Cardiovascular information for improving biometric recognition

by

Paloma Tirado Martín

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Electric, Electronic and Automation Engineering

Universidad Carlos III de Madrid

Advisor(s):

Raul Sánchez Reillo
Judith Liu Jiménez

<Defense Month>

A mis padres, mi hermano y mi abuela.

ACKNOWLEDGEMENTS

Mentiría si dijera que no llevo muchísimo tiempo imaginando y deseando el momento de escribir estas palabras, y por fin ha llegado. Inicialmente, dar las gracias a Raúl por ofrecerme esta oportunidad, y a Judith por tus charlas esperanzadoras.

No ha sido fácil, mi amigo el síndrome del impostor aún no se ha ido, y para colmo nos tocó una pandemia mundial que hizo más solitarios estos dos últimos años de tesis. Sin embargo he tenido la suerte de contar con el apoyo de mis padres, a los que he dedicado mi TFG, mi TFM y ahora mi tesis doctoral. ¿A quién se lo iba a dedicar si no? Si estoy aquí es por vosotros, por vuestra paciencia, vuestro cariño y la confianza que siempre habéis depositado en nosotros. Por hacerlo siempre lo mejor que podéis, y porque tiene mérito haber aguantado a dos doctorandos en casa, con lo monotemáticos que nos ponemos. También he tenido la suerte de contar con mi hermano, que es una de las razones por las que llegué a plantearme este camino. Me has ayudado a aprender desde mi primer bucle for, hasta cómo enfocar problemas más enrevesados de mi tesis, técnicos y no tan técnicos. Contar contigo ha sido una gran baza que no todo el mundo tiene la suerte de tener, porque lo de pasar por esto no lo entiende cualquiera, y tener tu ayuda me ha sacado a flote más de una vez. También se lo dedico a mi abuela, porque sé la ilusión que te habría hecho poder leer esto y verme arreglada hablando en inglés delante de mucha gente.

Cómo no mencionar a Lidia, Irene y Laura. Por mis chapas, los audios disponibles en Spotify, y mis crisis existenciales por videollamada. Por alegraros casi tanto como yo de saber que esto está finiquitado, por estar siempre ahí y saber que puedo contar con vosotras. A mis compañeros de LoL, que hicieron que no me diese tanta pena quedarme confinada en casa, y en especial a Miguel, por escucharme estos últimos meses y aguantar estoicamente. A todos los gutitos con los que compartí despacho, desayunos, charlas de lingüística y conversaciones aleatorias. Por esos viajes por Corea, Japón y Montreal, que no habrían sido lo mismo sin vosotros.

PUBLISHED AND SUBMITTED CONTENT

- P. Tirado-Martin, R. Sanchez-Reillo, "BioECG: Improving ECG Biometrics with Deep Learning and Enhanced Datasets,". *Applied Sciences*, vol. 11, no. 13, 2021[1].
  - Role: Conceptualization, methodology, software, validation, formal analysis, investigation, writing, review and editing.
  - Wholly included in the thesis.
  - Chapters 5 and 9.
  - The material from this source included in this thesis is not singled out with typographic means and references.
- P. Tirado-Martin, J. Liu-Jimenez, Sanchez-Casanova, and R. Sanchez-Reillo, "QRS Differentiation to Improve ECG Biometrics under Different Physical Scenarios Using Multilayer Perceptron," *Applied Sciences*, vol. 10, no. 19, 2020 [2].
  - Role: Conceptualization, methodology, software, validation, formal analysis, investigation, writing, review and editing.
  - Wholly included in the thesis.
  - Chapters 5 and 8.
  - The material from this source included in this thesis is not singled out with typographic means and references.
- P. Tirado-Martin, R. Sanchez-Reillo, and J. Park, "Effects of Data Reduction when using Gaussian Mixture Models in Unidimensional Biometric Signals," in *2018 International Carnahan Conference on Security Technology (ICCST)*, 2018, pp. 1-5 [3].
  - Role: Conceptualization, methodology, software, validation, formal analysis, investigation, writing, review and editing.
  - Wholly included in the thesis.
  - Chapter 6.
  - The material from this source included in this thesis is not singled out with typographic means and references.

OTHER RESEARCH MERITS

- P. Tirado-Martin, R. Blanco-Gonzalo, A. Alvarez-Nieto, and A. Romero-Diaz, "Image processing techniques for improving vascular hand biometrics," in *2017 International Carnahan Conference on Security Technology (ICCST)*, 2017, pp. 1-5 [4].
  - Role: Formal analysis, investigation, writing, review and editing.
- J. Sanchez-Casanova, J. Liu-Jimenez, P. Tirado-Martin, and Sanchez-Reillo, "Unsupervised and scalable low train pathology detection system based on neural networks," *Heliyon*, vol. 7, no. 2, e06270, 2021 [5].
  - Role: Conceptualization and review.
- P. Fernandez-Lopez, J. Sanchez-Casanova, P. Tirado-Martín and J. Liu-Jimenez, "Optimizing resources on smartphone gait recognition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 31-36 [6].
  - Role: Conceptualization and review.

ABSTRACT

The improvements of the last two decades in data modeling and computing have lead to new biometric modalities. The Electrocardiogram (ECG) modality is part of them, and has been mainly researched by using public databases related to medical training. Despite of being useful for initial approaches, they are not representative of a real biometric environment. In addition, publishing and creating a new database is none trivial due to human resources and data protection laws.

The main goal of this thesis is to successfully use ECG as a biometric signal while getting closer to the real case scenario. Every experiment considers low computational calculations and transformations to help in potential portability. The core experiments in this work come from a private database with different positions, time and heart rate scenarios. An initial segmentation evaluation is achieved with the help of fiducial point detection which determines the QRS selection as the input data for all the experiments.

The approach of training a model per user (open-set) is tested with different machine learning algorithms, only getting an acceptable result with Gaussian Mixture Models (GMM). However, the concept of training all users in one model (closed-set) shows more potential with Linear Discriminant Analysis (LDA), whose results were improved in 40%. The results with LDA are also tested as a multi-modality technique, decreasing the Equal Error Rate (EER) of fingerprint verification in up to 70.64% with score fusion, and reaching 0% in Protection Attack Detection (PAD).

The Multilayer Perceptron (MLP) algorithm enhances these results in verification while applying the first differentiation to the signal. The network optimization is achieved with EER as an observation metric, and improves the results of LDA in 22% for the worst case scenario, and decreases the EER to 0% in the best case. Complexity is added creating a Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) based network, BioECG. The tuning process is achieved without extra feature transformation and is evaluated through accuracy, aiming for good identification. The inclusion of a second day of enrollment in improves results from MLP, reaching the overall lowest results of 0.009%–1.352% in EER.

Throughout the use of good quality signals, position changes did not noticeably impact the verification. In addition, collecting data in a different day or in a different hour did not clearly affect the performance. Moreover, modifying the verification process based on attempts, improves the overall results, up to reach a 0% EER when applying BioECG.

Finally, to get closer to a real scenario, a smartband prototype is used to collect new databases. A private database with limited scenarios but controlled data, and another local database with a wider range of scenarios and days, and with a more relaxed use of the device. Applying the concepts of first differentiation and MLP, these signals required

the Stationary Wavelet Transform (SWT) and new fiducial point detection to improve their results. The first database gave subtle chances of being used in identification with up to 78.2% accuracy, but the latter was completely discarded for this purpose. These realistic experiments show the impact of a low fidelity sensor, even considering the same modifications in previous successful experiments with better quality data, reaching up to 13.530% EER. In the second database, results reach a range of 0.068%–31.669% EER. This type of sensor is affected by heart rate changes, but also by position variations, given its sensitivity to movement.

RESUMEN

Las mejoras en modelado de datos y computación de las últimas dos décadas, han llevado a la creación de nuevas modalidades biométricas. La modalidad de electrocardiograma (ECG) es una de ellas, la cual se ha investigado usando bases de datos públicas que fueron creadas para entrenamiento de profesional médico. Aunque estos datos han sido útiles para los estados iniciales de la modalidad, no son representativos de un entorno biométrico real. Además, publicar y crear bases de datos nuevas son problemas no triviales debido a los recursos humanos y las leyes de protección de datos.

El principal objetivo de esta tesis es usar exitosamente datos de ECG como señales biométricas a la vez que nos acercamos a un escenario realista. Cada experimento considera cálculos y transformadas de bajo coste computacional para ayudar en su potencial uso en aparatos móviles. Los principales experimentos de este trabajo se producen con una base de datos privada con diferentes escenarios en términos de postura, tiempo y frecuencia cardíaca. Con ella se evalúan las diferentes selecciónes del complejo QRS mediante detección de puntos fiduciales, lo cual servirá como datos de entrada para el resto de experimentos.

El enfoque de entrenar un modelo por usuario (open-set) se prueba con diferentes algoritmos de aprendizaje máquina (machine learning), obteniendo resultados aceptables únicamente mediante el uso de modelos de mezcla de Gaussianas (Gaussian Mixture Models, GMM). Sin embargo, el concepto de entrenar un modelo con todos los usuarios (closed-set) demuestra mayor potencial con Linear Discriminant Analysis (Análisis de Discriminante Lineal, LDA), cuyos resultados mejoran en un 40%. Los resultados de LDA también se utilizan como técnica multi-modal, disminuyendo la Equal Error Rate (Tasa de Igual Error, EER) de la verificación mediante huella en hasta un 70.64% con fusión de puntuación, y llegando a un sistema con un 0% de EER en Detección de Ataques de Presentación (Presentation Attack Detection, PAD).

El algoritmo de Perceptrón Multicapa (Multilayer Perceptron, MLP) mejora los resultados previos en verificación aplicando la primera derivada a la señal. La optimización de la red se consigue en base a su EER, mejora la de LDA en hasta un 22% en el peor caso, y la lleva hasta un 0% en el mejor caso. Se añade complejidad creando una red neural convolucional (Convolutional Neural Network, CNN) con una red de memoria a largo-corto plazo (Long-Short Term Memory, LSTM), llamada BioECG. El proceso de ajuste de hiperparámetros se lleva acabo sin transformaciones y se evalúa observando la accuracy (precisión), para mejorar la identificación. Sin embargo, incluir un segundo día de registro (enrollment) con BioECG, estos resultados mejoran hasta un 74% para el peor caso, llegando a los resultados más bajos hasta el momento con 0.009%–1.352% en la EER.

Durante el uso de señales de buena calidad, los cambios de postura no afectaron

notablemente a la verificación. Además, adquirir los datos en días u horas diferentes tampoco afectó claramente a los resultados. Asimismo, modificar el proceso de verificación en base a intentos también produce mejoría en todos los resultados, hasta el punto de llegar a un 0% de EER cuando se aplica BioECG.

Finalmente, para acercarnos al caso más realista, se usa un prototipo de pulsera para capturar nuevas bases de datos. Una base de datos privada con escenarios limitados pero datos más controlados, y otra base de datos local con más espectro de escenarios y días y un uso del dispositivo más relajado. Para estos datos se aplican los conceptos de primera diferenciación en MLP, cuyas señales requieren la Transformada de Wavelet Estacionaria (Stationary Wavelet Transform, SWT) y un detector de puntos fiduciales para mejorar los resultados. La primera base de datos da opciones a ser usada para identificación con un máximo de precisión del 78.2%, pero la segunda se descartó completamente para este propósito. Estos experimentos más realistas demuestran el impact de tener un sensor de baja fidelidad, incluso considerando las mismas modificaciones que previamente tuvieron buenos resultados en datos mejores, llegando a un 13.530% de EER. En la segunda base de datos, los resultados llegan a un rango de 0.068%–31.669% en EER. Este tipo de sensor se ve afectado por las variaciones de frecuencia cardíaca, pero también por el cambio de posición, dado que es más sensible al movimiento.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In the last two decades, several fields such as mathematics and computing have experienced huge advances. The interdisciplinary efforts done between disciplines such as neural networking and high performance hardware, have expanded the limits of existing machine learning solutions. These advances have been the basis of the rising of new biometric traits.

While conventional modalities are based on features that do not change with time, such as fingerprints, iris or even DNA, the most recent modalities have a tendency to focus on less obvious patterns than those in a fingerprint image, yet can be used to distinguish one individual from the rest. The nature of these patterns and their complexity make them difficult to forge or replicate, and add convenience depending on the application.

The field of ECG biometrics has been present since then, in the early 2000s, and has been proven as a good biometric signal. However, the lack of standardization and difficulties in publishing databases complicate the replication of real environments. In this thesis we present an attempt to recreate these environmental conditions, going from low to more complex situations and solutions. Moreover, we observe the differences and consequences that changing the heart-rate, position, or collection device have in recognition.

The main core of the thesis is achieved with a private database. The data collection was achieved with a professional ECG device, allowing to assume that the obtained results only depend on the process, and not on the quality of the data. This database has a variety of scenarios, representing cases from sitting down, standing and after exercise, including collections in different days. These characteristics give a closer representation to real case scenario that has not been previously observed in literature.

Once the main observations are achieved with this database, two new databases are introduced. These databases get closer to the real scenario by using a smartband prototype, adding a level of uncertainty to the results, as they could be affected by the signal quality. In this case, the databases differ on the collection protocol, adding the assessment of how controlled the collection must be to get acceptable results.

This thesis begins with an introduction to biometrics in chapter 2, in order to further understand biometrics generally, followed by 3.1 which discusses the characteristics and advantages of ECG biometrics. Then, chapter 4 deeply describes the applied databases, as well as the prototype components. The following chapter 5 explains and details the different tools and techniques that were needed throughout the different experiments. The first experiments are collected in chapter 6, where we focus on the suitability of the main database in recognition, and extended the final results. In chapter 7, we show the capabilities of this modality to improve fingerprint verification. The initial verification

results get improved in chapter 8 by using more complex classification, while assessing different approaches and observing their behavior. Chapter 9 collects the experiments related to the use of Deep Learning for recognition purposes, as well as the evaluation of the effects of including an extra day of enrollment. The last chapter regarding experiments is chapter 10, which applied the obtained knowledge from previous chapters to the case of using a smartband device in recognition. Finally, the last chapter 11 summarizes the conclusions of this thesis and includes extra considerations for future works.

2. INTRODUCTION TO BIOMETRICS

The words "biometrics" or "biometry" are defined in the dictionary as *the process by which a person's unique physical and other traits are detected and recorded by an electronic device or system as a mean of confirming identity*. Therefore, the biometric trait or modality refers to the characteristic that allows human recognition. Based on these traits, the recording and feasibility vary.

2.1. Biometric signals, applications and modalities

Initially, biometrics were mainly used in criminal investigations. However, nowadays the improvements in the involved technologies have spread the usage of biometric traits to other fields such as forensic purposes (i.e.: body identification, criminal investigation, kinship determination), governmental applications (i.e.: identity documents, border control) and commercial transactions (i.e.: ATMs, building access control, phone locking, online payments).

Every physiological or behavioral characteristic in humans has the potential of being used as a biometric trait. The physiological traits collect all those characteristics that are intrinsic to the human body, whereas the latter consist on the different ways to do common actions. To be suitable for biometrics, they need to fulfill specific requirements: they need to be present in every human (universality) but different enough among individuals (uniqueness). The characteristic must be permanent over time with respect to the matching criterion (permanence) and possible to measure quantitatively (collectability) [7]. Fingerprint, face, hand geometry, palm print, vein, iris, ear, ECG and DNA are physiological modalities, while dynamic and static signature, gait, voice or keystroke form part the behavioral group.

In addition, when designing a biometric system, the chosen trait must be considered based on the cost of collecting that data and its convenience, as well as the environment the system would take part on. These and other factors can impact how easily accepted the system is in people's daily lives (acceptability), how difficult is to be forged or accessed by fraudulent users (circumvention) and how fast and accurate the system is (performance) [7]:

Even though some characteristics can be used in biometrics, not all of them have been equally researched or reached to the same performances. In Table 2.1 there is a comparison on the most widely spread biometric traits with a comparison among their different requirements:

Table 2.1: Comparison of different biometric traits according to [7]. H: High. M: Medium. L: Low.

	Universality	**Distinctiveness**	**Permanence**	**Collectability**	**Performance**	**Acceptability**	**Circumvention**
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand geometry	M	M	M	H	M	M	M
Hand vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Palmprint	M	H	H	M	H	M	M
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

2.2. Biometric systems

Biometric systems and their different approaches are countless. However, their software usually follows similar stages for their performance. There are different schemes in literature to represent these stages, which vary according to the procedure: enrollment, verification or identification. These three schemes are represented independently in [7], and have been merged and modified accordingly to summarize in one represented in Figure 2.1.

The three processes require the user-sensor interaction to collect the raw data. The data is processed with heterogeneous tools in the feature extraction stage, to finally achieve a reference that represents the user's information. However, the differences in each path aims to represent follows:

- Enrollment: the goal of this stage is providing the reference information. From this data, there are two approaches:
  - Distances: the features are stored in the references for further distance comparison. Each user has their own reference. The number or type of

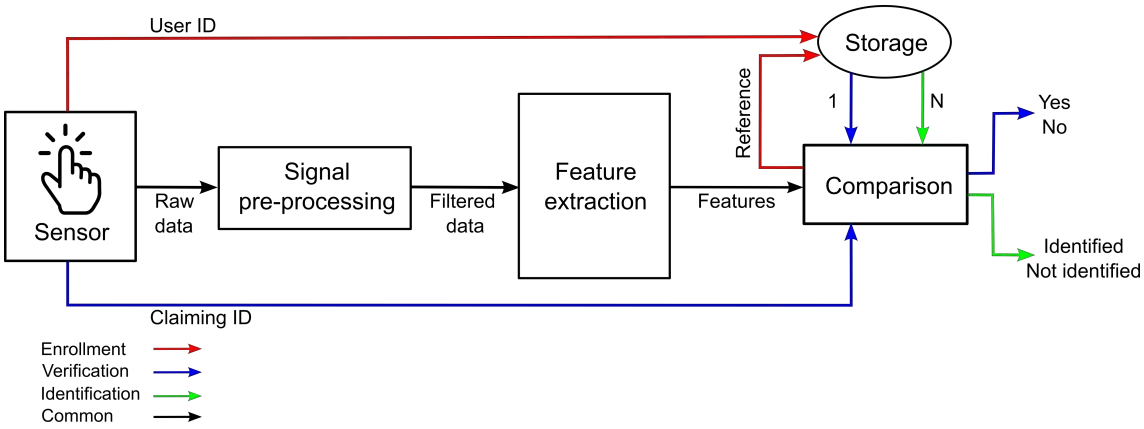


Figure 2.1: Main stages for a biometric system in enrollment, verification and identification. Arrows coming out of storage are retrieved references, whereas 1 indicates only one retrieved reference and N refers to all the stored references.

references for one user is defined by design.

  - Models: the features are applied as training data to create a model. If only one model characterizes the whole set of enrolled users, it is called closed-set, as it does not allow new users to come in without re-training the model. If, on the contrary, every user has their own model, the system is open-set, as enrolling new users in the system would not require extra changes.

- Recognition: analogously to the enrollment, the recognition process needs to specify how many samples are part of a recognition attempt. This parameter is determined by design or requested by the system on-the-go. When more than one sample is considered as part of one attempt, the criteria for the decision needs to be properly designed to enhance the performance. There are two types of recognition:

  - Verification: confirms whether or not the user is who he/she claim to be in the provided ID. This procedure only requires the retrieval of one reference from the database, or only one comparison for the model. This process summarizes into a binary decision: valid or not.
  - Identification: searches for the match (if any) of the user's reference, so the ID information is not required. The reference retrieval consists of as many references as users are enrolled, carrying out a comparison for each one of them. The final decision is an ID assignation or no assignation if the system considers that the user's data is not enrolled.

2.2.1. Sensor

Every biometric system starts with a sensor that collects the biometric data of the user. These sensors are specific to every biometric trait, and need to be fast and feasible in the acquisition. Data should preferably be consistent independently of the circumstances of

its collection. In addition, to improve its acceptability, the sensor should be easy to use by the average user of the system.

Depending on the modality, the sensor would have a specific degree of intrusiveness. In modalities such as iris, the user takes part on more inconvenient recognition processes. On the contrary, fingerprint or face biometrics result in very easy and non-intrusive collections, obtaining more acceptability.

These issues are often related to the trade-off between security and convenience. In very secure environments is preferable to use inconvenient but more feasible processes, such as those involved in iris. However, in scenarios where convenience is more relevant than security, such as phone unblocking, the recognition process needs to be as simple as the one in fingerprint, which provides slightly less security than iris, but is compensated with a non-intrusive sensor.

2.2.2. Signal pre-processing and feature extraction

The pre-processing block optimizes the signal for the following procedures. The main goal is to increase the Signal-to-Noise Ratio (SNR), providing more information about the collected data and getting rid of unwanted information. This stage is usually formed by frequency filters or more sophisticated systems to detect more complex noise in the signal.

The feature extraction module has the objective of obtaining signal features that are discriminant and help distinguishing the individual from any other potential user. The goal in this stage of a biometric system is to facilitate the next comparison process. The feature extraction stage can have some or all the following steps:

- Filtering: average, median or frequency filters.
- Transformations: Wavelets, Fourier Transform, thresholding.
- Feature detection: location of specific features such as minutiae in fingerprint or face feature location for face biometrics.
- Feature selection: discarding of the least relevant features to reduce data and avoid mistakes in classification.

2.2.3. Comparison

Up to this stage, enrollment and recognition data are affected by the same procedures, with no difference. However, when proceeding to compare, the data gets affected differently. For distances in enrollment, the obtained reference/s go to storage with the corresponding user's ID. In the case of modeling, the data is modeled with the remaining

users (closed-set) or independently (open-set) and the model is stored. These metrics are further referred in ISO 19795 [8].

Each attempt in recognition is specified in design, as they can be formed by one or more samples, dealing with the distances or scores in different ways. For identification, the new data gets compared against the reference data from all the users already enrolled in the system. Some identification systems require meeting a given criterion like a threshold; others consider a given number of best compared candidates regardless of threshold. Usually, the valid decision is determined by the reference ID that results in the highest similarity or lowest distance. To characterize an identification system, the False Negative Identification Rate (FNIR) and False Positive Identification Rate (FPIR) are usually obtained based on the threshold. The rate of correct identifications is given by the accuracy in Equation (2.1).

$$Accuracy = \frac{Num.of correct identifications}{Num.of comparisons} \tag{2.1}$$

In the case of verification, data is only compared against a given ID previously specified, as represented in red arrows in Figure 2.1. As the corresponding result is obtained after comparing against one user, there are no comparative criteria to take the decision like it happens in identification. This issue is solved by using a threshold value that determines whether the sample is valid or not. The two types of errors are represented in rates, where the false negative rate is the False Non-Match Rate (FNMR) and the false positive rate is the False Match Rate (FMR). The system's performance has analogous metrics, False Accept Rate (FAR) and False Rejection Rate (FRR), that follow Equations (2.2 and 2.3). These equations depend on both metrics FNMR and FMR, but also Failure-to-Acquire Rate (FTAR). When the FTAR is unknown or not considered, FRR = FNMR and FAR = FMR. These values vary depending on the selected threshold value, providing as many results as values are given to the threshold. The minimum error is given when FNMR = FMR, which is called the Equal Error Rate (EER). These results can be summarized in graphs of FNMR vs FMR or in Detection Error Trade-off graphs, based on probabilities as showed in Figure 2.2. However, the final threshold for the real system is determined in design based on these evaluations. Depending on the application of the system, the trade-off between FNMR and FMR needs to be selected. Having very discriminant thresholds may lead to more attempts that result inconvenient, and on the contrary it may result in a very inefficient result.

$$FAR = FMR(1 - FTAR) \tag{2.2}$$

$$FRR = FTAR + FNMR(1 - FTAR) \tag{2.3}$$

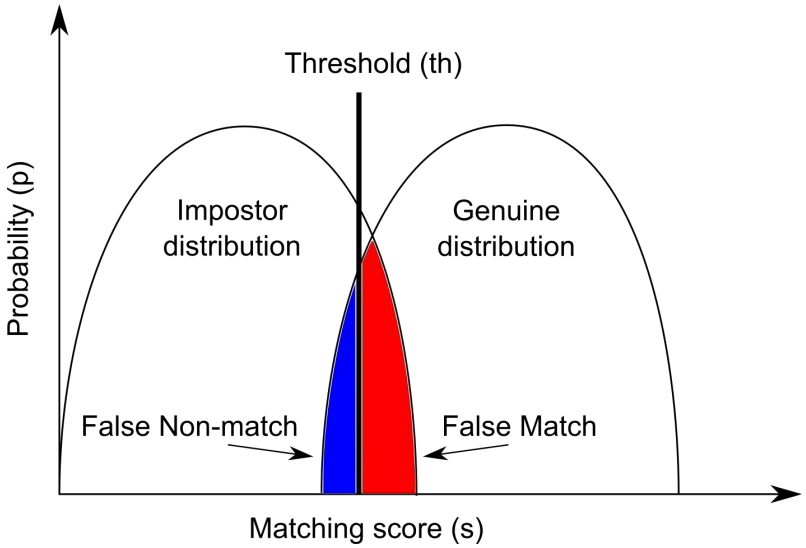


Figure 2.2: Error rates in a biometric system as given in [7].

3. ECG AND ITS USE AS A BIOMETRIC MODALITY

3.1. Introduction to the ECG

The complex functionality of the heart can be summarized as a periodical contraction that pumps the blood through the body. The blood gets oxygenated in the lungs and flows into the general circulation, providing the rest of the body with the required oxygen. This pumping activity propagates through the heart muscle cells, whose main function is generating and propagating electrical currents.

The cardiac impulse begins in the sinoatrial (SA) node and travels towards the atrioventricular (AV) node, producing the atrial depolarization resulting in the contraction of the heart. When the AV node is reached, there is a brief pause and the signal disseminates through the bundle of His, which is formed by left and right branches. The left branch propagates the impulse towards the ventricles through the Purkinje fibers, producing the ventricle contraction [9].

The Electrocardiogram (ECG) is a graph that represents this electrical heart activity with respect of the time. The most typical representation is formed by 3 waveforms which are represented in Figure 3.1 and have the following meaning:

- P wave: corresponds to the atrial depolarization and is the result of the superposition of both right and left atrium. Its repolarization is overlapped by the following QRS complex. The PR interval lasts 0.12 to 0.20 s [9].
- QRS complex: represents the current that causes the ventricular contraction or depolarization, and it is more noticeable in the ECG as it implies more voltage than the atrial depolarization. A normal QRS has a duration up to 0.12 s [10].
- T wave: belongs to the ventricular repolarization. The QT interval has a duration of 0.35 to 0.43 s.

3.1.1. Lead-based systems

The first ECG records were taken in by Willen Einthoven, in a work that was completely published in 1906 [11]. Einthoven modified the string galvanometer to measure the heart's electrical activity. The electrical currents were conducted through a short and thin silver-coated quartz filament between two electromagnets. The filament produced an electromagnetic field that was strong enough to move a string that could be captured in a photographic paper. He created the famous Einthoven's triangle, where leads I, II and III were defined. The 12-leads were derived from these points, forming the most

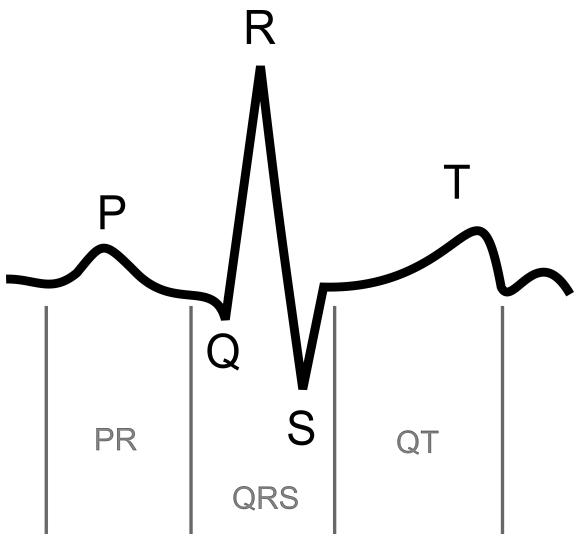


Figure 3.1: Scheme of an ECG waveform and its relevant intervals.

common technique for ECG acquisition for medical purposes. The different leads are useful because, even though the general ECG pattern is constant, some of its waveforms are transformed depending on where it is measured from. As a result, the ECG gets more precision as some heart issues get better reflected in certain leads. The angle acquisition is represented in Figure 3.2 and is classified into two types [12]:

- Limb leads: require four sensors on right arm (RA), left arm (LA), right leg (RL) and left leg (LL). RL behaves as a ground.
  - I, II and III: also called standard bipolar leads. They measure voltages in pairs LA-RA, LL-RA and LA-LL respectively.
  - aVR, aVL, and aVF: named after augmented unipolar leads. These lead do not require ground references, as they represent relative voltages with respect to the extremities.
- Chest leads (V1, V2, V3, V4, V5, and V6): also denominated precordial leads. Six sensors are placed on different parts of the chest, requiring very precise positioning with respect to the ribs. An extra sensor acts as a ground reference.

