FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

ALEX BARROS DOS SANTOS

A MACHINE LEARNING FRAMEWORK FOR ECG BIOMETRIC SYSTEM

DM: 11/2020

UFPA / ITEC / PPGEE
Guamá University Campus
Belém-Pará-Brazil

2020

ALEX BARROS DOS SANTOS

A MACHINE LEARNING FRAMEWORK FOR ECG BIOMETRIC SYSTEM

Dissertation submitted to the Judging Committee at the Federal University of Pará as part of the requirements for obtaining a Master's Degree in Electrical Engineering in the area of Applied Computing.

Advisor: Eduardo Coelho Cerqueira

Co-Advisors: Denis Lima do Rosário

BELÉM-PARÁ-BRAZIL

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S237m Santos, Alex

A machine learning framework for ECG biometric systems / Alex Santos. — 2020.

62 f. : il. color.

Orientador(a): Prof. Dr. Eduardo Cerqueira

Coorientador(a): Prof. Dr. Denis Lima do Rosário

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2020.

1. Biometria. 2. Machine Learning. 3. Eletrocardiografia . 4. Redes de Computadores. 5. Dispositivos Vestíveis. I. Título.

CDD 621.3

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"A MACHINE LEARNING FRAMEWORK FOR ECG BIOMETRIC SYSTEMS"

AUTOR: **ALEX BARROS DOS SANTOS**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 28/02/2020

BANCA EXAMINADORA:

Eduardo Cerqueira
Prof. Dr. Eduardo Coelho Cerqueira
(Orientador – PPGEE/UFPA)

Denis Lima do Rosário
Prof. Dr. Denis Lima do Rosário
(Co-Orientador – PPGEE/UFPA)

Adriana Rosa Garcez Castro
Prof.ª Dr.ª Adriana Rosa Garcez Castro
(Avaliadora Interna – PPGEE/UFPA)

Thaís Lira Tavares dos Santos
Prof.ª Dr.ª Thaís Lira Tavares dos Santos
(Avaliador Externo ao Programa – GERCOM/UFPA)

VISTO:

Prof.ª Dr.ª Maria Emília de Lima Tostes
(Coordenadora do PPGEE/ITEC/UFPA)

ALEX BARROS DOS SANTOS

A MACHINE LEARNING FRAMEWORK FOR ECG BIOMETRIC SYSTEM

Dissertation submitted to the Judging Committee at the Federal University of Pará as part of the requirements for obtaining a Master's Degree in Electrical Engineering in the area of Applied Computing.

Approved at: __/__/____

Masters Dissertation Examination Board

Prof. Ph.D. Eduardo Coelho Cerqueira
(Federal University of Pará - Advisor)

Prof. Ph.D. Denis Lima do Rosário
(Federal University of Pará- Co-Advisor)

Prof. Ph.D. Adriana Rosa Garcez Castro
(Federal University of Pará - Internal Member)

Ph.D. Thais Lira Tavares Dos Santos
(Federal University of Pará - External Member)

Dedico este trabalho a Deus e minha amada família.
I dedicate this work to God and my beloved family.

Acknowledgements

I would like to express my sincere thanks to:

Our Almighty God Who has always covered me with blessings throughout my life and all the places I have been.

My family (father Abdias, mother Vânia and sister Amanda) who always supported me at all the times. Thanks for believing that education would bring us here. My wife Fernanda who help me to survive during some tough months. Once again we will pretend that "dog days are over".

Professor Eduardo Cerqueira for the opportunity and confidence deposited in me. Professor Denis Rosário for guidance, time and effort dedicated so that we could get things done before the deadline. I'm sorry about the deadlines we missed. Thanks for the friendship and for keeping push me to improve.

I would like to acknowledge the financial support obtained from RNP-NSF for the project HealthSense: Assessing and Protecting Privacy in Wireless Wearable Sensor-generated Medical Data. Particularly, the partnership between the Federal University of Para and Federal University of Parana made this project happens.

And last but not least, for my friends from GERCOM, especially to Thais, Lucas, Iago, Paulo, Nagib, and friends from UFPR who were part of HealthSense team. We had so many meetings on Fridays afternoon, I hope we can meet more time to drink and cheers in the future. Also a big thanks to Marcelo and Renato, two young guys who helped me a lot into this world of Python and Machine Learning. Thanks for allowing me to be part of your team.

"Even a man on a sinking ship can be happy when he clambers aboard a lifeboat (...) Compare yourself to who you were yesterday, not to who someone else is today"
Jordan B. Peterson

Abstract

Abstract of Master's Thesis presented to UFPA as part of the requirements for obtaining a Master's Degree in Electrical Engineering.

A Machine Learning Framework for ECG Biometric System

Advisor: Eduardo Coelho Cerqueira
Co-advisor: Denis Lima do Rosário
Key words: 1. Biometric. 2. Machine Learning. 3. Electrocardiogram. 4. Computer Networks. 5. Wearables.

The new environment of IoT and the deployment of 5G networks have been generating a huge amount of data. Developers are creating new applications and redesigning other ones completely. Also, a society greater concern with health increases the demand for health services provided with the usage of wearable devices that are getting cheaper. Moreover, the applications require more data protection and privacy. Thus, biometrics has become one of the primary mechanisms for protecting information used by users in all kind of systems and applications. This work investigates the use of an ECG signal in biometrics systems approaching machine learning techniques. This signal is a new alternative not only to increase current safety standards by providing the individual's continuous authentication but also to assess health with cardiac monitoring already well established in medicine by evaluations. In this context, this master's thesis proposes some processing steps to data sets, improving its quality that allows it to be used as a reliable source of biometric data. We define techniques for extracting signal considering mobile application constraints and design a structure that allows the use of ECG as a biometric signal in a scalable and heterogeneous environment considering different machine learning techniques such as Support Vector Machine, Random Forest and Neural Networks. The set of our proposed feature extraction, processing steps of data set and a machine learning model are the main contributions of this work.

Contents

List of Figures

List of Tables

List of Abbreviations

AF	Atrial Fibrillation
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
CA	Continuous Authentication
CCNL	Computational Clinical Neurophysiology Laboratory
CDAC	Clinical Data Animation Laboratory
CNN	Convolutional neural network
DNN	Deep Neural Network
DT	Decision Tree
DWT	Discrete WT
ECG	Electrocardiogram
EEG	Electroencephalogram
EER	Equal Error Rate
EMG	electromyography
EOG	electrooculography
EOG	electrooculography
GSR	galvanic skin resistance
HP	Hewlett-Packard
IMDs	Implantable Medical Devices
IoT	Internet of Things
LTU	Linear Threshold Unit
MGH	Massachusetts General Hospital
MLP	Multilayer Perceptron

NCA	Non-Continuous Authentication
NNC	nearest neighbor classifierSVM
PCA	principal component analysis
PPG	Photoplethysmogram
PTB	Physikalisch-Technische Bundesanstalt
RF	Random Forest
ROC	Receiver Operating Characteristic
SaO2	oxygen saturation
SIMCA	soft independent modeling of class analogy
SVM	support vector machine
TLU	Threshold Logic Unit
WT	wavelet transform

CHAPTER 1

Introduction

1.1 Overview

In the last few years, wearable technology reached the realms of science fiction, and computing devices stopped to be an item used only at our homes and places of work or carry in our bags and pockets. Nowadays, we can wear those devices and use them connected to the Internet, providing a variety of new applications. Connectivity has become easy and increasingly crucial for many of us and access to data and knowledge has been made affordable, manageable and convenient. Comparatively speaking, wearable devices are facing the same challenge cellphones did in the 2000s. They are becoming popular among some tech users very fast and arousing curiosity among the regular ones [8].

However, wearable devices are not a new phenomenon. They have been around for many years but have been expensive and cumbersome and were mostly used by people engaged in research and development projects [9]. The pioneering wearable computers were introduced in the 1980s, but recently with sophisticated features, performing a variety of tasks and allowing the enhancement of operations of some professions they established itself as a new trend product. Interestingly, it is computer and software companies that are developing wearable devices, not manufacturers of traditional wearable products [9].

Wearable devices have great potential for health monitoring and fitness. For instance, the US alone spends almost 4 trillion dollars on healthcare each year [10]. Wearable devices provide a lower cost for the treatment, greater mobility for the patient and, potentially, improved physiological data for the physicians to analyze when attempting to diagnose the condition [11, 12, 13]. Ideally, these systems should be able to carefully, conveniently and robustly monitor patients at the same time, they perform their daily activities (such as eating, sleeping, walking, and natural communication with parents)

without interfering significantly with their comfort [14, 15]. In this way, wearable devices can sense, collect, and upload physiological data continuously, provides opportunities to improve the efficiency of healthcare systems and the quality of life [8].

Over the last few years, wearable devices appeared in the market offering different functionalities and wearing options at the same time, a significant volume of scientific effort has been done highlighting various research challenges related to wearable devices [16, 17, 18]. In this way, there is a big market to be explored that motivates companies to look for the huge reward that can be gained by tapping into such markets. A portion of the health budget has been directed towards the acquisition of wearable devices and this market has increased very quickly recently as can be seen in Table 1 [7]. The global shipments of wearable devices have almost doubled in the last year.

Company	3Q19 Shipments	3Q18 Shipments	Market Share	Year-Over-Year Growth
1.Apple	29.5	10.0	35.0%	195.5%
2.Xiaomi	12.4	7.4	14.6%	66.1%
3.Samsung	8.3	3.2	9.8%	156.4%
4.Huawei	7.1	2.3	8.4%	202.6%
5.Fitbit	3.5	3.5	4.1%	0.5%
6.Others	23.8	16.9	28.1%	40.4%
Total	**84.5**	**43.4**	**100.0%**	**94.6%**

Table 1: Top 5 Wearables Companies by Shipment Volume, Market Share and Year-Over-Year Growth, Q3 2019 (shipments in millions)[7]

The market has been categorized mainly in ear-wear, wristbands, and smartwatches [8]. The ear-wear presented the highest year-over-year growth and amassed the most market share compared to all other wearable products. Wristbands maintained their popularity within the market, ahead of smartwatches and other wearable devices, driven by its straightforward value proposition and lower price points. In the case of smartwatches, price-sensitive customers took advantage of the lower prices for devices capable of many, if not most, of the features of its newer cousins (wristband).

Despite their price, smartwatches are the product with more features and capabilities to win over the costumers with in-built 4G and LTE connection they don't have to be cellphone dependent anymore, and some specialists believe that sometime, in the future, they might threaten the smartphone market itself. The customers now recognize the value of being able to complete a wide range of tasks on the device, including receiving notifications, messaging, accessing smart home controls, among others. The big traditional watch brands like Rolex and Patek Philippe, for instance, are taking a look at Apple, Samsung, and others as they are threatening a well-established market of luxury accessories. For example, traditional watches fell to just 56% of the value of total U.S. watch sales last year smartwatches outsold classic watches, accounting for 55% of total

sales for the holiday period in the 4th quarter of 2018. Also, according to some market research [19], one in four Americans (23%) aged 18 to 34 owned a smartwatch in 2018. Hence, owning a smartwatch is no longer an assumption in the future. It is going to be a premise soon.

As those devices are used to do so many tasks and store a significant amount of personal data, the concern of how our data can be exposed starts to gain attention. They have GPS in-built, usually can receive all kinds of messages like old SMS, incoming calls, e-mail messages, and notifications from applications. Wearable sensors collect biometrics and/or environmental interaction data, such as pulse rate, Electrocardiogram (ECG), Electroencephalogram (Electroencephalogram), Photoplethysmogram(PPG), O2 saturation, blood glucose levels, accelerometer data for remote controllers or impact assessments for sports players, exercise metrics (*e.g.*, running distances), visual data, weather information, and others [20]. In this way, on one side, the wearable devices are providing a wide range of new applications and can be used as one sensor to collect a new biometric sign (we will explain it in the next sections). They will be a possible point of failure if the manufactures don't consider some security issues such as the wireless communication between the device and the Internet and the hardware/software aspects.

According to McAfee [21], 43% of people surveyed feel that they have a lack of control over their personal information. Only 37% of individuals use an identity theft protection solution. However, 61% of respondents are more concerned about cybersecurity than they were five years ago. With the boom of the Internet of Things (IoT) devices, the home network is being asked to handle more devices than ever before, and it is becoming harder to manage for the owners of those devices. Then, 52% of respondents were either unsure or had no idea how to check to see if their connected devices and apps are secured. Considering the growth of the wearable market, the concerns of privacy and security linked to them, this master's thesis explores the use of biometric as a crucial factor of improvement in security and personal data protection.

1.2 Motivations and challenges

We are living a new era of wearable devices and connectivity when more data than ever have been produced. There is an evident lack of security and we are almost without control of our data. In this context, some security methods have to be redesigned to meet new requirements of privacy. Many different application systems support Internet access for general users like Websites, smartphones, safes, cars, houses, buildings, banks, and airports, all of them rely on any identification or authentication systems to protect and guard ourselves, our information, or our belongings. Several still depend on traditional systems based on cards, keys, or passwords. A security measures found on something you have (cards or keys) or something you know (password). However, in the last decades, researchers have focused on avoiding the problems of traditional systems: they can be lost, stolen, discovered, or copied.

Biometrics presents the perfect opportunity to achieve that goal, as they are focused on intrinsic characteristics of the person, requiring their physical presence, and minimizing the probability of success of possible impostors. Then, security started to be based on something you are. A biometric system aims to either identify or authenticate a person based on a measurement of one or several biometric traits. Many human traits have been proposed and studied for identity recognition, especially face, fingerprints, voice, and iris [22]. However, fingerprint and voice proved to be vulnerable years ago. The last trait that arrived in smart devices was the biometric-based on the face and initially, it sounds to be secure enough. However, some weakness was already discovered. For instance, Apple's FaceID has a liveness detection feature, which prevents someone who is sleeping to have his phone unlocked by putting it in front of his face while he's sleeping. To hack this feature, researchers bypassed a victim's FaceID and logged into their phone only by putting a pair of modified glasses on their face. The attack itself is severe, given the bad actor would need to figure out how to put the glasses on an unconscious victim without waking them up. Other types of attacks are usually based on generated fake data, such as pictures or videos of the victim [23].

In this sense, security systems based on physiological features such as ECG, PPG, and EEG have the advantage of assuring user's mobility with a higher certainty of an individual's liveness. Individually, ECG excels with other physiological traits in some aspects, and also provides universality, uniqueness, hidden nature, permanent, and simple acquisition [24]. In this sense, ECG has been used by many researchers as biometric identification, since it has features that are unique to an individual, such as, statistical, morphological, and wavelet features. Furthermore, it brings some health information and the applicability for continuous recognition systems [6]. Table 1.2 shows the benefits and drawbacks of biometric traits for authentication [25].

Furthermore, since Apple launched a novel feature of the Apple Watch (Apple Inc.) series 4 that enables consumers to record a rhythm strip and assist with self-diagnosis of Atrial Fibrillation (AF), an embedded ECG, several others manufactures announced this feature in upcoming devices. The benefits of having an ECG strapped to a wrist at all times are significant - for those with heart issues, being able to record an on-demand ECG can be handy. With this new feature in a wrist, people don't need a doctor to know about their heart rate, pulse rate, among others. They just need an ECG smartwatch. These smartwatches monitor every heartbeat of yours. So, you can analyze how fit you are inside every day at any time. Usually, this information shocked people to know, one in every four deaths is caused by heart disease each year. One person dies every 37 seconds in the United States from cardiovascular disease [26].

ECG records electrical activity generated by the heart in a non-invasive way. The advantages of ECG are the inherent liveness biometric, being possible to capture only in a living individual. This signal is less vulnerable to changes in illumination and pose, unlike fingerprint and face, for instance. It is hard to counterfeit because sophisticated human body functions generate the signals [1]. ECG signals are capable of providing the individual's clinical information, which brings such significant value for consumers

Trait	Benefits	Drawbacks
ECG	Universality, Hidden Nature, Simple Acquisition	Requires Contact Variability over time
EEG	Universality Hidden nature	Expensive equipment Vulnerability to noise Variability over time
Face	Easily measurable Affordable equipment	Easy circumvetion Depends on face visibility and lighting
Fingerprint	High Performance Permanent over time	Requires Contact
Gait	Easy to measure Affordable equipment	Low performance Variability over time
Iris	High Performance	Expensive equipment
Palmprint	High measurability Permanent over time	Requires Contact
PPG	Easy to acquire Hidden Nature Affordable Equipment	Low Performance Variability over time
Voice	Affordable Equipament	Low Performance

Table 2: Main benefits and drawbacks of biometric traits

interested in a health monitoring application. The extraction of appropriate features, as well as the classification procedures, are both crucial issues for ECG as a biometric, that's the reason we decided to investigate them [27].

An ECG signal consists of three main components associated with different heart activities: atrial depolarization (P wave), ventricular depolarization (QRS complex), and ventricular repolarization (T wave) [28]. The literature presents three feature types for ECG signal, namely, fiducial, non-fiducial, and hybrid. Precisely, fiducial features extract time-domain characteristics from the ECG waveform, which is computed as time intervals, amplitudes, angle and dynamics (R-R intervals) based on the trait points in the ECG signal. On the other hand, non-fiducial features apply a transformation function to the characteristics points. Finally, and hybrid features combine fiducial and non-fiducial methods. We consider fiducial approaches since they exclusively use as features the measurements of fiducial landmarks of the ECG signal in the time domain. These measurements vary widely throughout the state-of-the-art. We have observed that some decisions impact the accuracy of this authentication process, such as the method employed for feature extraction, the method of artificial intelligence involved in the classification, and where the identity database is stored (in the cloud or in a local system, for instance).

1.3 Objectives

This master's thesis introduces a particular feature selection, using only fiducial points related to amplitude and time that can be found directly from the signal acquired, without any complicated processing, since wearable devices have limited computational resources. We also investigate some of the most used machine learning algorithms for user identification, and we analyze the accuracy of these algorithms to classify people in continuous authentication and identification scenario. Evaluation results show the potential of the proposed solution, which has reached accuracy higher than 98%.

Thus, the objectives of this work include:

- Extraction of ECG features with simple manner to best fits the applicability for individual's identification
- Analysis of different machine learning algorithms to identify individuals based on ECG data in the time domain
- Propose a set of steps to improve the quality of data set resulting in improvements of accuracy and other metrics during authentication and identification process

1.4 Text Organization

The rest of this master's thesis is organized following the ordering described below:

- Chapter 2: Presents an overview of technologies related to the biometric system. Our biosignal of study, ECG, is better explained and how its characteristics can be used to identify a person. The overview of all stages of general biometric systems is detailed and some of the well-known Machine Learning Algorithms are shortly explained.
- Chapter 3: This chapter brings some related works of biometric systems using ECG. Authors have performed many different approaches to this signal and also tried to validate it using various machine learning algorithms. The evolution in this area is demonstrated and even open issues.
- Chapter 4: Presents our proposal for work. Some data sets, the pre-processing, segmentation and other necessary steps are presented. After the proposal, some simulations are used to investigate the results of Drive DB and Challenge DB.
- Chapter 5: Concludes the current mater's thesis, suggests the expected future works and presents the published articles associated with this research.

