiological Measurement* 32.3 (2011), p. 369. ISSN: 0967-3334. DOI: 10.1088/0967-3334/32/3/008.

[21] Mohamed Elgendi. “On the Analysis of Fingertip Photoplethysmogram Signals”. In: *Current Cardiology Reviews* 8.1 (Feb. 2012), pp. 14–25. ISSN: 1573-403X. DOI: 10.2174/157340312801215782. pmid: 22845812.

[22] *Toward Generating More Diagnostic Features from Photoplethysmogram Waveforms*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5871966/ (visited on 10/04/2018).

[23] Reza Firoozabadi, Eric D. Helfenbein, and Saeed Babaeizadeh. “Efficient Noise-Tolerant Estimation of Heart Rate Variability Using Single-Channel Photoplethysmography”. In: *Journal of Electrocardiology* 50.6 (Nov. 2017), pp. 841–846. ISSN: 00220736. DOI: 10.1016/j.jelectrocard.2017.08.020.

[24] S. Nemati et al. “Monitoring and Detecting Atrial Fibrillation Using Wearable Technology”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Aug. 2016, pp. 3394–3397. DOI: 10.1109/EMBC.2016.7591456.

[25] Christina Orphanidou et al. “Signal Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring”. In: *IEEE Journal of Biomedical and Health Informatics* (2014), pp. 1–1. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2014.2338351.

[26] I. Silva, J. Lee, and R. Mark. “Photoplethysmograph Quality Estimation through Multichannel Filtering”. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Aug. 2011, pp. 4361–4364. DOI: 10.1109/IEMBS.2011.6091082.

[27] Andrius Sološenko et al. “Modeling of the Photoplethysmogram during Atrial Fibrillation”. In: *Computers in Biology and Medicine* 81 (Feb. 2017), pp. 130–138. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2016.12.016.

[28] Chenggang Yu et al. “A Method for Automatic Identification of Reliable Heart Rates Calculated from ECG and PPG Waveforms”. In: *Journal of the American Medical Informatics Association : JAMIA* 13.3 (2006), pp. 309–320. ISSN: 1067-5027. DOI: 10.1197/jamia.M1925. pmid: 16501184.

[29] Eric W. Weisstein. *Discrete Fourier Transform*. URL: http://mathworld.wolfram.com/DiscreteFourierTransform.html (visited on 02/26/2019).

[30] Rajet Krishnan, Balasubramaniam Bala Natarajan, and Steve Warren. “Two-Stage Approach for Detection and Reduction of Motion Artifacts in Photoplethysmographic Data”. In: *IEEE transactions on bio-medical engineering* 57.8 (Aug. 2010), pp. 1867–1876. ISSN: 1558-2531. DOI: 10.1109/TBME.2009.2039568. pmid: 20172800.

[31] J. D. Wander and D. Morris. “A Combined Segmenting and Non-Segmenting Approach to Signal Quality Estimation for Ambulatory Photoplethysmography”. In: *Physiological Measurement* 35.12 (2014), p. 2543. ISSN: 0967-3334. DOI: 10.1088/0967-3334/35/12/2543.

[32] W. Karlen et al. “Photoplethysmogram Signal Quality Estimation Using Repeated Gaussian Filters and Cross-Correlation”. In: *Physiological Measurement* 33.10 (2012), p. 1617. ISSN: 0967-3334. DOI: 10.1088/0967-3334/33/10/1617.

[33] Q. Li and G. D. Clifford. “Dynamic Time Warping and Machine Learning for Signal Quality Assessment of Pulsatile Signals”. In: *Physiological Measurement* 33.9 (2012), p. 1491. ISSN: 0967-3334. DOI: 10.1088/0967-3334/33/9/1491.

[34] Mohamed Elgendi. “Optimal Signal Quality Index for Photoplethysmogram Signals”. In: *Bioengineering* 3.4 (Sept. 22, 2016). ISSN: 2306-5354. DOI: 10.3390/bioengineering3040021. pmid: 28952584.