The Vectorcardiogram (VCG) was derived from the 12-leads and was highly researched between the 1950s to mid 1980s [14]. The Frank leads system is the most famous VCG acquisition. This system obtains 3 orthogonal leads: X, Y and Z with 8 sensors represented in Figure 3.3. Two of the sensors are positioned on the back, which is uncomfortable for ambulatory acquisition and one the reasons why this system has lost popularity in the last decades [14].

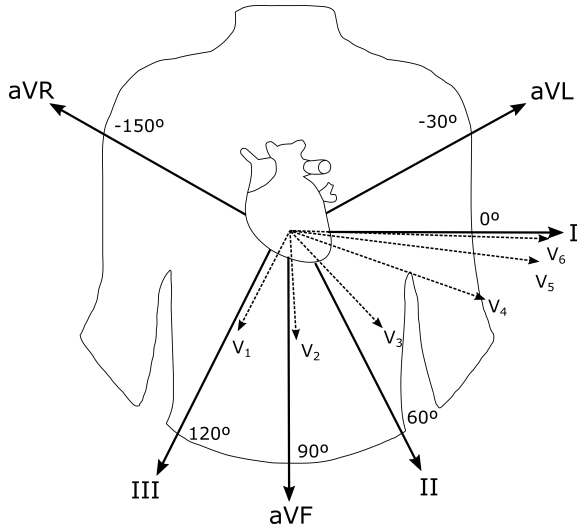


Figure 3.2: Representation of the 12-lead vectors. Solid arrows represent the limb leads and the discontinuous arrows correspond to the chest leads [13].

3.1.2. Medical monitoring

Even though the 12-lead is the most well-known type of ECG acquisition, these leads are limitless. Multichannel ECG (MECG) provides different leads that may differ from those in the 12-leads. Sensors are placed based on body mapping by using t-shirts, belts or vests. Both techniques are applied for short-term ECG monitoring which allows to observe conditions that are frequently present in the ECG. However, if the cardiac pathology is only observable in specific periods of time, the diagnostic could be done through ambulatory monitoring. These acquisitions usually last up to 24-48 hours and can be done out of the hospital, as they are usually carried out with a portable Holter monitor, represented in Figure 3.4. Nonetheless, these devices are uncomfortable for the user, as they require wires and wearing the device for the long data acquisition [16]. Alternatives for in-home and ambulatory recording have arisen due to the inconvenience of Holter monitors. Depending on the aim of the monitoring process, different devices are already commercialized to facilitate the process [17]:

- Patch ECG monitors: allow long-term recordings and are wireless. In the case of Zio AT patch [19], recordings can last up to 14 days with one lead and it provides alerts and reports that are transmitted daily to the doctor.
- External loop recorders (ELR): only record segments with specific duration fixed by the patient. The sensor system-on-chip by Imec [20] has 3 channels and it is implemented with instant transmission and can be worn up to 30 days.
- External event recorders: are activated by the patient after a symptomatic event. These type of monitors do not demand long lasting batteries or big storage capacity, as they do not provide continuous monitoring. Alivecor's Kardia [21] is one of the

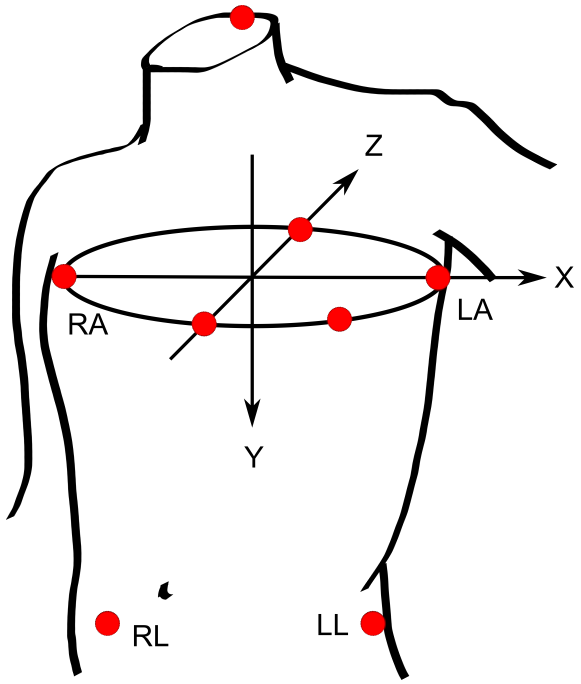


Figure 3.3: Representation of the Frank lead vectors and sensor placement. X goes from right arm to left arm. Y goes from neck to feet. Z goes from front to back [15].

most well-known monitors in this category, which requires an attachable sensor that communicates with a smartphone, in order to send and visualize data.

These devices must minimize the measured noise and its impact. The noise added when measuring the heart's activity with a sensor is produced by three main sources: baseline wanders and drift, power-line interference and muscle artifacts [22]. The baseline wander is related to the impedance variation between the electrode and the skin, which is a result of the user's breathing or subtle movements [23]. This type of noise translates into abrupt movements drastically affect the signal. Power-line interference ranges between 50 and 60 Hz whereas the baseline wander is usually between 0.2 and 0.5 Hz. Moreover, the hearts is surrounded by other muscles and also produce electrical pulses. This independent activity also propagates to the skin, adding undesired information. The muscles provide high frequency artifacts with frequencies above 100 Hz [24]. Other noises are originated by human mistakes, such as sensor displacement when interchanging electrodes, which produce reverse amplitudes; or not precised positioning in the chest leads, as the rib cage distortions the signal [25].

The ECG was first used as a biometric trait in the early 2000s [26]. For 20 years it has experimented a popularity increase provoked by the improvements in processing and computing techniques. This section discusses ECG signals' suitability and characteristics in the field of human recognition.

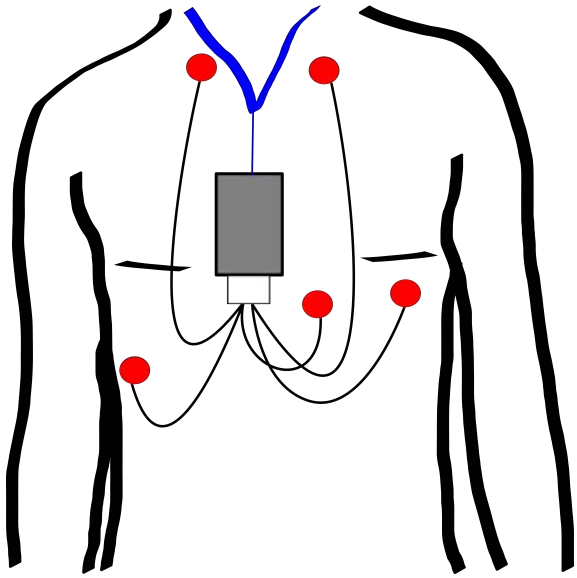


Figure 3.4: Representation of a Holter monitor with 5 sensors (represented in red). The gray area is the monitoring device, which is carried by the user for the whole monitoring period [18].

3.1.3. Suitability as a biometric trait

The monitoring of the heart's activity was first intended to be helpful in medical diagnosis. However, the nature of the ECG meets the requirements of a potential biometric signal. Every alive individual can produce an ECG signal thanks to the functioning heart. It can also be measured with non-intrusively and precisely with electrocardiographs. The information provided by the ECG is variant among individuals due to the different factors that take place in the electrical propagation of the heart's activity. Age, gender, weight and the shape of the chest are some factors that make each ECG unique [27]. Furthermore, the ECG has been proven as a long term stable signal up to several years [28].

3.1.4. Advantages and challenges

A biometric system based on ECG provides extra security: the user must be alive in order to produce the required electrical impulses, which complements the modality with an intrinsic life-proof detection. In addition, data is not easy to collect without the user's cooperation because it is not visible and requires specific sensors to be observed. Even if the biometric data is finally acquired without permission, its replication is an added issue which complicates forgery. However, this last characteristic also presents a challenge for correct recognition. The ECG has constant width and amplitude-wise transformations caused by the heart rate variations and other physiological scenarios, such as pathologies or the user specific conditions. These condition make more difficult to detect the repetitive pattern in every cycle, isolating the circumstantial variations.

3.2. State of the art

This section presents and discusses the approaches that can be found in the existing literature. They are divided according to the different stages involved in the process as indicated in Figure 2.1.

3.2.1. Sensors

Section 3.1 referred to the types of ECG acquisitions, which involve several sensors and a specific and precise placement. This requirements would not be user-friendly in a biometric recognition environment, as it requires expertise and time for the user. Therefore, it is convenient to simplify the sensor or capture device to only collect one lead, choosing the most representative one. The most used in biometrics is the type I lead, as it only involves sensor placement in both arms, involving fingers and/or wrists [29]–[32].

The device also must be precise enough to capture the P-QRS-T waves with a clear resolution. The fastest part of the cycle is the QRS complex, which usually takes 0.12 s, meanwhile P and T are slower waveforms. The selection of the sampling frequency does not gravitate to a clear value. Some approaches establish this value in 1 KHz [26], [33]–[35], whereas successful results have also been obtained under lower sample frequencies such as 300 Hz [36], 360 Hz [37] and even 125 Hz [38].

The sample frequency selection tends to select low frequencies in these type of applications as it directly impacts the processing and storing technology. Higher frequencies require more powerful hardware regarding those issues, and limit the performance in devices with lower capabilities such as mobile phones. This choice does not benefit approaches that focus en individual cycle (fiducial) characteristics, as the resolution is more relevant in these cases. However, global (non-fiducial) features are not as affected by the resolution. As mobile devices are proven to add noise and decrease the signal quality [39], selecting non-fiducial features would benefit the system in these cases, as they do not consider specific data points that may have gotten distorted.

The selection of capture devices in literature is heterogeneous, as the selection varies between using commercial devices or prototypes. The most common commercial devices are the following, which are divided into non-portable and portable devices:

- Non-portable devices:
  - BioPac MP150 [40], which is used for one of the databases in this thesis.
  - BioPac Remote Monitoring [41] for remote ECG acquisition.
  - Vernier EKG [42], which allows 3-lead ECG and surface EMG.
- Portable devices:

- AliveCor's Kardia [21] which operates on smartphones and does 30 s signal collection.
- ECG Savvy [43] is a long term event recorder with medical purposes.
- FitnessShirt [44], a vest with incorporated sensors for ECG an respiration recording.

Prototype designs are frequently based on disposable electrodes with default configuration. When referring to type I lead or its modifications, devices are employed using the thumbs, the wrists or hand palms. Even though these devices are not as optimized as the commercial devices, they are flexible and can be adapted to the point of being portable.

The recent increment in the number on works related to ECG biometrics have increased the interest on developing portable ECG devices, leading to commercial releases focused on mobile measurements. The ECG Check by Cardiac Designs [45] is formed by two plates to place the fingers of both hands, as observed in Figure 3.5, and connects to the smartphone through Bluetooth. The goal of this solution is monitoring the ECG to detect potential cardiac problems, but does not allow to extract the raw data that has been collected.

Figure 3.5: Example of use with ECG Check by Cardiac Designs [45].

Another commercial ECG capture device is the Nymi's band [46] in Figure 3.6, which is a wearable smartband whose objective is the user's recognition. The device gets linked to a smartphone or computer, allowing to unblock the device. As it happens with the ECG Check, this smartband does not allow to obtain the ECG or any parameter in the authentication system.

Figure 3.6: Nymi wearable smartband [46].

3.2.2. Existing databases

Some authors do not include data collection in their works, as it is a non-trivial procedure with complex logistics. Thanks to the publicly available ECG databases, the field of ECG biometrics was created and has been able to improve since then. These databases are collected in PhysioNet [47], which provides large collections of physiological and clinical data. The most well-known databases in literature are summarized in Table 3.1. These acquisitions are mainly characterized by long recordings, as the collection focuses in education about cardiovascular diseases.

Depending on the database, the collection reports miss some data in the database description, as the type of capture device or sensor used. Most of the information collects long and continuous recordings of the user, which invalidate the option of developing a recognition system. The option of validating these databases is also discarded, as well as comparing the obtained results with other specific databases, due to significant differences in the procedure.

The fact of considering users with cardiovascular diseases implies limitations when applied to biometric recognition. The common population are not affected by cardiovascular diseases, resulting in a lack of representation in these public databases. Moreover, there are possibilities that the system could only be detecting the pathology instead of the user.

As a relatively new biometric trait, ECG does not have quality standards to be used in human recognition. The only found work regarding the standardization of the quality assessment was done in [48] and has not been confirmed by the scientific community. Moreover, the public databases have very heterogeneous parameters and the results are not really comparable.

From all the public databases, there is only one specifically developed for human recognition purposes: the ECG-ID database [49]. However, this database is not homogeneous regarding the number of samples per user and the time of capture, which can vary between the same day to up to 6 months between visits. In addition, it does not include specific scenarios that could modify the heart rate.

Table 3.1: Summary of characteristics for some of the public ECG databases available. P: Prototype. C: Commercial. -: Unknown.

Database	Users	Records	Lead	Device	Duration	Notes
ST-T [50]	79	90	2 leads	-	2 h	Myocardial ischemia
L-T ST [51]	80	86	Lead combinations	C	21-24 h	ST segment changes
MIT-BIH [52]	47	48	Mostly lead II + V1	C	30 min	Mixed (40% arrhytmia)
PTB [53]	290	549	12-leads + Frank leads	P	38.4 - 104.2 s	Several pathologies
ECG-ID [49]	90	310	Lead I	-	20 s	No pathologies. Aimed for biometrics.

It is common to observe works that use the same public database. The MIT-BIH has been broadly used, but the number of users vary among works. Some authors used 20 users [54], [55] or 18 [56] and some cases apply the entire database [57] or mix it with others like PTB [58]. In fact, the PTB database is also a common selection in public databases. However, the number of subjects selected from it is also heterogeneous, ranging from 13 [59], 14 [35] and 20 [33] to 74 [28].

Custom databases are recurrent in literature, as they allow to focus on specific contexts. Some authors did the capturing using prototypes [54], [55], [60]–[63], others achieved it with commercial devices [38], [39], [64]–[67] and others selected both approaches [68], [69].

3.2.3. Signal pre-processing and feature extraction

The data heterogeneity in the published results is high, so there are not clear, common paths among different research works. However, there is a clear differentiation between two types of features: fiducial and non-fiducial.

Fiducial features are those based on detecting certain points in the signal regarding the cycle shape, i.e.: P wave or Q, R and S points. These points act as references for further calculations, such as amplitudes or temporal distances between them. On the contrary, non-fiducial features are based on the entire segment of data, and manipulates the information through specific transformations such as Wavelets or covariance matrices. These approaches are based on the assumption that all cycles have the same type of pattern.

Non-fiducial features usually result in better performances than fiducial features. However, some authors proved otherwise in [70], which is part of an earlier work in the field. The different results in these approaches show the lack of universality in the databases.

The earliest approaches in this biometric trait date from the early 2000s, and applied fiducial features. The features were based on different variations of correlations i.e.: correlation matrix [26], average correlation [71] or stepwise canonical correlation [54], [55]. Even though these correlation-based features did not lose presence in the following years, new techniques for feature extraction appeared. The Discrete Cosine Transform (DCT) is applied both for feature extraction and signal denoising [33], [35], [58], [70], [72]. Wavelet transforms have been similarly applied for signal conditioning and feature extraction [37], [73], [74]. Other approaches throughout literature regarding noise removal, Moving Average (MA) filters [75] and Convolutional Neural Networks (CNNs) [76] are two of the alternatives. However, the most recurrent tool is the band-pass filter [77].

In this stage, some works also apply data reduction and discriminant analysis such as Linear Discriminant Analysis (LDA) [54], [55], [59], Independent Component Analysis (ICA) [37] or Principal Component Analysis (PCA) [32].

3.2.4. Comparison

The comparison stage is not different in terms of the numerous alternatives there are in literature. The earlier approaches consisted on template comparison based on different distance metrics. One of the most recurrent ones is the Euclidean distance [28], [35], [54], [78] but Hamming and Wavelet distances are also considered [39], [60], [79].

Another common solution is based on data modeling. To achieve this, two algorithms have been the most selected by authors [28], [80]: Support Vector Machines (SVMs) [57], [62], [67], [81], [82] and k-Near Neighbors (k-NN) [38], [56], [58], [66], [82]. However, in the last decade, Artificial Neural Netowrks (ANNs) have increased their popularity [65]. One of the simplest algorithms in the field is the Multilayer Perceptron (MLP) [56], [65], [83], which evolved into more sophisticated architectures such as Deep Neural Networks (DNNs), which have the potential to solve problems in SVM and k-NN [77].

Since 2017 the research on ECG biometrics has leaned towards the use Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are applied in several ways, such as multiscale 1D or 2D CNN [84]–[86] and the most common approach for RNNs is the Long Short-Term Memory (LSTM) networks, given the memory characteristics that benefit ECG classification [87]. In these cases, feature transformation becomes expendable, even though the classification can be achieved with extracted features.

Considering all the possible parameters that have taken part in literature, determining the average recognition results in ECG biometris is not possible. Some works focus on verification while others aim for identification, and some works report reaching 100% of accuracy in the latter. Typically, Equal Error Rates (EER) in verification range between 0% [88] to 14.3% [89] depending on the circumstances. The more the evaluations

approach to a real biometric environment, the higher the EER tends to be. Including users, lowering the data fidelity and the processing, are parameters that impact these results.

3.3. Conclusion

There are motley solutions regarding ECG biometrics. Each one approaches the issue differently, from the used capture device to the data segmentation process. These lead to different results that cannot be compared with each other by these values, but they require more contextualization.

A desirable approach should consider public databases in order to facilitate comparisons between works. However, the current public databases are not representative enough of a biometric environment, and the developed systems cannot be considered totally feasible for a real application. Those databases that try to be more representative of a biometric context, are not public due to data privacy restrictions. Therefore, these results cannot be tested nor improved by other authors, decreasing its feasibility. In addition, there are not commercial devices suitable for recognition that allow to obtain the ECG signal, which also complicates the development of realistic databases.

As a result of the different data that takes part in literature, the techniques and tools to solve this issue are also variable. It is not possible to confirm which type of classification, fiducial detection or features works best for ECGs, as the results rely heavily on collection parameters. It would be logical to obtain worse results from a wearable prototype if it considers less data than another public database collected with a professional device. However, these results cannot be checked by the scientific community, and can only be used as suggestions for other data or capture devices.

In conclusion, the results of the state of the art must be carefully assessed to consider if they are suitable for our purposes or not. Even though some approaches may be reported as precise and useful in some cases, that could not happen in another specific conditions. It is clear that the field of ECG biometrics requires some specific guidelines and frameworks to allow comparisons between works. In addition, the work towards data publication regarding ECG biometrics should also be prioritized, even though the existing legislation do not facilitate this task.

4. SENSORS AND DATABASES

The present chapter collects all the related information to the three private databases that were used in this thesis. The BMSIL and BSMIL-SB databases were externally collected,while the GUTI database was collected locally.

4.1. BMSIL database

The BMSIL database was collected in the Biomedical Signal and Information Laboratory (BMSIL) at Seoul National University (SNU). The original database collected a variety of bio-signals such as Electroencephalogram (EEG) and Ballistocardiogram (BCG). However, the present work has only required two of the signals: ECG and fingerprint.

4.1.1. Hardware

The following, is a summary of the features and configurations of every hardware device used according to the collected signal:

- Fingerprint:
  - Device: Hamster I by Nitgen.
  - Resolution: 248x292 pixels (500 dpi).
- ECG:
  - Device: ECG100C with MP150 by Biopac.
  - Electrode displacement: left wrist (V+), right wrist (V-), right wrist (GND).
  - Electrode type: wet Ag-AgCl.
  - Signal bandwidth: 0.5 - 35 Hz.
  - A/D converter: 16 bits within ±10 V and 1000 gain.

4.1.2. Acquisition protocol

The BMSIL database is formed by 105 healthy users with gender and age proportions referred in Table 4.1. The age range gets wider as the age increases because the number of participants decreases as age increases. The database was collected in a university environment, where the majority of the population are students. Considering under-graduate and post-graduate students, these ages fluctuate between 18 and 30 years. This age gap was divided into two to add more resolution. The next range encompasses

an age gap of 10 years when the last range is of 20, as the representation keeps decreasing. The female to male ratio is close to half the total, but the distribution based on age and gender is heterogeneous. There is a predominance of people from 25 to 30, followed by people from 18 to 24, getting a total of more than 78% participation in the database, whereas people older than 41 only constitute less than 5% of the total data.

Table 4.1: Number of users based on age range and gender for the BMSIL database. The age ranges get wider as the number of users in the range decreases. Proportions are represented in % with respect to the total users and are rounded to two decimals.

Gender	**Female**		**Male**		**Total**	
Age range	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**
18-24	23	21.90	16	15.23	39	37.14
25-30	21	20.00	23	21.90	44	41.90
31-40	6	5.71	11	10.48	17	16.19
41-60	4	3.81	1	0.01	5	4.77
Total	54	51.42	51	48.58	105	100

For the fingerprint data, every user provides two captures of all fingers from left and right hand, in a total of 10 fingers. The ECG signals were acquired simultaneously, but independently of the fingerprint data. The data collection was divided in two different stages, providing two subsets extracted from the BMSIL database. These sets of data follow the same general scheme, as the collection is taken in two different days (D1, D2) with a separation of one day to two weeks. In each day, there are two different visits (V1, V2), specially selected to affect ECG signals. Each of the visits is repeated 5 times with a 70 s duration of proper signal acquisition and extra adjusting time, depending on the case. For all the signal acquisitions, the first and last 5 seconds are truncated, reducing each signal to 60 s reducing the probability of having motion artifacts.

S1

This first subset (S1) was planned to observe the day-to-day variations. In this case, visits 1 and 2 only differ in the user having their eyes open or closed. However, recent works with this database did not take these conditions into consideration [90]. Therefore, visits 1 and 2 are considered as visits taken under the same conditions, providing two different sets of data. The steps for this acquisition process are the following:

1. Sit and adjust: 20 seconds.
2. Visit 1 (V1): resting, sitting down 70 s.
3. Visit 2 (V2): resting, sitting down 70 s.
4. Adjust: 30 s.
5. Repeat 2-4.

The data was collected from 50 of the users. The corresponding age and gender proportion is specified in Table 4.2. This initial stage also reflected a predominance of the 25-30 age range, with a 54% of the data in S1. However, the male proportion was greater than it was in the entire BMSIL database, with a 60%.

Table 4.2: Number of users based on age range and gender for subset S1. Proportions are represented in % with respect to the total users in S1.

Gender	**Female**		**Male**		**Total**	
Age range	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**
18-24	6	12	4	8	10	20
25-30	10	20	17	34	27	54
31-40	4	8	8	16	12	24
41-60	0	0	1	2	1	2
Total	20	40	30	60	50	100

S2

The second subset of the BMSIL, S2, was collected to focus on ECG variations related to position and heart rate variation. With the difference of S1, visits in this case were not taken one after the other in each repetition. On the contrary, repetitions on the first visit were finished before collecting data related to the second visit of the day. Visit 1 on each day provided information about the user being resting while sitting down. Visit 2 in the first day focuses on changing the user's position from sitting down to standing up. In the second day, the user exercises for 5 minutes on a stepper to increase their heart rate up to 130 bpm, being their data acquire after sitting down. For clarification, the followed steps were:

- Day 1 (D1):
    1. Sit and adjust: 20 s.
    2. Visit 1 (V1): resting, sitting down 70 s.
    3. Adjust: 20 s.
    4. Repeat 2 and 3.
    5. Stand and adjust: 20 s.
    6. Visit 2 (V2): stand up 70 s.
    7. Adjust: 15 s.
    8. Repeat 6 and 7.
- Day 2 (D2):
    1. Sit and adjust: 20 s.
    2. Visit 1 (V1): resting, sitting down 70 s.

3. Adjust: 20 s.
4. Repeat 2 and 3.
5. Exercise: 5 min.
6. Sit and adjust: 20 s.
7. Visit 2 (V2): sitting down after exercise 70 s.
8. Adjust: 15 s.
9. Repeat 7 and 8.

These protocols resulted in two sitting down visits, one standing visit and another one after exercising. These procedures were achieved with the remaining 55 users in the BMSIL database, with gender and age proportions collected in Table 4.3, where the female to male proportion were the opposite as in S1, with almost 62% of females. In this subset the age range of 25 to 30 loses presence in advantage of the range 18 to 24, which composes more than half of the database.

Table 4.3: Number of users based on age range and gender for the subset S2. The age ranges get wider as the number of users in the range decreases. Proportions are represented in % with respect to the total users and are rounded to two decimals.

Gender	**Female**		**Male**		**Total**	
Age	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**	**Num. Users**	**Proportion (%)**
18-24	17	30.91	12	21.82	29	52.73
25-30	11	20.00	6	10.91	17	30.91
31-40	2	3.63	3	5.45	5	9.09
41-60	4	7.27	0	0	4	7.27
Total	34	61.81	21	38.18	55	100

For easier understanding, Table 4.4 summarizes the different experiments in every of the subsets, as well as the naming criteria that is followed in the rest of the manuscript.

Table 4.4: Summary of days (D1, D2) and visits (V1, V2) for every BMSIL subset of data (S1, S2). R: Resting, sitting down (S1). R1: Resting, sitting down in D1 (S2). R2: Resting, sitting down in D2 (S2). S: Standing up. Ex: after exercise.

	Day 1 (D1)		**Day 2 (D2)**	
	Visit 1 (V1)	**Visit 2 (V2)**	**Visit 1 (V1)**	**Visit 2 (V2)**
S1	R	R	R	R
S2	R1	S	R2	Ex

4.2. ECG Smartband Databases

To avoid the inconvenience of operating with devices that require long periods of adjusting and placement, wearable or mobile devices are key in biometrics. Nowadays, smartbands

are commonly worn and its usage has been widespread in the last years. However, it is not applied to ECG, as its acquisition is mainly required for medical purposes and as a consequence, demands more precision. To the best of the author's knowledge, there is only one commercial smartband that collects ECG data, the Nymi's band [46], and it does not allow to retrieve the raw data as it is managed in software for personal verification. Due to the lack of commercial devices that facilitate raw data retrieval, the smartband data was collected with help of specific hardware, which has not been commercialized, completing two different databases and protocols.

4.2.1. Hardware

The goal device must be wireless and compatible with widely spread Operative Systems (OS) such as Windows, Android or iOS. As this sensor was conceptualized as a smartband, the only measurable lead was the type I, which only consider the limbs. In order to measure the voltage between left and right arms, the smartband had to contain two different sensors, each one being in contact with each limb. The process could be achieved by locating one of the sensors in contact with the left wrist, and an external sensor, on the opposite side, which would require to position the right finger on it.