CHAPTER 2

Theoretical Reference

This Chapter presents some of the main concepts about healthcare systems focusing on the ECG signal. It is our candidate to be used as a biometric for user authentication in the following years since it has such an aggregate value that other signal did not have yet. The ECG carries health information at the same time as unique characteristics as a biomarker. Section 2.1 describes how Biometric signals can be used for user authentication. Section 2.2 introduces the main components of the most important Machine Learning techniques are also discussed, as well as its role in biometric System architecture.

2.1 Biometric System for User Authentication

Offering efficient healthcare systems is one of the most critical social and economic challenges nowadays. Healthcare administrators, clinicians, researchers, and other practitioners are encountering increasing pressure generated by the growing expectations from both the public and private sectors [29, 30]. In this context, wearable sensors, such as smartwatches, fitness tracker, and smart glasses, lead healthcare applications to a new era by allowing non-invasive diagnosis of vital and non-vital functions of the human body [16, 17, 12, 18]. Wearable devices can sense, collect, and transmit physiological data continuously, where they are equipped to radio communication with a gateway that has a connection to the Internet. Examples of data are body temperature, ECG, EEG, PPG, pulse rate, respiration rate (percentage of breathing), and blood pressure, which provide opportunities to improve the efficiency of healthcare systems and the quality of life [8].

Wearable devices provide a lower cost for the treatment, greater mobility for the patient and, potentially, improved physiological data for the physicians to analyze when attempting to diagnose the condition [11]. Ideally, these systems should be able

to carefully, conveniently and robustly monitor patients, while they perform their daily activities (such as, eating, sleeping, walking, and naturally communicating with parents) without interfering significantly with their comfort [14]. Therefore, over the last few years, wearable devices appeared in the market offering different functionalities and wearing options, and at the same time, a significant volume of scientific effort has been done highlighting various research challenges related to wearable devices.

Wearable devices collect patient's biological signs and send the collected data across the Internet to a healthcare system (for post-process), as shown in Figure 1. In this scenario, a method to detect legitimate traffic is essential for security purposes, for both the patient and the hospital. Hence, biometrics plays a crucial role in security by providing individual characteristics as a manner of access control and identification of humans by their characteristics or trait, such as provided by fingerprints, retina/iris, and facial features [25]. In this context, the wearable device on the forearm acquires the biometrics signal, processes it to extract the target feature, and uses it for both identification and authentication. In a nutshell, the authentication process starts with wearable devices, *e.g.*, on the forearm, acquiring the biometrics signal; next, they process this signal to extract a target feature; and then they use it for both user identification and authentication.

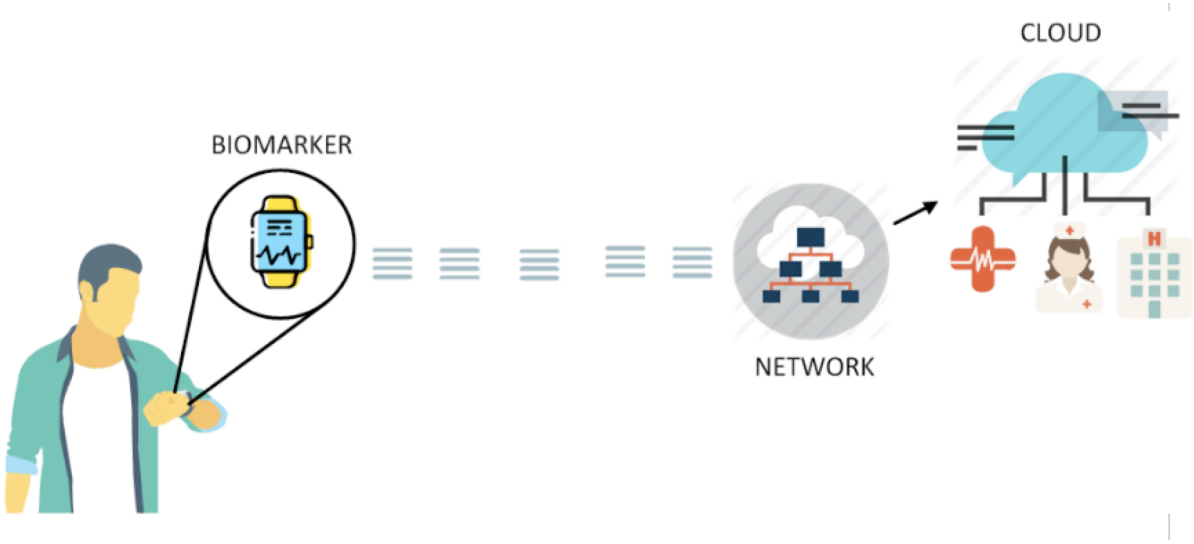

Figure 1: Wearable device collecting patient's biological signs and sending the collected data across the Internet to a healthcare system (for post-process)
Source: Author

According to Lumini *et al.* [31], biometrics refers to technologies used to measure human physical or behavioral characteristics, which are sub-divided into six classes, namely, hand region, facial region, ocular region, medico-chemical, behavioral, and soft [32], as shown in Figure 2. It provides not only an alternative to IDs or PINs (knowledge-based schemes) to authenticate someone into a system, also a continuous authentication method. We are looking for biomarkers that can measure medico-chemical signs, *e.g.*, ECGs, EEGs, EMG, and respiration rate.

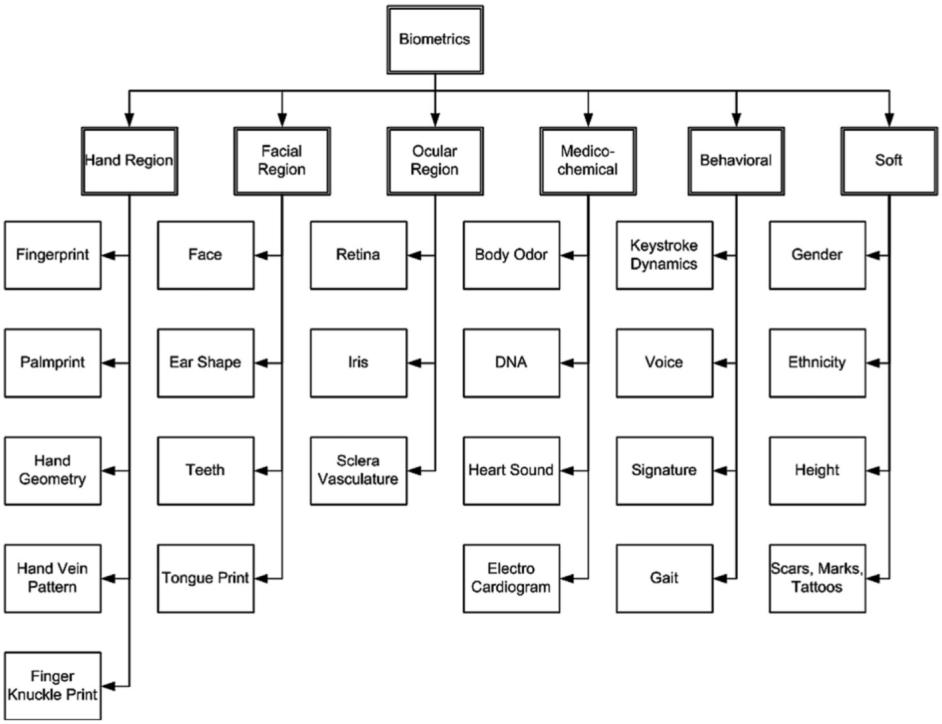

Figure 2: Biometric classes
Source: Unar *et al.* [32]

Many researchers define a typical biometric recognition system with at least two stages, namely: enrollment and recognition/verification stages [32, 33, 34]. In the enrollment stage, the biometric system acquires a live sample of the biometric trait of an individual, extracts a salient or discriminatory feature set from it, and stores the extracted feature set in a database (often referred to as a template), along with an identifier associating the feature set with an individual. During the recognition or verification stage, the system once again acquires the biometric trait of an individual, extracts a feature set from it, and compares this feature set against the templates in the database to determine a match or to verify a claimed identity.

There is a conceptual difference between identification and verification that should be clarified. Biometric identification is a One-to-Many matching of the captured biometric sample against all stored templates to determine a person's identity even without his/her knowledge or consent [31]. On the other hand, biometric verification considers a One-to-One matching of the captured biometric sample against the template of the person he/she claims to be. The identity claim is accepted as genuine if the degree of similarity is sufficiently high or as impostor otherwise. Figure 3 shows the four main components of biometric systems for biometric verification: Acquisition, Feature Extraction, Matching, and Decision Modules.

Among existing medico-chemical signs, ECG is handy for biometric authentication, since it has features that are unique to an individual, such as, statistical, morphological, and wavelet features. ECG also provides universality, uniqueness, hidden nature,

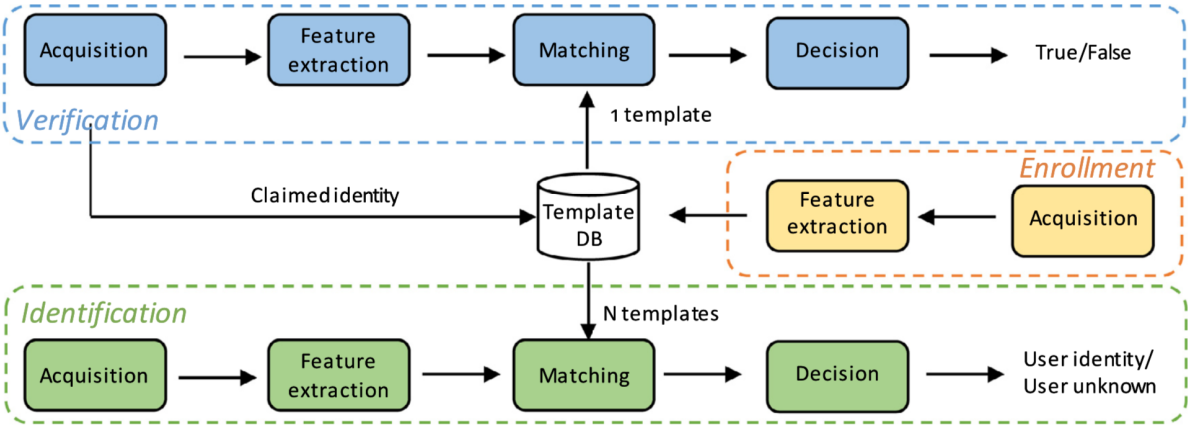

Figure 3: General flow of a biometric system
Source: Lumini *et al.* [31]

permanent, and simple acquisition [24, 25]. ECG records electrical activity generated by the heart in a non-invasive way. The ECG has a wide range of possible applications, such as, measuring the heart rate, examining the rhythm of heartbeats, diagnosing heart abnormalities, emotion recognition, and biometric identification [35]. Figure 4 depicts a typical cardiac cycle: at step 2, the P wave shows the status of activation of the atria. It is not a forceful muscular contraction. Thus, the P wave is small in size and its duration ranges from 60ms to 120 ms. Step 3 (PQ stretch) measures the time that elapses from the time when the atria begin to be activated until the moment in which the ventricles are activated. At step 4 there is the QRS complex, which is the most important in terms of characterization of the signal. The QRS complex represents the depolarizations of the septum, of the left ventricle apex (the long and narrow wave) and of the basal and rear regions of the left ventricle. The alterations of this complex suggest the presence of heart diseases, such as arrhythmia, fibrillation, and heart attack. At step 5, there is the ST stretch, which corresponds to the time interval where the ventricles contract and return to rest, and it is approximately in the baseline of the ECG signal. The typical duration of ST stretch ranges from 230 ms to 460 ms, and it can reveal ischemic problems. At the end of the cycle, there is the T wave, which represents the repolarization of the ventricles (the time when the ventricles have finished their activation stage, and they are ready for a new contraction). Its analysis can provide information on cardiac hypertrophy, heart failure, and ischemic heart disease.

As each heart has its characteristics such as size, position in the breast and age, the waveforms of the ECG are not equal. The waveforms of different patients have some similarity in the case of pathologies. It is the principle of ECG in use for so long in medicine. Based on the uniqueness of ECG signal for an individual, some works have been developed to identify a specific individual from a group of candidates by using a one-lead ECG such as Hoekema *et al.* [36], Steven *et al.* [37], and Shen *et al.* [38]. Thus, ECG can be an approach to identify users and security issues. ECG data is handled con-

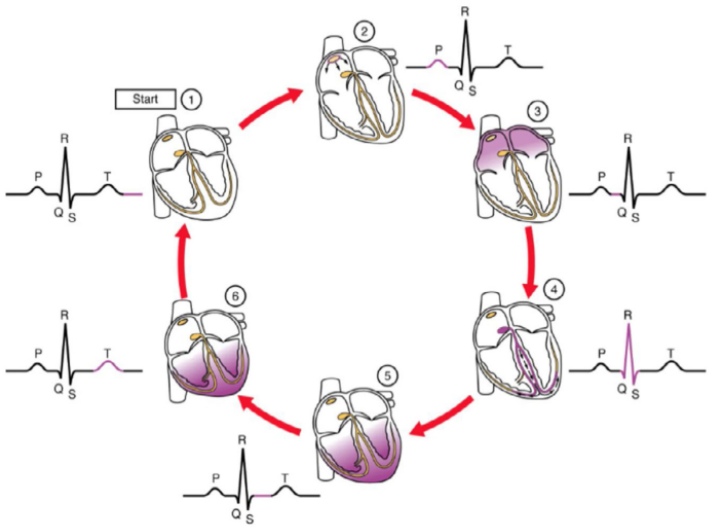

Figure 4: A sketch of a common cardiac cycle with the associated waves of an ECG signal (one-lead)

Source: Berkaya *et al.* [35]

sidering the following steps: preprocessing, feature extraction, feature selection, feature transformation, and classification, explained in the following.

2.1.1 Preprocessing of ECG signal

In the preprocessing step, the primary goals are to reduce the noise and artifacts to determine the fiducial points (P, Q, R, S, T), as well as to avoid amplitude and offset effects to compare signals from different patients. Typical types of noise are Power line interference, Baseline wanders, Electrode contact noise, Electrode motion artifacts, Muscle contractions (Electromyography noise), Electrosurgical noise, and Instrumentation noise.

In cardiovascular disease classification, noise causes a physician to make wrong assessments about patients and also reduces the diagnostic correctness. On the other hand, in biometric systems, noise reduces the accuracy and can provide a mismatching or even become the identification process impossible for anyone at all. Existing methods for automatic arrhythmia identification and biometric can enhance the whole system's quality.

2.1.2 Filtering

Filtering blocks are used in the preprocessing stage to deplete artifact signals from the ECG signal [35]. Many types of filtering can be used, such as lowpass, bandpass, highpass, and adaptive. Bandpass filtering is widely used to reduce muscle noise, baseline wander, power line interference, and low- and high-frequency noise components, as well as to limit ADC saturation and address anti-aliasing [35]. Rahman *et al.* [39] proposed adaptive filters based on the normalized signed regression LMS, and normalized sign–sign

algorithms. The adaptive filter ensures a non-distorted signal waveform during noise removal. They can be a powerful tool for the rejection of narrowband interference in a direct sequence spread spectrum receiver.

The choice of filter structure is an essential part of the system [40]. A nonlinear adaptive filter does not necessarily have a linear relationship between the input and output at any moment in time [68]. A nonlinear adaptive filter has more signal processing capabilities. However, non-linear adaptive filters require more complicated calculations. Thus, the actual usage is still most often the linear adaptive filter.

2.1.3 Re-sampling, digitalization and artifact removal

It is one of the most critical stages of new developing researches. First, there are not databases big enough to validate all the identification systems for ECG, as it exists for the fingerprint. Second, the ECG acquisition process has more sensitivity due to different equipment used. Thus, the re-sampling or down-sampling method can be used not only to preserve the consistency of databases to reduce the memory requirements and computational cost [41] but also to integrate different databases for the same application or research. This stage starts after filtering an ECG signal. The digitalization uses many frequencies *e.g.*, 125 Hz, 200 Hz, 250 Hz, 257 Hz, 360 Hz, and 1 kHz. Different types of Wavelet transform (WT) - based downsampling has been used for this purpose, and the Discrete WT (DWT) is supposed to be one of the popular approaches [35].

For denoising and baseline wandering removal, different types of WTs were usually applied to ECG signals. For example, an undecimated WT or Stationary WT (SWT) and DWT. The SWT denoising technique is an advanced signal processing approach used to it [42]. Normalization, applied in the preprocessing, avoids amplitude and offset effects, linear predictive models estimation, Pan and Tompkins algorithm, or Hilbert transform-based peak-finding.

2.1.4 Feature extraction and selection

Thus is the core of the biometric system, and the literature proposes various feature extraction techniques to expose the unique information from ECG signals for different purposes, such as analysis and classification. Those features can be used individually or in combination with other features. The most used are P-QR Morphological features S-T complex features for an ECG signal corresponding to the locations, durations, amplitudes, and shapes of particular waves or deflections inside the signal. Typically, an ECG signal has a total of five significant deflections, including P, Q, R, S, and T waves, plus a minor one, namely, the U wave, as shown in Figures 5 and 4.

Researchers consider various attributes of the QRS complex as the features such as R wave duration, P+ amplitude, QRS p-p amplitude, R wave amplitude, ST amplitude, T+ amplitude, QRS wave area, ST and so on [35]. Other features extracted from the raw

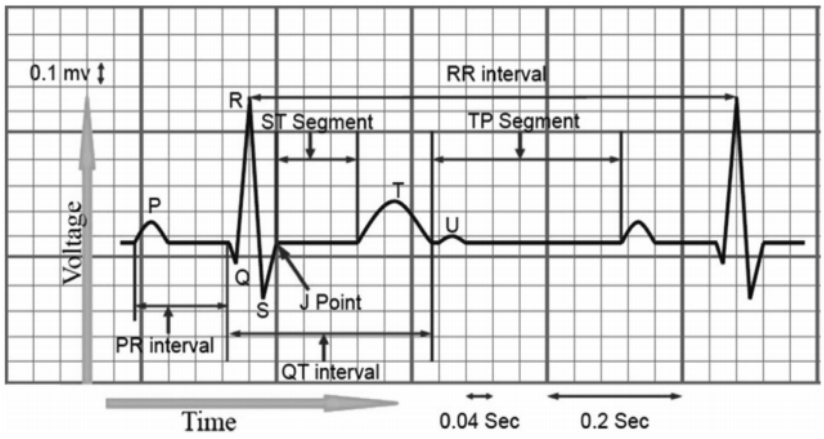

Figure 5: Standard fiducial points in the ECG (P, Q, R, S, T, and U) together with clinical features

Source: Berkaya *et al.* [35]

signal are the statistical ones. These features are widely used for ECG analysis, providing an effective way of analyzing the level of complexity and the type of distribution that any time-series exhibits. In the case of ECG recordings, these features would help to discriminate patient-specific and disease-specific variations, or both of them in such a way that achieves better classification performance. Researchers extract statistical attributes such as energy, mean, standard deviation, maximum, minimum, kurtosis, and skewness.

The frequency-based techniques are the most popular wavelet feature extraction techniques for representing ECG signals for classification purposes. It considers the fact of ECG signals to be intrinsically non-stationary. Hence, WTs are a useful tool for the analysis of ECG signals and frequency-based feature extraction, with its powerful time-frequency localization property [43]. The WT is a linear transform that provides the time-frequency representation of the signal. WTs use wavelets (*i.e.*, scaled and shifted versions of a mother wavelet) to decompose the signal into simpler elements. We mention some recent studies on ECG analysis that utilize wavelet-based features as [44, 45].