[35] Yongbo Liang et al. “An Optimal Filter for Short Photoplethysmogram Signals”. In: *Scientific Data* 5 (May 1, 2018). ISSN: 2052-4463. DOI: 10.1038/sdata.2018.76. pmid: 29714722.

[36] *Calculate Your BMI - Standard BMI Calculator*. URL: https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm (visited on 02/24/2019).

[37] *1.3.6.6.3. Cauchy Distribution*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3663.htm (visited on 02/25/2019).

[38] *1.3.6.6.9. Lognormal Distribution*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm (visited on 02/25/2019).

[39] *Lognormal Distribution - MATLAB & Simulink - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/lognormal-distribution.html (visited on 02/25/2019).

[40] *1.3.6.6.8. Weibull Distribution*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3668.htm (visited on 02/25/2019).

[41] *Weibull Distribution - MATLAB & Simulink - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/weibull-distribution.html (visited on 02/25/2019).

[42] *Gamma Probability Density Function - MATLAB Gampdf - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/gampdf.html (visited on 02/25/2019).

[43] *1.3.6.6.11. Gamma Distribution*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366b.htm (visited on 02/25/2019).

[44] *1.3.6.6.7. Exponential Distribution*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3667.htm (visited on 02/25/2019).

[45] *Exponential Distribution - MATLAB & Simulink - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/exponential-distribution.html (visited on 02/25/2019).

[46] *1.3.5.14. Anderson-Darling Test*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm (visited on 02/25/2019).

[47] *Anderson-Darling Test - MATLAB Adtest - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/adtest.html?s_tid=doc_ta (visited on 02/24/2019).

[48] Eric W. Weisstein. *Bonferroni Correction*. URL: http://mathworld.wolfram.com/BonferroniCorrection.html (visited on 02/24/2019).

[49] Megan Goldman. "1 Why Is Multiple Testing a Problem?" In: (), p. 5.

[50] *Kruskal-Wallis Test - MATLAB Kruskalwallis - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/kruskalwallis.html?s_tid=doc_ta (visited on 02/24/2019).

[51] *7.4.1. How Can We Compare Several Populations with Unknown Distributions (the Kruskal-Wallis Test)?* URL: https://www.itl.nist.gov/div898/handbook/prc/section4/prc41.htm (visited on 02/25/2019).

[52] *Distance between Signals Using Dynamic Time Warping - MATLAB Dtw - MathWorks Italia*. URL: https://it.mathworks.com/help/signal/ref/dtw.html (visited on 02/24/2019).

[53] H. Sakoe and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (Feb. 1978), pp. 43–49. ISSN: 0096-3518. DOI: 10.1109/TASSP.1978.1163055.

[54] K. Paliwal, A. Agarwal, and S. Sinha. "A Modification over Sakoe and Chiba's Dynamic Time Warping Algorithm for Isolated Word Recognition". In: *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 7. Paris, France: Institute of Electrical and Electronics Engineers, 1982, pp. 1259–1261. DOI: 10.1109/ICASSP.1982.1171506.

[55] James Gleick. "Chaos: Making a New Science". In: (), p. 119.

[56] M. Brennan, M. Palaniswami, and P. Kamen. “Do Existing Measures of Poincare Plot Geometry Reflect Nonlinear Features of Heart Rate Variability?” In: *IEEE Transactions on Biomedical Engineering* 48.11 (Nov./2001), pp. 1342–1347. ISSN: 00189294. DOI: 10.1109/10.959330.

[57] Norbert Marwan et al. “Recurrence-Plot-Based Measures of Complexity and Their Application to Heart-Rate-Variability Data”. In: *Physical Review E* 66.2 (Aug. 6, 2002). ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.66.026702.

[58] *One-Way ANOVA - MATLAB & Simulink - MathWorks Italia*. URL: https://it.mathworks.com/help/stats/one-way-anova.html (visited on 02/25/2019).

[59] *1.3.5.9. F-Test for Equality of Two Variances*. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm (visited on 02/25/2019).