This smartband concept is not unfamiliar nowadays, as the usage of this type of devices is getting more and more accepted in society. In addition, it requires the user to do a conscious verification, so the data cannot be measured without consent. However, the process itself requires full arm mobility so it would not be recommended in specific environments such as driving or situations where the user may be carrying weight, such as grocery bags or hand luggage.

Electrodes

The electrodes must have good conductivity and reduce impedance with the skin. They must be small enough to fit in the device and not be uncomfortable to wear on the wrist. The Figure 4.1 represents an example of commercial electrodes applied in ECG, surrounded by an adhesive that facilitates its fixation. However, the use of the smartband required to apply pressure on the external sensor, which fixes the internal one to the skin, not requiring the addition of the adhesive.

Amplifier

The ECG signal was amplified with an operational amplifier which was set in a simple configuration represented in Figure 4.2. The chosen amplifier was a LM358-N by Texas Instruments.

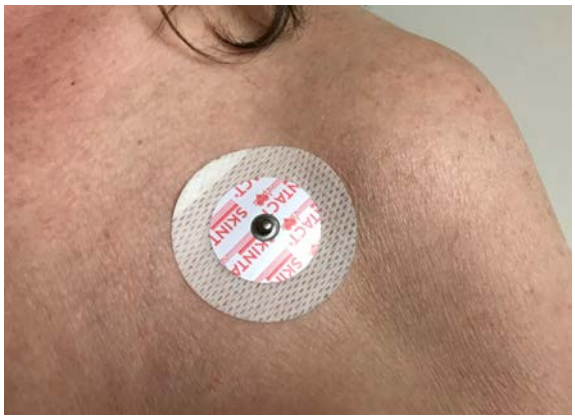


Figure 4.1: Example of a commercial ECG electrode [91].

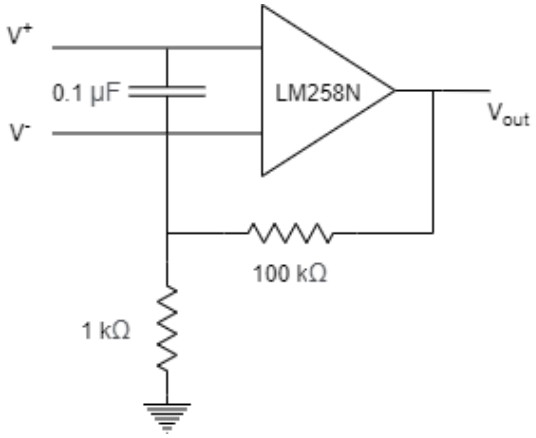


Figure 4.2: Simple amplification scheme.

Wireless communication

The data transfer must be wireless and universally compatible, therefore only WiFi and Bluetooth (BT) are considered. The 4.0 version is considered low energy (BT LE), with improved security connections in BT 4.2 [92]. This factor constraints the microcontroller unit (MCU) selection, as discussed in the following paragraphs.

A/D converter

Databases throughout literature implement sampling frequencies from 250 kHz to 1000 kHz. Sampling at higher frequencies allow to extract data with higher precision, so 1000 kHz is preferred as long as the MCU allows it. There is not a special requirement for the A/D converter, which would be determined by the chosen MCU.

Microprocessor

In addition to the previous characteristics, the MCU is preferred to require small batteries for an ergonomic smartband adaptation and convenience. Therefore, the selected MCU

must have a low energy consumption. The chosen MCU is the model C2540R2F by Texas Instruments with, but not limited to, the following features:

- ARM Cortex-M3 at 48 MHz.
- 12 bits A/D converter.
- Maximum sampling frequency of 200 kHz.
- Ultra-Low Power Sensor with possible autonomy of 2KB SRAM.
- AES-128 security module.
- Low energy characteristics:
  - Power from 1.8-3.8 V.
  - Active-Mode RX: 5.9 mA.
  - Active-Mode TX at 0 dBM: 6.1 mA.
  - Active-Mode TX at +5 dBm: 9.1 mA.
  - Active-Mode MCU: 61 μA/MHz.
  - Active-Mode Sensor Controller: 0.4 mA + 8.2 μA/MHz.
  - Standby: 1.1 μA.
  - Shutdown: 100 nA.
- Bluetooth Low Energy compatibility with 4.2 and 5.0 versions.

Battery and power supply

The device was estimated to be used 4 times a day on average, with a total 12 seconds of activity and 40 seconds in stand-by, hibernating the remaining time. Considering the BT requirements, power should provide 410 mAh. Battery dimensions are also limited to those in the device, which are 17x35 mm^2. Considering these constraints, the available commercial devices are limiting. The final choise was a battery by Adafruit with 350 mAh and dimensions 36x20 mm^2. A final charging circuit was added, based on the USB-C technology. The circuit is also specifically made by Adafruit for the selected battery. Both components can be observed in Figure 4.3.

Final prototype

The final prototype was assembled by Upines, based in South Korea. The result has 12x35 mm^2 dimensions. The main module is represented in Figure 4.4. Electrodes and battery were attached with the addition of a plastic band to fix the device to the wrist and make it easier to use. The final prototype and a usage example are shown in Figure 4.5.

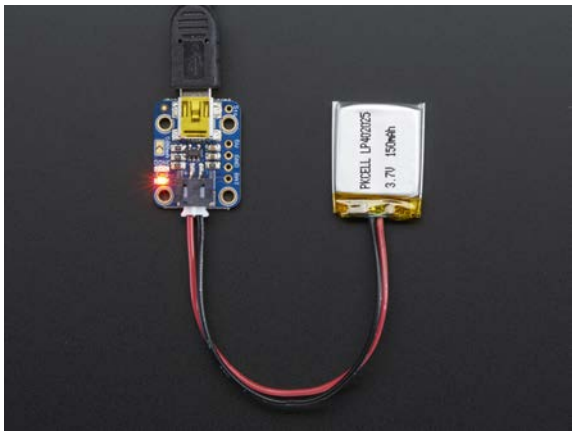


Figure 4.3: Battery and charging circuit connected [93].

Figure 4.4: Main module for the prototype, without electrodes and battery.

4.2.2. Acquisition protocols

Two different databases were acquired using the smartband prototype: the BMSIL-SB and the GUTI database. The stages in each protocol differ substantially and are further explained in the following paragraphs.

BMSIL-SB Database

The first database collected with the smartband prototype was also collected by BMSIL, and included fingerprint acquisition with the same device and protocol as in section 4.1. The database was formed by 206 users from 18 to 68 years old, with age and gender distributions summarized in Table 4.5. Users are generally healthy, with the exception of 4 cases of Premature Atrial Contraction (PAC), which represents less than 2% of the database. In this case, the presence proportion of users from 18 to 30 is even more noticeable than observed in the previous database.

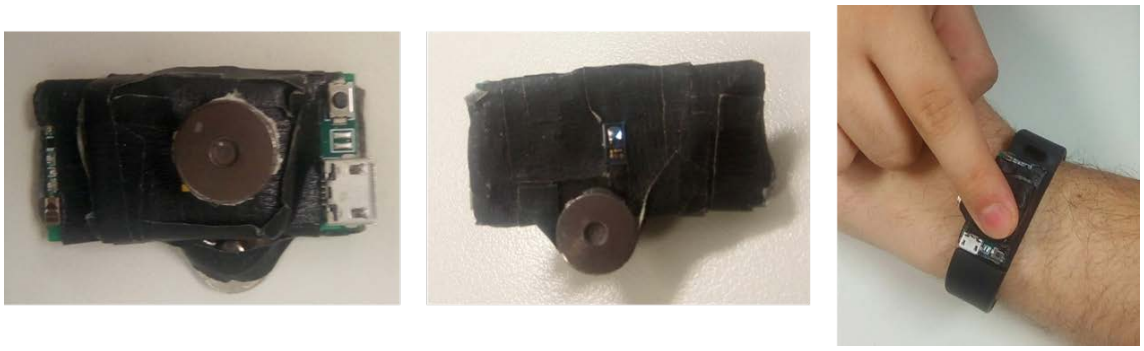


Figure 4.5: Final prototype and usage example.

Table 4.5: Number of users based on age range and gender for the BMSIL-SB database. The age ranges get wider as the number of users in the range decreases. Proportions are represented in % with respect to the total users and are rounded to two decimals.

Gender	**Female**		**Male**		**Total**	
Age	**Num. users**	**Proportion (%)**	**Num. users**	**Proportion (%)**	**Num. users**	**Proportion (%)**
18-24	53	25.73	47	22.82	100	48.55
25-30	30	14.56	52	25.24	82	39.80
31-40	8	3.88	8	3.88	16	7.76
>40	5	2.43	3	1.46	8	3.89
Total	96	46.60	110	53.40	206	100

In opposition to the initial BMSIL database, the BMSIL-SB database was focused on obtaining data while resting and after exercise, without different day acquisitions. The proper ECG capturing in each session lasted 9 s. The number of sessions per user varies, ranging between 4 to more than 8. The majority of users presented from 5 to 8 sessions, as observed in Table 4.6. The different scenarios are defined as follows:

- Rest: the left hand, where the smartband is placed, is steady while the right hand's index is executing the action. The technician evaluates when the user is calmed and quiet to capture the signal, and the process is repeated accordingly.
- Exercise: the user exercises during 5 minutes in a stepper, rising between 40% and 50% the heart rate. The signal is captured afterwards, when the user sits down, with a steady hand. This process is done under time constraints to avoid the heart rate reaching its resting state.

Table 4.6: Distribution of number of sessions among the users, with their corresponding proportion.

Num. sessions	4 sessions	5-8 sessions	> 8 sessions
Num. users	2	188	16
Proportion (%)	0.97	91.26	7.77

GUTI Database

This database was collected in the University Group for Identification Technologies (GUTI). The main motivation behind the development of this database was to extend the strong points in BMSIL database in smartband while adding extra considerations. The BMSIL database only considered scenarios in different days for resting experiments, not making possible to compare how the change of position or the heart rate evolve through time.

The GUTI database proposed an acquisition protocol repeated in two different days (D1, D2), with at least 15 days of separation. Each day consisted on two visits (V1, V2) with a minimum separation of 2 h. All the data collection was achieved with a technician that evaluates the correct performance of the experiments. An oximeter is employed before collecting the signals. In the case of resting and standing, the pulse needed to be stable to consider that the user was relaxed. For the exercise phase, the heart rate was required to be a minimum of 120 bpm. Initially, the exercise required the user to workout on the stepper. However, it ended up not being limited as some users had more tolerance and required stronger cardio activities to reach the heart rate goal.

Each type of scenario had a collection of 5 ECG signals with 9 s duration, with the smartband positioned on the left wrist:

1. Sit: the user is sitting down while resting the left wrist on a table. The right index finger is positioned on the external sensor, providing a gentle pressure, and avoiding any movement.

2. Walking: the user walks calmly for 20 m, without rising the heart rate, to ensure that the user was standing for a while before capturing the signal. The user proceeds to place the right index finger onto the external sensor while trying to have the left hand steady and horizontal.

3. Exercise: the user works out, measuring the pulse from time to time until reaching the required minimum of 120 bpm. The user stands up while having the signal captured, trying to stay in the same positions as in the standing experiment.

The database contains data of a total of 72 different users, with no specified age group, gender nor cardiac diseases. However, as it was collected in University Carlos III of Madrid, the population was also prominent with younger users, between 18 to 30 years old. Once the user came to the given appointment for the data collection, the three scenarios were carried out with no exception. D1 had a 100% completion for both visits. However, four users were missing from D2V1 and an extra user missed D2V2, having a total of 5 users without D2 completion, as reflected in Table 4.7.

Table 4.7: Number of users that completed the data collection, based on the different days (D1, D2) and visits (V1, V2). 67 users completed the whole process, a 93.05% of the initial users.

Visit	D1V1	D1V2	D2V1	D2V2
Users	72	72	68	67

4.3. Conclusion

The available databases in this thesis are characterized heterogeneously. The initial BMSIL database was collected with professional devices and provided reliable data to further study ECG biometrics under different circumstances and periods of time. The addition of fingerprint data allowed the possibility of fusing it with ECG. The smartband databases complemented the previous one, considering features that resulted in scenarios that were closer to a realistic biometric environment, even though data might be less reliable.

Altogether, the collected databases provided enough information for a structured study, going from more ideal cases (i.e.: data collected with professional devices) and narrowing the solution to a more specific, constrained problem (i.e.: ECG information acquired by sensors with lower fidelity).

5. METHODS

The applied tools in this thesis are further discussed in this chapter, following the scheme in Figure 2.1. for a detailed description of the employed sensor, refer to chapter 4.

5.1. Signal pre-processing

The raw signal retrieved by the corresponding sensor is considered a raw ECG signal. As referred in section 3.1.2, ECG main noise sources are baseline wander (0.2-0.5 Hz), power-line interference (50-60 Hz) and muscle artifacts (around 100 Hz). One of the most common approaches in literature are the band-pass filters, which were the selection in pre-processing throughout the whole thesis. This tool was applied based on [90] which used the tool in the BMSIL database. The band-pass filter is a 5^{th}-grade Butterworth filter with cutting frequencies of 1 and 35 Hz.

This filtering is common for all chapters that require pre-processing: chapters 6, 8, 9 and 10.

5.2. Feature extraction

The QRS complex segmentation has been a common approach at this stage for both fiducial and hybrid features, defining the waveform as the most information-dense part of the ECG signal. Firstly, this section details the fiducial point detection methods applied for the QRS segmentation. Transformations and feature selection techniques are explained in the following subsections.

5.2.1. Fiducial point detection

The QRS segmentation is easier to achieve once the R peak is detected. This point is generally the most prominent in every P-QRS-T waveform, as observed in Figure 3.1, making it convenient to detect. The length of the QRS complex can be also determined by the Q and S point detection, but this problem is complex and adds further issues in the topic resulting in data samples with different lengths. Therefore, the segmentation range was fixed with the detected R peak as a reference point.

BMSIL algorithm

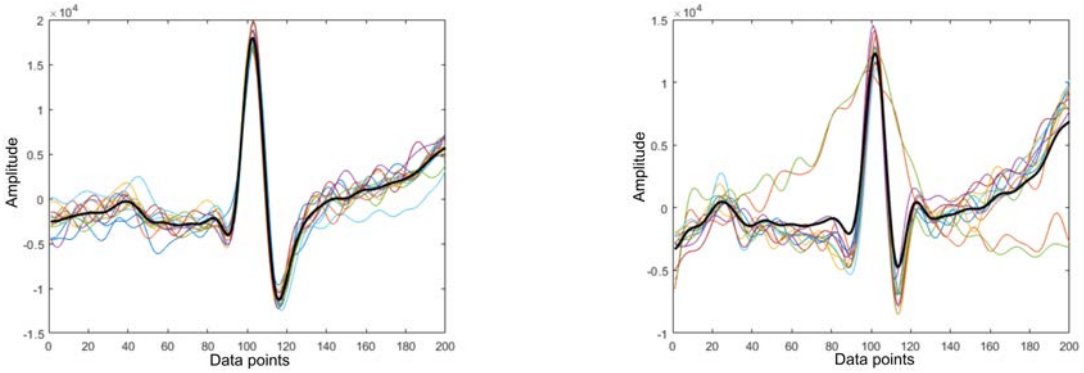
An alternative algorithm was developed by the BMSIL considering high quality signals in the BMSIL database. The alternative algorithm aimed to be easy and deal with signals that do not come from medical monitoring. The fiducial point detection also needed to be simple for biometric purposes, avoiding the use of complex methods to make it easier to scale. In [90] the R peak detection in signals with these characteristics was achieved with the same database. The pre-processed signal gets differentiated to obtain first and second derivatives, helping with the detection of the R peak. Finally, outliers are discarded by thresholding the correlation coefficients.

Pan-Tompkins algorithm

The Pan-Tompkins algorithm [94] was a really successful algorithm that detected the QRS complex in real time. The databases applied to develop the algorithm had high quality signals from Holter monitoring, and included pathologies.

The Pan-Tompkins algorithm's good performance in literature leads to check its performance in lower quality signals, such as those in BMSIL-SB and GUTI databases. Two ECG signals are represented for the most relaxed scenarios of each database: one reflecting an example of the best performance, and another one showing less ideal results. All the examples come from original ECG with clear QRS complexes.

The examples corresponding to the BMSIL database are in Figure 5.1, where 5.1a shows a correct performance, properly locating the R peaks in the center, with not observable mistakes. On the contrary, Figure 5.1b shows two errors in detection, misinterpreting T-waves as R peaks, while correctly detecting the remaining peaks.



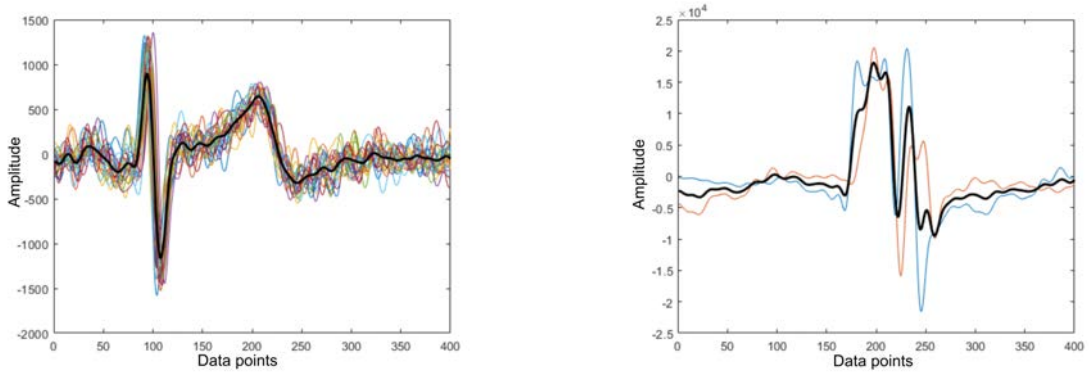
(a) Signal with ideal detection.

(b) Signal with some incorrect detections.

Figure 5.1: Extreme performance cases for the Pan-Tompkins algorithm for the BMSIL-SB database, considering good signals.

Similarly, Figure 5.2 shows different performances of the Pan-Tompkins algorithm in the GUTI database. In the best case, Figure 5.2a shows a correct QRS detection, with

no errors. However, the signal is shifted to the left, finding the T-wave in the center of the QRS. This fact informs us that the Pan-Tompkins algorithm is detection the T-wave maxima as the R peak. Even though this misinterpretation leads to a proper QRS complex detection, it is not what it is meant to do. Probably this issue is what leads to poor performances such as those observed in Figure 5.2b, which represents the result in an original signal with similar quality as the previous one.



(a) Signal detection centered in the T-wave. **(b)** Signal with poor detection.

Figure 5.2: Extreme performance cases for the Pan-Tompkins algorithm for the GUTI database, considering good signals.

The different nature of these databases can be observed in Figures 5.1a and 5.2b. The BMSIL-SB database has more stable complexes. It was also observed that this database provided signals with lower number of cycles (heart rate) in the most relaxed scenario. This implies that users for the BMSIL-SB database were more relaxed than those in the GUTI database, probably in relation to the different protocol followed prior to the signal acquisition.

The different peak detection algorithms are applied in those experiments that require detecting the QRS complex. The BMSIL algorithm is applied to the BMSIL database in chapters 6 and 8, 9. The Pan-Tompkins algorithm is one of the peak detection alternatives for BMSIL-SB and GUTI databases to test its performance in smartbands in chapter 10.

5.2.2. Transformations

The segmented data usually requires specific processes to increase the following classification performance. In this thesis we mainly focus on computations that do not require complex computations, for potential viability in devices with lower processing capabilities.

Differentiation

The differentiation is a simple transformation that emphasizes abrupt changes in the signal. First and second differentiation are calculated in the BMSIL peak detection

algorithm, and also used as features in chapters 8 and 10.

Wavelet Transform (WT)

The WT allows to analyze a signal into different frequencies at different resolutions, known as multi-resolution analysis. The windowing is done with functions called wavelets, and their scaling allows to obtain the different resolutions: narrow wavelets increase time resolution, whereas wider wavelets improve the frequency resolution.

The discrete version of WT is the Discrete Wavelet Transform (DWT). The DWT is a decomposition, which passes the signal through low and high pass filters, providing coefficients at every level. The low pass portions give approximation coefficients, and high pass portions refer to detail coefficients, which are downsampled. The whole process results in sets of approximation and detailed coefficients. However, this transformation is time variant, which translates into different results when the original signal presents a significant movement.

The Stationary Wavelet Transform (SWT) was developed to solve the time variant property of the DWT. The SWT dispenses with downsampling, and filters the same way it happens in the case of DWT. Level 1 results in the same number of data points as the original signal. The process is achieved for the j^{th} level by filtering with the coefficients in the level (j – 1) [95]. The Figure 5.3 represents the scheme for the 1D SWT decomposition, where the original signal length needs to be divisible by 2^j.

The 1D SWT is employed in chapter 10 as one of the transformations to the signal, in order to improve the classification with smartband databases.

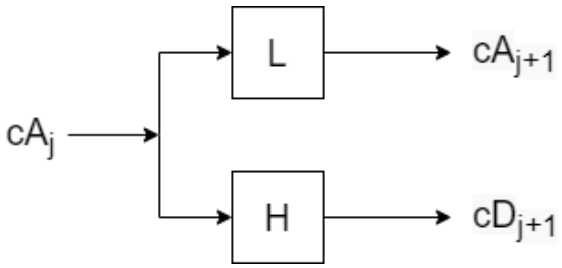


Figure 5.3: Approximation (cA) and detailed (cD) coefficients of level j for SWT. L and H represent the low and high pass filters, respectively [95].

Discrete Cosine Transform (DCT)

The DCT converts a time sequence into the sum of cosine functions in different frequencies and amplitudes. The DCT allows to represent signals with a lower number of coefficients as it stores more signal energy in each one of them. Its role results has been very common in audio and ECG signal compression, among other multiple applications.

The DCT is applied to the ECG data in chapter 6 to observe the difference in performance with other types of features.

5.2.3. Feature selection

In some circumstances, the available features give out large amounts of information. Some of the extracted data could be redundant, noisy or confusing for classification. In addition, when dealing with great databases, the data must be minimized to lower the processing time and complexity. These issues get solved through feature selection mechanisms. The ones applied in this Thesis are briefly explained in the following paragraphs.

Principal Component Analysis (PCA)

The PCA algorithm establishes a new set of coordinates based on the covariance matrix. If the original data has n samples and m features (n, m), it gets into transformed data of dimensions (n, l) where l < m. The algorithm evaluates the relevance of every available feature, providing a ranking for feature selection.

This tool is applied in chapter 6 for data reduction and to observe how the performance is impacted by the discarding of certain features.

Infinite Feature Selection (IFS)

The IFS algorithm ranks a set of given features from most to least relevant. The method is based on graphs that consider every feature distribution as a node, V. Every possible pair of distributions is modeled by E. The final graph G is represented as an adjacency matrix A, which represents the energy terms between the different feature pairs. The calculation of every element of the matrix is given by Equation (5.1), where $0 \leq \alpha \leq 1$ and $\sigma_{ij} = \max(\sigma^{(i)}, \sigma^{(j)})$ being σ the standard deviations over the samples of each feature distribution. Finally, the coefficient c_{ij} is defined in Equation (5.2), where the Spearman operator refers to the Spearman's rank correlation coefficient [96].

The approach in IFS is applied in chapter 10 as an attempt to reduce the dimensions and improve the classification algorithm.

$$a_{ij} = \alpha\sigma_{ij} + (1 - \alpha)c_{ij} \tag{5.1}$$

$$c_{ij} = 1 - |Spearman(f^{(i)}, f^{(j)})| \tag{5.2}$$

5.3. Comparison

The main event in the comparison stage is related to the classification of new data in comparison to previously acquired references, which eventually leads to a decision.

The classification algorithms applied in this thesis belong to Machine Learning and Deep Learning tools. In general, the data is used as a closed-set, which implies that it is assumed that no new data is going to enroll in the system. This leads to modelling the whole database as one, using supervised algorithms. However, in experiments dealing with open-set experiments, an extra unsupervised algorithm is applied.

5.3.1. Classifiers

The explored algorithms are detailed below, following a chronological order of use and complexity. These classifiers depend on internal parameters or hyperparameters that need to be adjusted for an optimal solution. For this hyperparameter optimization, the used data is isually divded into the following subsets:

- Development set: it is formed by the data that helps creating the model. In this thesis, this set comes from the data collected during enrollment. The development proportion is defined by d, where $0 < d \leq 1$. At the same time, this set usually divides into two:
  - Train set: it is the bigger division of the development set. It is the proper information to form the model and its proportion in the development set is usually 0.8 or 80%.
  - Validation set: it is the remaining data and always represent a smaller percentage of the data than the train set. It helps to check whether the training is achieving good results with new data. As a consequence, it usually is 20% of the development set.
- Test set: it is the new data that comes to the trained model. In biometrics this set is formed by the new data to be recognized. As a consequence, when the test set comes from the same set of data, its proportion is $1 - d$.

Dynamic Time Warping (DTW)

DTW is a distance based algorithm that measures similarities between two time-dependent signals. The main advantage of this algorithm is on the capability of dealing with signals with different lengths. However, it can only happen if their starting and finishing points get mapped. DTW is useful for sequence alignment and similarity measurement through the warping path.

The lack of constraints for the sample length is very valuable in the problem of ECG biometrics, as the heartbeat width varies depending on the user and conditions. If the starting and finishing point of each of the QRS complexes is known, DTW helps transforming data into same length samples.

The DTW algorithm is applied in chapter 6 to determine the most suitable segmentation in ECG classification.

Support Vector Machine (SVM)

SVMs calculate the hyper-plane that best separates two sets of data. The result is achieved by finding the hyper-plane whose distance to the closest point from both groups is maximized. If the hyper-plane is not found, the data gets projected to another dimension with the help of kernel functions, creating a new feature space. The applied kernel is linear, as this type of kernel requires lower computational costs than other alternatives such as Radial Basis Function (RBF) or polynomial kernels.

In practical cases, data is not linearly separable. Some incorrect classifications are allowed with a penalty factor, considering a soft-margin with formulation in Equation (5.3). The goal of the SVM classifier is to minimize this soft-margin, where λ is the box constraint, n represents the number of samples, y_i is the given label, w the classifier weights, x_i the available data to create the classifier. In the case of multiclass classification, SVMs take the one-vs.-one approach, producing $n_{classes} \cdot (n_{classes} - 1)/2$ classifiers [97].

$$\left[\frac{1}{n}\sum_{i=1}^{n} max(0.1 - y_i(w^T x_i - b))\right] + \lambda\|w\|^2 \tag{5.3}$$

SVMs are used in chapter 6 as one of the first approach for a closed-set recognition with ECG.

k-Near Neighbors (k-NN)

The k-NN algorithm is a simple approach that finds the k closest pattern points (nearest neighbors) stored in the model. The distance is calculated based on different approaches such as Euclidean or Hamming distance, among others. The first one is selected in the present thesis.

This algorithm is applied to chapter 6 similarly to SVMs, as it is one of the initial approaches to study the viability of ECG recognition.

Linear Discriminant Analysis (LDA)

The LDA classification is a generalization of Fisher's linear discriminant, which finds the lineal combination of features to help separate two or more classes. It can be used as a tool in dimension reduction, but the purpose in this thesis is only as a classifier.

The generalization depends on the mean and co-variance of each class as observed in Equation (5.4), where C is the number of classes with the same variance Σ, and each class has the mean μ_i. The separation of each class is given by the Equation (5.5). The co-variances and means are usually unknown in real cases, so they get estimated by training. In addition, in the case of non-linear classification the problem can be extended with the use of kernels as it happened with SVMs. In real situations, sometimes the software cannot invert the covariance matrix, so it does the pseudo inverse.

$$\Sigma_b = \frac{1}{C}\sum_{i=1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T \tag{5.4}$$

$$S = \frac{\vec{w}^T \Sigma_b \vec{w}}{\vec{w}^T \Sigma \vec{w}} \tag{5.5}$$

The LDA algorithm is one of the main algorithms applied in this thesis, and takes part in chapter 6, and its results are the ones required for those experiments in chapter 7.