After the features extraction process, usually, because of computing restrictions, it is necessary to select only the most relevant features. Feature selection is a set of techniques to determine a subset of the relevant features for building robust learning models by removing the most irrelevant and redundant features[35]. The primary goal of feature selection is threefold: improving the performance of classification, providing a faster and more cost-effective learning process, and a better understanding of the underlying process that generates the data. Feature selection methods are grouped into three categories, called filter, wrapper, and embedded methods, as introduced by Berkaya *et al.* [35].

As described by Odinaka *et al.* [28], ECG consists of three main components associated with different heart activities: atrial depolarization (P wave), ventricular depolarization (QRS complex), and ventricular repolarization (T wave). Most of the problems associated with feature extraction methods relate initially to the correct detection of the ECG characteristic points, *i.e.*, P, Q, R, S, J, T, and U such as exemplified in Figure 5.

Odinaka *et al.* [28] refer that features are derived from these characteristic points. Thus, once the modeling correctly identifies the characteristics points, the method can calculate the features.

Fiducial features are calculated as time intervals, amplitudes, angle and dynamics (R-R intervals) based on the characteristics points in the ECG signal [28]. They considered the morphological aspects in the time domain with low computational complexity. Non-fiducial features consider a transformation method applied over segments of these relevant points. In this sense, the R peak is the primary point of ECG segmentation and alignment [28]. Hybrid feature extraction methods consider a combination of both.

Karpagachelvi *et al.* [46] provided a comprehensive survey of ECG feature extraction techniques. The authors evaluated 17 research articles about feature extraction techniques. In this context, Wavelet Transform (WT) and variations of WT such as orthogonal and bi-orthogonal wavelet, discrete wavelet, and quadratic wavelets, received attention from many works. Additionally, methods proposing innovative algorithms for P, QRS, T detection. and R-R interval detection was in evidence in Karpagachelvi *et al.* [46] survey. Statistical methods, matched filters, cepstrum coefficients, and chaos theory were less frequent.

2.1.5 Classification

Literature reports the application of various machine learning techniques to classify data extracted from ECG. The feature extraction and selection are related to the problem we want to solve. Individually, they are the descriptors that categorize the problem. To select the classifier, we must take into consideration both the problem we want to solve and the features that better describe it. Finally, the combination of classifier and feature selection will influence the classification accuracy.

In ECG scenario, machine learning algorithms, *e.g.*, Artificial Neural Networks (ANNs), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Bayesian classifiers, can be used to classify the signal acquired as a health or non-health individual. The received ECG signal fulfills different purposes, *e.g.*, training of a machine learning model, classification according to a trained model, template stored in a specific database [35], and also biometric system application.

2.2 Machine Learning Algorithms

Machine learning is a subset of Artificial Intelligence (AI) which computer systems use to effectively perform a specific task without using explicit instructions. Machine learning algorithms build a mathematical model of training data to make predictions (regressions) or decisions (classification/ pattern recognition) without being explicitly programmed.

Machine Learning systems can be classified according to the amount and type of supervision they get during training. The major categories are supervised learning, unsupervised learning, semisupervised learning, and Reinforcement Learning. In this mater thesis, to simplify we will not cover semisupervised and reinforcement learning. Basically, in supervised learning, the training set you feed to the algorithm includes the desired solutions, called labels as in Figure 6. The agent observes some example input–output pairs and learns a function that maps from input to output. In other words, data composed of examples of the desired answers. For instance, a model that identifies fraudulent credit card use would be trained from a dataset with labeled data points of known fraudulent and valid charges. Most machine learning is supervised.

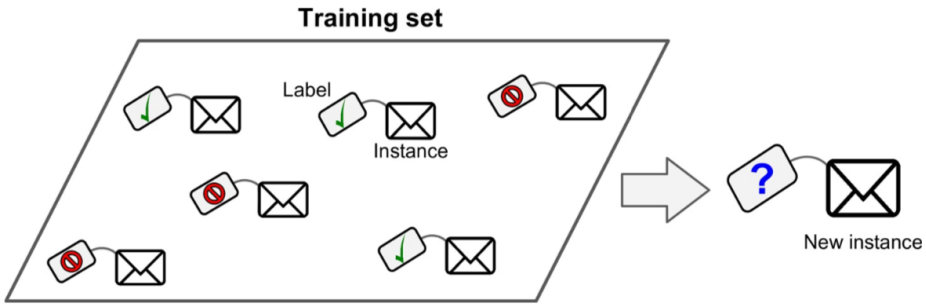

Figure 6: A labeled training set for spam classification (an example of supervised learning)
Source: Wei-Meng Lee [47]

A typical task is a classification. The traditional example is the spam filter that can be trained with many example emails along with their class (spam or ham), and it must learn how to classify new emails. Another typical example is to predict the price of a car or a house, given a set of features (mileage, age, brand, etc., size, location) called predictors. This sort of task is called regression. To train the system, you need to give it many examples of cars or houses, including both their predictors and their labels (*i.e.*, their prices).

On the other side, there is unsupervised learning, in which the system tries to learn without a teacher. Unsupervised learning algorithms are used on data with no labels, and the goal is to find relationships in the data. The agent learns patterns in the input even though no explicit feedback is supplied. Suppose the scenario where there is a database with customer's information of a national store. We may want to run a clustering algorithm to try to detect groups of similar customers. At no point do we tell the algorithm which group a customer belongs to it finds those connections without our help? For example, it might notice that 40% of our customers are males who love comic books and generally read our blog in the evening, while 20% are young sci-fi lovers who visit during the weekends. If we use a hierarchical clustering algorithm, it may also subdivide each group into smaller groups. This may help us target our posts for each group.

A related task is dimensionality reduction, in which the goal is to simplify the data without losing too much information. One way to do this is to merge several corre-

lated features into one. Yet another crucial unsupervised task is anomaly detection—for example, detecting unusual credit card transactions to prevent fraud, catching manufacturing defects, or automatically removing outliers from a dataset before feeding it to another learning algorithm. In this master's thesis, SVM, DT, RF, and Neural Networks were used in the classification stage. There is not a better choice, it depends on the data and the goal of the application. Further, it is an overview of each of them.

2.2.1 Support Vector Machine

The main idea behind SVM is to draw a line between two or more classes in the best possible manner like in Figure 7 [47]. A key term in SVM is support vectors, which are the points that lie on the two margins. Once the line is drawn to separate the classes, you can then use it to predict future data. For example, given the snout length and ear geometry of a new unknown animal, you can now use the dividing line as a classifier to predict if the animal is a dog or a cat. In the real-world, divide one or more classes using theses vectors are not that simple as in Figure 7.

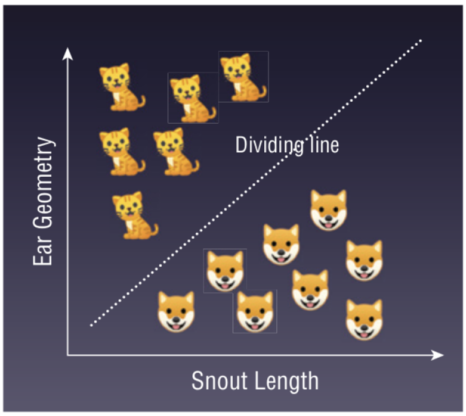

Figure 7: Using SVM to separate two classes of animals
Source: Wei-Meng Lee [47]

Even if the task would be to separate only the green form blue, where the line should be? For SVM, the correct line is the one that has the widest margins (with each margin touching at least a point in each class). In Figure 8, d1 and d2 are the widths of the margins. The goal is to find the largest possible width for the margin that can separate the two groups. Hence, in this case, d2 is the largest. Thus the line chosen is the one on the right.

Each of the two margins touches the closest point(s) to each group of points, and the center of the two margins is known as the hyperplane. In SVMs our optimization objective is to maximize the margin. The margin is defined as the distance between the separating hyperplane (decision boundary, the line separating the two groups of points) and the training examples that are closest to this hyperplane, which are the so-called support vectors. We use the term "hyperplane" instead of "line" because in SVM we

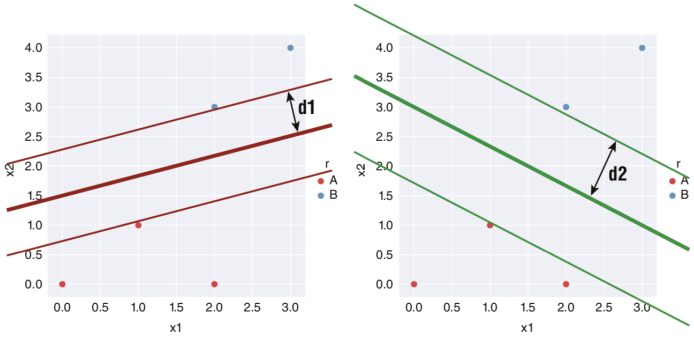

Figure 8: A plot considering the RR Interval and mean R's amplitude features of three subjects

typically deal with more than two dimensions, and using the word "hyperplane" more accurately conveys the idea of a plane in a multidimensional space. To find the best margin to fit the support vectors, a *Constrained Optimization* problem should be solved, usually is used the Lagrange Multipliers technique. It is beyond the scope of this dissertation to discuss how to find the margin and math behind the Lagrange operators.

2.2.2 Decision Tree

DTs are versatile Machine Learning algorithms that can perform both classification and regression tasks, and even multioutput tasks. A DT has become a popular method of classifying machine learning due to its versatility for applications in many problems, from object identification to medical diagnostics such as digital analysis, iris or even ECG. The concept is to use the tree structure to divide features into different classes based on probabilistic criteria and numerical limits. Features or attributes define the class. There are many algorithms for developing a DT, such as ID3 (Iterative Dichotomiser 3), C4.5 (alternative to ID3), Cart (Classification and Regression Tree) and Chaid(Chi-Squared Automatic Interaction Detector).

It is shown that the DT technique delivers better performance compared to the SVM technique for building the regression model. They are very powerful algorithms, capable of fitting complex datasets. Decision Trees are also the fundamental components of Random Forests, which are among the most powerful Machine Learning algorithms available. Indeed, we will use Random Forest in our biometric applications instead of simple decision trees. To better understanding how DT works, let's take a look at how it makes predictions in Figure 9.

Supposing you want to classify someone based on his ECG signal. DT starts at the root node (depth 0, at the top): this node asks whether the mean amplitude of R Peaks is smaller or greater than 0.578582 mV. After this, then you move down to the root's next child node (depth 1). If the person under evaluation has the feature mean R greater than 0.578582 mV, the right leaf node will check the value of mean amplitude S's peak, depending on the interval of this feature the DT can already end the classification or

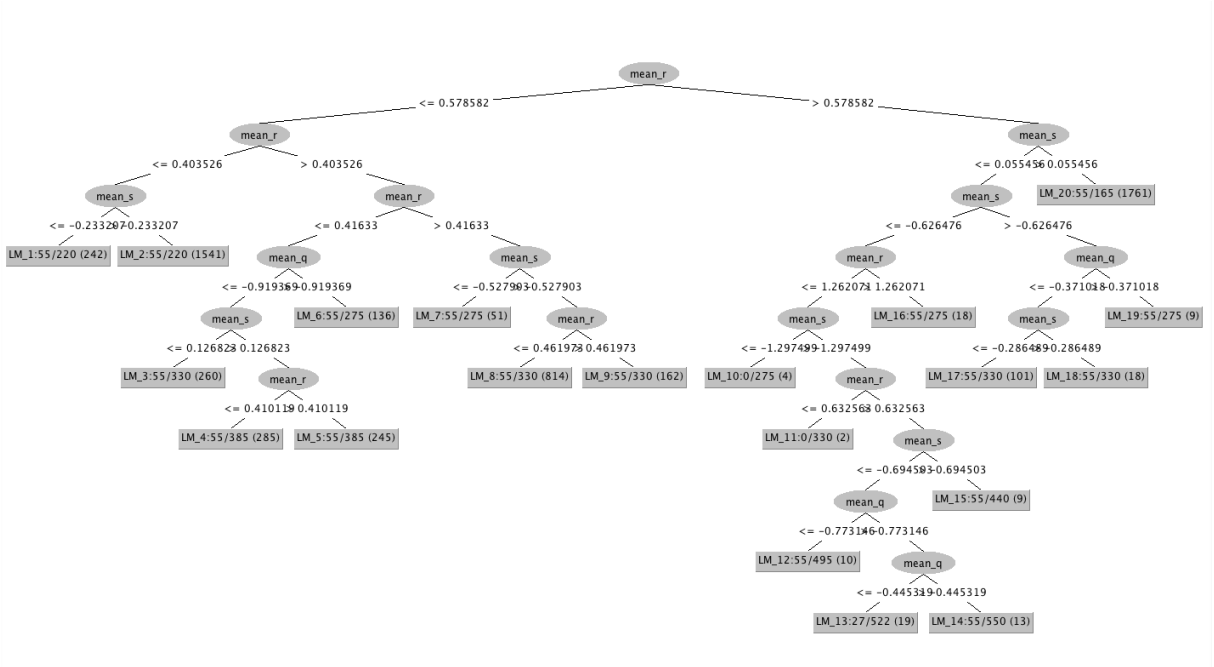

Figure 9: A decision tree used for classification of three persons based on ECG features
Source: Extracted from Weka during our simulations

continue. So, after the model was generated, you just have to take a look at the decision tree and see in what leaf your instance will finish. It's that simple.

DT has the advantage of being easy to implement and interpret when compared to other classification methods. However, depending on the chosen features and the way the data is divided into the tree, the model may lose its generalization due to overfitting. There are two common ways to deal with overfitting: limit the number of divisions or allow division only if there is a minimum number of data points in the tree branch.

An essential aspect of the DT configuration is how to define the importance of each feature to maximize the classification results. The algorithm takes the characteristic that best represents the class and places it at the root. There are some indexes proposed in the literature to identify the most relevant characteristics: the Gini coefficient and Entropy Index. the Gini coefficient is a measure of the importance of the variable to the data set and the entropy index is a measure of uncertainty associated with the variable. The Gini Index is calculated according to Eq. 2.1.

$$Gini(s) = 1 - \sum_{j} p_j^2 \tag{2.1}$$

considering,

- s the data set
- p_j is the ratio of class j instances among the training instances in the data set s

In addition to common decision tree approaches, some more specific decision tree structures are often used for ECG classification. A more complex approach is to use the random forest, where several decision trees are trained with subsets of data and will be explained below.

2.2.3 Random Forest

RF works aggregating the predictions of a group of predictors called an ensemble; thus, this technique is called Ensemble Learning, and an Ensemble Learning algorithm is called an Ensemble method [48]. The principle is based on the logic if aggregate the predictions of a group of predictors (such as classifiers or regressors), you will often get better predictions than with the best individual predictor. For example, you can train a group of DTR classifiers, each on a different random subset of the training set. To make predictions, you just obtain the predictions of all individual trees, then predict the class that gets the most votes. Such an ensemble of DT is called a Random Forest, and despite its simplicity, this is one of the most powerful Machine Learning algorithms available today [48].

This methodology ensures that each tree is slightly different from each other. Thus, each tree can return a different result for a data set. The voting system can calculate the direct or weighted vote. Specifically, direct voting counts how many trees have classified a given feature under a specific class. Weighted voting returns the proportion of elements belonging to a particular class.

RF performs better than DT in two critical aspects: anomaly detection and overfitting. Due to the training process, outliers will be present in some of the trees, but not in all of them. Thus, the voting system ensures that anomalous results are isolated. The voting system also minimizes the effect of overfitting in relation to the individual DT. However, both RF and DT have problems extrapolating data. Specifically, the attribute values in the validation set must be within the training set value limits. Untrained or out-of-bounds attributes lead to unpredictable results when included in the validation set.

The RF algorithm accepts that the number of trees grows as a configurable parameter. There is no better value and the limit must be the storage capacity to save the DT. However, a larger number of decision trees do not necessarily reflect in the classification results. One approach is to start with a few trees and gradually increase their number until the benefits are not worth the increases. The Figure 10 describes the structure of an RF.

2.2.4 Artificial Neural Networks

ANN is a mathematical model that is inspired by biological neural networks, which aims to solve both linear classification and non-linear classification problems with various network structures and learning algorithms. They were first introduced back in

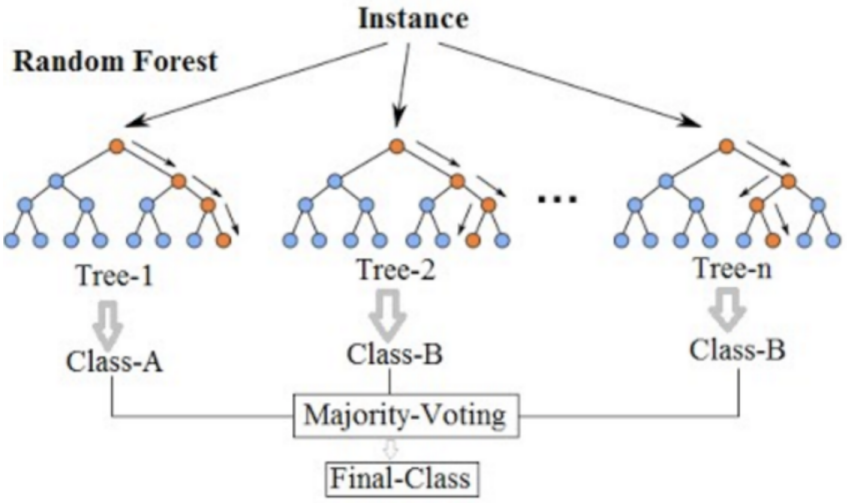

Figure 10: Random Forest structure

1943 by the neurophysiologist Warren McCulloch and the mathematician Walter Pitts. In their paper, "A Logical Calculus of Ideas Immanent in Nervous Activity," the authors presented a simplified computational model of how biological neurons might work together in animal brains to perform complex computations using propositional logic [48]. The classifiers have its properties and capabilities depending on the data structure (feature) used as input. Even if the classifier performs within the expectations, the ECG issues remain present. Heart rate variability due to mental, emotional, and physical changes, issues relating to sensor placement, scalability to larger populations, and the time-varying nature of the ECG signal.

ANNs are the base of Deep Learning. They are versatile, powerful, and scalable, and have been used in highly complex Machine Learning tasks, such as classifying billions of images (*e.g.*, Google Images), powering speech recognition services (*e.g.*, Apple's Siri), or recommending the best videos to watch to hundreds of millions of users every day (*e.g.*, YouTube) [48]. They have good success in the 1960s, however, the promises of truly intelligent machines didn't become true and the ANNs were put aside. They had a new revival in the 1980s and 1990s but other techniques were invented and seemed to offer better results and stronger theoretical foundations.

Nowadays, there is a huge quantity of data available to train neural networks, and They frequently outperform other ML techniques on very large and complex problems. The increase in computing power since the 1990s now makes it possible to train large neural networks in a reasonable amount of time. ANNs seem to have entered a virtuous circle of funding and progress. Amazing products based on ANNs regularly make the headline news, which pulls more and more attention and funding toward them, resulting in more and more progress, and even more amazing products [48].