Gaussian Mixture Models (GMM)

GMMs are an approach normally applied for clustering, as it is an unsupervised learning algorithm. These models assume that the original data is formed by the combination of a fixed number of multivariate Gaussian distributions. GMMs are the only unsupervised method applied in the present thesis.

The prior distribution of the vector of parameters in the mixture model, θ, is modeled as in Equation (5.6). K is the number of components in the mixture model, and each of them are characterized differently by weights ϕ_i, means μ_i and covariance matrices Σ_i. The posterior distribution of the available data, $p(\theta|x)$ is also a Gaussian mixture model, represented in 5.7 where the $\tilde{\phi}_i$, $\tilde{\mu}_i$ and $\tilde{\Sigma}_i$ are updated with the Expectation-maximization (EM) algorithm. The initial values of θ are randomly assigned and iterate until reaching convergence [98].

$$p(\theta) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\mu_i, \Sigma_i) \tag{5.6}$$

$$p(\theta|x) = \sum_{i=1}^{K} \mathcal{N}(\tilde{\mu}_i\ \tilde{\Sigma}_i) \tag{5.7}$$

GMMs are used as an alternative algorithm for open-set ECG verification in chapter 6.

Multilayer Perceptron (MLP)

The MLP algorithm is one of the most simple NN algorithms. MLP networks are applied in supervised learning and they have three main parts: input, output, and hidden layers, as represented in Figure 5.4. The input layer is formed by nodes or neurons that represent the different input features $\{x_i|x_1, x_2, \ldots, x_n\}$. Every feature is labeled with its corresponding class, $\{y_i|y_1, y_2, \ldots, y_n\}$. In the case of only having one hidden layer, the output layer gives the function in Equation (5.8) [99]:

$$f(x) = W_2 g(W_1^T x + b_1) + b_2. \tag{5.8}$$

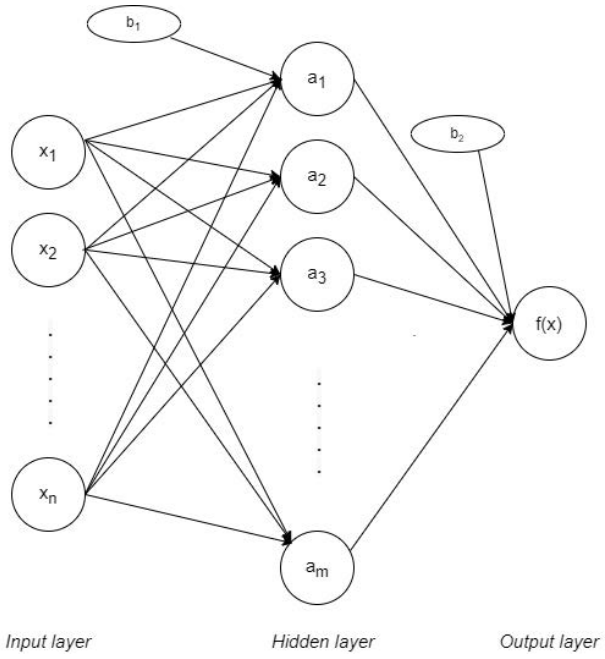


Figure 5.4: MLP with one hidden layer.

Where W_1 represents the sets of weights applied to every feature in the input layer. These weights vary between them, in the way that every feature x_i has m different weights: one per node in the following hidden layer. On the same way, W_2 represents the weights applied in the hidden layer, at nodes $\{a_j|a_1, a_2, \ldots, a_m\}$. Value b_1 is the bias in the hidden layer while b_2 is the bias on the output layer. The activation function is represented by $g(\cdot)$. The most common functions are identity (or no activation function), logistic, hyperbolic tangent (tanh) and rectified linear unit function (ReLU). All the corresponding functions are represented in Table 5.1.

As this structure works for supervised learning, the weights need to change in every connection after the data is processed to decrease the processed error. In this case,

Table 5.1: Most common activation functions.

Name	**Identity**	**Logistic**	**Tanh**	**ReLU**
Formula	$g(x) = x$	$g(x) = \frac{1}{1+e^{-x}}$	$g(x) = tanhx$	$g(x) = max(0, x)$

it is done by back propagation, which comes from the Least Mean Squares (LMS) algorithm. These weights can be updated differently, depending on the approach for their optimization. The most common optimizer is of Stochastic Gradient Descent (SGD). Its formula depends on a factor called learning rate, which ensures the weights converge quickly.

The MLP is the algorithm that chapter 8 is based on, aiming to improve those results from GMMs and LDA in chapter 6.

Convolutional Neural Networks (CNN)

The purpose of this neural networks is summarizing the segmented data by extracting the most relevant features. It reduces the amount of data to interpret by the following units, easing the procedure and reducing its complexity [100].

The CNN has specific properties for one-dimensional signals, but the concept is similar to 2D CNN. The main difference is related to how the sliding window moves through the data. Two-dimensional sliding windows need to specify their width and height, as they slide horizontally and vertically. In the case of 1D convolutions, the only required value is how many features are taken into consideration in every sliding iteration. The hyperparameters that affect the 1D CNN output size are:

- Kernel (k): the number of samples that are used in every iteration for the convolution.
- Filters (f): the number of sliding windows involved in the process, which translates into the number of extracted features.
- Strides (s): the number of positions the window slides each time.

The final dimensions are summarized in Figure 5.5, where b is the batch size, k the kernel size, f the number of extracted filters and how many units each filter strides. The output dimensions per batch correspond to a 3D matrix, as the process in every batch is done with f number of filters (b, o, f). The value o is the output width determined by Equation (5.9).

$$o = \frac{W - k}{s} + 1 \tag{5.9}$$

CNNs take part in the initial approaches in chapter 9 and as a classfier and feature extractor for the final BioECG network in the same chapter.

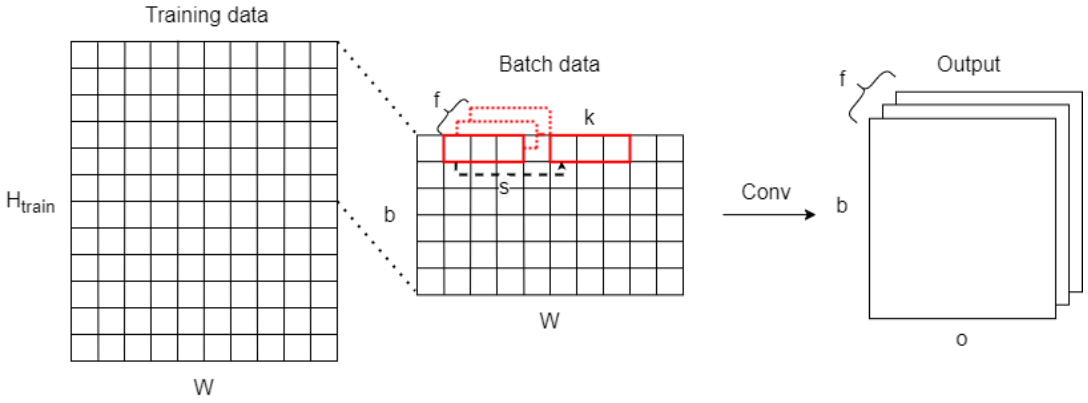


Figure 5.5: Scheme for CNN layer, where H_{train} is the number of samples to train and W the number of features.

Long-Short Term Memory (LSTM)

LSTMs are a type of Recurrent Neural Network (RNN), as they also have chained recurrent modules. However, the different LSTM cells are more complex than those in standard RNNs. The specific structure is represented in Figure 5.6, where every rectangle represents a fully connected layer with their corresponding activation, sigma (σ) and tanh. Input data in the timestep t is represented by x_t. Similarly, the current cell state and outputs are represented by C_t and h_t.

The current cell state, C_t, depends on minor linear interactions related to the previous cell state, C_{t-1}. The LSTM gates are formed by a sigmoid (σ) layer and a pointwise multiplication so the outputs are kept between 0 (discarded information) and 1 (valid information). The forget gate, f_t, operation is in Equation (5.10), where W_f represents the corresponding weight matrix for that gate and stays unaltered through time. This calculation determines which information is required to be kept based on the previous output and the current information, plus b_f which represents a bias. Similarly, i_t is obtained with the same process and different weights and bias, W_i and b_i, as seen in Equation (5.11). This gate is known as the input gate; it selects which values get updated. The output i_t gets combined with a vector of candidate values, $\bar{C}_t$, obtained with a tanh layer, weights W_C and bias b_C, as observed in Equation (5.12).

The previous cell state C_{t-1} updates resulting in Equation (5.13). C_t then a tanh pushes it to values between -1 and 1 before getting multiplied by the output of another sigmoid gate. This part leads to the final output as formulated in Equations (5.14) and (5.15), using weights W_o and bias b_o. This process is done recurrently as many times as timesteps there are.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5.10}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5.11}$$

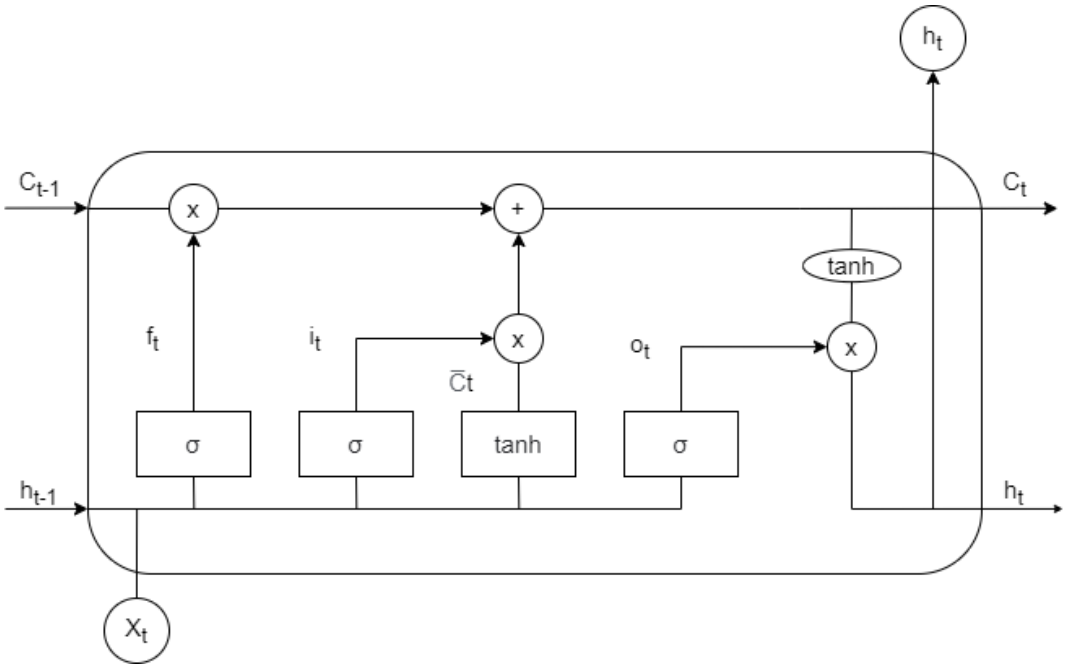


Figure 5.6: LSTM cell composition.

$$\bar{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5.12}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \bar{C}_t \tag{5.13}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5.14}$$

$$h_t = o_t \cdot tanh(C_t) \tag{5.15}$$

To implement a multilayered LSTM, the output sequence of the LSTM cell in a given timestep, h_t, is returned and fed into the next layer. Figure 5.7 is an unrolled two-layered LSTM, where T represents the maximum number of timesteps. The last layer does not require retrieving all the hidden cell outputs but only the output in the last timestep. In the case of Figure 5.7, the final output corresponds to h'_T.

As a result, there are two key hyperparameters that define LSTM networks:

- Hidden neurons (n): number of hidden neurons in the LSTM cell gates.
- Hidden layers (L): number of LSTM layers to connect.

LSTM networks are the second most important part of the BioECG network developed in chapter 9, which is in charge of classifying the data retrieved by the CNN.

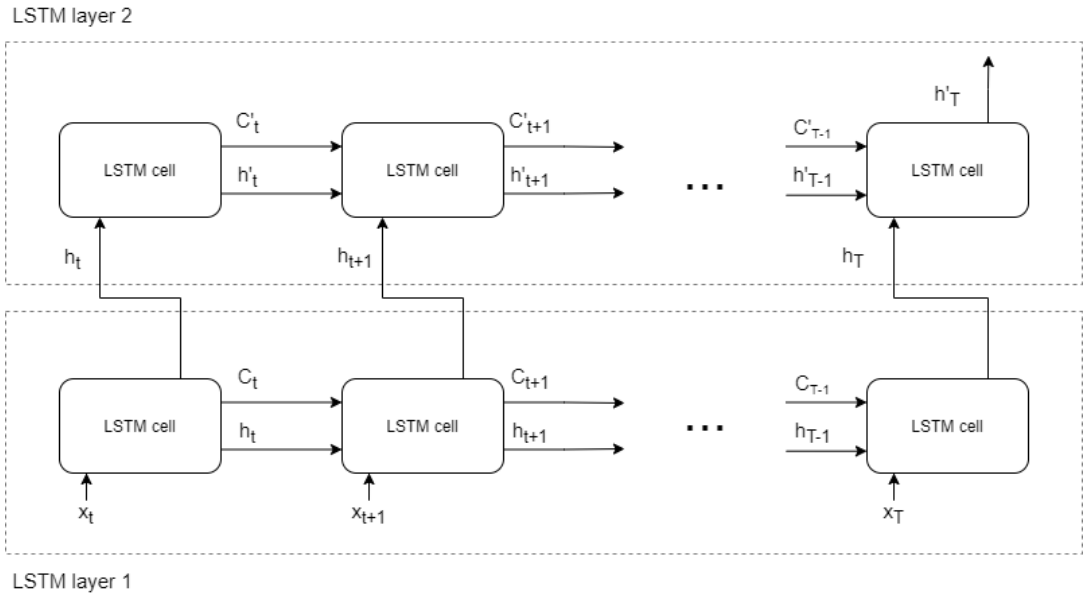


Figure 5.7: Scheme for an unrolled LSTM with two layers.

5.3.2. Metrics

Every type of recognition is associated to different metrics: accuracy for identification and EER for verification. The present Thesis focuses on verification, which is related to EER. However, identification metrics are frequently applied in model training to observe their suitability for this process. These accuracies are obtained considering each sample individually applying the Equation (2.1). On the contrary, for verification, other approaches are taken.

For generalization, the verification data matrix has height of H_{user} samples and a width W that corresponds to the number of available features, as represented in Figure 5.8. At user level, every attempt is formed by H_A samples. The data is divided accordingly to Figure 5.8, with no repetition between attempts, and obtaining as many attempts as the verification data allows. If the division is not an integer, the remaining samples are not used, flooring the result. The number of attempts is $N_A = \lfloor \frac{H_{user}}{H_A} \rfloor$. Each attempt data is then transformed into a score matrix, with the same height H_A, but with a width of U, which corresponds to the number of users in the stored model.

Considering this general scheme, two paths are considered to determine the metrics:

- Verification with one attempt: the EER is calculated with every attempt individually. The scores of every user, U, are averaged along the columns, obtaining a score average vector of length U. Once this calculation is done for the same attempt number in all the different users U, the final score matrix is obtained for that given attempt. Doing this process recursively with all different attempts, gives a vector of EER with as many results as number of total attempts, N_A.

- Verification with all attempts: the EER is calculated considering all the attempts.

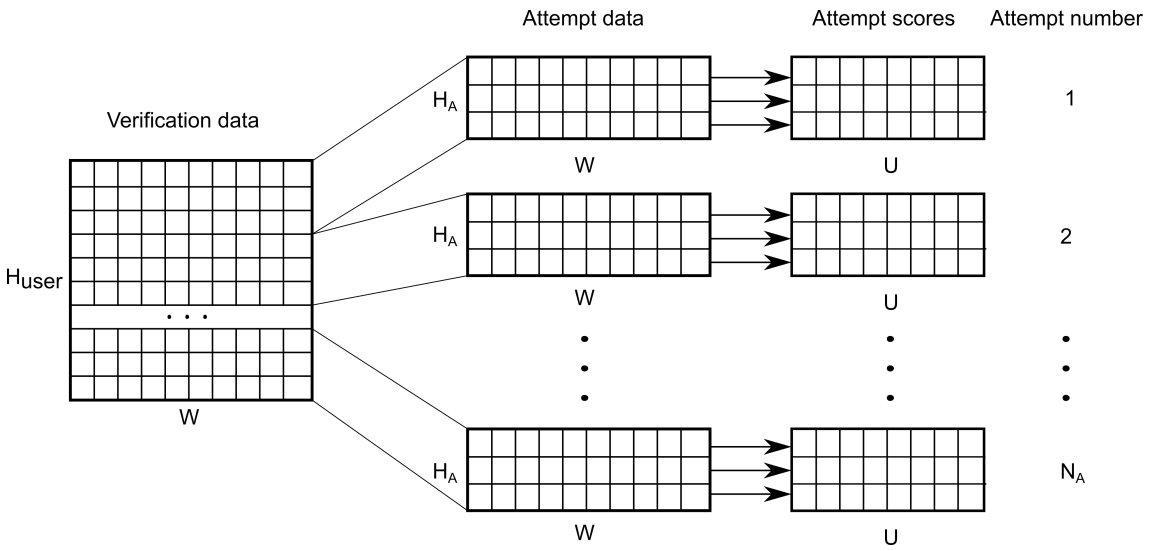


Figure 5.8: Division for different attempts and variable length H_A in verification.

The mean scores of the samples in each attempt, are averaged as a single result. This process results in a squared score matrix of dimensions (U, U) which leads to a single EER calculation.

Once the scores are obtained, the EER can be obtained. In this thesis, FMR and FNMR are calculated for 100 different thresholds in steps of 0.01. The EER is theoretically calculated when FNMR = FMR. However, values from these curves are not continuous in real case scenarios, so the EER has to be estimated from the available data. This is achieved by finding the two values of the threshold, th_1 and th_2, that correspond to the two closest points of FNMR and FMR vectors, as represented in Figure 5.9. These values provide two values for each curve, allowing to characterize the straight lines that pass through these values. This estimation finally provides the estimation where the two curves meet, the EER.

Both of these approaches are used as an extended and more realistic evaluation of the performances obtained in chapter 8 and 9. However, it is important to remark that final performances are obtained with $H_A = 1$ and using all the available samples, as these are the conditions that provide the FNMR vs. FMR graphs, and the extended verification techniques only provide a behavior insight on the different scenarios.

5.4. Conclusion

The present chapter has collected all the methods applied in this Thesis for the different stages of the developed systems.

Due to the different characteristics of the three available databases for this work, the resources are multiple and always focus on trying to best analyze and improve the recognition process through ECG.

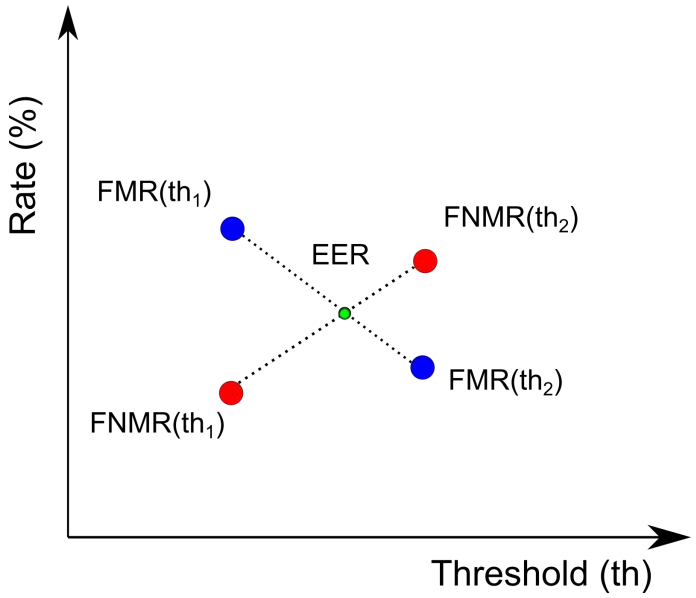


Figure 5.9: Graphic representation of the EER calculation from FNMR and FMR data.

6. VIABILITY OF HUMAN RECOGNITION WITH THE BMSIL DATABASE

The initial stages of this thesis start with the viability study of the BMSIL database. The preliminary experiments focus on determining the best QRS segment, and how to approach the enrollment regarding the number of stored references, used score and the algorithm for open-set and closed-set. These experiments were defined more in depth in [101]. The open-set approach with GMM, closed-set with LDA and their extended verification are specifically done for this thesis.

6.1. Reference selection with DTW

The signal is pre-processed and R peaks are detected using the BMSIL peak detection algorithm referred in section 5.2.1. From this starting point, there is the need to specify the segment delimitation. How do the P and T wave affect the recognition? Is it better to provide only information related to the QRS complex? Is it worth to relay on fiducial detection for the segmentation? How many stored patterns need to be stored to improve the performance?

The R peak detection is achieved with the BMSIL peak detection algorithm referred in section 5.2.1. To help in different versions of the segmentation, P and T wave detection are also detected to determine the QRS boundaries. The P and T waves are observed in a frequency range between 0 and 10 Hz, so a Butterworth filter with cut-off frequencies 2-10 Hz retrieves more obvious P and T waves. The final fiducial detection is achieved by finding local minima and maxima and it is represented in Figure 6.1. QRS start and QRS end points correspond to the QRS complex defined in Figure 3.1. Q' is the local maximum associated to the Q point, and S' is the local maximum that corresponds to the S point, which in this case is the same as the QRS end. After the fiducial detection, 14 segmentation criteria are defined in Table 6.1 using the temporal interval durations referred in section 3.1.

The results are obtained using only the first session as the enrollment data and treating it as the development set. Experiments are achieved using D1V1 and D1V2, as they represent the most relaxed state in both S1 and S2 (sitting and standing). As the available segments may have different lengths, DTW is calculated for every available segment provided by the same user. This approach performs an open-set classification. For the reference selection, there are two approaches based on the mean DTW of each segment:

- Single reference: the segment with the lowest mean distance of all is the selected reference.

Table 6.1: Different segmentation versions based on theoretical interval time criteria and fiducial point detection. T and P refer to theoretical T and P wave duration, while T' and P' refer to detected wave duration.

Version	Start	End
1	Middle point R and previous R	Middle point R and next R
2	QRS start	QRS end
3	Q	S
4	0.1 s before R	0.1 s after R
5	QRS start	0.43 s after QRS end
6	0.2 s before R	QRS end
7	0.43 s before QRS end	QRS end
8	QRS start	0.2 s after QRS end
9	QRS start	T' wave end
10	P' start	QRS end
11	T' duration before QRS end	QRS end
12	QRS start	P' duration after QRS end
13	P' start	T' end
14	0.2 s before QRS start	0.43 s after QRS end

- N references: select the N references with lowest distance, and store them. This number needs to be optimized.

The recognition experiments are done using D2V1 and D2V2 as new visits. The decision is also taken by calculating the DTW against the stored reference/s. There are two ways to approach the recognition based on the number of references:

- Single references: every segment is compared against the reference, obtaining a distance.
- N references: distances are calculated against the N references, where $1 < N \leq 10$. There are four final available values to determine the performance:
  - Best: lowest distance.
  - Worst: highest distance.
  - Median: median of the N distances.
  - Average: mean of the N distances.

After testing the possibilities with DTW and the different references in both enrollment and recognition, machine learning classification is selected searching for a performance improvement. Two approaches are taken for classification with machine learning algorithms: closed-set and open-set. These experiments also deal with three types of features and their sorting order in PCA:

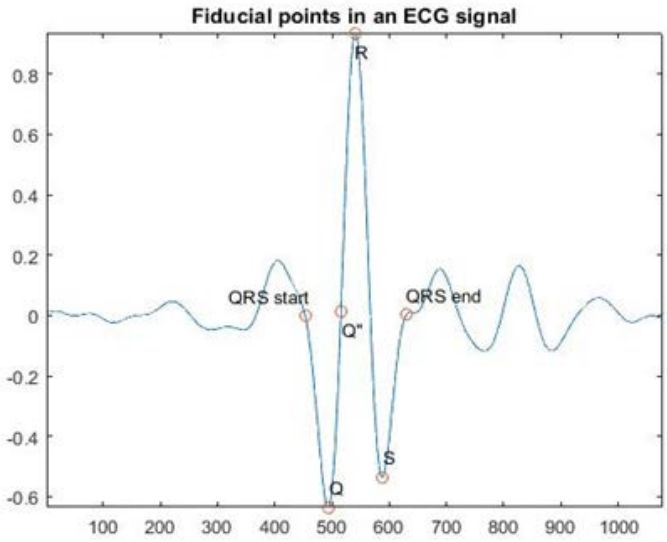


Figure 6.1: Example of the fiducial points used in [101].

1. Time: the whole segment of data. The version may vary according to Table 6.1.
2. DCT: DCT transformation is applied to the time segment data.
3. Metrics: related to time distance between fiducial points and their amplitudes.

6.2. Open-set recognition with GMMs

In the open-set experiments, every user needs to be modeled separately. This implies a large number of training processes and storage. To reduce these limitations, the database is reduced. After heuristically observing tendencies, users 10 to 30 are selected for this experiment.

Previous experiments related to the open-set environment limited the BMSIL database to 20 users. In addition, the obtained results were related to identification. The goal for this experiment is to improve performances with the complete database in verification. For this purpose, GMM is proposed as the classification algorithm. The cycle segmentation corresponds to version 4 in Table 6.1, which corresponds to 0.1 s duration at each side of the detected R peak ($W = rng_1 + rng_2 = 200$).

6.2.1. Model convergence

This experiment was developed using Mathworks and Netlab's GMM Matlab libraries [102]. The GMM's convergence is relevant to be observed in both of the libraries. For a user to be enrolled in the system, the algorithm needs to converge. The number of converged model is tested using data from one session of D1V1, with d = 1. Experiments in section 6.2 provided extra results with PCA for the different applied features (time,

DCT and metrics). This information applies to data for further dimension reduction, which allows to observes how the convergence varies.

To solve convergence issues, the number of Gaussians is fixed to k = 10 while using DCT and metric features over version 4 segments. The convergence is similarly tested observing the different features independently.

6.2.2. Verification

For verification, results are achieved with D1V2 as new data, which in S2 is a standing position, resulting in mixed data for the entire BMSIL database. The score for the EER calculation is the one retrieved by the last sample. The classification is done using Netlab's implementation, as previously referred. Two main parameters are tested in verification: number of Gaussians (k), and the number of attributes with the selection order given by the PCA algorithm.

6.3. Closed-set recognition with Machine Learning: an initial approach

For the closed-set, the idea is modeling using D1V1 as training data by using SVM, k-NN and LDA algorithms, using d = 1. These algorithms require samples of fixed length. Therefore, the selected versions are 4 and 14, which only depend on time criteria and provide fixed segments.

Section 6.2 obtains the accuracy of LDA in a closed-set experiment, and it leads to the attempt of using this algorithm in a verification environment. The enrollment is done with D1V1 data from one session with no available data reduction. In this case, all the first sessions of the remaining BMSIL visits are used for verification: D1V2, D2V1 and D2V2. The first experiment with LDA approaches verification in a way that differs from the general procedures in this thesis: the stored score for the EER calculation is the one corresponding to the last detected cycle of the attempt. Given this situation, the results are further extended with LDA.

6.4. Closed-set verification with LDA

Initial results with LDA have proven this algorithm as a potential solution for ECG biometric recognition. However, some aspects can be improved: the BMSIL database provides 4 different types of visits whose information is collected in different and relevant scenarios. Even though these experiments have been partially covered, it is not done consistently and extensively. The main focus of this section is working with LDA scores, as the previous experiments only use the score from one of the samples in the recognition visit and only considering the first session for training and testing. These issues give a lot of uncertainty as these result could be just a consequence of a coincidence, and not be

representative from all the provided information. In addition, available information is not applied to the problem.

Considering the number of detected samples in each session of the visit, c, and n number of sessions per visit, each user has H_{user} samples per visit. Composing a matrix with all the samples, M_{user} has dimensions (H_{user}, W), where W = rng_1 + rng_2 = 200 in the BMSIL database, using version 4 segmentation. Multiplying H_{user} by the development set proportion, d, and flooring the result gives the different set dimensions with a constant number of features, W. Table 6.2 shows the different samples of each data per user, when using n = 5 and c = 50.

Table 6.2: Available samples per user in every set using D1V1 as enrollment. Development samples are the sum of train and validation.

d	Test	Development	Train	Validation
0.8	50	200	160	40
0.5	125	125	100	25

To evaluate the performance, different type of attempts are considered by evaluating the number of samples that take part. Two proportions of development set are observed: d = 0.5 and d = 0.8. If the verification data proceeds from D1V1 or the remaining scenarios, number of attempted samples vary according to Table 6.3.

6.4.1. Verification with one attempt

The enrollment data is obtained from the D1V1 visit. For the validation of the results, the system uses the accuracy as the assessment metric. However, when final results are achieved, there is an EER calculation. The accuracy of the test set results for D1V1 allows to observe how well data is modeled in the best case scenario. If the results are good, it can be considered as a good model for the remaining visits, allowing to observe how this type of data behaves in both identification and recognition.

Table 6.3: Number of attempts, N_A, depending on the available verification samples and samples per attempt H_A.

H_A \ Samples	50	125	250
1	50	125	250
5	10	25	50
10	5	12	25
15	3	8	16
20	2	6	12
25	2	5	10
30	1	4	8

Table 6.4: EER (%) for both enrollments and type of segment.

Enroll	**D1V1**		**D1V2**	
Recognition / **Version**	**D2V1**	**D2V2**	**D2V1**	**D2V2**
1	12.26	26.40	22.82	26.66
2	22.10	25.06	26.08	25.00
3	24.01	26.27	26.44	26.71
4	19.02	22.37	22.77	23.59
5	19.07	22.02	21.54	21.98
6	23.81	28.05	26.25	27.31
7	26.40	29.89	27.88	29.46
8	19.44	23.01	19.50	22.25
9	18.15	20.97	20.9	23.01
10	20.59	24.96	24.47	25.06
11	22.90	27.30	24.70	27.09
12	20.82	24.43	23.44	25.15
13	16.81	20.05	20.74	21.19
14	19.80	22.63	21.23	20.27

As described in section 5.3.2, every attempt provides different scores and mated and non-mated data, as a consequence. To summarize how the system performs when having as many EER results as attempts, the mean EER and standard deviations are calculated.

6.4.2. Verification with all attempts

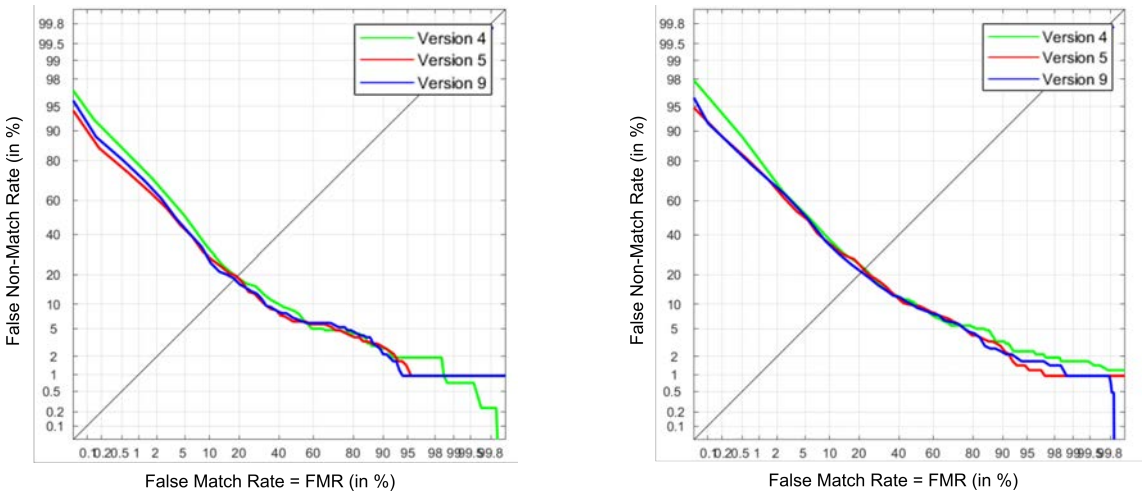
The process for the metric calculation is described in section 5.3.2, by obtaining a single score from averaging the scores from all the attempts with different number of samples, H_A. It results in a single EER, that reflects the system's performance using the specific attempts.

6.5. Results

The results in this section correspond to those experiments previously detailed in the present chapter.

6.5.1. Reference selection with DTW

With the single reference enrollment for each segment version in Table 6.1, the EER is calculated. Results are summarized in Table 6.4 for both types of enrollments.



(a) Recognition with D2V1.

(b) Recognition with D2V2.

Figure 6.2: DET performance with D1V1 in enrollment and segments 4, 5 and 9.

Observing results when D2V1 is the recognition data, the best type of segmentation are 1, 4, 5, 8, 9, 13 and 14, in no specific order. Every type of version is related to the inclusion of different parts of the ECG cycles. Based on this, DET graphs are used for further conclusions.

Segmentation types 4, 5, and 9 include the QRS complex. Version 4 involves QRS time criterion for the delimitation, version 5 refers to including the T wave with time criterion and version 9 also includes the T wave, but using fiducial detection criterion. Figure 6.2 collects DET graphs for D1V1 as training and D2V1 and D2V2 in recognition, Figures 6.2a and 6.2b, respectively. Differences between time and fiducial criteria can be observed by comparing only 5 and 9 segments.

Version 4 is now compared against segments which include the P wave segment, with time constraint and fiducial detection, corresponding to versions 6 and 7, respectively. Results are represented in Figure 6.3. Differences between the two different criteria in P detection can be observed between versions 6 and 7.

Analogously, the same procedure is considered against version 4 with fiducial and time criteria when including the whole P-QRS-T segment which belongs to version 13 and 14. The different DET graphs are collected in Figure 6.4 for D2V1 and D2V2 (Figures 6.4a and 6.4b).

Results conclude in version 13 as the optimal approach for the cycle segmentation. Recalling Table 6.1, this option segments the P-QRS-T complex with the P and T wave fiducial detection.

Different number of references from D1V1 are stored following the version 13, collecting the different types of metrics referred in the design. Results are summarized in Table 6.5 and the corresponding DET represented for 1, 4 and 10 patterns in Figure 6.5

There is also the option to optimize how many cycles are considered a single

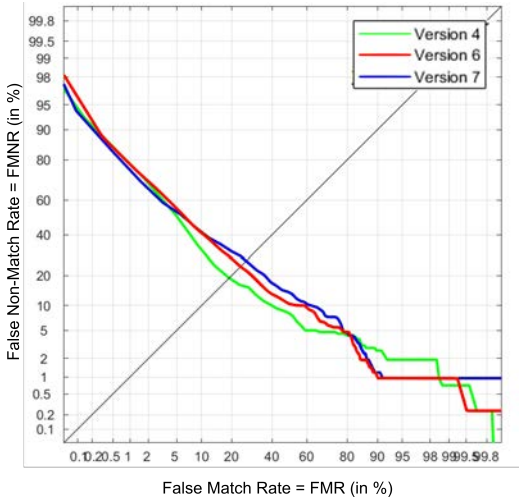
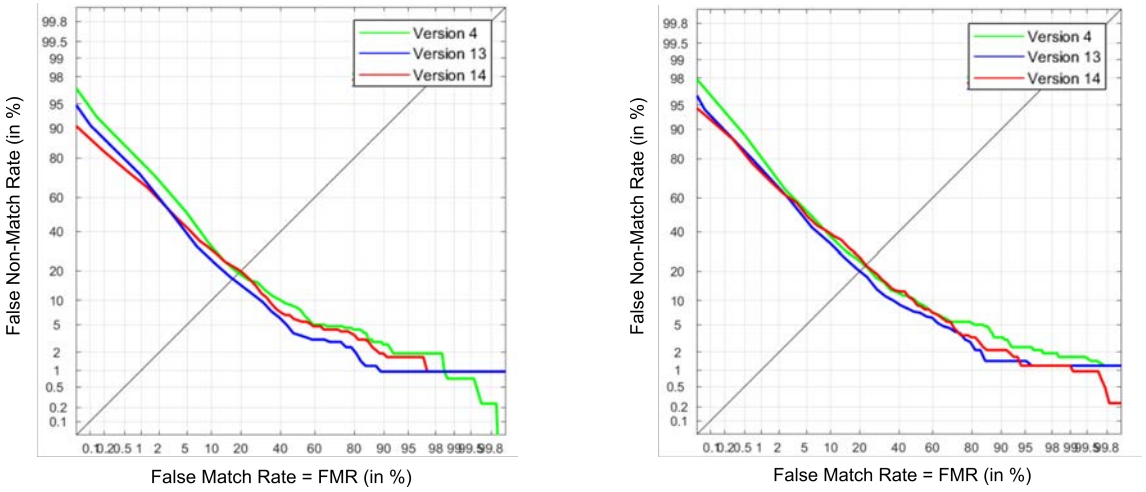


Figure 6.3: DET performance with D1V1 in enrollment and segments 4, 6 and 7 for recognition with D2V1.



(a) Recognition with D2V1.

(b) Recognition with D2V2.

Figure 6.4: DET performance with D1V1 in enrollment and segments 4, 13 and 14.

Table 6.5: EER (%) with D1V1 as enrollment and D2V1 in recognition for version 13 segmentation and different number of references, N.

N / Distance	1	2	3	4	5	6	7	8	9	10
Mean	16.81	17.02	17.15	17.00	16.65	16.78	16.76	16.71	16.71	16.61
Best	16.81	15.83	15.26	14.93	13.89	13.53	13.34	13.13	13.02	12.91
Median	16.81	17.02	17.35	17.02	17.41	16.81	17.11	16.78	16.8	16.75
Worst	16.81	17.41	17.45	17.34	17.27	17.30	17.14	17.09	16.98	16.99

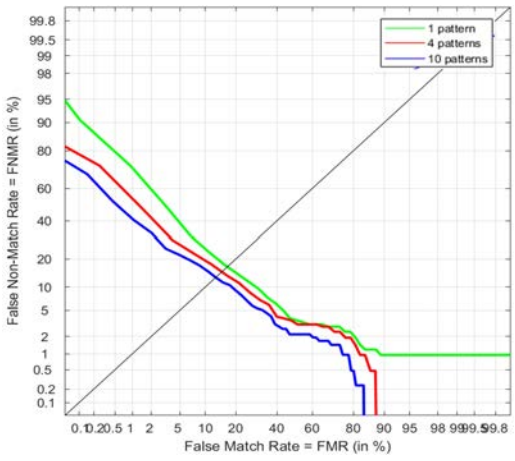


Figure 6.5: DET performance for version 13, with D1V1 in enrollment, D2V1 for recognition and up to 10 stored references.

recognition attempt. The procedure is similar to the comparison against more than one reference. Considering a single reference, all the recognition cycles are compared to it using DTW with the same metrics as previously referred. Best results are detected selecting the lowest distance of all the references. For all the metrics, the performance improves as the number of stored references increase. Storing more referencess decreases the EER from 16.81% for D2V1 and version 13 in Table 6.4 to 12.91% in Table 6.5. If the number of references keeps increasing, the EER should continue improving until reaching a certain number. The stored references are selected based on how well they represent the remaining cycles in the enrollment signal, and getting the lowest distance among a large range of references, could mean that the obtained value is the one belonging to the least representative signal. This issue would cause the EER to increase again. In addition, including more references would multiply the number of comparison and storage capability of the system

Results are plotted in Figure 6.6. Again, the results are analogous to those in the enrollment optimization, as the higher the number of cycles, the best performance considering the best distance. When using average and median, results are similar and do not improve. On the contrary, they decrease as they represent the general performance of all the presented cycles. With the best distance, the EER reaches values under 10%.

The obtained DET graphs in Figure 6.7 show the different evolution under several number of cycles using the best distance, for both D2V1 and D2V2. The improvement in Figure 6.7a is obvious when increasing the number of cycles. However, in Figure 6.7b the impact of this variation is not significant. D2V2 refers working out for the S2, which may be the reason why even dealing with the lowest distance still provides lower performances.

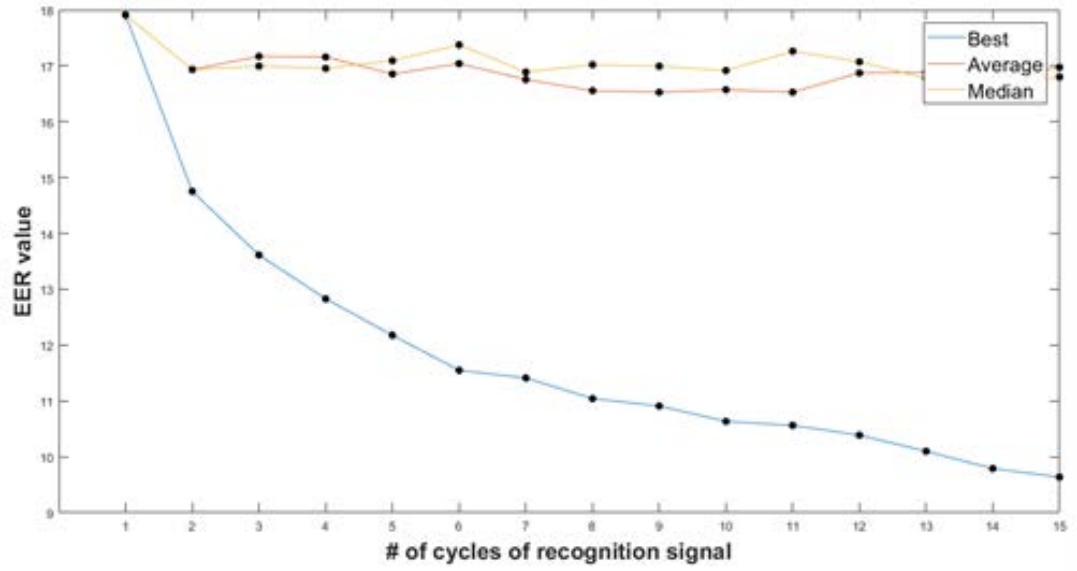
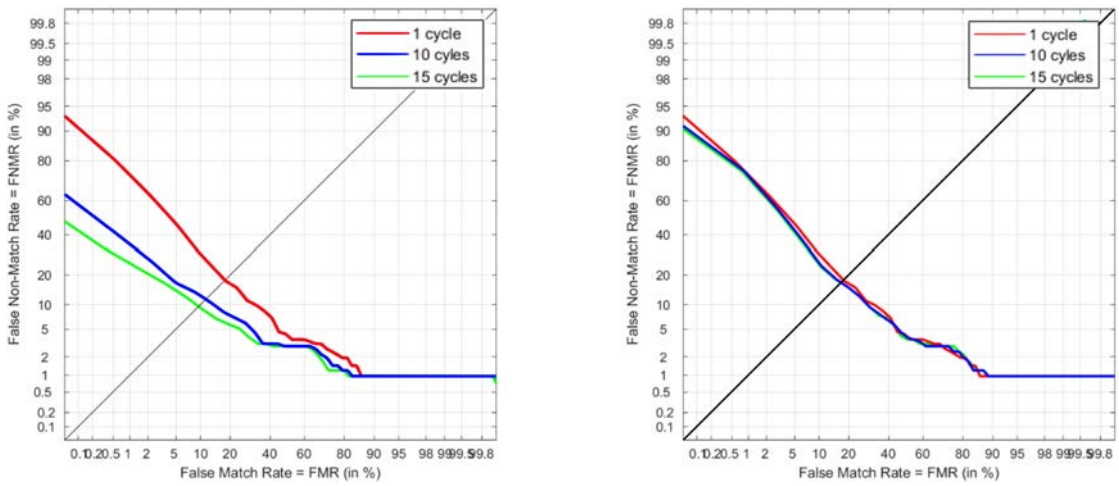


Figure 6.6: Evolution of the EER (%) based on the number of cycles and measures for single pattern enrollment with D1V1 and recognition with D2V1.



(a) Recognition with D2V1.

(b) Recognition with D2V2.

Figure 6.7: DET performance with 1, 10 and 15 cycles in recognition using the best distance.

6.5.2. Open-set recognition with GMMs

The parameter values for the best obtained results under every experiment is collected in Table 6.6. Difference of performance between LDA and other algorithms is obvious. Varying the parameters λ and k in SVM and k-NN respectively, did not result in accuracies above 47%. The performances with the reduced database in an open-set experiment are summarized in Table 6.7.

Table 6.6: Best accuracies for the closed-set experiment.

Classifier	SVM	k-NN	LDA
Features	Version 4: time	Version 14: DCT - time	Version 4: DCT - time
Accuracy (%)	41	47 - 46	97.9 - 97.7

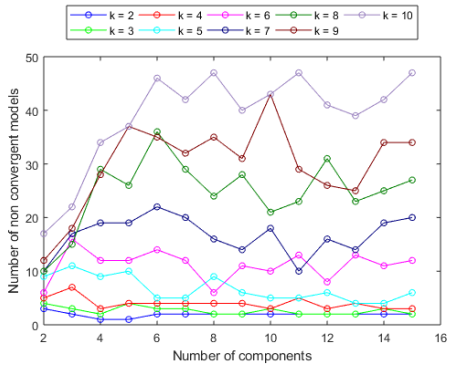
Table 6.7: Best accuracies for the open-set experiment with 20 users.

Classifier	SVM	k-NN	LDA
Features	Version 4: time	Version 4: time	Version 4: DCT - time
Accuracy (%)	83	88	91 - 91

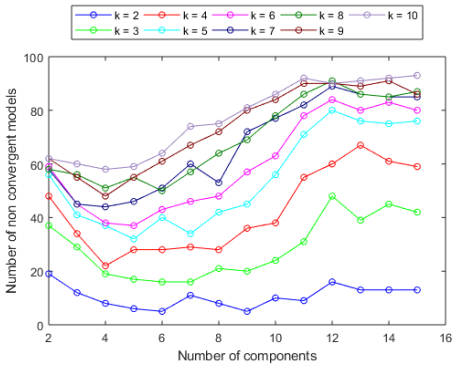
When using all time data from version 4 segmentation, k = 2 for the GMMs, to reduce computational costs for the first approach. Both toolboxes have a maximum number iteration of 100. The number of training models by Mathworks is noticeably higher than results with Netlab: 85 (80.9%) and 17 (16.19%). The convergence in Netlab must be increased, as Mathworks cannot be used for this thesis' goals due to its low convergence. The number of non-convergent models for the different number of features and Gaussians (k) is represented in Figure 6.8.

Figure 6.8a collects results from DCT features, observing an acceptable number of non-convergent results with the lowest values of k, specially between 2 and 4, where the number of features does not really affect convergence. Greater numbers of Gaussians behave worse when dealing with more than 3 features. Analogously, results are plotted in Figure 6.8b for metric features. The behavior noticeably varies from results in DCT, where lower k improve in convergence (lower number of non-convergent models) to a certain point, when they start to spike around 4 o 6 features. In this case, increasing the number of Gaussians from 2 to 3, implies a huge decrease in convergence, where the percentage of non convergent models is almost doubled.

In both cases, the lower number of Gaussians, the more converged models. For values from 2 to 5, the EER is calculated for both types of features, and represented in Figure 6.9. These graphs do not provide the whole data, as we also have to consider the number of converged model for every k and number of features. In the case of Figure 6.9a, between 4 and 6 Gaussians provide good performances, and spike at 7. When the number of components is higher than 9, k = 2 performs better. According to Figure 6.9b, the best values of EER are achieved with the highest number of features and Gaussians. However,

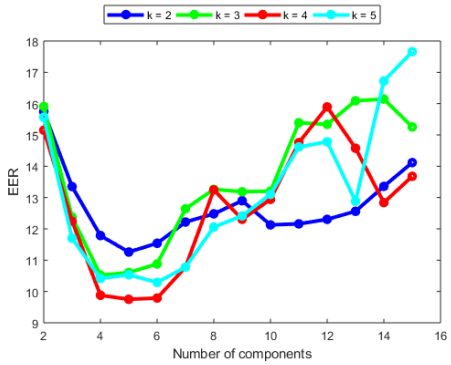


(a) DCT.

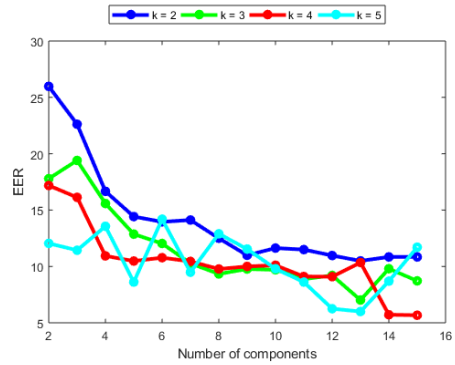


(b) Metric.

Figure 6.8: Number of non-convergent models, based on the number of components (features) and Gaussians (k) when training with D1V1 with DCT and metric features.



(a) DCT.



(b) Shape.

Figure 6.9: EER (%) based on the number of components (features) and Gaussians (k) when testing with D1V2 with DCT and metrics features.

Table 6.8: Results for the parameter combinations that best converge for DCT and metric features.

Feature	Num. of features	Num. of Gaussians	EER (%)	Non-convergent models
DCT	4	2	11.78	1
	5		11.26	
Metrics	6		13.95	5
	9		10.99	

it is easier to achieve better performances when you are dealing with a lower number of models to compare to: the higher number of features and Gaussians, the lowest number of convergent models.

To observe this trade-off more clearly, Table 6.8 collects the two best EER results for the parameter combinations that provide the highest converged models. For both types, the best results for convergence are when k = 2.

The best results for DCT and metrics are close: 11.26 and 10.99% in EER. However,

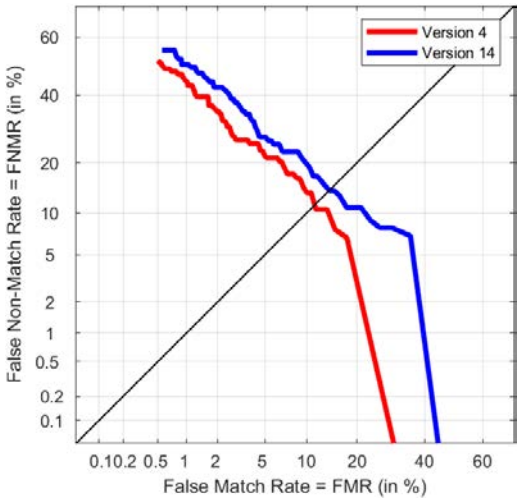


Figure 6.10: DET performance for versions 4 and 14, with DCT features, k = 2 and 5 features.

results in DCT are more significant, as there is only one use that did not reach convergence. Same parameter combinations are tested in the Version 14 segmentation in DCT features. Results are represented in Fiigure 6.10. According to the graph, version 4 still performs better than version 14.

6.5.3. Closed-set recognition with Machine Learning: an initial approach

Using LDA in the whole database as a verification problem provided the DET and FNMR vs. FMR plot. The observed EERs in the DET from Figure 6.11 gives 7.465%, 7.704%, 7.951% and 8.096% EERs for D1V1, D1V2, D2V1 and D2V2 visits, respectively. Figure 6.12 shows the corresponding thresholds, which are around 0.04, giving the idea that mated comparisons do not score high and data is not clearly separated. It is remarkable that the EER for D1V1 is not close to 0, or results in a lower EER given that is the most similar data for the enrollment. One issue that could affect these results is that the verification with this visit also requires less comparisons, as it is formed by the remaining available data from enrollment, which is half the data from the remaining visits.

6.5.4. Extended closed-set verification with LDA

The LDA training achieved good results with the remaining test data in D1V1, reaching accuracy of 99.72% and 99.64% for d = 0.8 and d = 0.5, respectively. Both proportions result in two different models, leaving remaining test sets of 50 and 125 samples per user. Both results are similar so the selected value the enrollment proportion is the lowest one, d_{enr} = 0.5. The average EER and their standard deviations are represented in Figure 6.13, based on the number of samples per attempt, H_A. To make observations more precisely in terms of the mean EER value, Table 6.9 collects all the specific values.

In Figure 6.13a, the higher EER is obtained under D1V1 which is very remarkable,

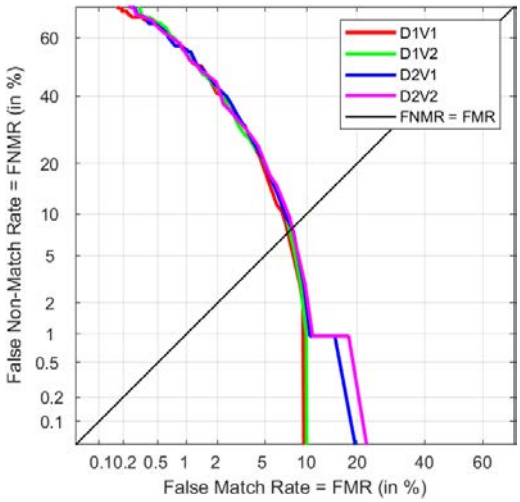


Figure 6.11: DET performances for LDA using one sample per attempt considering all attempts with D1V1 as enrollment and $d_{enr} = 0.5$.

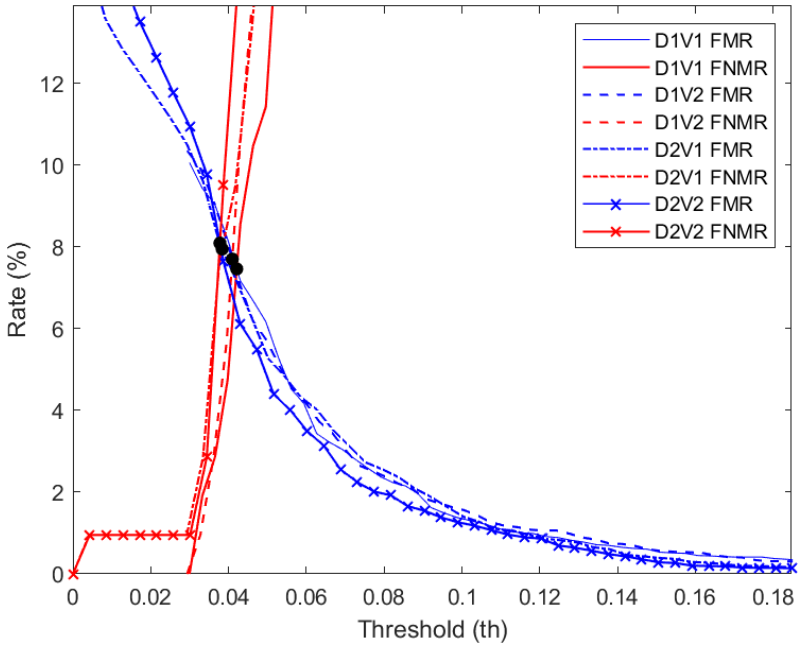
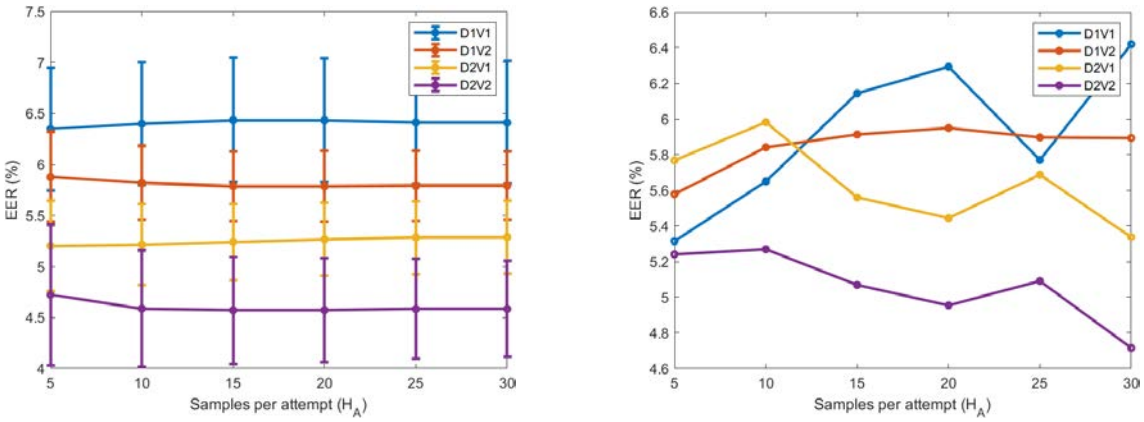


Figure 6.12: FNMR and FMR curves for LDA using one sample per attempt attempts with D1V1 as enrollment and $d_{enr} = 0.5$.