The Perceptron is one of the simplest ANN architectures, invented in 1957 by Frank Rosenblatt. It is based on a slightly different artificial neuron(figure 11 called a

Threshold Logic Unit (TLU), or sometimes a Linear Threshold Unit (LTU): the inputs and output are now numbers (instead of binary on/off values) and each input connection is associated with a weight. The TLU computes a weighted sum of its inputs (z = w1 x1 + w2 x2 + + wn xn = xT w), then applies a step function to that sum and outputs the result: hw(x) = step(z), where z = xT w.

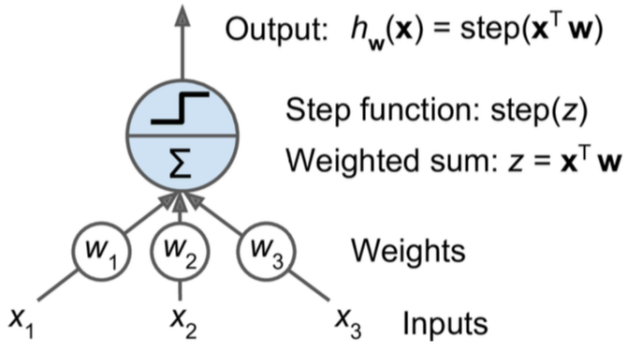

Figure 11: Threshold logic unit
Source: Aurélien Géron [48]

A Perceptron is simply composed of a single layer of TLUs, with each TLU connected to all the inputs. When all the neurons in a layer are connected to every neuron in the previous layer (*i.e.*, its input neurons), it is called a fully connected layer or a dense layer. To represent the fact that each input is sent to every TLU, it is common to draw special passthrough neurons called input neurons: they just output whatever input they are fed. All the input neurons from the input layer. Moreover, an extra bias feature is generally added (x0 = 1): it is typically represented using a special type of neuron called a bias neuron, which just outputs 1 all the time. A Perceptron with two inputs and three outputs is represented in Figure 12. This Perceptron can classify instances simultaneously into three different binary classes, which makes it a multi-output classifier.

2.2.5 Multilayer perceptron

An MLP is composed of one (passthrough) input layer, one or more layers of TLUs called hidden layers, and one final layer of TLUs called the output layer (see Figure 12. The layers close to the input layer are usually called the lower layers, and the ones close to the outputs are usually called the upper layers. Every layer except the output layer includes a bias neuron and is fully connected to the next layer. When an ANN contains a deep stack of hidden layers, it is called a Deep Neural Network (DNNDeep Neural Network).

The field of Deep Learning studies DNNs, and more generally models containing deep stacks of computations. Even so, many people talk about Deep Learning whenever neural networks are involved (even shallow ones). For many years researchers struggled to find a way to train MLPs, without success. But in 1986, David Rumelhart, Geoffrey Hinton, and Ronald Williams published a groundbreaking paper that introduced the

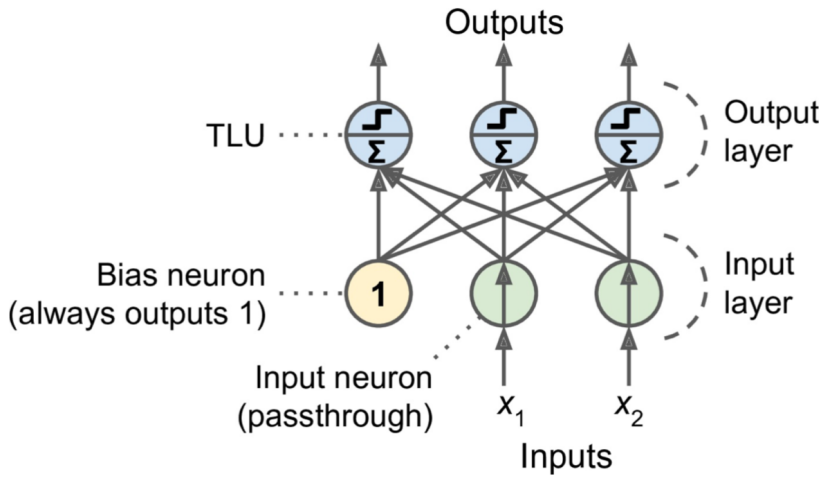

Figure 12: The Architecture of a Perceptron with two input neurons, one bias neuron, and three output neurons

Source: Aurélien Géron [48]

backpropagation training algorithm, which is still used today. In short, it is Gradient Descent using an efficient technique for computing the gradients automatically: in just two passes through the network (one forward, one backward), the backpropagation algorithm is able to compute the gradient of the network's error with regard to every single model parameter. In other words, it can find out how each connection weight and each bias term should be tweaked in order to reduce the error. Once it has these gradients, it just performs a regular Gradient Descent step, and the whole process is repeated until the network converges to the solution.

2.3 Chapter Conclusions

This Chapter provided good insight into the theme of the ECG signal and its use in biometric. Also, a brief review of some foundations of Machine Learning Techniques was provided. All these topics presented in this chapter serves as a basis for a better understanding of this dissertation. Now it is time to move towards the related work from the literature and understand how machine learning is used with biometric systems to detect patterns and provide more security to applications.

CHAPTER 3

Related Works

In this chapter, it is presented the main related works surrounding authentication systems using biometric signs. Also, after presenting the main types of machine learning techniques and the explanation of the electrocardiogram signal in Chapter 2, it is appropriate to explain in more detail the several different variations of machine learning configuration for biometric systems.

Specifically, passwords and code lockers can be compromised while biometric-based systems require a unique characteristic to identify the individual. In this sense, security systems based on physiological features such as ECG, PPG, and EMG obtained from wearable devices have the advantage of assuring user's mobility with a higher certainty of an individual's liveness. However, wearable devices require regular calibration and the right positioning to work properly. The combination of biometric and physiological characteristics can provide a better authentication method in a two-step strategy. In this context, biometric features such as the face, fingerprints, or iris recognition, can be used for initial authentication, while physiological features such as ECG, PPG, or EMG can be used to enhance the authentication accuracy and keep the section continuously authenticated. However biometric seems to be something new, the first scientific paper on automated fingerprint matching was published by Mitchell Trauring in the journal Nature in 1963 [33]. Let's take a look at some works in this context.

3.1 Automatic Human Identification using ECG

The paper *"ECG Analysis: A New Approach in Human Identification"* published in 2001 by Biel *et al.* [1], to the best of my knowledge, was one of the first to explore ECG as a biomaker. The author used soft independent modeling of class analogy (SIMCAsoft independent modeling of class analogy) as a classifier and special equipment used for the

measurements, one SIEMENS Megacart. The information from the SIEMENS ECG was transferred and converted to a usable format. As all measurements were done with 12-lead rest ECG recordings which are composed of six limb leads and six chest leads, and from the digitalized signal, 30 features could be generated by the equipment, the authors had 360 (30 x 12 leads) features available for each person to be identified. The results from the conversion will be 360 (30 x 12 leads) features for each person to be identified. All these features are normally used for help with clinical diagnoses. ECG measurements were done on 20 persons. The ages of the persons vary between 20 to 55 years old. Both women and men have been participating in the experiment. For a minor group, ten ECG measurements have been done. For the rest of the group, four or five ECG measurements were done. All measurements are done within a time period of six weeks.

The classification was made, in the third step, using SIMCA. The first step in modeling was to build a principal component analysis (PCA) model for each class. PCA is a mathematical transform that is used to explain variance in experimental data. In order to use classifiers for the identification of a particular subject, the model has to be *"taught"* the specific features of the subject. This is accomplished by presenting a set of measurements for each subject to the classifier. Based on these data sets, a statistical model for each person is constructed. Thus a group of statistical models is built and stored. Once this training process is done, the SIMCA classifier can be presented new ECG data from one randomly selected person belonging to the original group of subjects. The classifier then picks the person providing the best match in the group as output.

Some considerations, the authors calculated a correlation matrix that showed a strong correlation between the different leads for a specific feature. So, they decided not to use all the 12 leads, choosing the limb leads because it is easier to attach the corresponding electrodes compared to the chest leads. Usage of just limb leads will reduce the number of features to 180. They executed more features reductions ending with only 7 features and getting 45 out of 50 correctly classification, but the best results were using 10 features, achieving 100% of accuracy. The firsts lesson learned was: It is not necessary to use all 12 leads of ECG, It is possible to identify someone using fewer features (around 10) and the electronic must evolve because ECG couldn't be used as biometric if we still using SIEMENS Megacart as acquisition system(Figure 13 shows the equipment used in back 2001).

Figure 13: A Siemens Megacart ECG Equipament used in the experiment [1]

More recently, other authors such as Camara *et al.* [2] published some works related to our scenario, they considered a scenario of an air traffic control tower. They focused on the continuous availability of biosignals, that can performing an advanced form of authentication, called Continuous Authentication (CA). This variant is different from Non-Continuous Authentication (NCA). In NCA, the user is authenticated once at time T, for example when s/he is logged in a system with authentication checking. On the contrary, in a CA setting the user is authenticated every period time Ti, thus ensuring the continued presence of the user. The benefit of biosignal-based CA approaches is that users cannot transfer their privileges to other parties. Despite this benefit, one drawback is that biosignals evolve over time and maybe slightly different from time to time. As a consequence, the authentication mechanism should be continuously enhanced and not static as time goes by.

In that work[2], the cardiac signal was acquired by an Implantable Medical Devices (IMDs) (e.g., a pacemaker or an implantable cardioverter-defibrillator), or perhaps by external sensors attached to the body of the individual, the scenario is described in Figure 14. Once ECG signals are recorded, they need to be preprocessed before feature extraction. To do this the ECG signal is split into time windows and, for each window, a set of numerical features is extracted. Then, the similarity module filters samples discarding those that do not seem to come from the user. Finally, the samples are classified using a classifier such as a decision tree, a support vector machine, the nearest neighbor algorithm, and so on. They used some transforms, avoiding use the time-domain approach to get features of the signal. All samples passed into similarity module which discards bad ECG samples, that is, samples that were considered to be too far from the reference model (outliers). In our opinion this could be efficient but have to be used carefully, It might introduce some bias to the system. The experiments were performed using the recordings of 10 individuals from the MIT-BIH Normal Sinus Rhythm Database [49]. The individuals chosen didn't show any relevant medical problems and were observed during a long time period. The best results were the accuracy of over 93.5%, achieving 97.4% and 97.9% depending ON some parameters.

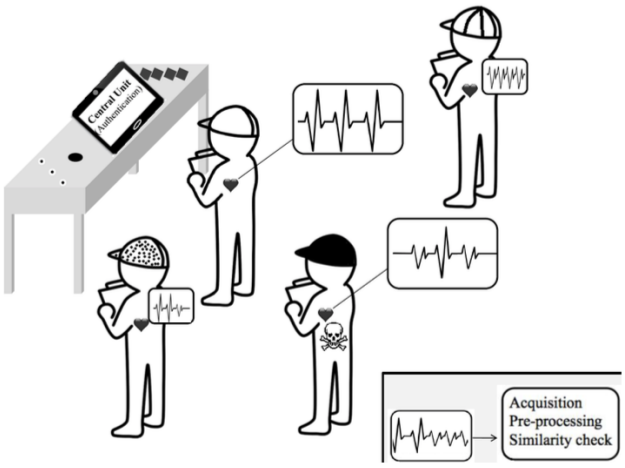

Figure 14: Example of an scenario for continuous ECG-based stream authentication. [2]

The next work, published by Zhang and Wu [50] considered an authentication scenario using ECG collected from two fingers electrodes in association with a smartphone application. They selected 85 subject's ECG records from widely used public databases on physionet [51]. 50 subjects from PTBDB, 19 subjects form MITDB and 16 subjects from NSRDB. Most of the subjects selected were healthy. The system application considers a pre-processing filtering stage, a fiducial detection with a feature extraction stage, and a classification stage. The filtering eliminates the noise associated with power-line, interference, muscle movement, and high-frequency noise. The fiducial detection module works as a pre-stage to feature extraction. The authors considered non-fiducial feature extraction computing complex, and thus decided to implement the model using waveforms divided by fiducials (WF). The classification stage refers to two biometrics methods: authentication and identification. During authentication, the system compares the user's biometry against a stored template computing the Euclidian distance as a comparison metric. For identification, the system fulfills a classification using Support Vector Machine (SVM) and Neural Networks (NN). For both cases, a voting mechanism required more than half of the voters to validate the testing subject. Compared with other related works, the research from Zhang and Wu [50] achieved 97.55% of accuracy and performed authentication in 4 seconds.

Although the fiducial approach is more popular, there are some papers [52] that proposed a system combining fiducial and non-fiducial features (hybrid approach) to enhance accuracy when authenticating a large number of users. The authors considered the PQRST peaks as the main fiducial features for the purpose of authenticating individuals. Specifically, the segments PQ, QR, RS, and ST duration, PQ, PT and SQ amplitudes determined by wavelet transform. For non-fiducial based features, the authors defined the ECG signal as a matrix (X), obtained the Gramian matrix multiplying $X^T X$, and finally obtained the features from the eigenvalues and eigenvectors from the Gramian matrix. The investigation considered the second set of non-fiducial features as the spectrum of the ECG signal generated by the Fast Fourier Transform. The ECG pattern recognition considered incremental training of a Linear Discriminant Analysis (LDA) algorithm, using the exposed fiducial and non-fiducial features. Zhang *et al.* [52] considered the dataset MIT-BH which included 100 samples containing 200 ECG signal files per sample file, *i.e.*, 20.000 ECG signals. The research concluded that identification based on fiducial features and Fast Fourier Transform presented low accuracy, in the order of 70% - 75%. On the other hand, the implementation of a hybrid approach achieved 99%. Finally, the authors expressed their concern that increasing the number of features will lead to increased computational effort as mentioned in [1], and the proposed scheme improves the efficiency.

Recently, the Convolutional neural network (CNNConvolutional neural network) has gained a lot of space, being one of the major deep learning algorithms, is now gaining tremendous attention leveraging its power in automatically learning the intrinsic patterns from the data. In [53], inspired by the observation that the ECG stream can be seen as a 1D-image, they propose a novel wavelet domain multi-resolution convolutional neural network approach (MCNN) for ECG biometric identification. Avoiding heavy feature

engineering efforts, and also let the CNN capture more hidden patterns from data and learn a high-level abstraction. In that study, eight diverse datasets were considered which included not only different electrode placement methods (chest and wrist) but also various heart health conditions (with and without cardiac abnormalities), the total of subjects was 220 people. The ECG stream was firstly blindly split up into signal segments with an equal length of two seconds without leveraging any heartbeat location information, which is not only immune to diverse morphological/beat-to-beat interval variability but also tolerant to signal artifacts that are usually major challenges in non-blind segmentation approaches.

The authors applied the wavelet transform (WT) approach to each 2-second window to obtain a multi-resolution time-frequency representation for the raw ECG signal. To explore the topology of parallel 1D-CNN, They designed a CNN named as CNN1-8, the former four correspondings to CNN with one to four stages without dropout regularization, and the last four with dropout operation (25% dropout). The result was 96.5% identification rate for normal datasets, 90.5% for abnormal, and 93.5% for all datasets. This technique of getting random windows from ECG segments is also used in our work to increase data representation when we are working with small number of samples.

Furthermore, in [3] a similar use of CNN as [53] was performed. The contribution was to introduce a method to binarize ECG templates that permits to reduce the matching time and apply template protection techniques. Only a few studies have considered the application of deep learning strategies for ECG analysis, but they focus on the classification of heartbeats in healthy and non-healthy using techniques such as CNNs, autoencoders, or deep belief networks. Despite the small number of studies using CNN, the work in [3] wasn't the first they believe. Their approach extracts a set of m QRS complexes from ECG samples of short duration and joins them in the signal V. In closed-set identification, a CNN processes V and indicates who is the closest registered user. In identity verification and periodic re-authentication, the CNN processes V to obtain a biometric template T. A simple distance measures, such as Euclidean or cosine distances is used to compute the matching score.

In my opinion, CNN has its power well known in computer vision application and perhaps, could show its value in biometric as well, but the architecture still seems to be very complex to be processed in wearable devices with computer power constraints. In [3], the architecture of CNN was composed of six convolution layers that use ReLu (Rectified Linear Units) neurons, three max-pooling layers, three LRN (Local Response Normalization) layers, one dropout layer, a fully connected layer, and a Soft-max layer (for training and closed-set identification). Figure 15 shows the schema of a deep learning approach.

They considered two public datasets to extract training and test sets, namely, E-HOL-03-0202-003 (Intercity Digital Electrocardiogram Alliance—IDEAL) database and PTB Diagnostic ECG Database. Some data from the first data set were discarded due to being corrupted by noise and artifacts, from the second they only considered the 52 healthy volunteers. They achieved 100% for 52 healthy volunteers, but mainly, they prefer

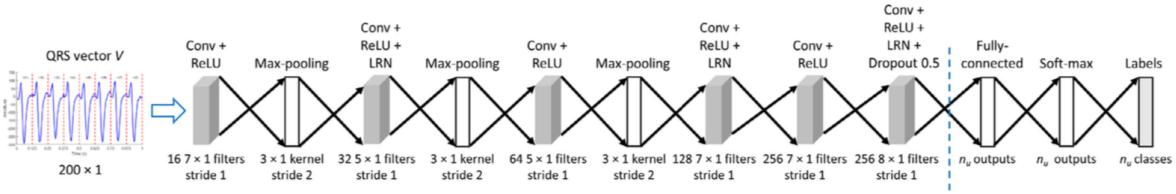

Figure 15: Schema of the proposed deep learning approach from [3] based on CNN.

to evaluate in terms of the Receiver Operating Characteristic (ROCReceiver Operating Characteristic) curve and the Equal Error Rate (EEREqual Error Rate).

Last, but not least, I decided specifically comment the work of Zhang *et al.* [4] because It brings the use of pre-trained networks to our horizon, they are basically, a neural network already trained to perform a specific task (e.g. classification) on specific a data set (e.g. a set of images) and presents recognized good results. They use the beneficial characteristics of the Google Inception Net [54] and residual neural network (ResNet)[55] to their proposed architecture, Figure 16 demonstrated the system diagram. Also, It is not required any reference point detection or time-consuming handcrafted feature engineering efforts. This work differs from [53] because It doesn't take into consideration the domain knowledge, this stage is replaced by a deep CNN layer.

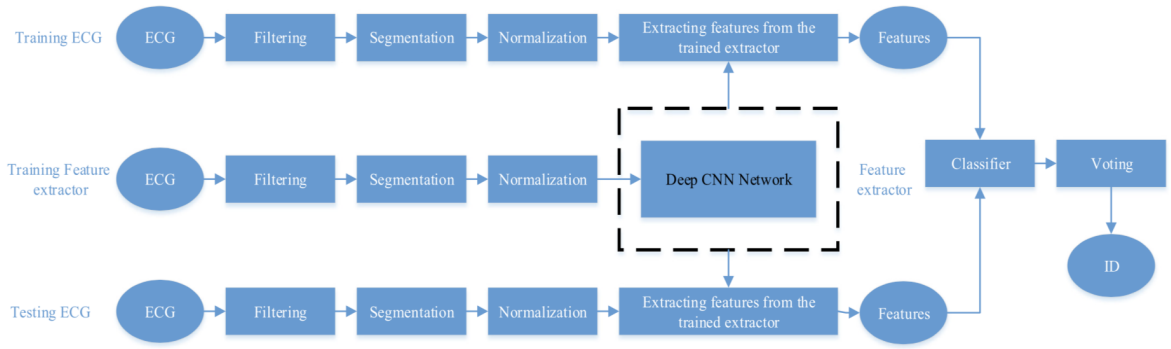

Figure 16: The system diagram of the proposed architecture proposed in [4]. Source: Zhang et al [4]

The deep CNN network was used as the deep feature extractor, being able to extract distinctive deep features. Then, the nearest neighbor classifier (NNC) and linear support vector machine (SVM) classifier are applied to evaluate the generalization ability of the trained deep feature extractor. Four public datasets were tested including PTB (PTB Diagnostic ECG Database), CEBSDB (Combined measurement of ECG, Breathing and Seismocardiograms), NSRDB (MIT-BIH Normal Sinus Rhythm Database), and MITDB (MIT-BIH Arrhythmia Database). Their results were 97.7% (NNC), 98.7% (SVM)) on the three datasets. This work, unlike most approaches do not require any reference point detection. However, in my opinion, the feature dimension of 400 still a great

number, especially to be used in mobile and wearable applications. Another problem, it takes much more time to train due to complex architecture as demonstrated in Figure 16.