(a) Considering a single attempt.

(b) Considering all attempts.

Figure 6.13: EER (%) average results for both extended verification alternatives with D1V1 as enrollment and $d_{enr} = 0.5$. Consdering D1V1 experiments contain half the samples as the remaining.

as the best results are obtained with verification using D2V2. This could be, initially, provoked by two factors: firstly, the verification samples for D1V1 are half the samples in the remaining visits, involving less attempts and secondly, that mated and non-mated data could be less separated in the different attempts in the case of D1V1. In addition, performances throughout the different number of samples per attempt is not really remarkable, as standard deviations are similar along along the x-axis. When $H_A = 20$ the average EER ranges between 4.571% and 6.433%.

When considering all attempts in Figure 6.13b, the patterns with respect to H_A are similar in the second day, obtaining lower performances in the case of resting. This alternative uses all the possible attempts, so choosing the highest value of H_A does not really affect to the recognition process, considering $H_A = 25$ as the most suitable value for D1V1 when not dramatically impacting over the remaining experiments. Table 6.9 also collects the threshold ranges where the EER is calculated for every visit. The threshold fixation could impact the performance in one way or another, impacting the false positive and false negative ratios. The wider ranges of the threshold value are given by the D2V2 as it could be expected by the nature of the data.

Table 6.9: EER (%) average results considering the two types of extended verification considering H_A. D1V1 is the enrollment and $d_{enr} = 0.5$.

Visit	D1V1		D1V2		D2V1		D2V2	
Type / H_A	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts
5	6.349±0.602	5.315	5.877±0.442	5.578	5.201±0.444	5.768	4.721±0.691	5.241
10	6.400±0.600	5.649	5.819±0.365	5.841	5.214±0.398	5.982	4.586±0.572	5.269
15	6.435±0.610	6.144	5.789±0.343	5.913	5.241±0.371	5.560	4.570±0.524	5.069
20	6.433±0.606	6.293	5.787±0.350	5.949	5.265±0.359	5.444	4.571±0.511	4.955
25	6.413±0.606	5.770	5.790±0.345	5.897	5.283±0.361	5.687	4.582±0.489	5.091
30	6.411±0.600	6.419	5.792±0.337	5.893	5.287±0.361	5.337	4.543±0.610	4.715

6.6. Conclusions

The initial stage of this chapter provided a reduction of the segmentation approaches for the ECG segmentation based on the R peak detection with an specific algorithm. Through the use of DTW the best choices were reduced to two: version 4, which refers to a temporal windowing of 0.2 s, whose center is the detected R peak; and version 14, which starts 0.2 s prior to the theoretical QRS starting point and finishes 0.43 s after the theoretical QRS end.

After narrowing down the segmentation criteria, several Machine Learning algorithms were preliminary tested and their results are summarized in Table 6.10. The performance evaluation was mainly based on accuracy and enrollment with the entire D1V1 set, considering identification with the same scenario in another day (D2V1), and using only the score from the last sample comparison. The closed-set with the entire BMSIL database was only doable through the use of LDA classification, as SVM and k-NN algorithms did not retrieve good results considering the ideal conditions of the enrollment an recognition. The open-set approach required a noticeably higher memory management, and the BMSIL database was then reduced to 20 users. The three algorithms performed better, but for SVM and k-NN was not possible to know if it was caused by the shrinking of the database.

Table 6.10: Summary of the initial results for identification for the different algorithms tested in the present chapter. The parameter d refers to the proportion of the visit used for enrollment.

Classifier	Type	Users	Segmentation	Features	Enroll	Recognition	Accuracy (%)
SVM	Closed-set	105	Version 4	Metrics	D1V1 (d_{enr} = 1)	H_A = 1. Last attempt.	41
k-NN			Version 14	Metrics/DCT			47/46
LDA			Version 4	Metrics/DCT			97.7/97.9
SVM	Open-set	20	Version 4	Metrics			83
k-NN				Metrics			88
LDA				Metrics/DCT			99/91

More complex approaches were achieved based on improving results in open-set with all users BMSIL database, as well as to further research the potential of LDA in closed-set experiments. The final results are collected in Table 6.11. This time, to narrow down the complexity of identification, results were evaluated for verification. GMMs were successfully tested considering variations in the position and same position data collected in the same day (D1V2), obtaining the best result under DCT features and training with the entire enrollment visit. The convergence only failed in one user, which implied less than 1%, and allowed to perform the open-set task and an acceptable verification result considering the conditions.

Results for closed-set the best results were achieved using 50% of the visit for enrollment, allowing to test the verification under the remaining data. The tested algorithm was LDA, as its EER with reduced data in identification was desirable, as summarized in Table 6.11. In this case, no data reduction nor transformation is demanded,

using all the time points of the selected QRS segmentation. Results are improved with respect to GMMs even when including more complex scenarios (such as those for D2V2) and reducing the number of attempts.

Table 6.11: Summary of the best results for verification and the different algorithms tested in the present chapter. The parameter d refers to the proportion of the visit used for enrollment. In identification, the metric is accuracy, when in verification it refers to the EER. Visit where X and Y can be substituted by 1 or 2.

Classifier	Type	Users	Segmentation	Features	Enroll	Recognition	EER (%)
GMM	Open-set	104	Version 4	DCT	D1V1 (d_{enr} = 1)	D1V2. H_A = 1. All attempts.	11.26
LDA	Closed-set	105	Version 4	Time	D1V1 (d_{enr} = 0.5)	DXVY. H_A = 1. All attempts.	7.465–8.096
						DXVY. H_A = 20. One attempt.	4.571–6.433
						DXVY. H_A = 25. All attempts.	5.091–5.897

This chapter has proven the viability of ECG recognition under the BMSIL database characteristics, which imply including exercising data and visits in different days. Moreover, it settles an starting point for further improvements and study of other factors and algorithms.

7. MULTIMODAL VERIFICATION

In chapter 6 the ECG has been tested as a potential good modality, even more when the user is resting. However, classic biometric modalities still provide better performances, such as fingerprint recognition which has verification performances with EER = 0.1% or lower, depending on the case scenario [103]. However, this modality is susceptible to attacks, and the increase of its use in mobile biometrics has increased the interest on Presentation Attack Detection (PAD) and fusion.

The ECG as a biometric trait provides advantages that are not usual in fingerprint-based systems, as they are more difficult to access, provide liveness detection and it is a continuous signal. The addition of this modality could lead to an improvement of the fingerprint performance, helping with the rejection of false presentations.

7.1. Fingerprint performance

The BMSIL database contains data related to fingerprints of every user that took part in it. Each finger was collected twice to provide a reference and a comparison sample. The Innovatrics [104] algorithm analyzes the quality of the fingerprint image and extracts its reference based on the detected minutiae. The DET and EER get calculated considering only data when the reference is the right index in the first samples. The scores range from 0 and 1000, and get normalized between 0 and 1 to have the same range as those scores in ECG.

7.2. Score fusion

Score fusion consists on merging both modalities' results into one to enhance the system's performance. For this purpose, weighed sum in Equation 7.1 is proposed, where A is the weight for ECG scores (s_{ECG}) and B is for fingerprint (s_{fp}). This concept is represented in Figure 7.1 when using the scores obtained from with LDA using the last sample verification in chapter 6.

$$s_{fusion} = A \cdot s_{ECG} + B \cdot s_{fp} \tag{7.1}$$

7.3. PAD

The score fusion does not allow to detect attacks. The sequential comparison of both modalities could produce better results discarding forgeries, as ECG allows to detect if

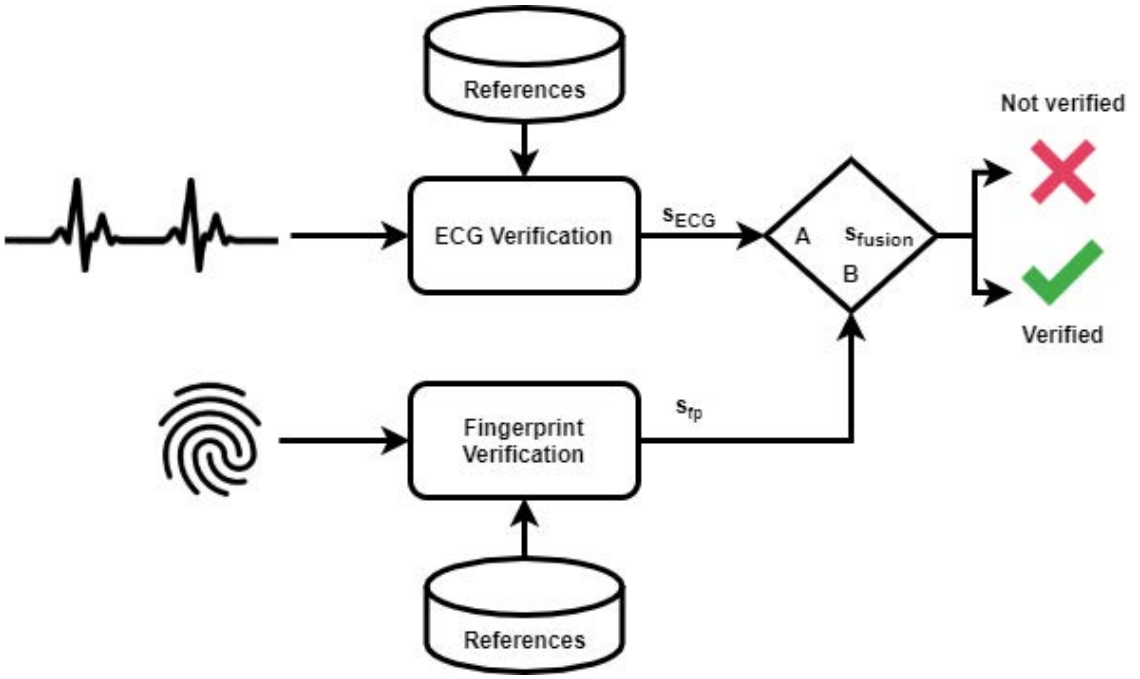


Figure 7.1: Score fusion scheme for ECG and fingerprint.

the sample comes from an alive user, and how likely it is for the following fingerprint to be the correct one. As it has been seen, the ECG is not as accurate as fingerprint, but establishing a minimum threshold could potentially help determining which user is suitable to continue with the fingerprint verification. This threshold is obtained from those scores obtained with LDA in chapter 6

The followed steps are represented in Figure 7.2, where the ECG is initially required, and the scores s_{ECG} are filtered with a threshold. If the obtained score for the ECG is higher than the threshold, the verification continues asking for the fingerprint data, and considering it the final system's performance. On the contrary, if the threshold condition is not met, the system stops considering the data as an attack.

7.4. Results

This section collects the results of the PAD and fusion approaches for ECG and fingerprint.

7.4.1. Fingerprint performance

The fingerprint database has an EER of 0.218%. The DET representation is in Figure 7.3. One must consider that these performances in mobile fingerprint biometrics usually have lower performances, as the computation and sensors are more limited.

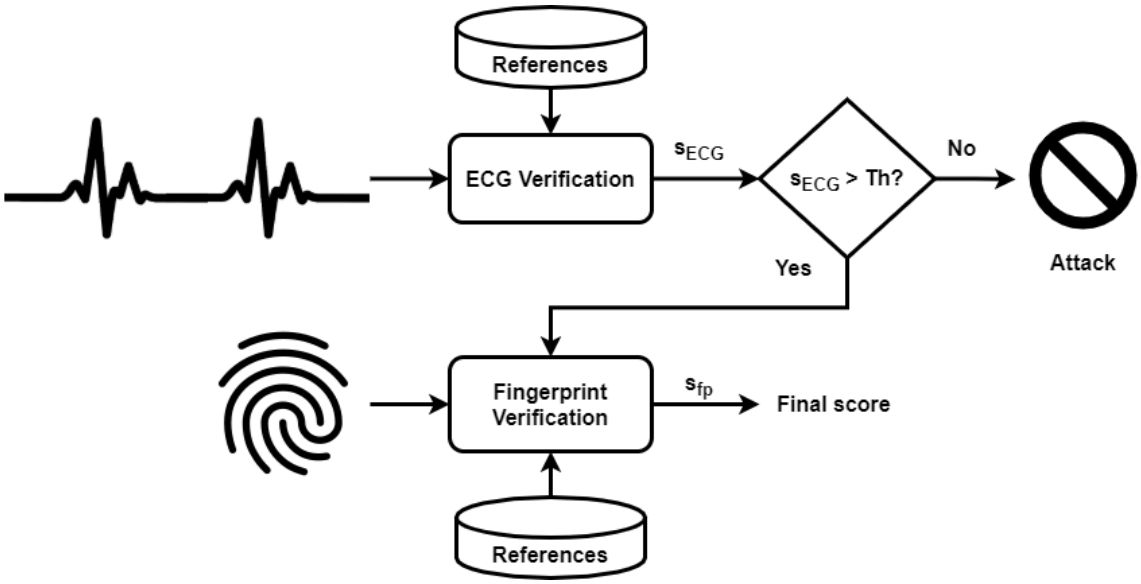


Figure 7.2: PAD scheme for ECG and fingerprint.

7.4.2. Score fusion

The observed results are summarized in Table 7.1 under the different obtained values of A and B. The best improvement is achieved under A = 1 and B = 2, giving more relevance to those results in fingerprint. Compared to the fingerprint results by themselves, they were improved from 41.28% to 70.64%. These results are easily observed in the DET from Figure 7.4 and the distributions in Figure 7.5. In all three distributions, the non-mated scores are well differentiated from the mated ones. This behavior correspond to accurate true negatives or very low FNMR rates. The scores in mated comparisons are generally more heterogeneous.

Table 7.1: EER (%) results for the different verification data and A and B combinations in score fusion, as well as the improvement with respect to the initial fingerprint performance.

Verification	A	B	EER (%)	Improvement (%)
D1V2	1	1	0.092	57.80
D2V1			0.119	45.41
D2V2			0.137	37.15
D1V2	2	1	0.101	53.67
D2V1			0.128	41.28
D2V2			0.147	32.57
D1V2	1	2	0.064	70.64
D2V1			0.101	53.67
D2V2			0.128	41.28

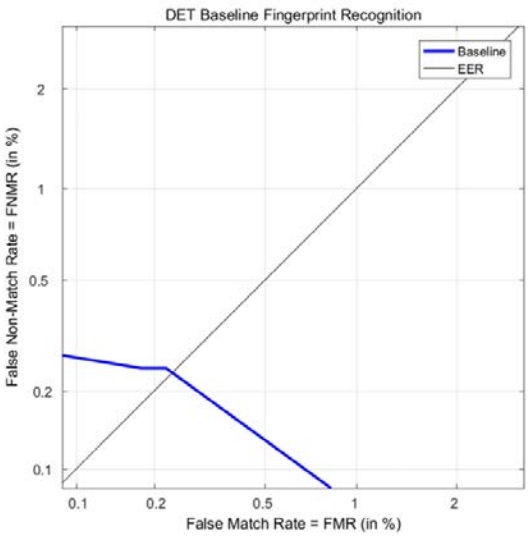


Figure 7.3: Fingerprint performance in the BMSIL database.

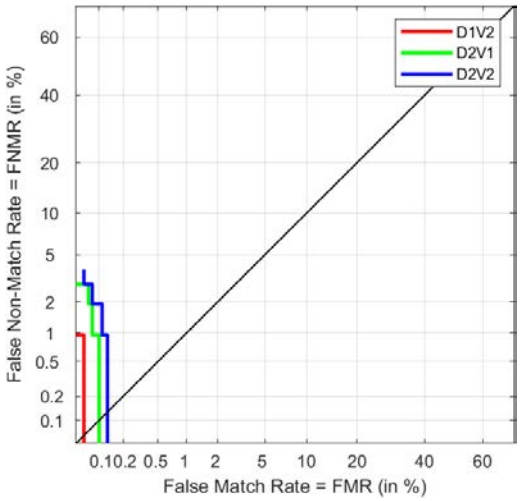
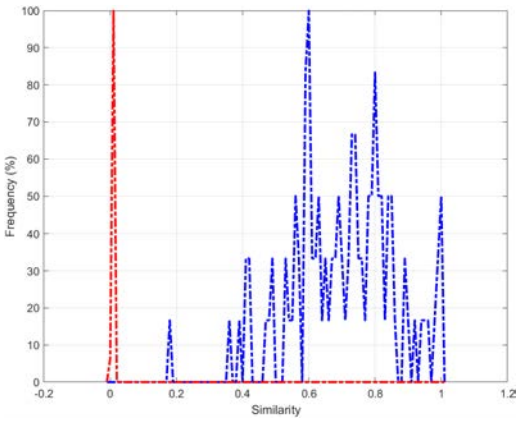


Figure 7.4: DET graph for the different performances with A = 1 and B = 2 in score fusion.

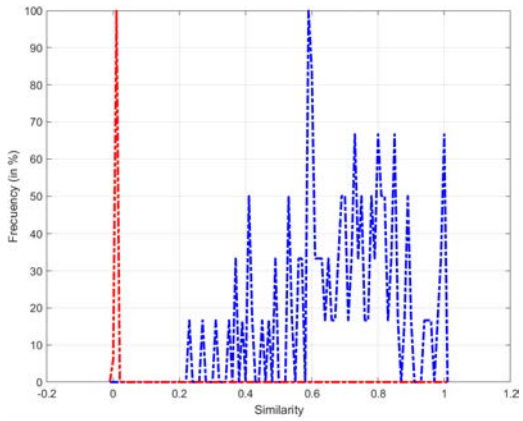
7.4.3. PAD

To observe the different results based on the threshold for ECG, several thresholds are tested to observe the performance. The percentage of discarded samples for mated and non-mated comparisons are collected in Table 7.2 for the different type of verification data.

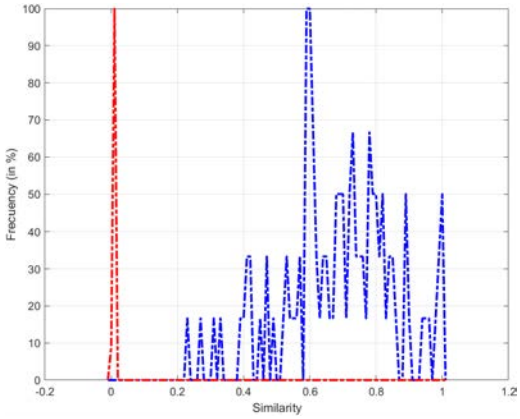
Logically, the higher the threshold, the highest number of discarded non-mated comparisons. However, it also implies a higher mated discarding, which is not desirable as it would imply that a genuine user would require more attempts to be verified. The best trade-off is obtained under the score threshold of 10^{-10}. As the non-mated discarding is 99.222%, it implies that only 0.778% of possible attacks are not detected as such. According to ISO/IEC 30107-3, this results is the Attack Presentation Classification Error Rate (APCER). Similarly, the genuine comparisons detected as attacks is the



(a) D1V2.



(b) D2V1.



(c) D2V2.

Figure 7.5: Distribution graphs for A = 1 and B = 2 in score fusion. Red belongs to non-mated scores, and blue to mated scores.

1.905%, considered the Bonafide Presentation Classification Error Rate (BPCER). After this thresholding, the final alid scores are those provided by fingerprint comparisons, so the system summarizes in 0% EER.

7.5. Conclusions

This chapter has proven the potential of ECG as a multi-modal approach with a conventional biometic trait as fingerprint. The fusion of these two signals strengthens the outcomes that each one have individually. A score fusion approach is slightly affected by the ECG scores, as fingerprint already provides enough precision on its own. However, they are still improved with help of ECG up to 70%. These improvements could be even higher considering mobile recognition, as the fingerprint performance usually decrease due to limited computation and sensor quality. The multi-modal approach has been more productive when applying it as a PAD scheme. This approach characterizes the final system with liveness detection and produces an initial discarding, avoiding presentation attacks. This process results in detecting more than 99% of the attacks, and the valid

Table 7.2: Percentage of discarded mated and non-mated comparisons for different threshold in the PAD scheme.

Verification	Threshold	Discarded mated (%)	Discarded non-mated (%)
D1V2	10^{-10}	1.905	99.222
D2V1		3.809	99.286
D2V2		4.762	99.066
D1V2	10^{-5}	4.762	99.725
D2V1		6.667	99.744
D2V2		7.619	99.615
D1V2	0.1	6.667	99.890
D2V1		8.571	99.872
D2V2		9.524	99.853
D1V2	0.4	6.667	99.908
D2V1		10.476	99.881
D2V2		13.333	99.863
D1V2	0.9	10.476	99.945
D2V1		12.381	99.927
D2V2		14.286	99.918

fingerprint scores give a system with 0% EER.

These results are a good initial approach for the use of both modalities, but further research on this topic would be required to finally confirm its benefits. In addition, the database did not provide elaborated Presentation Attacks, such as those with fake fingerprints, so the correct performance of ECG in PAD is not really observed, as results could be highly improved. However, the application of these approaches could be really relevant once applied in mobile biometrics.

8. ECG VERIFICATION USING MULTILAYER PERCEPTRON

Chapters 6 and 7 have dealt with the complete BMSIL database. Throughout the achieved results, we can infer that data collected under scenarios different to the one in enrollment provoke lower performances. This chapter is initially planned to be focused on the S2 part of the BMSIL database, as it is the one with different acquisition scenarios, both in time and physical conditions. However, it is further extended to the entire BMSIL database as the experiments evolve.

The present chapter aims to improve the previous results using MLP algorithm implementation by *scikit-learn* in Python 3 [99]. In addition, considering the potential use in mobile devices, the feature transformations are reduced to the minimum to avoid extra computations. It is based on the process that was carried out in [2]. However, even though the experiments are executed in the same way, the results have been re-calculated after error correction of the developed code.

8.1. Input data

The data pre-processing is formed by the 1-35 Hz band-pass filter referred in previous chapter 5. The R peak detection is carried out with the BMSIL algorithm in section 5.2.1. This algorithm requires the first and second derivative calculations to proceed with the peak detection. Therefore, using the differentiated QRS complexes does not add computation complexity, as they were already calculated. Figure 8.1 shows the data pre-processing and preparation for the three differentiation matrices: ND (Non-Derivative), FD (First Derivative) and SD (Second Derivative). The number of samples per user, H_{user} is determined by the number of cycles detected in each session of the visit. In this case, as it happened in chapter 6, the number of cycles per session is fixed to 50, c = 50. All the five available sessions are used for this chapter so n = 5. Therefore, the available data per user is given by $H_{user} = c \cdot n$, which results in 250 samples. The variable W refers to the cycle length or number of features in the sample. Experiments in previous chapters have shown that segmenting in 0.1 s with the R peak in the center is a good approach. Therefore, considering the sample frequency, $W = rng_1 + rng_2$, where $rng_1 = rng_2 = 100$. The FD and SD matrices have the same length as ND, as the segmentation is obtained from the differentiated signal, not after the segmentation.

8.2. Hyperparameter optimization

The hyperparameter optimization process is done independently for each of the differentiation types. In this section the type of matrix is not referred, as the process

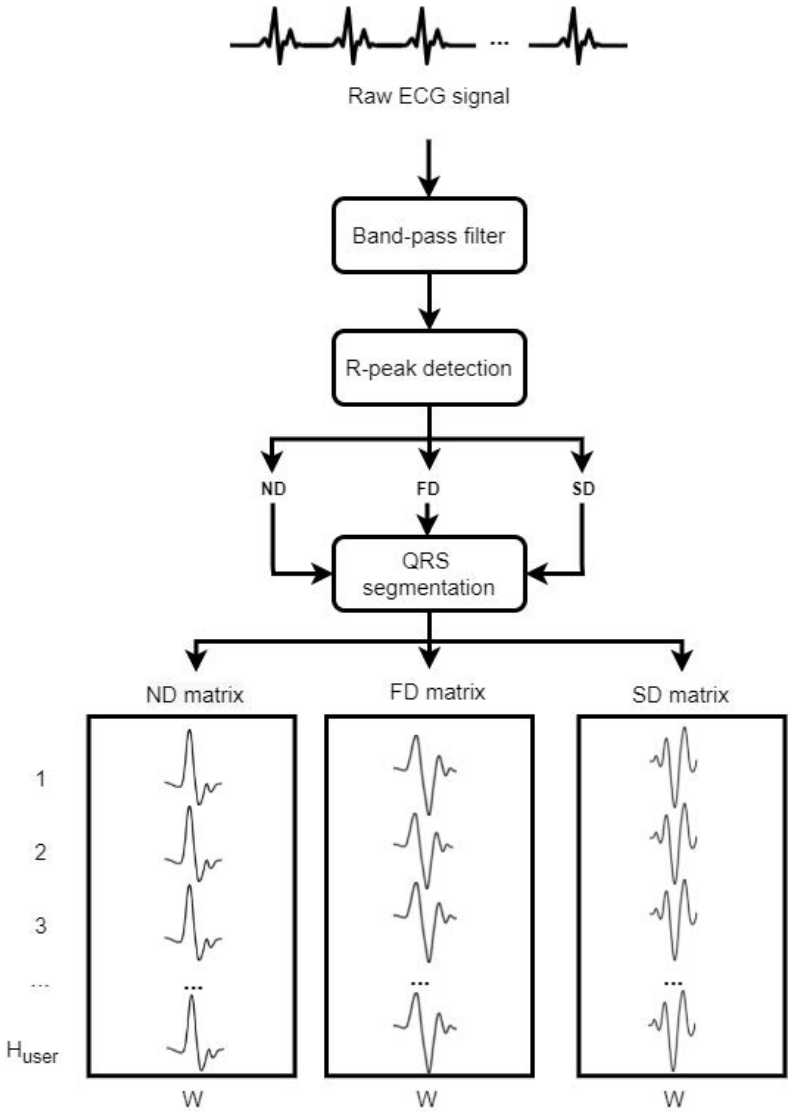


Figure 8.1: Scheme of pre-processing and data preparation for one user.

is common for ND, FD and SD. However, the final hyperparameter sets are likely to be different.

This process aims to find the set of hyperparameters for the MLP algorithm that best fit the training data. To avoid large computational costs as a result of the large number of possible hyperparameters, some of them get fixed based on previous knowledge. The possible values that every hyperparameter can get are defined and random combinations chosen and tested using the Random Search and Exhaustive Grid methods. The best set is then retrieved to proceed the evaluation of the different performances.