This chapter of related works doesn't intend to be a survey of all biometric works published recently. As this master's thesis was developed under a research project called HealthSense, in the beginning, some surveys were read and we would like to mention the work of Odinaka *et al.* [28] to talk about fiducial or non-fiducial choice. They surveyed fifty studies dedicated to human identification, where 66% of the surveyed articles employed non-fiducial features, 26% applied fiducial features, and 8% of the research works used the hybrid. Regarding the classification method, 44% of the research work selected k-Nearest Neighbor (kNN) or Nearest Center (NC) algorithms, 16% implemented Neural Networks (NN), and 16% used Linear Discriminant Analysis (LDC). Finally, 12% of investigations achieved accuracy higher than 99% and 20% of the surveyed articles achieved 100%. This work gave us some directions of what look for, but we believe that some statistics are out of date as I mentioned some new papers using CNN, for instance. One of our future works is to produce an updated survey in this context.

3.2 Chapter Conclusions

After taking a closer look at some papers related to biometric systems using ECG in this chapter It is possible to figure out that there are many challenges to be overcome, especially considering our focus of application: wearable devices. Most of the existing works don't consider the constraints of power processing, energy or memory available in small devices. Most of them faced the problem of data set with a small number of subjects available, that is one point that we expect to be overcome soon because of the increasing number of new devices with ECG Sensor as mentioned in Chapter 1. Probably, a lot of public data set will be published in the next years, allowing the research in this topic advances faster.

Then, to conclude this chapter, we summarize some important points analyzed previously from related works, considering the characteristics that were also important for my work. Table 3.2 gathering the data.

Paper	**# of subjects**	**# of features**	**Acquisiton method or Database**	**Classifier**	**Metrics**	**Observation**
[1]	20	10 to 360	Siemens Megacart	PCA built in SIMCA	100%	Fiducial Features and Feature Reduction
[2]	10	256	IMDs	DT, SVM, NN	93.5% 97.9%	Non-fiducial
[50]	85	9	PTB, MITDB, NSRDB	SVM, NN, LIBSVM	96.6% 97.7%	Used voting mecanism
[52]	100	10 + Feature space	MIT-BH	PCA + LDA	99%	Used hybrid approach
[53]	220	Wavelet Transform	CEBSDB, WECG, NSRDB, MITDB, VFDB, AFDB, STDB, FANTASIA	1D-CNN	96.5% 90.5% 93.5%	Used 1D-CNN
[3]	237	V vectors	E-HOL-03-0202-003 and PTB	1D-CNN	100%acc EER=3.81% EER=3.37%	Prefered use EER and ROC
[4]	319	400	PTB, CEBSDB, NSRDB, MITDB	SVM, NNC	96.5%-99.1% 95.4%-(7.8%	Used Inception and ResNet

Table 3: Summary of Selected Related Works.

CHAPTER 4

ECG-Based authentication and identification

Among existing biometrics, ECG is very useful for user identification, and thus it is required to identify unique characteristics of ECG. As most of the identification systems, our proposition has five main steps: acquisition, signal preprocessing, feature extraction, matching, and classification. Generally, vital signals are captured by a sensor and preprocessed to remove noise. In our scenario, the ECG signal might be captured mainly through wearable devices, but It could be captured by devices designed exclusively for work as a biometric acquisition sensor. Anyway, after the signal was captured, It must be divide into segments. In this stage, we start our discussion. How the segment should be divided? How long each segment must be? Should It be divided considering only the duration of the presence of peaks/features? Since the size and the approach to get the segment was defined. How many segments are necessary to generate a reliable template for someone?

Even without answering these questions, peak detection is one of the most important due to the relevance of R's Peak between other features. Then, after segmentation, the feature extraction step is applied. The resulting features are processed to form a template that is compared to the authorized user template. Finally, classification is applied to distinguish genuine and imposer vital signal data. Figure 17 could be used to have an idea of each step included in our approach for the biometric system.

In this chapter, the proposal is described, starting with the description of some data sets investigated. One of the big challenges to investigate new biometric systems is the design of a reasonable data set. The physionet database [51] was the main source of inspiration and five sets were used. They differ in number of subjects, healthy or not, type of acquisition sensor used, inside or outside the hospital environment, and so on. Then, our results are presented base on two data sets: Driver DB and Challenge 2018.

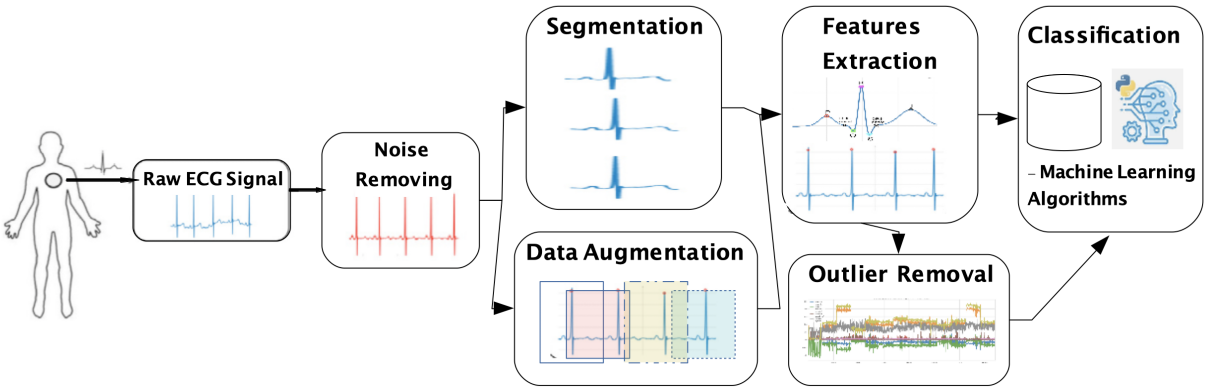

Figure 17: Overview of all processing steps in our proposal (Adapted figure from [5]).

4.1 Proposal

In the context of data analysis, sometimes the biggest challenge isn't to develop a model or choose the classifier, the challenge is to find a good data set, closer to real-world situations. During this master's thesis, many data set were evaluated and some were tested in the biometric scenario. In the following paragraphs, a brief description of some important data set is made.

The first data set used was originally part of a publication by authors Jennifer Healey and Rosalind Picard [56] from Hewlett-Packard (HP) Cambridge Research Laboratory, the data set is known as Stress Recognition in Automobile Drivers dataset. The objective of the study for which these data were collected was to investigate the feasibility of automated recognition of stress based on the recorded signal, which includes ECG, EMG (right trapezius), GSR (galvanic skin resistance) measured on the hand and foot, and respiration. It has a collection of multiparameter recordings from healthy volunteers, taken while they were driving on a prescribed route including city streets and highways in and around Boston, Massachusetts. It was used in our previous work [6], this data set was chosen because it was collected using an interesting methodology that allows us to extrapolate to other scenarios. For example, operators of trains, airplanes, flight controllers and other high-risk jobs require continuous monitoring of the users, where it is necessary to be in good condition of healthy but also, check if the users of such valuable and dangerous machines have the authorization.

This work [56] used 5-min intervals of data from well-defined segments of rest, city, and highway driving to detect three general stress levels: low, medium, and high. After that, the authors evaluated how individual physiological features vary with driver stress at each second of the drive, including those segments of the drive between the rest, city, and highway segments. Specifically, the ECG electrodes were placed in a modified lead II configuration to minimize motion artifacts and to maximize the amplitude of the R-waves, which are the most important fiducial point. The ECG was sampled at 496 Hz and initially, there were 27 drivers according to the authors, some drivers completed the

course only once and for others, some signal was missing. In our study[6] we considered data from 14 drivers of 17 available with 1 hour of continuous acquisition. Driver number 3 has only 1 minute of ECG, drivers number 13 and 14 have the same information for ECG signal, so, we used one of them and driver 17 the data wasn't collected continuously.

The MIT-BIH Arrhythmia Database [57], a dataset of standard test material used since 1980 in innumerable scientific works for the evaluation of arrhythmia detectors and classifiers. It has recordings from 47 subjects, specifically 25 men aged 32 to 89 years and 22 women aged 23 to 89 years. They used dedicated equipment in acquisition designed and built in the BIH Arrhythmia Laboratory, including the tape-drive controllers and the analog-to-digital converter (ADC) interfaces. The digitization rate was 360 samples per second. Each ECG record of the MIT-BIH Arrhythmia Database includes two leads originating from different electrodes. The most common leads in the database are the modified limb lead II (MLII), and V1, obtained by placing the electrodes on the chest.

ECG-ID Database [58] has experimental studies involved 90 volunteers and ECG records were made in the sitting position. The author use single-lead ECG in Lead I position. It was chosen, according to the authors because it is easily measured and it is not sensitive to minor variations in electrode locations. The data collected for this study comprise the ECG-ID Database, consisting of 310 I-lead ECG recordings from 90 individuals, each 20 seconds long, sampled at 500 Hz with 12-bit precision. This is an interesting data set because It was originally collected to evaluate biometric aspects, most of the data sets available are related to medical or unhealthy situations.

Next, the Physikalisch-Technische Bundesanstalt (PTB) from the National Metrology Institute of Germany Diagnostic Database was analyzed [51]. It contains 549 records from 290 subjects (aged 17 to 87, 209 men, and 81 women; ages were not recorded for one female and 14 male subjects). The signal was sampled at 1000 Hz with a 16-bit resolution with 0.5 V, 16 input channel, and 0–1 kHz bandwidth. It was collected from Benjamin Franklin University Hospital, Cardiology Department of Cardiology, Berlin, and contains the ECG data of various healthy and diseased patients.

The last one is The PhysioNet Computing in Cardiology Challenge 2018[51], called *"You Snooze You Win"*. Data were contributed by the Massachusetts General Hospital's (MGH) Computational Clinical Neurophysiology Laboratory (CCNL), and the Clinical Data Animation Laboratory (CDAC). The dataset includes 1,985 subjects that were monitored at an MGH sleep laboratory for the diagnosis of sleep disorders. The challenge dataset could be used in future works because they captured a variety of physiological signal recorded as they slept through the night including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiology (ECG), and oxygen saturation (SaO2). Also, It has one of ECG biggest data set publicly available including data from almost 2 thousand subjects, It provides the opportunity to evaluate how the biometric system based in ECG performs when the number of people increases. Most of the papers don't consider the system must be trained again when new users are included. However, this data set also has its weakness, as It was collected to evaluate sleeping conditions, It doesn't represent the real world situation related to wearable

Data Set Name	Number of subjects	Duration of Segments	Healthy or diseased	Sample Rate	Sensor Position	Hospital or outside
DriverDB	16	65 to 93 min	Healthy	496 Hz	Modified Lead II	Collected while driving a car
MITDB	47	60 s	Patients from BIH Arrhythmia Laboratory	360 Hz	Modified Lead II	Collected in hospital
ECGID	90	20 s	Healthy	500 Hz	Lead I	Not in a hospital but at sitting position
PTBDB	290	High variability of duration	Mix of healthy volunteers and patients with heart diseases	1000 Hz	Lead II	Collected in hospital
Challenge 2018	1983	60 min for most of them	Supposed to be healthy (Patients from MGH sleep Lab)	200 Hz	Lead II	Collected in hospital (sleeping)

Table 4: Summary of data sets investigated.

devices users like noise due to mobility and heart variation related to stress, fatigue or physical exercises. The table below summarizes the data sets used in this master thesis.

Despite all details mentioned, two data sets were investigated deeper: Drive DB and Challenge 2018. The first one because it was collected in a real situation of drivers using the car as a regular day. So, I believe that data set can prove the potential of ECG as biomarkers. The second one was used to study the sleeping, however, It is not in our knowledge that It was used before in biometric investigation. Many papers have mentioned its concerns about what happens when the data set grows, but for a while, this wasn't investigated yet.

4.1.1 Noise Removing

The first step as showed in Figure 17 is the noise removing stage. The pre-process techniques are used on top of these data to get filtered ones with higher quality. The most mentioned noise is the baseline drift (also called the baseline wander). It is a low-

frequency artifact in the ECG that arises from breathing, electrically charged electrodes. Typically, a complete baseline wander removal requires that the cut-off frequency of the high-pass filter be set higher than the lowest frequency in the signal. The majority of baseline wander removal techniques have in common that they cancel the low-frequency components of the signal. The frequency of the baseline wander high-pass filter is usually set slightly below 0.5 Hz. During ECG data gathering, the baseline drift could be also occurred by certain movements of an applicant besides the low-frequency noise. Therefore, these filtering techniques may use if we do not know the proper indicator of the cut-off frequency or expecting certain movements of an applicant (HOS case).

To avoid this, in our first work with Drive DB, we used the technique mentioned in [5]: curve fitting technique. Curve fitting is the process of constructing a curve which finds the best fit to a series of data points and polynomial curve fitting finds an exact fit to the data more smoothly. The ECG baseline could be adjusted by subtracting the fitted curve data from the original ECG data. This process should be done in the whole data, having more effect depends on the noise in each segment. Figures 18 and 19 are from driver 1 of Driver DB mentioned before.

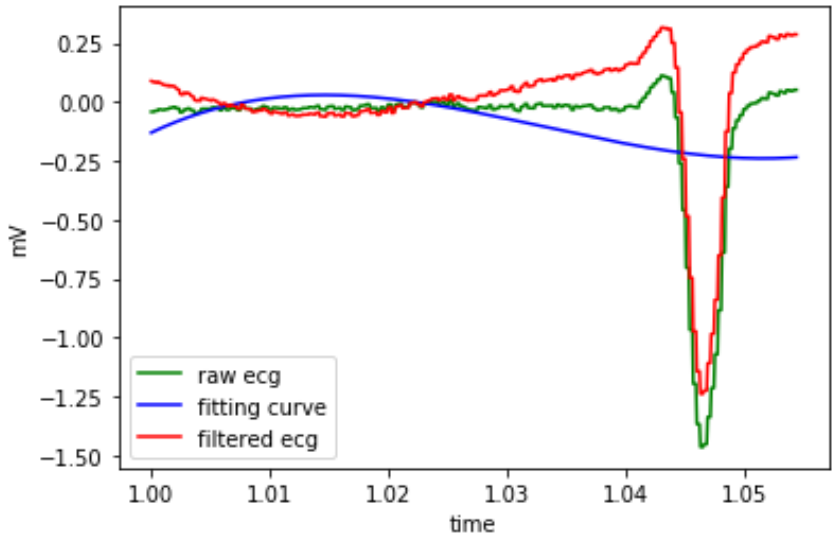

Figure 18: Baseline adjustment by using the polynomial curve fitting order 3rd for one second.

In the second data set(Challenge 2018) 4-order Butterworth bandpass (1–40Hz) filter is utilized for each ECG record to remove baseline wander. As this data was collected in a controlled environment, It is expected that the signal has less noisy and more quality than the first, collected while the person was driving, doing some movements and so on. If the 4-order Butterworth bandpass would be applied without care, It could remove not only the noise, but also some useful information. Figure 20 shows how the Butterworth bandpass is applied.

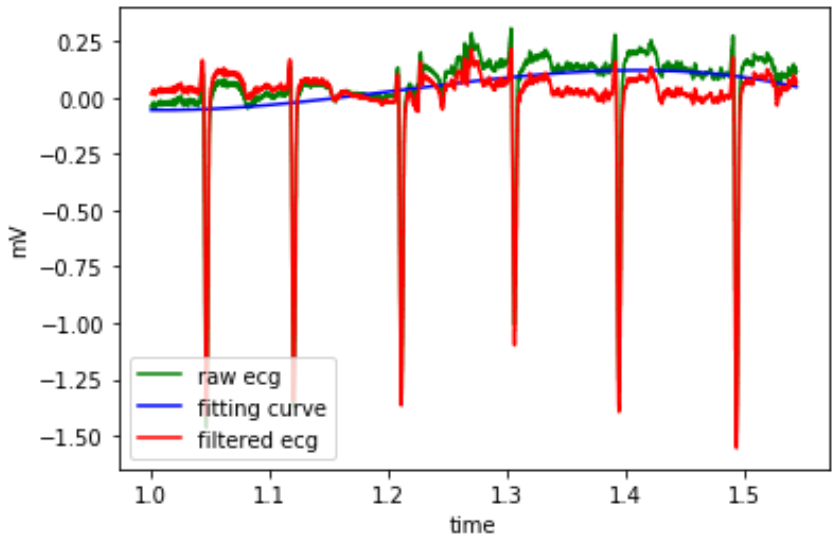

Figure 19: Baseline adjustment by using the polynomial curve fitting order 3rd for ten second.

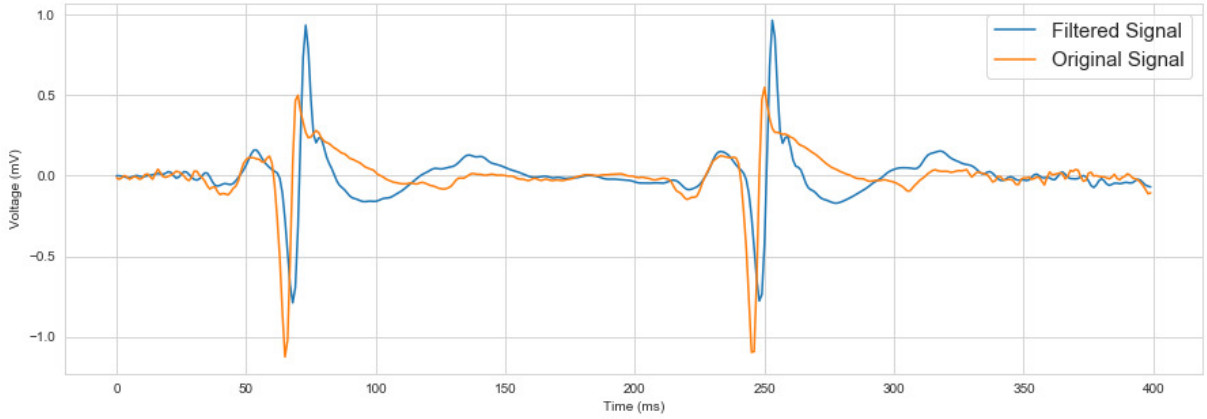

Figure 20: Baseline adjustment by using the butterworth bandpass.