8.2.1. Fixed hyperparameters

Fixing some of the hyperparameters facilitates the following tuning process. The number of layers is set to one, as it usually performs properly and avoids extra slowdown [105]. The used optimizers are Stochastic Gradient Descent (SGD), Adam [106]; which is SGD

based, and a quasi-Newton optimizer, L-BFGS. However, preliminary trials showed very low performance results with all of them except for Adam. These results allowed to fix the optimizer to Adam, discarding the selection of a specific learning rate, as it is only required for SGD optimizers. As a consequence, there is a learning rate step, which is the quantity to upgrade the weights and it is fixed to the *scikit-learn* default value. Finallly, the number of iterations without change is heuristically fixed to 10, and refers to the number of epochs that need to present no change to consider convergence. The fixed hyperparameter values are summarized in Table 8.1.

Table 8.1: Summarization of the fixed MLP hyperparameters.

Hyperparameter	Description	Value
Hidden layers	Number of hidden layers.	1
Solver	Function used for weight updating.	Adam
Learning rate	Function used to update the learning rate that takes part in the optimizer.	Not applicable with adam optimizer.
Learning rate step	Step size for the learning rate updates.	0.0001
No change iterations	Number of iterations with no relevant change to consider convergence.	10

8.2.2. Tuning process

After fixing some of the hyperparameters, there are still remaining values that need to be determined by hyperparameter tuning. The number of hidden layers is set to one, however, the number of nodes needs to be determined. In addition, the activation function can be any of those in Table 5.1, and it is optimized in tuning. The alpha (α) parameter belongs to the L2-regularization component $\alpha\|W\|_2^2$, where α or alpha is the penalty term and $\|W\|_2$ represents the Euclidean norm of the weights. The value of alpha avoids overfitting it is also included in tuning. The final value to set is the tolerance, which determines the improvement in the loss that needs to be improved in order to keep iterating.

The tuning process follows the steps in Figure 8.2, where H_{db} refers to all the available samples for the tuning process, considering all users. In this case, as all the different users provide the same number of samples, $H_{db} = H_{user} \cdot U$, being U the number of users in the database, which is 55 in the case of S2. The whole process is based on the EER metric instead of accuracy, by using a customized Python function that calculates the EER one sample per attempt, considering all attempts (section 5.3.2). The possible array of values for the different hyperparameters is collected in Table 8.2. The activation functions are those mentioned in Table 5.1, and the numerical values try to represent extreme values in small steps to observe how determining each hyperparameter is.

The different parts of the tuning process have the following purposes:

- Stratified Shuffle Split: divides the data into development set with H_d samples, and

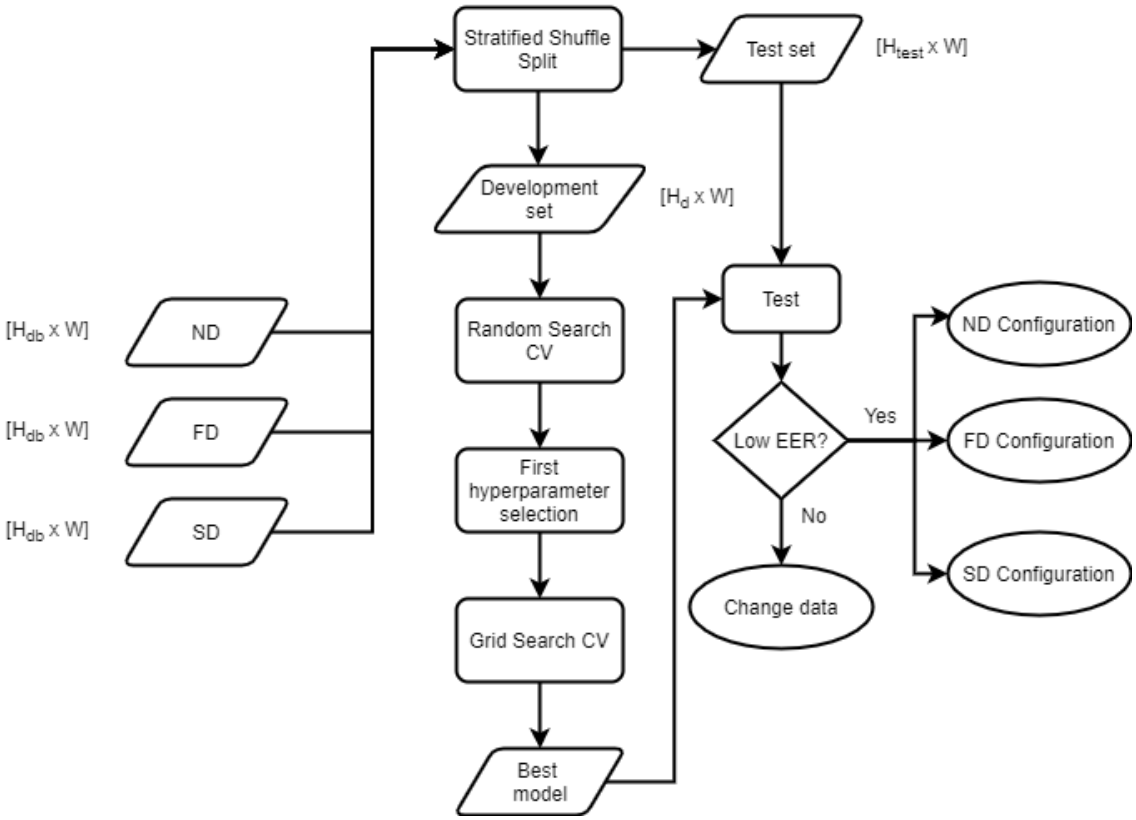


Figure 8.2: Hyperparameter tuning steps.

Table 8.2: Possible values for the remaining hyperparameters.

Hyperparameter	**Possible Values**
Hidden layer size	1, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250
Activation	Identity, logistic, tanh, ReLU
Alpha	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 5, 10, 50, 100
Tolerance	0.01, 0.05, 0.1, 0.5

test set, with H_{test} samples. The development size proportion specified by d. This division shuffles the samples so they do not have temporal correlation and keeps the category proportion, which is the same for all the users.

- Random Search CV: a determined number of hyperparameter combinations are evaluated using cross-validation (CV) with 5 folds, assessing 50 for this experiment. This process implies dividing the development set into training and validation sets, with a 80%-20% ratio.

- First hyperparameter selection: the best 3 results from the Random Search are selected, and their hyperparameter values fed into the following step to narrow down the possible combinations.

- Grid Search CV: similar process to Random Search, but evaluating all the possible

combinations based on the given value possibilities. The combinations are reduced as a consequence of the previous steps, and returns the final best result.

- Test: the best set of hyperparameter values is then used for training the final model using the whole development set. It then gets tested with the test set to observe its performance. As the results belong to the same data scenario, the EER should be low to confirm the correct modeling. This testing process provides intra-class information, as further explained in the following section.

8.3. Optimization of the design

At this point, we must differentiate two similar parameters: d and d_{enr}. The proportion of visit in the data used in the development set for tuning is defined by d. However, the parameter d_{enr} refers to the proportion of data used for enrollment. This differentiation is included in this chapter to avoid having extreme performances that could be only related to the size of the development set.

Once the hyperparameter tuning is achieved using proportion d, there are more required parameters to characterized the system. These parameters are not part of the classification model, but affect the data and have impact in its optimal configuration. In this case, we focus on three aspects: type of differentiation, enrollment size (d_{enr}) and characteristics of attempts in verification. The process consists on narrowing it down from first to last until the final system characterization.

8.3.1. Differentiation

Obtaining the test results with the final hyperparameter set for each of the three differentiation already gives information about the inter-class variability, as they are obtained under the same conditions and visit, D1V1. These results are considered the baseline, as they are a result of the best possible conditions. However, as the S2 provides 3 extra visits in different scenarios and days, it is interesting to further observe the differences in those cases. The test stage in Figure 8.2 is also represented in Figure 8.3, when the used model is the best one obtained in the Grid Search, therefore $d = d_{enr}$ for this scenario.

From now on, d_{enr} is used instead of d because we are not referring to the tuning process anymore, but the final model training with the given enrollment information. When $d_{enr} = 1$, there is no remaining information for testing with that data, so the DET for D1V1 cannot be calculated. If $d_{enr} < 1$, the test set with D1V1 has H_{test} samples and the training set, H_{enr} samples.

The inter-class variations are represented when obtaining DETs for those experiments that do not belong to the enrollment data D1V1. Comparing these results to those in D1V1 gives information about the data differences that could produce a lower performance. In

these cases, the whole information is used for testing by using one sample per attempt, considering all the attempts. This results in testing sets of H_{db} samples.

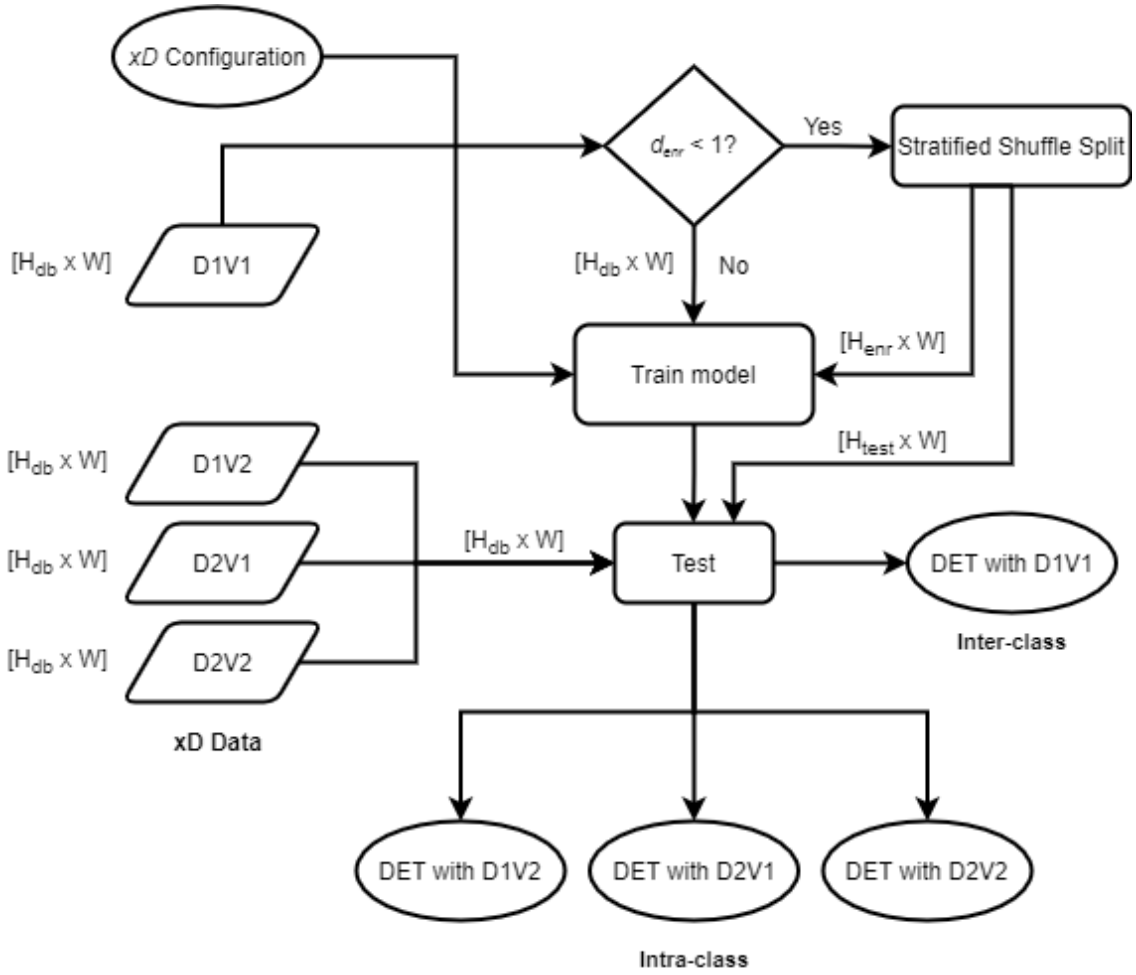


Figure 8.3: Steps to obtain the DETs for all the possible visits.

8.3.2. Enrollment size

Once the best differentiation is determined, it can be used to observe how changes in enrollment size affect the result. They are tested for values of 0.3, 0.5, 0.7 and 0.9. As the number of used samples may change the behavior of the network, the same values are used for d in tuning, resulting in four different sets of hyperparameters.

Every possible enrollment proportion is used in all the four possible hyperparameters combinations, determining the final model that will represent the system.

8.3.3. Extended verification

After considering the best differentiation and enrollment size, we can further observe the effects of different approaches in attempts. So far, the results are obtained using EER

considering the one sample per attempt with all the possible available attempts. This section follows the alternatives previously referred in section 5.3.2, considering 5, 10, 15, 20, 25 and 30 samples per attempt in both options.

8.4. Results

Once the optimization is achieved for S2, the same process is carried out with the entire BMSIL database following the same steps.

8.4.1. Optimization of the design

Differentiation

The first stage for the optimization is getting the system tuned to compare all types of differentiation. In this case, tuning is done with d = 0.5 for D1V1 in S2, using the remaining data for testing. The retrieved EER for these features would allow to assess which one performs better. The visit D1V1 provides data in resting conditions while sitting down, which is expected to provide more stable and feasible data. The mean EER results and the corresponding hyperparameter values obtained from the Random Search CV are summarized in Table 8.3 and come from 50 different combinations. The results in this initial search already show how potentially different the performances of ND, FD and SD are going to be.

Table 8.3: Best hyperparameter values in Random Search CV and their EER (%) for ND, FD and SD.

Signal	Hidden Layer Size	Activation	Alpha	Tolerance	Mean EER (%)
ND	350	ReLU	0.01	0.5	0.040
	350	Identity	1	0.5	1.212
	700	ReLU	1	0.05	1.223
FD	500	ReLU	0.0001	0.05	0.175
	700	ReLU	0.005	0.05	0.505
	400	Identity	0.01	0.1	0.593
SD	500	Identity	0.005	0.01	3.347
	350	Identity	0.0001	0.05	5.280
	350	Identity	0.01	0.01	6.218

The different values for the hyperparameters are fed into the Random Search CV, obtaining the best three sets of values for each one of the differentiated matrices, as seen in Table 8.4. The right column has the EER results for the remaining test set of D1V1, which is equivalent to saying that $d_{enr} = d$.

Table 8.4: Best hyperparameter values in Exhaustive Grid Search and their EERs for ND, FD and SD. Each final set of values has the corresponding testing result.

Signal	Hidden Layer Size	Activation	Alpha	Tolerance	Mean EER (%)	Test EER (%)
ND	350	Identity	1	0.05	0.761	0
FD	400	Identity	0.0001	0.1	0.128	0
SD	500	Identity	0.005	0.01	3.151	2.290

The results in Table 8.4 show a good separation between data in ND and FD, being lower in the case of SD. These results imply good training and sufficient inter-class variation for the verification process. However, comparing these results with the remaining visits is insightful of the intra-class variability between different physiological scenarios. The results are numerically summarized in Table 8.5 and graphically represented in Figure 8.4. Generally, the best performances are obtained with FD, as the EER are the lowest for all the visits. ND represents a huge increase in the scenario with exercise, D2V2. In terms of reducing the EER in different scenarios, SD successes, as the EER in visits D2V1 and D2V2 decrease. However, the differences in performance between D1V1 and the remaining visits show the lack of generalization with this data.

Table 8.5: EER (%) performances when $d_{enr} = 0.5$.

	D1V1	D1V2	D2V1	D2V2
ND	0	3.636	4.810	11.970
FD	0	3.636	3.112	5.454
SD	2.290	7.273	6.003	5.454

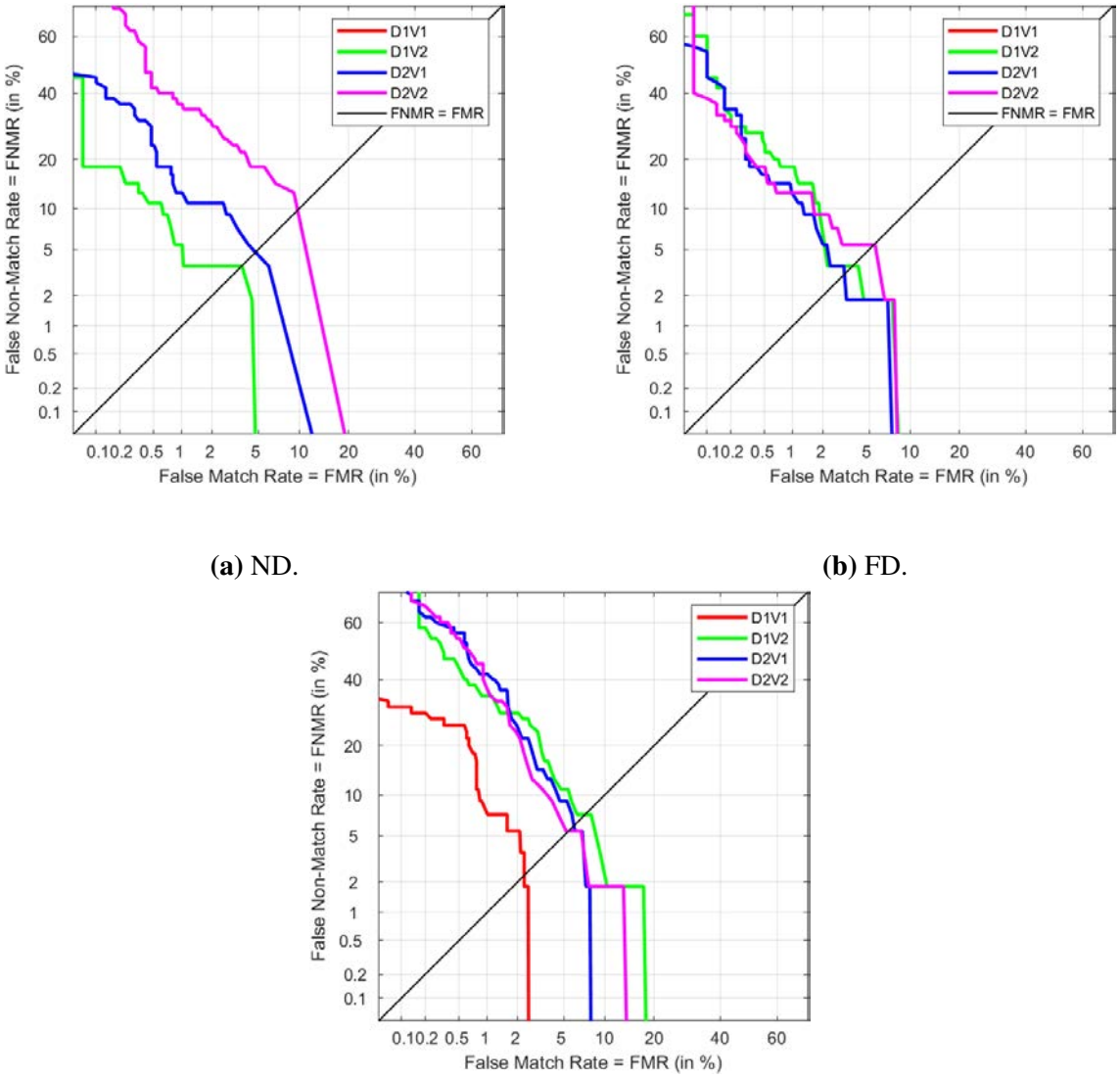
Enrollment size

Once that the FD is determined as the best differentiation, the tuning is achieved for every value of d. The obtained hyperparameters under every value of d are the ones in Table 8.6. The difference is patent to those initial results in Table 8.4, as the hyperparameter selection is more heterogeneous throughout the different sizes. Finally, the best EER is obtained for d = 0.9.

Table 8.6: Best hyperparameter configurations for FD obtained under different values of d and the mean EER (%) obtained in validation.

d	Hidden Layer Size	Activation	Alpha	Tolerance	Mean EER (%)
0.3	700	Tanh	0.0005	0.05	0.081
0.5	400	Identity	0.0001	0.1	0.054
0.7	700	ReLU	0.01	0.01	0.058
0.9	500	ReLU	0.0001	0.05	0

Considering the previous configurations, the results for the each enrollment size are



(a) ND.

(b) FD.

(c) SD.

Figure 8.4: DET graphs for each differentiation after tuning and training with d = d_{enr} = 0.5.

collected in Table 8.7 and with their DETs representation in Figure 8.5. Within the same visit D1V1, using less than half of the data for training would not be advisable, as the EER is not optimal, being this information the only one visible in Figure 8.5a. There is no real difference between using 0.5 or 0.7 proportions in enrollment for the same day, as the results are very similar, as observed in Figure 8.5c. However, in the case of D1V2, increasing the enrollment size to 0.9 noticeably lowers the EER in 40%. In the case of D2V1, the improvement is clear when $d_{enr} > 0.3$, more specifically using 0.7. Finally, the worst performances are obtained under D2V2 data, with no difference between 0.5 and 0.9 enrollment ratios and surprisingly reaching the lowest EER when $d_{enr} = 0.3$.

Table 8.7: EER (%) performances for the different enrollment proportions with the parameters obtained when tuning with d = 0.9.

Visit / d_{enr}	D1V1	D1V2	D2V1	D2V2
0.3	0.101	3.636	3.636	4.796
0.5	0	3.636	3.112	5.454
0.7	0	3.636	2.647	5.454
0.9	0	2.515	2.681	5.454

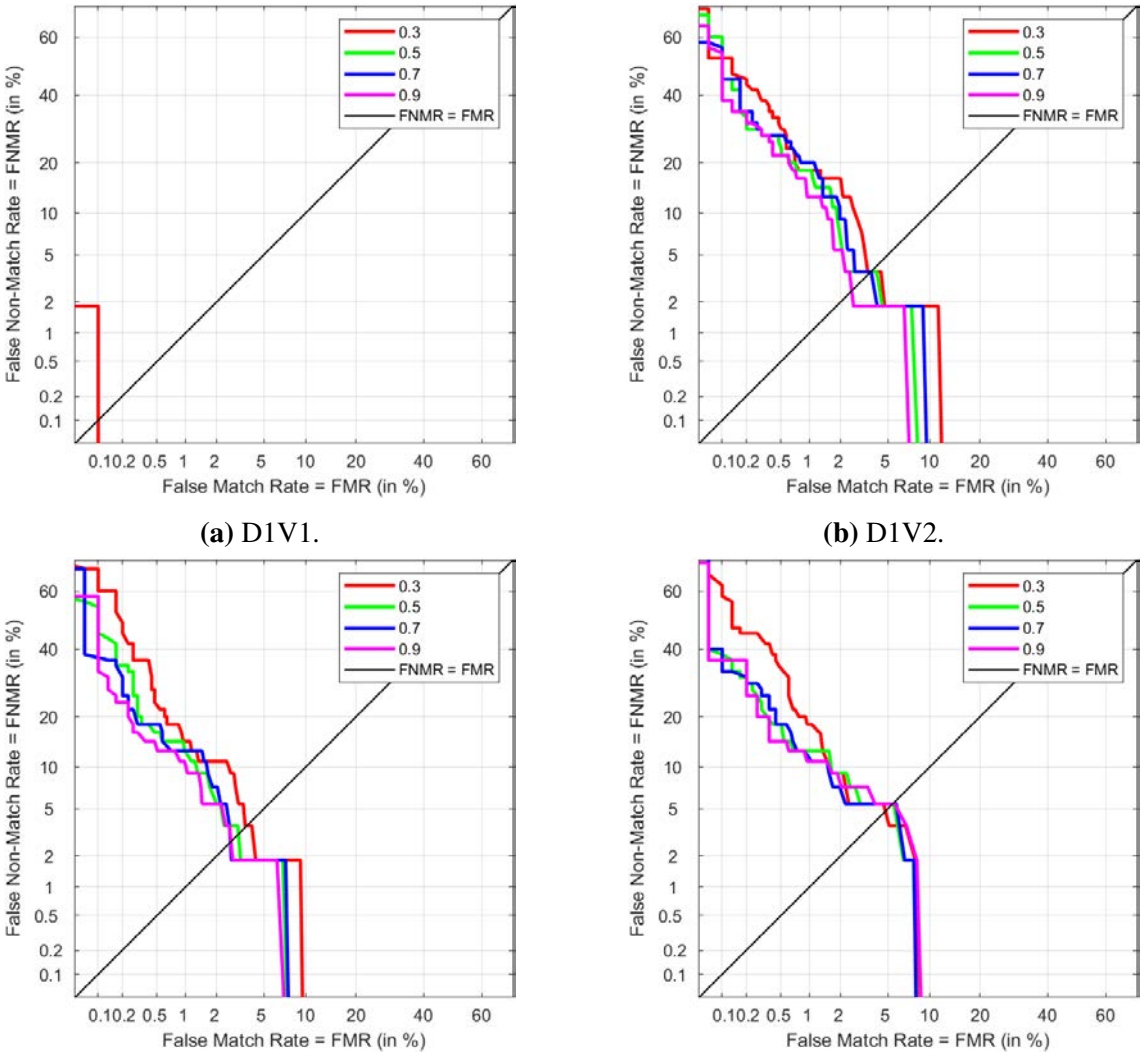
Extended verification

By using FD and d_{enr} = 0.9 with the initial verification approach, the FNMR and FMR plots result in those in Figure 8.6 where D1V1 is not plotted as the EER is 0. To extend the information and refer to more realistic context, different types of attempts are evaluated with all the visits. The results are graphically represented in Figure 8.7, where 8.7a considers one attempt at a time, reflecting the mean EER and Figure 8.7b collects results considering the average score of all the possible attempts.

Table 8.8: EER (%) average results considering the two types of extended verification considering H_A. D1V1 is the enrollment and d_{enr} = 0.9.

Visit	D1V1		D1V2		D2V1		D2V2	
Type / H_A	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts
5	0	0	0.094±0.071	0.034	0.373±0.507	0.202	0.473±0.597	0.640
10	0	0	0.085±0.057	0.034	0.337±0.382	0.236	0.425±0.601	0.438
15	0	0	0.082±0.052	0.067	0.318±0.355	0.269	0.318±0.255	1.010
20	0	0	0.084±0.046	0.067	0.309±0.346	0.269	0.356±0.363	1.413
25	0	0	0.039±0.247	0.067	0.393±0.393	0.202	0.310±0.219	0.673
30	-	-	0.076±0.035	0.067	0.282±0.320	0.303	0.299±0.244	0.774

Results for one attempt are similar throughout the different values of H_A. The EER increases as the experiments change more significantly. In Figure 8.7a the standard deviation is higher for experiments in the second day, more significantly in D2V2. This implies that the selected samples are really relevant in verification for this visit, as probably some of them are more stable than others, resulting in best performances. In the case of H_A = 15, the EER average goes ranges between 0% to 0.318%. In the experiment considering all the possible attempts, in Figure 8.7b, huge variations also occur in the case of D2V2, where the EER drops noticeably when using 10 samples in each attempt but spikes when increasing H_A. It is remarkable, as this data is the less stable one, which could mean that each score for every attempt gets heavily compensated in some cases. In general, results are less constant in all visits when varying the samples per attempt. However, in general, the EER average is noticeably decreased in comparison to data only from one attempt. In the case of D2V2 for 10 samples, the EER goes from 0.034% to 0.438%. However, verification time must be considered, as using 10 samples per attempt allows to execute 25 attempts with the available data, which is not a convenient number



(a) D1V1.

(b) D1V2.

(c) D2V1.

(d) D2V2.

Figure 8.5: DET graphs for the each enrollment size after tuning with d = 0.9.

of attempts in a real case scenario.

8.4.2. Results with the entire BMSIL dataset

The initial work presented in [2] only considered the S2 subset as the goal was only focused on physiological changes. However, to facilitate further comparisons between algorithms, the results are extended to the entire BMSIL data, which includes S1 and S2 and elevates the number of users to 105. Considering FD the MLP model and verification results are obtained under the the same criteria as in the previous sections.