4.1.2 Segmentation

Specifically, ECG is a continuous signal captured from some device in contact with the body. In terms to be used as input into the biometric system, it has to be sliced in some segments. In this master's thesis, It was applied a blindly split up into signal segments with an equal length of three seconds without leveraging any heartbeat location information, similar to used in [53], avoiding the complicated signal fiducial characteristics extraction process. The feature extraction is performed, but only after the segments are calculated.

Based on our previous work [6] and [27], the time window chosen was of 3 seconds and after that, the feature set is extracted from the signal. Even some works use a small window of 2-second, as data sets have a different sample rate, and the challenge data set has the smallest sample rate among all of Table 4.1, 200 Hz, a three-second window was

chosen. So, for the challenge data set each segment would have at least 600 samples. We expect using this window size would be possible to get at least two heartbeats, allowing us to extract more features with less effort since the typical range of heart rate is from 40 to 208 beats per minute.

4.1.3 Feature Extraction and Selection

Feature extraction is the most important step for user identification. It extracts characteristics from the vital signal to be used as a template during the authentication process. An ECG signal has many characteristics that could be used as demonstrated [1], the smallest number should be chosen to avoid complexity or too much computing. As we want to evaluate the dataset to be used in the continuous authentication process, we divided it in peace of 3 seconds. This duration was chosen because with 3 seconds we guarantee at least two heartbeat, allowing us to use the feature R-R Interval. It is the interval from the onset of one R wave to the onset of the next one, one complete cardiac cycle. Inspired in our previous work [6], QRS beat detection is performed in some steps as described below.

The algorithm must detect all the local maximum points in 3 seconds period, and thus we compare the amplitude values, storing the peaks. The first value is temporally stored and compared with the next one to check if the signal is increasing or decreasing its amplitude, while it is increasing a temporary variable is replaced; when it starts to decrease, the two points ahead are compared to make sure it is downhill. If it is true, this point is stored as a local maximum. When the signal starts to increase again a new local search for the maximum is started. At the end of this step, there is an array of all peaks for every 3 seconds of ECG signal as shown in Figure 21.

In the second step, the algorithm finds the max R peak and using a threshold of 66% of the maximum value eliminates the points under this value. The remaining points are classified as R-Peak. The detection of R-peaks in an ECG acquisition is the primary goal of any algorithm for the automatic processing of ECG signals. After that, the first local maximum values back and ahead to the R-peak are classified as Q and S points, as shown in Figure 22.

At this point, We already found the Q, R, and S amplitude for each beat. In this sense, we compute to be used as feature the mean and standard deviation of each QRS points, figure out the QRS wave peak to peak amplitude and R wave duration. All features used are in Table 5.

Following the algorithm mentioned in Figure 17, after the extraction and selection of all features, It is time to generate a template and classification phase.

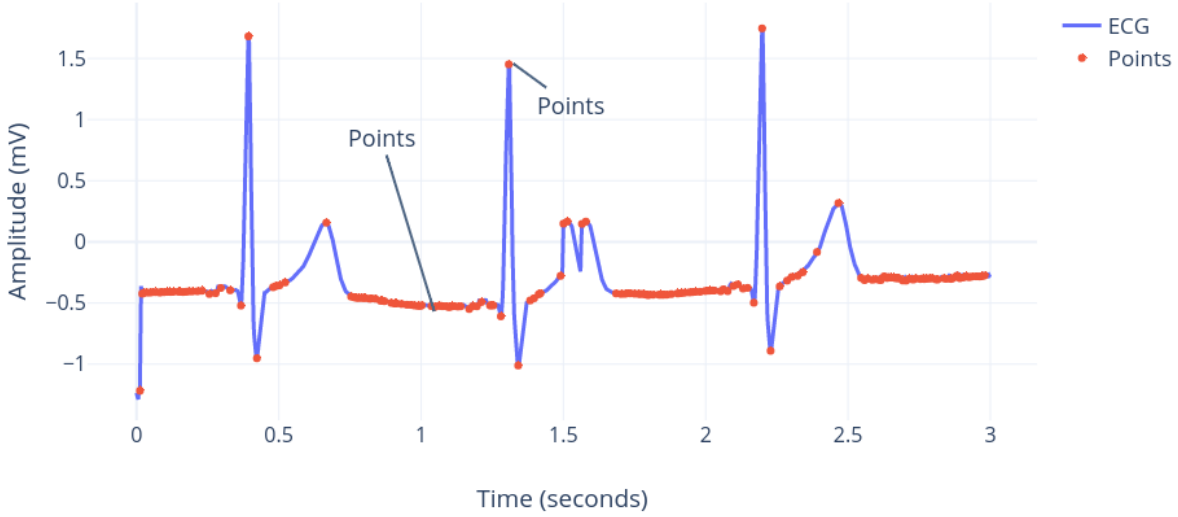

Figure 21: All Peaks Selected

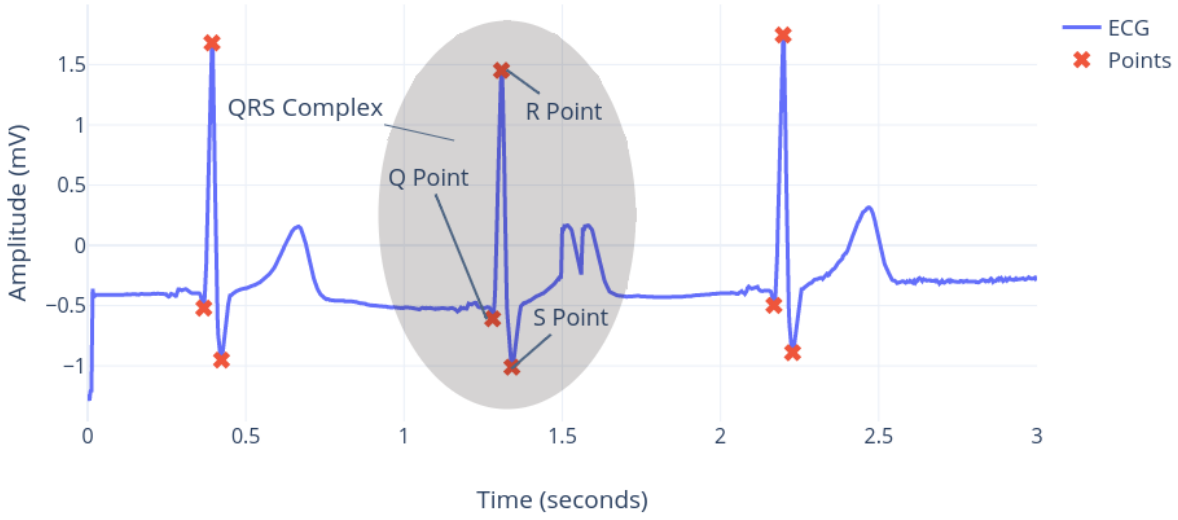

Figure 22: QRS filtered using floating threshold [6]

4.1.4 Underfitting and Overfitting

In statistics and machine learning the data is usually split into two subsets: training data and testing data (and sometimes to three: train, validate and test), and fit our model on the train data, to make predictions on the test data. During this process, the model can be underfitted or overfitted. They affect the predictability of our model,

Table 5: 6 features captured directly ECG (mean) and **3 derived**

No.	Features
1	Q wave amplitude(μ V)
2	R wave amplitude(μ V)
3	S wave amplitude(μ V)
4	R wave duration(ms)
5	QRS wave peak to peak amplitude (μ V)
6	R-R Interval
7	**Q wave Standard Deviation**
8	**R wave standard deviation**
9	**S wave standard deviation**

causing a lower accuracy and/or the loss of generalization capability.

Overfitting means that model we trained has trained "too well", performing too closely to the training dataset. Sometimes, It occurs when the model is too complex, having many features or variables compared to the number of samples available. So, the developer shouldn't increase the number of features without increasing the size of the dataset. This overfitted model will be very accurate on the training data but will probably be very not accurate on untrained or new data. It is because this model is not generalized. On the other hand, underfitted means that the model does not fit the training data and therefore misses the trends in the data. It also means the model cannot be generalized to new data. This is usually the result of a very simple model with a small number of features or variables. This model will have the poor predictive ability not only on training data but also to any other data.

Nevertheless, both situations must be avoided in data analysis. A good Train/test split and cross-validation help to avoid overfitting more than underfitting. The concept of cross-validation is explained next.

4.1.5 Cross-Validation

Usually, the whole dataset available is randomly divided into three or two subsets: training, validation, and test dataset. Training is the sample of data used to fit the model. The actual dataset that we use to train the model, It sees and learns from this data. The validation dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning the model. The test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. It provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets). Many times the validation set is used as the test set, but it is not good practice. The test should contain carefully sampled data that spans the various classes that the model would face. However, by partitioning

the available data into three sets, the number of samples which can be used for learning the model is drastically reduced, and the results can depend on a particular random choice for the pair of train and validation sets.

To solve this problem is used the cross-validation. In the basic approach, called k-fold cross-validation, the training set is split into k smaller sets. The following procedure is followed for each of the k "folds": A model is trained using k1 of the folds as training data[59]; the resulting model is validated on the remaining part of the data. The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop. This approach can be computationally expensive but does not waste too much data, which is a major advantage in problems such as inverse inference where the number of samples is very small.

In this master's thesis is used Scikit-Learn's K-fold cross-validation feature[59]. It randomly splits the training set into 10 distinct subsets called folds, then it trains and evaluates every 10 times, picking a different fold for evaluation every time and training on the other 9 folds. The result is an array containing the 10 evaluation scores. Figure 23 shows how it works.

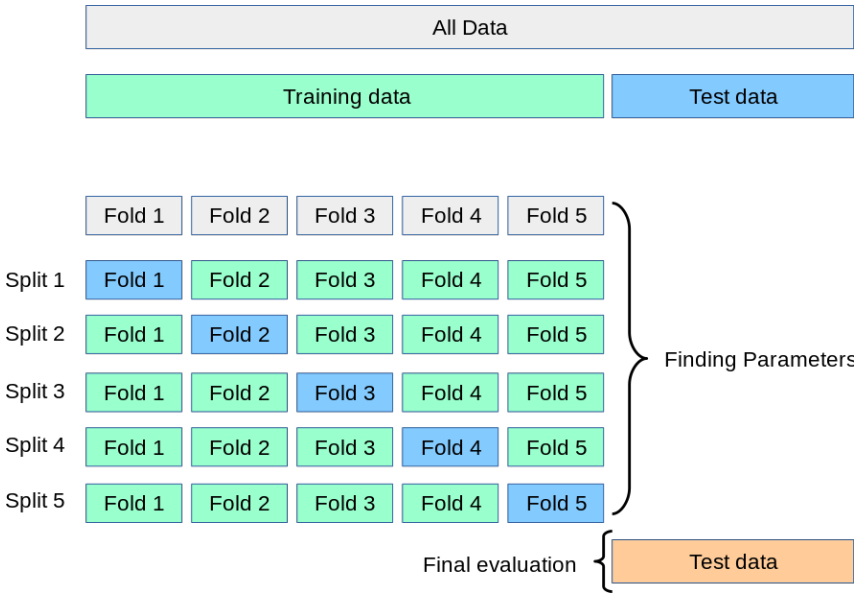

Figure 23: Cross Validation Process.

Cross-validation usually gives accuracy measures and its standard deviation. In this work, we try out many other models from various categories of Machine Learning algorithms (Support Vector Machines, Random Forest and Neural Networks without spending too much time evaluating the hyperparameters. This strategy is first to get the promising models, then try to find the best parameters. Other possible solutions for overfitting are to simplify the model, constrain it, or get a lot more training data. In this sense, in the next subsection, we explain our strategy used to get the data augmentation done.

4.1.6 Data Augmentation

Data augmentation is the process of increasing the amount and diversity of data. New data is not collected, actually, the present data is transformed. This process helps increase the amounts of data because sometimes it is not feasible to collect thousands or millions of data, so data augmentation the size of the dataset and introduce variability in the dataset. For each recording, an example of randomly chosen ECG windows is collected which usually includes a different number of heartbeats and highly different signal morphologies. The blind segmentation strategy can effectively avoid data-specific complicated heartbeat identification and segmentation techniques, but at the same time, also introduces a high variability to the ECG windows (number of heartbeats, the onset of the segment, etc.) and poses a big challenge to the following data representation and machine learning algorithms. With a 10 seconds dataset, we would get 5 segments with 2 seconds each. However, using data augmentation It is possible to select quite more segments randomly. It is necessary to evaluate the data collected to avoid false randomly segments.

4.1.7 Choosing Models and Tuning Parameters

After using the cross-validation process, we will get a shortlist of promising models. Then we will need to fine-tune them. A grid search is an option to avoid do this manually. There is a Scikit-Learn's GridSearchCV [59] to search those hyperparameters. All you need to do is tell it which hyperparameters you want it to experiment with, and what values to try out, and it will evaluate all the possible combinations of hyperparameter values, using cross-validation. Another way to fine-tune is to try to combine the models that perform best. The group (or "ensemble") will often perform better than the best individual model.

4.1.8 Summary of the proposal

In this section, the process of preparing the dataset before being used as input in a machine learning algorithm is described. Any data science project has some generalities in common as the preprocessing steps after getting the data. Sometimes, with data on the hand, the research does some analysis trying to discover some insights. In this stage, some visualization tools could help to find visual patterns. After the data is prepared for machine learning algorithms, the developer should have a strategy to evaluate and find the best model to train the data. Remember to fine-tune this model with hyperparameters and don't make any mistakes during the validation and test steps.

The following Figure 24 explains the steps after the data is already acquired.

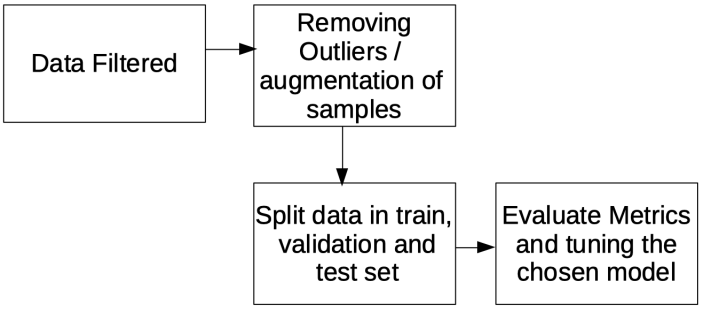

Figure 24: Steps before train the model.

4.2 Evaluation Methodology

In this section, an evaluation of the proposed methodology is performed. In the first part, the Driver Stress Data Set [56] is evaluated following the procedures mentioned in chapter 4. Some classifiers are applied to its data set and in the second step, one classifier is chosen to be explored deeper. In the second part, another data set is evaluated. In this case, the Challenge Data set was chosen. The idea of this part is to demonstrate how a system will perform under the growing number of users. Most of the papers in this area evaluate a small data set like us in the first step, mainly because of the difficulty to find public data with ECG information. The strategy used was to split the data set in the train and test subset. Using cross-validation to build the models, based on these results, compare the metrics to choose the best model for the data.

4.2.1 Metrics

Specifically, Accuracy is the relation between the correct classified instances of the problem over the total number of instances, as expressed on Eq. 4.1. So, It means the number of correct predictions made as a ratio of all predictions made. This is the most common evaluation metric for classification problems.

$$Accuracy = \frac{CorrectlyClassifiedInstances}{AllInstances} \tag{4.1}$$

The Recall is the true positive rate also called the sensitivity. It is the number of instances from the positive (first) class that predicted correctly. So, the precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the positive model (meaning they are correct), and false negatives are data points the model identifies as negative that are positive (incorrect), as expressed on Eq. 4.2.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.2}$$

Another metrics used is the precision, which is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are negative, or in our example, someone was label as an authenticated person when should be not authenticated. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant were relevant. It is expressed on Eq.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.3}$$

4.2.2 Evaluating Driver DB using Cross Validation

The database used in this step was Driver DB, there were 27 drivers according to the authors, some drivers completed the course only once and for others, some signals were missing. In our study, we considered data from 14 drivers of 17 available with 1 hour of continuous acquisition. Driver number 3 has only 1 minute of ECG, drivers number 13 and 14 have the same information for ECG signal, so, we used one of them and driver 17 the data wasn't collected continuously.

The data set was segmented in 1900 segments for each driver, building a balanced data set, this is very important to avoid a bias situation. The preview of the features collected could be seen in Figure 25. We used in this step 9 features extracted from the ECG signal as described in Chapter 4.

	mean_q	mean_r	mean_s	stdev_q	stdev_r	stdev_s	qrs_interval	rr_interval	rq_amplitude	qrs_interval.1	person
0	-0.723496	0.278612	0.067206	0.014606	0.006745	0.003261	72.580645	780.241935	1.002108	72.580645	drive16
1	-0.745529	0.283174	0.074638	0.010212	0.008582	0.003838	72.580645	721.774194	1.028703	72.580645	drive16
2	-0.727950	0.281287	0.059535	0.011583	0.012637	0.012770	71.908602	687.500000	1.009237	71.908602	drive16
3	-0.505013	0.285457	-0.206507	0.323151	0.011784	0.373444	64.516129	655.241935	0.790470	64.516129	drive16
4	-0.285455	0.301478	-0.485068	0.350774	0.014207	0.382756	54.435484	635.080645	0.586933	54.435484	drive16
...	...	...	...	...	...	...	...	...	...	...	...
26595	-0.796856	1.004739	-0.179860	0.066082	0.103765	0.125639	55.443548	623.991935	1.801595	55.443548	drive01
26596	-0.706270	0.942474	-0.271047	0.023319	0.069515	0.017781	60.147849	556.451613	1.648744	60.147849	drive01
26597	-0.697754	0.921037	-0.266242	0.010162	0.036736	0.014690	59.475806	653.225806	1.618791	59.475806	drive01
26598	-0.731650	1.002127	-0.305096	0.051388	0.050148	0.005644	60.819892	692.540323	1.733777	60.819892	drive01
26599	-0.739466	0.971679	-0.272621	0.006579	0.008387	0.008219	59.475806	622.983871	1.711144	59.475806	drive01

26600 rows × 11 columns

Figure 25: Data set Described.

For cross-validation, we used 10 fold configuration, and we experiment with the following classifiers: Decision Tree, Random Forest, Extra Trees, Support Vector Machine and Neural Networks. A voting classifier was also used. The idea behind the Voting Classifier is to combine conceptually different machine learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels. Such a

classifier can be useful for a set of an equally well-performing model to balance out their weaknesses [60]. In majority voting, the predicted class label for a particular sample is the class label that represents the majority of the class labels predicted by each classifier.

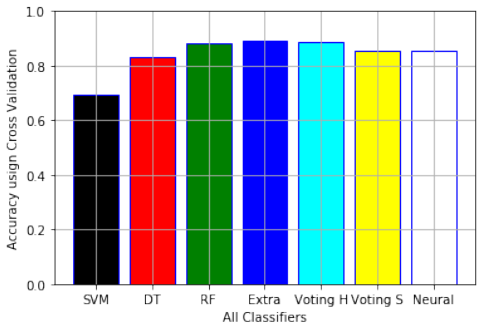

Figure 26: Accuracy using Cross Validation with K equal 10.

In these simulations, Extra Tree and Random forest got the best results in terms of accuracy, 89.11%, and 88.46%, respectively, as seen in Figure 26. Remember, those results are the mean of all ten simulations executed because of the cross-validation approach. The cross-validation process also gives us the standard deviation of those results, demonstrated in Figure 27. In this case, SVM showed the best results, however, its accuracy wasn't good enough to be used as main classified.

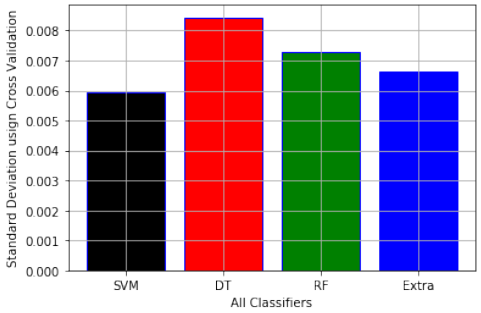

Figure 27: Standard deviation using Cross Validation with K equal 10.

It is possible to make trees even more random by also using random thresholds for each feature rather than searching for the best possible thresholds. A forest of such extremely random trees is simply called an Extremely Randomized Trees ensemble. This trades more bias for lower variance. Extra-Trees are usually much faster to train than regular Random Forests since finding the best possible threshold for each feature at every node is one of the most time-consuming tasks of growing a tree. We created an Extra-Trees classifier using Scikit-Learn's ExtraTreesClassifier class and RandomForestClassifier.

4.2.3 Choosing a promising model

After the train, the model using cross-validation, Random Forest and Extra-trees seems to be the most promising model for this data. We decide to train all models again

to evaluate the fitting time. It is important to evaluate these metrics because depending on the application target, the model has to be trained frequently and a long fitting time would be a problem. In our case, the system has to be ready to classify a person every time than a new one is registered. So, in real situations, the system will be updating its model with a high frequency. Figure 28 shows the time to fit all classifiers used. It is easy to see that almost classifiers spent the same time, except the Neural Networks. It was expected because of the natural way that Neural Networks works. We use 1000 epochs as parameters, so, the systems iterates 1000 times to update all the weights and found the best model. Reducing the number of epochs will reduce the time to fit, but probably will decrease the accuracy.

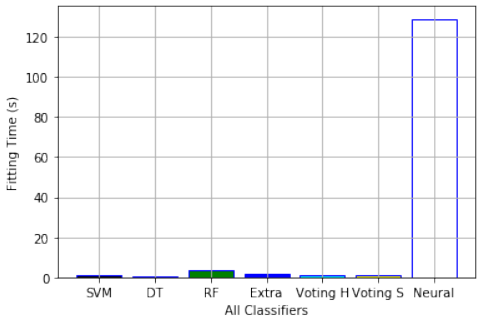

Figure 28: Fitting time of all classifiers.

We also calculated the precision and recall as mentioned before, and demonstrated in Figures 29 and 30 below.

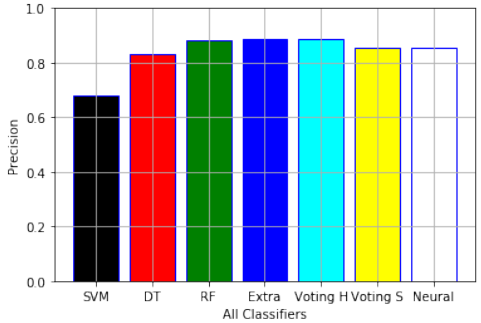

Figure 29: Precision for all classifiers.

After collecting these metrics, The random forest was chosen as the best classifier. Extremely Randomized Trees got almost the same results, and It differs in the sense that the splits of the trees in the Random Forest are deterministic whereas they are random. However, It could introduce some bias due to the fact the splitting of the subtree is made randomly and not based on some threshold as in the Random Forest.

Random Forest is a collection of Decision Trees trained with randomly selected data, which guarantees that each tree is slightly different from each other. Hence, each tree may return a distinct result for a given dataset. The RF algorithm classifies the data based on a voting system involving the results from the individual trees. Specifically,

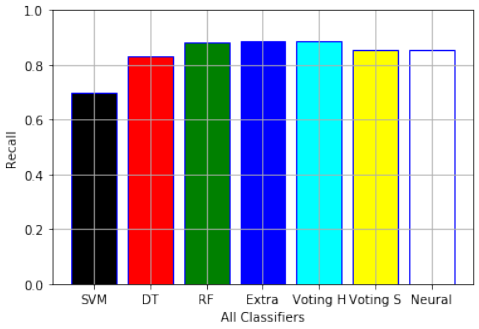

Figure 30: Recall for all classifiers.

direct voting count how many trees classified a given feature under a particular class. Weighted voting returns the ratio of elements belonging to a given class. RF presents a good performance in two critical aspects: anomaly detection and overfitting. It uses a voting system that helps to minimize the effect of overfitting concerning the individual decision tree.

Analyzing the results using different classifiers, we noticed that RF classification is promising in this application. RF implements a voting system to come up with a solution as an average of results provided by each Decision Tree [6]. All tests above were executed using a continuous authentication scenario. Each driver was labeled as originally in the dataset.

Taking in mind, this is an approach for continuous identification, the results above demonstrate the identification during a window of 3 seconds each, so, the security system should include an analysis about blocking the personal data only after some time of failed identification.

In the second part, we analyzed authentication, the selected driver was labeled as a legitimate driver and all the others were labeled as not drivers. Our result was 98.19% of average accuracy, 0,9805 of average precision and 0,9818 of average sensitivity.

The results of each driver are demonstrated in Table 6. These results are quite close to the works in Chapter 3. Considering the number of features extracted, the number of lead sensors used and the simplicity of the proposed method the result of 98.2% accuracy seems to be reasonable. In our opinion, biometric authentication and identification must start to consider continuous situations as the concern of privacy and security should increase every year due to data production expansion as IoT and smart cities become reality. Assessing the results considering 60 minutes of access, the average rejections would occur during only 1 minute in the worst case, which seems to be feasible to most applications. The results for authentication (detect someone against a group) were acceptable, but for identification (detect someone against each group's member) were good enough. So, in following we try to improve these results removing the outlier.

Table 6: Authentication results using 14 drivers

Drivers	Accuracy	Specificity	Sensitivity
D1	99.92%	0.999	0.999
D2	99.92%	0.999	0.999
D4	99.83%	0.998	0.998
D5	96.57%	0.963	0.966
D6	97.94%	0.979	0.979
D7	99.35%	0.993	0.994
D8	96.18%	0.959	0.962
D9	94.95%	0.944	0.95
D10	99.59%	0.996	0.996
D11	94.92%	0.943	0.949
D12	99.94%	0.999	0.999
D14	99.30%	0.993	0.993
D15	97.74%	0.977	0.977
D16	98.53%	0.985	0.985
Average	98.2%	0.9805	0.9818

4.2.4 Tuning the model

During the first step, Random Forest was trained using very simple parameters and only a few preprocessing was made to the data set. The parameters of RandomForestClassifier (scikit-learn) was n_estimators=100, max_features="auto",random_state=0. N estimator means the number of trees in the forest. Max Features means the number of features to consider when looking for the best split and Random State controls both the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node. During the tuning step, GridSearchCV was used with cross-validation set in 3. The parameters evaluated were max_depth (80, 90, 100, 110), min_samples_leaf(3, 4, 5), min_samples_split(8, 10, 12) and n_estimators(100, 200, 300, 1000). The maximum depth of the tree means nodes are expanded until all leaves are pure. The min_samples_leaf is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. The min_samples_split is the minimum number of samples required to split an internal node.

After the grid search, the best parameters found was max_depth=100, min_samples_leaf=3, min_samples_split=10, and n_estimators=200. However, even using the best parameters found the results didn't improve so much as demonstrated below.

To continue the improvement of the system, the approach chosen was to perform

	precision	recall	f1-score	support
0	0.99	1.00	0.99	223
1	0.98	0.98	0.98	227
2	0.62	0.59	0.60	217
3	1.00	1.00	1.00	219
4	0.88	0.85	0.86	249
5	0.90	0.88	0.89	217
6	1.00	1.00	1.00	227
7	1.00	1.00	1.00	235
8	0.69	0.77	0.72	213
9	0.82	0.86	0.84	220
10	0.94	0.98	0.96	255
11	0.76	0.72	0.74	239
12	0.66	0.61	0.63	226
accuracy			0.87	2967
macro avg	0.86	0.86	0.86	2967
weighted avg	0.86	0.87	0.86	2967

Figure 31: Metrics results after Grid Search.

a visual analysis of data looking for some pattern. Then, It was identified some outlier as it can be observed in Figure 32. All the observations with more than 3 times of standard deviation were eliminated. Those outliers can be generated by noise or even some error during the storage of data. Some of them are usually removed during the filtering step, but the remaining should be removed in some later steps to avoid mismatching classification.

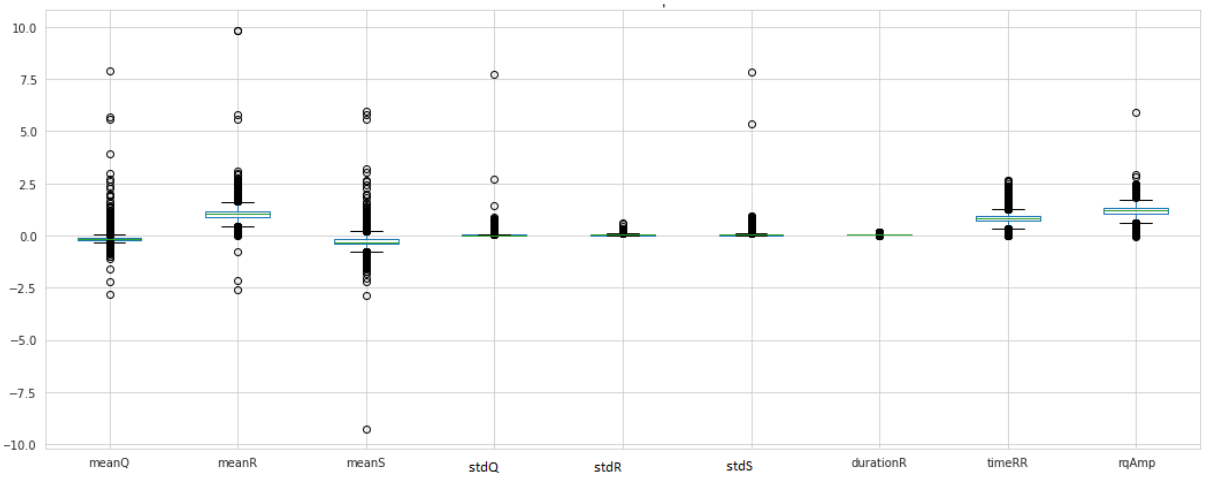

Figure 32: Using boxes to detect outlier.

After the removal of outliers, the data was divided again in train and test e submitted to the random forest. The parameters used was n_estimators=200, max_depth=100, min_samples_leaf=3, and min_samples_split=10). The results were significantly improved achieving 98% of accuracy against 86% using test data set and 88.46% with cross-validation approach. The metrics and the confusion matrix can be seen below.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	25
1	0.99	1.00	0.99	228
2	0.94	0.92	0.93	197
3	1.00	1.00	1.00	126
4	0.98	0.98	0.98	243
5	0.98	0.99	0.98	208
6	1.00	1.00	1.00	195
7	1.00	1.00	1.00	224
8	0.93	0.95	0.94	198
9	0.98	0.98	0.98	236
10	0.98	1.00	0.99	231
11	0.95	0.94	0.95	218
12	0.95	0.93	0.94	214
accuracy			0.97	2543
macro avg	0.98	0.98	0.98	2543
weighted avg	0.97	0.97	0.97	2543

Figure 33: Metrics Improvement after outlier removal.

4.3 Evaluating Challenge DB

The database used in this step was The PhysioNet Computing in Cardiology Challenge 2018[51], they captured a variety of physiological signals recorded as they slept through the night including electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiology (ECG), and oxygen saturation (SaO2). As our future works, we could use a multimodal approach to evaluate the contribution of other signals to the ECG signal.

Also, It is one of ECG biggest data set public available, almost 2 thousands subjects are included, It provides the opportunity to evaluate how the biometric system based in ECG performs when the number of person increases. Most of papers don't consider the system must be trained again when new users are included. However, this data set also has its weakness, as It was collected to evaluate sleeping conditions.

The simulations for challenge DB were divide in 4 groups related to the preprocessing of data collected. The simulations were divided on: Original Data, Data with increased number of examples, Data with increased number of examples and outlier removed, and then all stages plus the addition of new features. The Random Forest was used again with the same parameters used before, found using GridSearchCV. In this data set with almost 2000 persons, the results decrease as demonstrated in Table 7.

After that simulations, a data augmentation technique was applied. They are very common in image processing applications and It is use to artificially create variations in

```
[[ 25   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0 227   0   0   0   0   0   1   0   0   0   0   0]
 [  0   0 181   0   0   0   0   0   7   2   0   4   3]
 [  0   0   0 126   0   0   0   0   0   0   0   0   0]
 [  0   0   1   0 237   0   0   0   0   1   0   0   4]
 [  0   2   0   0   0 205   0   0   0   1   0   0   0]
 [  0   0   0   0   0   0 195   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0 224   0   0   0   0   0]
 [  0   0   1   0   0   0   0   0 189   0   1   5   2]
 [  0   0   1   0   0   4   0   0   0 231   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0 231   0   0]
 [  0   0   3   0   0   0   0   0   5   0   2 206   2]
 [  0   0   5   0   4   0   0   0   3   0   1   2 199]]
```

Figure 34: Confusion Matrix after outlier removal.

Table 7: First Results for challenge Data Base

	Precision	Recall	F1-score	Support
Accuracy	74.78%	74.78%	74.78%	74.78%
Macro avg	74.44%	75.16%	72.45%	8740.0
Weighted avg	77.24%	74.78%	73.83%	8740.0

existing images to expand an existing image data set. This creates new and different images from the existing image data set that represents a comprehensive set of possible images. This is done by applying different transformation techniques like zooming the existing image, rotating the existing image by a few degrees, shearing or cropping the existing set of images etc. In our application, a subsegment is randomly extracted from the ECG stream, allowing us to increase the size of data set. Using 60 seconds, now It is possible to get 100 segments of 2 seconds compared with 30 segments we got before. This helps to increase the performance of the model by generalizing better and thereby reducing overfitting. The results of second step are below in Table 8.

Table 8: Results with data augmentation for challenge Data Base

	Precision	Recall	F1-score	Support
Accuracy	75.74%	75.74%	75.74%	75.74%
Macro avg	76.58%	75.81%	74.31%	36311.0
Weighted avg	77.38%	75.74%	74.65%	36311.0

The third simulation was removing the outliers. This time in addition to the procedure done with Driver DB, the instances with less than 60 segments were also removed. Then, the number of subjects in database decreased to 1501. This approach is similar

to what most of biometric systems do today, They require good samples to create the template, if the samples don't have the minimum quality required, they are discarted and It is necessary to collect data again. The results of third step are below in Table 9

Table 9: Results with data augmentation and outliers removal.

	Precision	Recall	F1-score	Support
Accuracy	78.73%	78.65%	78.65%	78.65%
Macro avg	79.73%	77.71%	76.79%	32361.0
Weighted avg	79.87%	78.65%	77.43%	32361.0

In the last simulation, we evaluate the addition of some new features in the data set. The QRS Onset and Offset were calculated and Points P and S. The Onset and Offset are calculated after the Q and S points. Using Q as reference, the extractor must go back into the raw signal and calculate the greater slope between the new point and the Q. A window of 40ms was used. A similar procedure is done to find the offset, but the reference is S point and the extractor walks to the end of array. To find P and T, a window starting from QRS the onset/offset is used. The Figure 35 shows the position of QRS Onset and Offset.

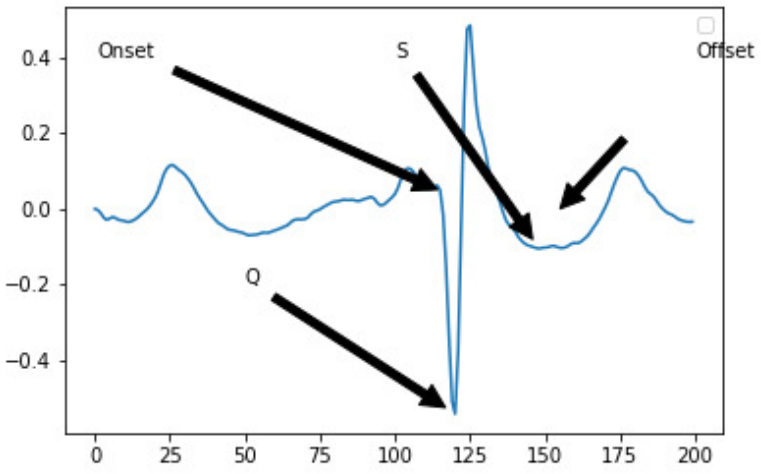

Figure 35: Adding QRS onset and offset.

Collecting these 4 news features is possible to generate mean p,mean t, mean qs distance, mean qt distance, mean qrs offset, and mean qrs onset. Total of 6 new features. The results using this new data set is below in Table 10.

Another way to see the results is expressed in Figure 37, a count of the metrics was performed. It is easy to see that the number of person with higher metrics improves after the data was transformed.

Table 10: Results with data augmentation, outliers removal and addition of new features.

	Precision	Recall	F1-score	Support
Accuracy	85.23%	85.23%	85.23%	38319.0%
Macro avg	84.6%	86.08%	84.8%	38319.0
Weighted avg	84.78%	85.91%	85.23%	38319.0

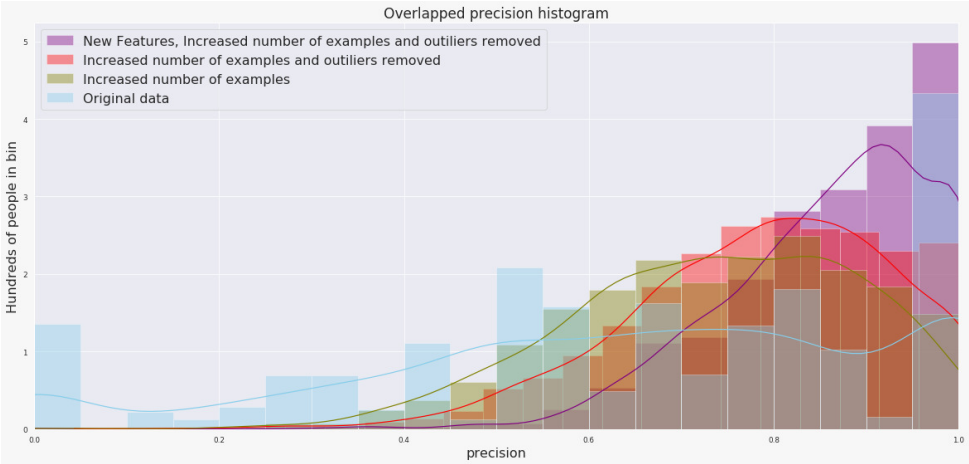

Figure 36: Cumulative results by instance.

Each step of our proposed processing of data contributed to overall improvement of system performance. It is possible to see that increasing the number of examples through our data augmentation process was the step that contributed more. It makes sense if you consider that the size of data set is one of the most important thing to train a model. The addition of new features was the second greater contributor. It is important to mention that the increasing of the number of features can also increase the complexity of the model, what can results in longer time of fitting. So, the idea is to keep the number of features low due the constraints of mobile/wearable applications.

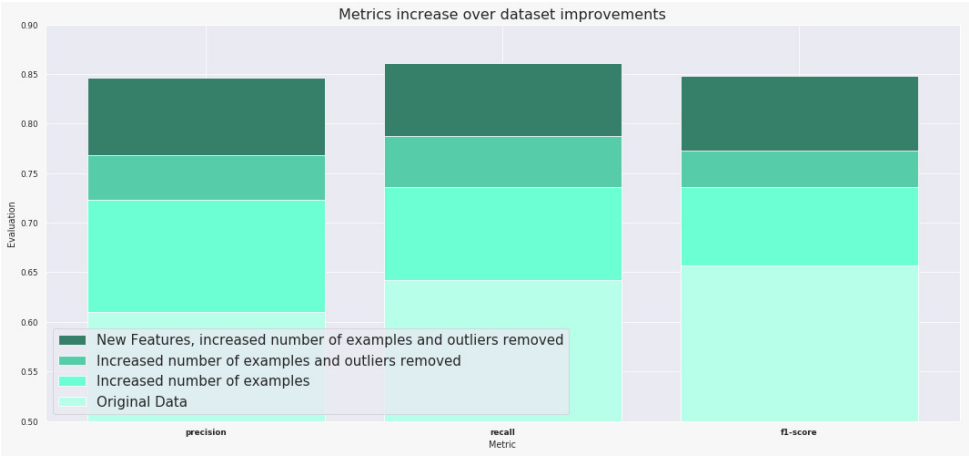

Figure 37: Contribution of each step in overall performance.

4.4 Chapter Conclusions

In this chapter, we demonstrated our proposal framework to use ECG personal data as input in biometric system. The model that seems to perfom better was the Random Forest. The parameters were chosen based on recognized approach called GridSearch using Cross Validation. Using Driver DB the accuracy was 98.2% for authentication and the results was improved from 88% to 98% of accuracy for identification after the outlier removal.

However, using Challenge dataset the scenario is quite more challenger. Increasing the number of person affected badly the accuracy of system. Our approach increased the results from 74% accuracy to 83.16%. Even these results don't seem too good, the ECG still can be used as biometric system, since other techniques are applied to reduce the person in the data set. Most of authentication system already apply this kind of technique. For instance, a time of live might be proposed. After sometime without authentication one user is removed from the list, helping the biometric system to increase the accuracy. Also, the ECG can be used as additional factor of authentication, considering some previous information, even with 80% of accuracy can help the system overall accuracy. All of this can be explored to evaluate, these topics are cited in the conclusion of this master's thesis as future work.

CHAPTER 5

Conclusions

In this master's thesis, an end-to-end framework to use the ECG stream as a biometric signal was proposed. As demonstrated in initial chapters, the wearable industry will gain relevant space in the near future, providing technology and resources to the health market growing. Even the users not interested in eHealth application will have their needs for security measures increased in the next years with IoT and 5G deployments. So, ECG is a promising signal to bring together eHealth and security lines.

There are some public data set available to do some research in this context, most of them weren't collect considering the biometric scenario, and the data collected in the clinical environment might not be similar to the data collected during activities. Based on these aspects, this thesis investigated the Driver DB data set. Using our proposed approach the results both for authentication and identification was good and compared with other works. The accuracy was 98.2% for authentication and the results were improved from 88% to 98% of accuracy for identification after the outlier removal. This simulation proves the feasibility of our proposal. As other biometric systems, during the acquisition of signals, some criteria have to be defined. ECG with low quality must be excluded from the system and the user must register itself again.

Furthermore, a bigger data set was evaluated, Challenge DB. In this case, the evaluation was to see the impact of our methodology in a data set with a higher number of subjects (almost 2 thousand against dozens).

In my opinion, ECG is going to be the next biometric signal to be used soon as the acquisition system(wearable) improves its quality. In the next section, some future works are discussed.

5.1 Future Works

As future works, the combination of other signals and information should be addressed. For example, to guarantee the true user is using a credit card, not the biometric should be used, but information about the user interests, GPS, or questions based on customer history can increase the security of the system. Considering only biometric systems, for instance, a system might reject the provided samples, due to poor quality or noisy signal [31]. Recognition performance is a measure of how well the system is able to match the biometric information from the same person correctly. In some cases the performance of a single biometric modality is insufficient, and thus methods combining biometric matches have attracted the academia interesting. This combination is called biometric fusion, which can be classified into two groups: unimodal biometric systems and multimodal biometric systems.

In our concern, the future of biometric systems is related to multimodal systems. They are also more resistant to spoofing attacks because it is more difficult for the attacker to spoof multiple biometric sources simultaneously. A biometric fusion system will use information from different sensors or using many samples/sources. So, The plan is to continue investigating biometric systems considering other signals and context information available through services and sensors in the city.

5.2 Published Works

Part of the results of this master's thesis was published at conference papers [6, 27]. Other works were also published [61]:

1. Cerqueira, E., Resque, P., Medeiros, I., Bastos, L., **Santos, A.**, Tavares, T., Rosário, D., Santos, A., Nogueira, M. Autenticação usando Sinais Biométricos: Fundamentos, Aplicações e Desafios. Jornada de Atualização em Informática 2019. 2aed.: SBC, 2019, p. 149-195.

2. **Barros, A.**, Rosário, D., Resque, P., Cerqueira, E. Heart of IoT: ECG as biometric sign for authentication and identification. In: 2019 15th International Wireless Communications and Mobile Computing Conference (IWCMC), 2019, Tangier. p. 307.

3. **Santos, A.**, Medeiros, I., Resque, P., Rosário, D., Nogueira, M., Santos, A., Cerqueira, E., Chowdhury, K. ECG-Based User Authentication and Identification Method on VANETs. In: the 10th Latin America Networking Conference, 2018, São Paulo. Proceedings of the 10th Latin America Networking Conference - LANC '18, 2018. p. 119.

References

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 3, pp. 808–812, June 2001. Cited 7 times in pages 12, 5, 24, 25, 27, 31 e 38.

[2] C. Camara, P. Peris-Lopez, L. Gonzalez-Manzano, and J. Tapiador, "Real-time electrocardiogram streams for continuous authentication." *Applied Soft Computing Journal*, 2018. Cited 3 times in pages 12, 26 e 31.

[3] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ecg: Convolutional neural networks for ecg biometric recognition," *Pattern Recognition Letters*, vol. 126, pp. 78 – 85, 2019, robustness, Security and Regulation Aspects in Current Biometric Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518301077 Cited 4 times in pages 12, 28, 29 e 31.

[4] Y. Zhang, Z. Xiao, Z. Guo, and Z. Wang, "Ecg-based personal recognition using a convolutional neural network," *Pattern Recognition Letters*, vol. 125, pp. 668 – 676, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865519302004 Cited 3 times in pages 12, 29 e 31.

[5] S. Kim, C. Y. Yeun, E. Damiani, and N. Lo, "A machine learning framework for biometric authentication using electrocardiogram," *IEEE Access*, vol. 7, pp. 94 858–94 868, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2927079 Cited 3 times in pages 12, 33 e 36.

[6] A. Santos, I. Medeiros, P. Resque, D. Rosário, M. Nogueira, A. Santos, E. Cerqueira, and K. R. Chowdhury, "Ecg-based user authentication and identification method on vanets," in *Proceedings of the 10th Latin America Networking Conference (LANC '18)*. New York, NY, USA: ACM, Oct. 2018, p. 119–122. [Online]. Available: https://doi.org/10.1145/3277103.3277138 Cited 9 times in pages 12, 5, 33, 34, 37, 38, 39, 47 e 56.

[7] M. FRAMINGHAM, "Worldwide wearables shipments surge 94.6% in 3q 2019 led by expanding hearables market," *International Data Corporation*, 2019. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS45712619 Cited 2 times in pages 14 e 3.

[8] S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, S. Khalifa, K. Thilakarathna, M. Hassan, and A. Seneviratne, "A Survey of Wearable Devices and Challenges," *IEEE Communications Surveys & Tutorials*, 2017. Cited 3 times in pages 2, 3 e 8.

[9] N. Sultan, "Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education," *International Journal of Information Management*, vol. 35, no. 5, pp. 521 – 526, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401215000468 Cited in page 2.

[10] D. M. Rabah Kamal and C. Cox, "How has u.s. spending on healthcare changed over time?" *KFF analysis of National Health Expenditure*, 2019. [Online]. Available: https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/#item-nhe-trends_total-national-health-expenditures-us-per-capita-1970-2018 Cited in page 2.

[11] T. Yilmaz, R. Foster, and Y. Hao, "Detecting vital signs with wearable wireless sensors," *Sensors*, vol. 10, no. 12, pp. 10 837–10 862, 2010. Cited 2 times in pages 2 e 8.

[12] K. Tehrani and A. Michael, "Introduction to Wearable Technology," 2017, [online] http://www.wearabledevices.com/what-is-a-wearable-device (accessed date: Oct. 2017). Cited 2 times in pages 2 e 8.

[13] P. Resque, S. Pinheiro, D. Rosário, E. Cerqueira, A. Vergutz, M. Nogueira, and A. Santos, "Assessing data traffic classification to priority access for wireless healthcare application," in *Latin-American Conference on Communications (LATINCOM)*. IEEE, 2019, pp. 1–6. Cited in page 2.

[14] Z. Zhu, T. Liu, G. Li, T. Li, and Y. Inoue, "Wearable sensor systems for infants," *Sensors*, vol. 15, no. 2, pp. 3721–3749, 2015. Cited 2 times in pages 3 e 9.

[15] P. Resque, A. Barros, D. Rosário, and E. Cerqueira, "An investigation of different machine learning approaches for epileptic seizure detection," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 2019, pp. 301–306. Cited in page 3.

[16] Q. Do, B. Martini, and K.-K. R. Choo, "Is the data on your wearable device secure? An Android Wear smartwatch case study," *Software: Practice and Experience*, vol. 47, no. 3, pp. 391–403, 2017. Cited 2 times in pages 3 e 8.

[17] P. J. Soh, G. A. Vandenbosch, M. Mercuri, and D. M.-P. Schreurs, "Wearable wireless health monitoring: Current developments, challenges, and future trends," *IEEE Microwave Magazine*, vol. 16, no. 4, pp. 55–70, 2015. Cited 2 times in pages 3 e 8.

[18] W. Zhou and S. Piramuthu, "Security/privacy of wearable fitness tracking IoT devices," in *Proceedings of the 9th Iberian Conference on Information Systems and Technologies (CISTI 2014)*. IEEE, 2014, pp. 1–5. Cited 2 times in pages 3 e 8.

[19] J. THOMPSON, "U.s. smartwatch sales are soaring," *hodinkee*, 2019. [Online]. Available: https://www.hodinkee.com/articles/us-smartwatch-sales-are-soaring-2019 Cited in page 4.

[20] M. L. Hale, K. Lotfy, R. F. Gamble, C. Walter, and J. Lin, "Developing a platform to evaluate and assess the security of wearable devices," *Digital Communications and Networks*, vol. 5, no. 3, pp. 147 – 159, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352864817302985 Cited in page 4.

[21] G. Davis, "Key findings from our survey on identity theft, family safety and home network security," *McAfee*, 2018. [Online]. Available: https://www.mcafee.com/blogs/consumer/key-findings-from-our-survey-on-identity-theft-family-safety-and-home-network-security/ Cited in page 4.

[22] J. Ribeiro Pinto, J. S. Cardoso, and A. Lourenço, "Evolution, current challenges, and future possibilities in ecg biometrics," *IEEE Access*, vol. 6, pp. 34 746–34 776, 2018. Cited in page 5.

[23] L. O'Donnell, "Researchers bypass apple faceid using biometrics 'achilles heel'," *Threatpost website*, 2019. [Online]. Available: https://threatpost.com/researchers-bypass-apple-faceid-using-biometrics-achilles-heel/147109/ Cited in page 5.

[24] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas, "Biometric authentication based on pcg and ecg signals: Present status and future directions," *Signal, Image Video Process.*, vol. 8, no. 4, pp. 739–751, May 2014. Cited 2 times in pages 5 e 11.

[25] N. Belgacem, R. Fournier, A. Nait-Ali, and F. Bereksi-Reguig, "A novel biometric authentication approach using ecg and emg signals," *Journal of medical engineering & technology*, vol. 39, no. 4, pp. 226–238, 2015. Cited 3 times in pages 5, 9 e 11.

[26] N. C. for Chronic Disease Prevention, D. f. H. D. Health Promotion, and S. Prevention, "Heart disease facts," *CDC*, 2019. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm Cited in page 5.

[27] A. Barros, D. Rosário, P. Resque, and E. Cerqueira, "Heart of iot: Ecg as biometric sign for authentication and identification," in *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, June 2019, pp. 307–312. Cited 3 times in pages 6, 37 e 56.

[28] I. Odinaka, P.-H. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh, "Ecg biometric recognition: A comparative analysis," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1812–1824, 2012. Cited 4 times in pages 6, 14, 15 e 30.

[29] E. A. Cudney, S. M. Corns, and S. K. Long, "Improving knowledge sharing in healthcare through social network analysis," *International Journal of Collaborative Enterprise*, vol. 4, no. 1-2, pp. 17–33, 2014. Cited in page 8.

[30] E. Omanović-Mikličanin, M. Maksimović, and V. Vujović, "The Future of Healthcare: Nanomedicine and Internet of Nano Things," *Folia Medica Facultatis Medicinae Universitatis Saraeviensis*, vol. 50, no. 1, 2015. Cited in page 8.

[31] A. Lumini and L. Nanni, "Overview of the combination of biometric matchers," *Information Fusion*, vol. 33, pp. 71 – 85, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253516300446 Cited 4 times in pages 9, 10, 11 e 56.

[32] J. Unar, W. C. Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects," *Pattern Recognition*, vol. 47, no. 8, pp. 2673 – 2688, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031400034X Cited 2 times in pages 9 e 10.

[33] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, pp. 80 – 105, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865515004365 Cited 2 times in pages 10 e 24.

[34] M. Chan, D. Estève, J.-Y. Fourniols, C. Escriba, and E. Campo, "Smart wearable systems: Current status and future challenges," *Artificial Intelligence in Medicine*, vol. 56, no. 3, pp. 137 – 156, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365712001182 Cited in page 10.

[35] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ecg analysis," *Biomedical Signal Processing and Control*, vol. 43, pp. 216 – 235, 2018. Cited 5 times in pages 11, 12, 13, 14 e 15.

[36] R. Hoekema, G. J. H. Uijen, and A. van Oosterom, "Geometrical aspects of the interindividual variability of multilead ecg recordings," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 551–559, May 2001. Cited in page 11.

[37] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "Ecg to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133 – 142, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320304002419 Cited in page 11.

[38] T. W. Shen, W. J. Tompkins, and Y. H. Hu, "One-lead ecg for identity verification," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, vol. 1, 2002, pp. 62–63 vol.1. Cited in page 11.

[39] M. Z. U. Rahman, R. A. Shaik, and D. R. K. Reddy, "Efficient sign based normalized adaptive filtering techniques for cancelation of artifacts in ecg signals: Application to wireless biotelemetry," *Signal Processing*, vol. 91, no. 2, pp. 225–239, 2011. Cited in page 12.

[40] A. Zaknich, "Principles of adaptive filters and self-learning systems," *Science & Business Media - Springer*, 2006. Cited in page 13.

[41] Y. Zou, J. Han, X. Weng, and X. Zeng, "An ultra-low power qrs complex detection algorithm based on down-sampling wavelet transform," *IEEE Signal Processing Letters*, vol. 20, pp. 515–518, 2013. Cited in page 13.

[42] L. Bastos, D. Ros, E. Cerqueira, A. Santos, M. Nogueira *et al.*, "Filtering parameters selection method and peaks extraction for ecg and ppg signals," in *2019 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE, 2019, pp. 1–6. Cited in page 13.

[43] E. Castillo, D. Morales, G. Botella, A. GarcÃa, L. Parrilla, and A. Palma, "Efficient wavelet-based ecg processing for single-lead fhr extraction," *Digital Signal Processing*, vol. 23, no. 6, pp. 1897 – 1909, 2013. Cited in page 14.

[44] Y. Özbay, "A new approach to detection of ecg arrhythmias: Complex discrete wavelet transform based complex valued artificial neural network," *Journal of Medical Systems*, vol. 33, no. 6, p. 435, Sep 2008. [Online]. Available: https://doi.org/10.1007/s10916-008-9205-1 Cited in page 14.

[45] K. Balasundaram, S. Masse, K. Nair, and K. Umapathy, "A classification scheme for ventricular arrhythmias using wavelets analysis," *Medical & Biological Engineering & Computing*, vol. 51, no. 1, pp. 153–164, Feb 2013. [Online]. Available: https://doi.org/10.1007/s11517-012-0980-y Cited in page 14.

[46] S. Karpagachelvi, M. Arthanari, and M. Sivakumar, "ECG feature extraction techniques - A survey approach," *CoRR*, vol. abs/1005.0957, 2010. Cited in page 15.

[47] W.-M. Lee, "Python® machine learning," *John Wiley Sons*, vol. First Edition, 2019. Cited 2 times in pages 16 e 17.

[48] A. Géron, "Hands-on machine learning with scikit-learn, keras, and tensorflow concepts, tools, and techniques to build intelligent systems," *O'Reilly Media*, vol. 2nd Edition, 2019. Cited 4 times in pages 20, 21, 22 e 23.

[49] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiological signals," *circulation*, vol. 10, no. 23, pp. e215–e220, 2000. Cited in page 26.

[50] Y. Zhang and J. Wu, "Practical human authentication method based on piecewise corrected electrocardiogram," in *Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on*. IEEE, 2016, pp. 300–303. Cited 2 times in pages 27 e 31.

[51] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals." *Circulation 101(23):e215- e220*, June 2000. Cited 4 times in pages 27, 32, 34 e 50.

[52] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, "Pea: Parallel electrocardiogram-based authentication for smart healthcare systems," *Journal of Network and Computer Applications*, vol. 117, pp. 10 – 16, 2018. Cited 2 times in pages 27 e 31.

[53] Q. Zhang, D. Zhou, and X. Zeng, "Heartid: A multiresolution convolutional neural network for ecg-based biometric human identification in smart health applications," *IEEE Access*, vol. 5, pp. 11 805–11 816, 2017. Cited 5 times in pages 27, 28, 29, 31 e 37.

[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014. Cited in page 29.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. Cited in page 29.

[56] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, June 2005. Cited 2 times in pages 33 e 43.

[57] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, May 2001. Cited in page 34.

[58] L. T.S., "Biometric human identification based on electrocardiogram," *Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation*, vol. Master's thesis, June 2005. Cited in page 34.

[59] A. in February/2020, "Cross-validation: evaluating estimator performance," *Scikit Learn Documantation*, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html Cited 2 times in pages 41 e 42.

[60] ——, "1.11. ensemble methods," *Scikit Learn Documantation*, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html Cited in page 45.

[61] E. CERQUEIRA, P. RESQUE, I. Medeiros, L. Bastos, A. Santos, T. Tavares, D. Rosário, A. Santos, and M. Nogueira, "Autenticação usando sinais biométricos: Fundamentos, aplicações e desafios." in *38 Jornada de Atualização em Informática (JAI) do XXXIX Congresso da Sociedade Brasileira de Computação (CSBC 2019)*. SBC, June 2019, pp. 149–195. [Online]. Available: http://dx.doi.org/10.5753/sbc.471.7.04 Cited in page 56.