The hyperparameters in tunning are specified in Table 8.9 with the corresponding testing results. In this case, the lowest testing results are obtained when d = 0.7. The parameters result in the FNMR vs. FMR graph plotted in Figure 8.8 where the EERs are 0%, 1.503%, 3.243% and 6.324% for D1V1, D1V2, D2V1 and D2V2, respectively.

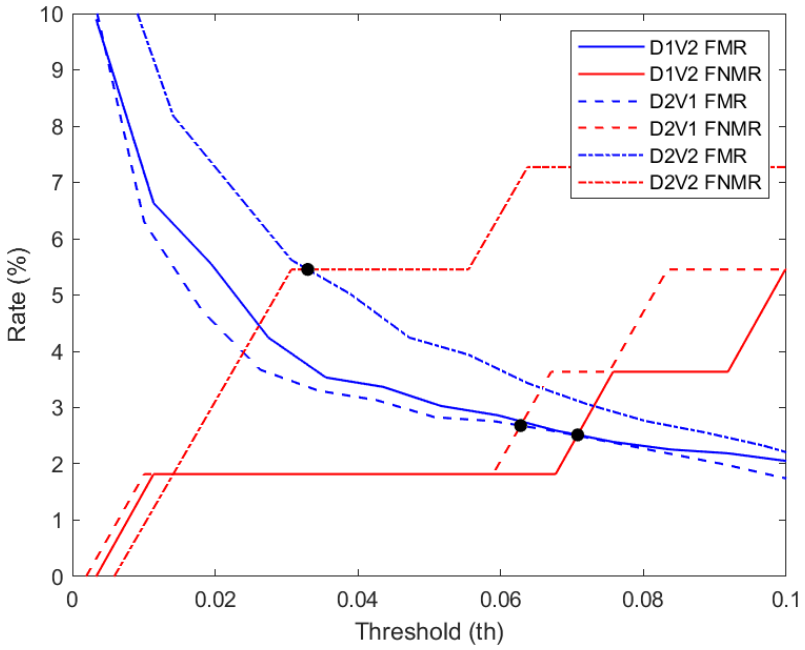
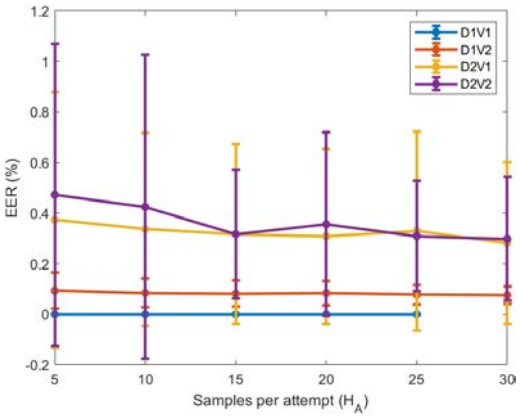


Figure 8.6: FNMR and FMR curves for D1V2, D2V1 and D2V2 with hyperparameters of tuning with d = 0.9 and same value of d_{enr}. The verification approach is using one sample per attempt with all the attempts. The EER points are marked in black.

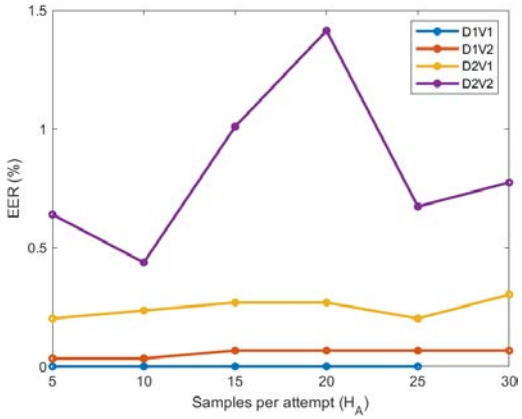
Table 8.9: Best hyperparameter configurations for FD obtained under different values of d and the mean EER (%) obtained in validation using the entire BMSIL database.

d	Hidden Layer Size	Activation	Alpha	Tolerance	Mean EER (%)
0.3	300	Tanh	0.0001	0.01	0.711
0.5	300	Identity	0.001	0.01	1.102
0.7	700	Identity	0.0005	0.01	0.158
0.9	400	Identity	0.0005	0.1	1.735

Results for the different types of verification are also summarized in Table 8.10 and their graphs plotted in Figure 8.9. In this case, there is a clear impact in comparison to those results with the S2 data. When verifying with one attempt the results are not really affected as the H_A varies, as represented in Figure 8.9a. The best result is obtained when H_A = 30 and is determined based on the visits from the second day, as experiments from the first day are always correctly classified. This is a consequence of adding the S1, as D1V2 data corresponds to resting and sitting acquisitions, probably making the system easier to classify. The EER average ranges between 0.831% to 1.886%. Nonetheless, the standard deviations are higher, even more for D2V2, so these results would depend heavily on the sample selection. When considering all the attempts, similar results occur in the first day, as the general EER is 0%, which is clear in Figure 8.9b. Best results are again obtained when H_A = 5, where EER has a range of 0–2.711%.



(a) Considering a single attempt.



(b) Considering all attempts.

Figure 8.7: EER (%) average results for both extended verification alternatives with D1V1 as enrollment and d_{enr} = 0.9. Considering D1V1 experiments contain half the samples as the remaining.

Table 8.10: EER (%) average results for the entire BMSIL database, considering the two types of extended verification considering H_A. D1V1 is the enrollment and d_{enr} = 0.7.

Visit	D1V1		D1V2		D2V1		D2V2	
Type / **H_A**	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts
5	0	0	0	0	1.542±3.111	0.467	2.711±5.309	0.348
10	0	0	0	0	2.255±3.515	0.412	4.039±6.077	0.275
15	0	0	0	0	2.441±3.508	0.247	4.591±6.301	0.366
20	0	0	0	0	2.377±3.408	0.265	4.588±6.278	0.284
25	0	0	0	0	2.393±3.414	0.302	4.590±6.270	0.320
30	0	0	0	0	2.325±3.412	0.229	4.346±6.182	0.247

8.5. Conclusions

In the present chapter the MLP algorithm has been successfully tested as a good approach for ECG biometric verification. The initial goal for this algorithm was to prove that changes in posture or heart rate do not interfere in the recognition. Additional, it provides good performances by using simple transformations. Based on the obtained results in chapter 6 for the QRS segmentation criteria, a hyperparameter tuning procedure has been established for MLP. The first differentiation was chosen as the most suitable one, retrieving the results in Table 8.11 for the S2 dataset.

The results for S2 gave EER values that ranged from 0% to 5.454% in the worst case scenario, which belonged to experiments regarding increasing heart rate. The extended approaches for verification using different samples per attempt, resulted in a lower EER than general results for LDA in the same verification approaches. Depending on these and the number of samples per attempt, the average EER generally decrease to values lower than 0.5% in both verification alternatives. Only considering one attempt provides more and more distant whenever the recognition scenario heavily differs from the enrollment one. This observation reinforces the requirements of having a quality assessment in the

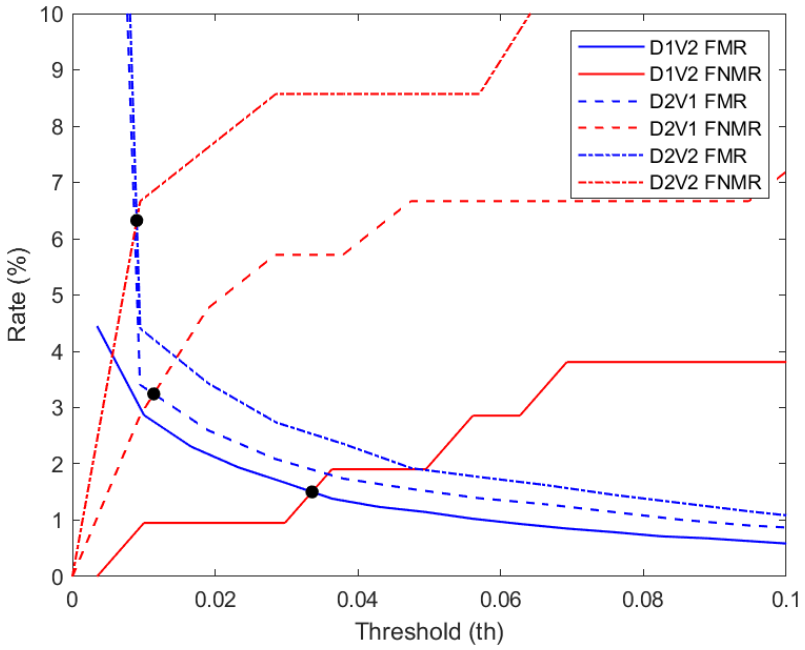


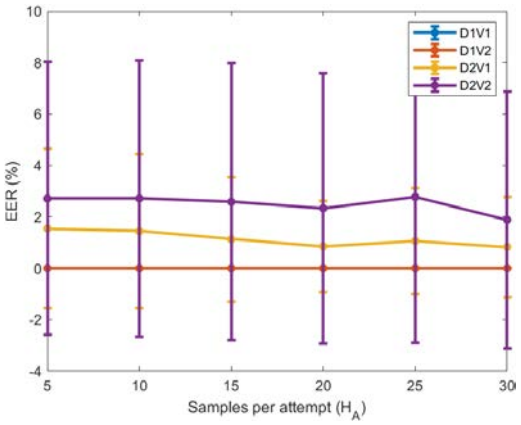
Figure 8.8: FNMR and FMR curves for verification using the entire BMSIL database and D1V1 as enrollment visit with $d_{enr} = 0.7$. The verification approach is using one sample per attempt with all the attempts. The EER points are marked in black.

recognition purposes.

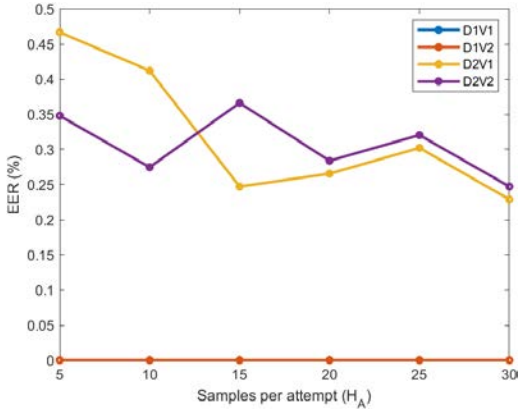
To facilitate the comparison with the results from chapter 6 concerning LDA, the procedure was repeated with the first differentiation using the BMSIL database in its entirety (S1+S2). Even though MLP required more enrollment length (70% vs. 50% in LDA), the improvement in EER is patent. The algorithm has proven to generalize better, as D1V1 experiments lay in a 0% EER, when LDA resulted in 7.465%. In addition, the worst case scenario got a maximum value of 6.324% while LDA had almost 2% more EER in that case. These initial results clearly impacted the two extended verification approaches, reaching a maximum of 1.886% when using a single attempt with 30 samples; and 0.247% considering all of the attempts.

Comparing the use of MLP with S2 and S1+S2, S2 provides less users so it requires more enrollment data to achieve the best results. The EER also is lower in general for S2, which could be provoked by the fact that the number of users is lower. However, it could be that using both S1 and S2 made the classification more confusing, as in the case of S1 the scenarios do not vary between visits.

The set of samples for verification and the approach have significantly affected the results in the MLP algorithm, which shows how different each comparison behaves. This could be a product of two issues: the lack of quality assessment in the verification samples and/or the impossibility of the algorithm to generalize under certain type of QRS complexes. In the former, there are no consensus as how to assess the quality of an ECG for verification. However, focusing in other complex algorithms is something doable and insightful for the available data.



(a) Considering a single attempt.



(b) Considering all attempts.

Figure 8.9: EER (%) average results for both extended verification for the entire BMSIL database. The enrollment visit is D1V1 and $d_{enr} = 0.7$. Considering D1V1 experiments contain half the samples as the remaining.

Table 8.11: Comparison of the final results for the best classifiers tried so far, LDA and MLP. The segmentation comprehends a window of 0.2 s with the R peak in the center. The recognition experiments are carried out with all the available visits, so the EER is represented as a range of percentages.

Classifier	Database	Features	Enroll	Verification	EER (%)
LDA	S1+S2	ND	D1V1 ($d_{enr} = 0.5$)	$H_A = 1$. All attempts.	7.465–8.096
				$H_A = 20$. One attempt.	4.571–6.433
				$H_A = 25$. All attempts.	5.091–5.897
MLP	S2	FD	D1V1 ($d_{enr} = 0.9$)	$H_A = 1$. All attempts.	0–5.454
				$H_A = 15$. One attempt.	0.082–0.318
				$H_A = 10$. All attempts.	0–0.438
MLP	S1+S2	FD	D1V1 ($d_{enr} = 0.7$)	$H_A = 1$. All attempts.	0–6.324
				$H_A = 30$. One attempt.	0–2.711
				$H_A = 30$. All attempts.	0–0.247

9. ECG RECOGNITION WITH DEEP LEARNING

In chapter 8 we proved that MLP is a more complex yet a better approach than LDA for ECG verification, even when having scenarios that vary the shape and acquisition of the ECG. The MLP algorithm is one of the simplest networks that are part of Machine Learning, but increases the complexity in LDA. However, more complex techniques have been developed and formed Deep Learning algorithms.

The present chapter has the goal of testing Deep Learning in similar conditions without using differentiation and experiment other type of enrollments. Given that more sophisticated algorithms are included, the experiments also aim for identification and not only verification. The whole implementation used Keras [107] in Python 3.

9.1. Initial approaches

The first goal of this chapter was complementing the results in MLP verification, in order to observe how well Deep Learning manages the different data. As a consequence, the employed data is the S2, using the same input data in section 8.1, but only considering ND differentiation.

9.1.1. LSTM and hardware limitations

Considering the characteristics of ECG and the decision of avoiding complex pre-processing, Neural Networks are a good approach as they are known to properly generalize from raw data [108]. LSTM networks were considered a good approach, as they belong to RNNs but also deal with long-term dependencies [109]. The early stages of the LSTM network relied on very simple structures, with only one hidden layer and nodes. However, the network optimization took days to end, with poor fitting results, and some times it ended running out ouf memory. The hardware was upgraded from an Intel®Core™i7-6700 CPU with 16 GB RAM without dedicated GPU to an i9-9900K CPU with 16 GB RAM and Nvidia GeForce RTX 2080 Ti GPU. Even the inclusion of a powerful GPU made the optimization impossible to achieve due to the required time and memory, so a LSTM-based network was not possible to design.

9.1.2. CNN

The CNN are commonly applied to image processing as they are 2D matrices. However, CNNs can also be one-dimensional and help extracting relevant features for classification. This type of networks reduce the quantity of data, probably avoiding the previous

problems with LSTM. Initially, only CNNs are used for this classification, and the process is further detailed in [110].

Network architecture

The network different layers are designed with a common and simple approach in Figure 9.1 [111][112]. The concept of 1D Convolutional layers was already defined in section 5.3.1. In this case, it is followed by a maximum pooling layer, which retrieves the maximum in the selecting window and usually performs better than average pooling [113]. The concatenation of these two layers are considered one hidden layer, and the design can use more of them, but they are not represented for simplicity. The dropout layer helps avoiding overfitting, as it discards samples with a given probability. Finally, the dense layer results in as many outputs as classes we want to predict when using a softmax activation.

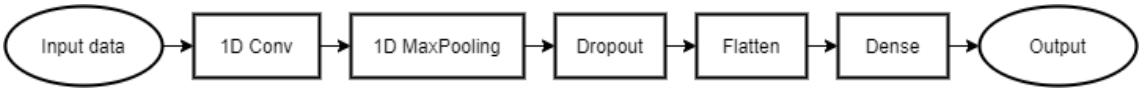


Figure 9.1: Basic layer architecture for CNN.

Hyperparameter optimization

The initial goal for the present chapter was to proceed with similar experiments to those in chapter 6. However, due to the differences in the used software, processing times and hardware limitations, some had to be changed accordingly as follows:

- Differentiation: the type of best differentiation initially was a parameter to further assess in this chapter, as it happened in chapter 8. The first experiments were achieved with ND, and the results were planned to be extrapolated to FD and SD. However, the processing time, amount of resulting data and analysis complexity experienced impeded achieving the same with the remaining differentiation. To be coherent and comparable to the results in MLP, these should have been achieved considering FD, but results with ND are sufficiently good and obtained under no transformations.

- Cross-validation: the hyperparameter tuning for MLP implemented cross-validation with 5 folds in all iterations, including the training of the final model. However, due to processing times required by the BioECG network in training, this step is avoided in its tuning. However, the final model is trained with a 5-fold cross-validation.

- Evaluation metric: the available toolboxes in *scikit-learn* allowed to train minimizing custom functions. This allowed to implement the EER as an evaluation

metric. However, to deal with neural networks we use *Talos* [114], which is a computer software that facilitates tuning in *Keras*. The capabilities of this software are not as wide as those in *scikit-learn*, limiting the options of creating a custom target function. As a result, this function is chosen from the available functions in software. In this case, in other to also pay attention to identification, we selected the accuracy.

As it has been done with MLP verification, this system also required hyperparameter fixation and further tuning. Table 9.1 collects all the fixed values and ranges given for tuning for the different hyperparameters. The fixed values were determined after heuristic observations with the inclusion of an Early Stopping criteria, which stops the training when there is no improvement, in order to avoid overfitting and long computations. The maximum value is determined by Epochs, which is 250. The loss function is the Cross-Entropy as defined in Equation (9.1) [111], where C is the number of classes, p(s) is the softmax result and t indicates whether the class is positive or not. In addition, the evaluation metric in this process is the accuracy, which is one of the biggest differences with the one carried out in section 8.2.2, where the metric was the EER.

Table 9.1: Values and set of values given to the CNN architecture.

Hyperparameter	**Value/s**
Conv activation	ReLU
Dense activation	Softmax
Optimizer	Adam
Loss function	Cross-Entropy
Dropout value	0.5
Epochs	250
Hidden layers	[1, 2, 3, 4]
Number of filters (f)	[32, 64, 128, 256]
Kernel size (k)	[2, 3, 4, 5]
Pooling size	[2, 3, 4, 5]
Batch size (b)	[32, 64, 128, 256, 512]

$$CE(t, p(s)) = -\sum_{i}^{C} t_i log(p(s)_i) \tag{9.1}$$

The tuning process follows the same steps as the one achieved in section 8.2.2. Different values of d are given to obtain more sets of hyperparameters for D1V1 of S2. This process determines the best model using the remaining data as test.

Enrollment size

After determining the set of hyperparameters, the values of d_{enr} are modified for training. To observe how differently these models may behave, each possible model is used for verification with all the enrollment proportions to determine the best one.

9.2. BioECG: design, optimization and recognition

After observing the performances with CNN and the limitations to use LSTM networks on their own, an extra network concept in introduced. The combination of CNN and LSTM architectures is commonly used through literature in human recognition, where feature extraction and classification are carried out in the same network [115]. This process is expected to decrease the amount of data to be processed by the LSTM network, making the corresponding calculations doable under the given hardware limitations. From now on, this combination is going to be named BioECG.

9.2.1. Architecture design

The general scheme for the BioECG network is represented in Figure 9.2. The different parts are defined as follows, when considering enrollments with D1V1 experiments:

- Input data: where H_{user} is the available data per user formed by 50 segmented QRS in 5 sessions, with 200 points length and the R peak in the 101^{th} index. The number of users is represented by U which is 105 for the whole BMSIL database. $U_1 = 50$ when only dealing with S1 and $U_2 = 55$ when applying S2. This input data may vary depending on the process: tuning is affected by the parameter d, and training by d_{enr}, which specify the proportion data used in each procedure. In the case of testing, $H_{user} = H_{test} = 250$.
- CNN: summarizes the data fed into the LSTM network.
- Batch normalization: helps with convergence and learning between layers when the input is fed in batches [116], keeping the same dimensions as the CNN output.
- LSTM(s): represents a single or multilayered LSTM network, as it was described in Figure 5.7. Considering L as the number of layers, n the number of neurons, b the batch size and o the output size, the possible configurations of this network are:
  - $L > 1$ and $l = 1$, the output dimensions were (b, o, 2n), as the number of hidden neurons is doubled.
  - $L > 1$ and $1 < l < L$, retrieves dimensions of (b, o, n).
  - $L > 1$ and $l = L$, resulted in (b, n) dimensions, as it was the last layer.

  - L = 1 the only layer was also the output layer. The output had dimensions (b, 2n).

- Dense: a densely connected layer with as many layers as users in the database, with softmax activation that allows to obtain the prediction probabilities per batch.

- Output: results in the final probabilities for all the batches. In the case of training, the final dimensions are (H_{test}, U)

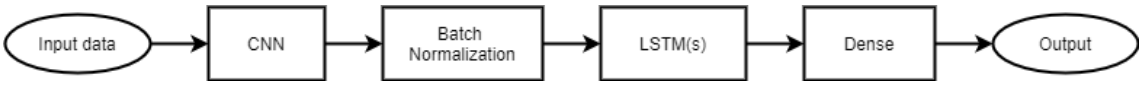


Figure 9.2: Scheme for the layers in the BioECG network.

9.2.2. Hyperparameter optimization

The hyperparameter optimization is achieved similarly as it was in section 9.1. There is an initial hyperparameter fixation based on previous knowledge and heuristic experimentation, and then it is followed by a Random Search and Exhaustive Grid. However, as a result of the constraints by hardware, there were two modifications:

1. There is no cross-fold validation in tuning, only in training, due to time limitations.

2. The enrollment sizes are not tested for every set of hyperparameters obtained with d. We assume d_{enr} = d all the time.

In this case, there is a preliminary tuning for the CNN network to obtain the best values of s, k and f. The final values were determined setting d = 0.5, b = 35, L = 2 and n = 32. Based on previous results, the adam optimizer is set as the most suitable option. The system is expected to have from 1 to 3 hidden LSTM layers, and neurons that change in a 2 exponent. The maximum number of epochs for early stopping is 500, with a 20 patience. All the possible values and the fixed ones are summarized in Table 9.2, where in the case of CNN hyperparameters the selected ones are in bold font.

9.2.3. Recognition analysis

The previous experiments that have used MLP and CNN for classification, only focused on the S2 database in verification. This was induced by the interest on classifying data under different physiological scenarios. However, the goal is to provide a further in-depth study in the case of BioECG. This implies not only assessing S2 but also S1, allowing to observe how the different conditions affect. The results collect EER and accuracy for both recognition purposes.

Table 9.2: Possible values and fixed hyperparameters for the BioECG architecture.

Hyperparameter	**Values**
Batch size (b)	[20, 35, 50, 100]
Kernel size (k)	[3, 5, **7**, 9]
Strides (s)	[3, **5**, 7, 9]
Filters (f)	[8, **16**, 32, 64]
Activation	ReLU
CNN layers	1
LSTM layers (L)	[1, 2, 3]
LSTM neurons (n)	[16, 32, 64, 128]
Epochs	500
Patience	20
Optimizer	Adam
Learning rate (η)	0.001

The recognition process considered two types of enrollments: one-day and two-days enrollments. These sets of data for enrollment are compared by using the same number of samples, but with different origin. The one-day enrollment refers to the development and enrollment proportions as in the previous section. This value translates into a given number of samples per user, $H_{user} \cdot d$. In the case of two-days enrollment, this value has to be constant, extracting data with proportion $d/2$ from each visit. In this case, data comes from D1V1 and D2V1 as they have the same physiological conditions in the entire BMSIL database. The total sum of samples result in the same quantity for both types of enrollment. The goal is observing how adding an extra enrollment day could impact the system, analyzing the trade-off between practicality and performance.

Due to the BMSIL characteristics, the recognition analysis is divided in two depending on the scenario. These experiments are carried out for both enrollments and further broken down into the following:

1. Same scenario.
    - Variations in the same day: D1V1, D1V2 for S1. D1V1 for S2.
    - Variations between days: D2V1, D2V2 for S1. D2V1 for S2.
2. Different scenario.
    - Different position: D1V2 for S2.
    - Different heart rate: D2V2 for S2.

9.2.4. Final configuration and extended verification

A final configuration is selected based on the recognition analysis, considering the values of d = d_{enr}. For this selection, the main goal is achieving the most reasonable EER, as the environment of use is undetermined. This final configuration is tested with S1, S2 and the entire BMSIL database (S1+S2) for identification and verification, using one-day and two-days enrollments. The results were obtained considering one attempt formed by as many samples as the available test data allows.

To observe the behavior under to a more realistic conditions, the verification gets extended results by applying the two different criteria described in section 5.3.2 which imply fixing the number of samples that belong to each attempt. The two approaches have been observed: considering one attempt and considering all the possible attempts. The obtained results have several issues to address: the difference between enrollments, difference of performance between the sets of data and how much the number of samples per attempt affect the result. Considering both types of enrollment, the available samples for testing under the different conditions are summarized in Table 9.3.

Table 9.3: Number of attempts per test set according to the number of samples H_A and the type of enrollment when d_{enr} = 0.5.

	One-day		**Two-days**	
Test set	**D1V1**	**Rest**	**D1V1 & D2V1**	**Rest**
Test samples / **H_A**	125	250	188	250
5	25	50	37	50
10	12	25	18	25
15	8	16	12	16
20	6	12	9	12
25	5	10	7	10
30	4	8	6	8

9.3. Results

9.3.1. Initial approaches

The final tuning values based on the enrollment proportion are in Table 9.4. As the number of samples for training increase, the system requires more epochs to finally reach an acceptable training. The hyperparameter values do not vary drastically between the different sizes of the development set. As the amount of data increases, more hidden layers help increasing the performance, and the batch size is also augmented. However, these trends are considered normal as more data requires more complexity.

Table 9.4: Final tuning values for every value of d using CNN.

d	0.3	0.5	0.7	0.9
Hidden layers	2		3	
Number of filters (f)	64			
Kernel size (k)	3			
Pooling size	2			
Batch size (b)	32			64
Final epochs	44	54	78	83

After observing the performances of all the possible combinations of enrollment sizes and hyperparameter sets, the tuning with d = 0.7 is considered the best. The result with respect to enrollment lengths are collected in Table 9.5, and the best option is using d_{enr} = 0.9. This enrollment is a 3 min 45 s procedure that results in theoretical performances of 1.730%, 3.523% and 10.194% EER for D1V2, D2V1 and D2V2 respectively. The different graphs are summarized in Figure 9.3.

Table 9.5: EER (%) with different values of d_{enr} after tuning with d = 0.7.

$\mathbf{d_{enr}}$	**D1V2**	**D2V1**	**D2V2**
0.3	3.636	7.585	8.948
0.5	2.034	6.214	11.434
0.7	1.818	6.604	14.070
0.9	1.730	3.523	10.194

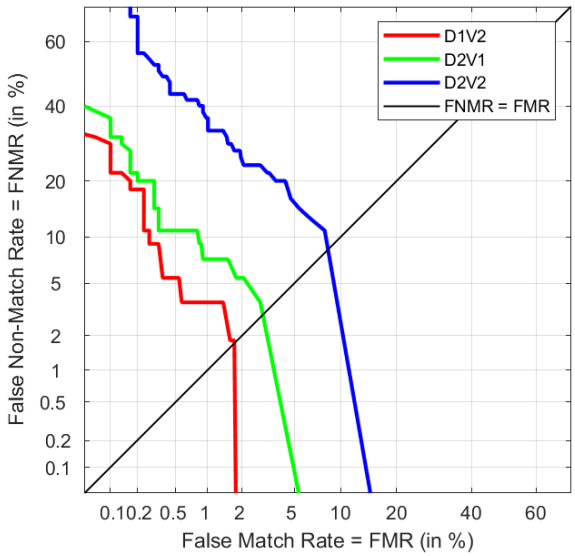


Figure 9.3: DET for test sets when training with d = 0.7 and d_{enr} = 0.9 of D1V1.

This selection is not the only solution, as it depends on the system's applications. In this case, the results for D1V2 and D2V1 have been noticeably improved while decreasing

performances in D2V2. This represents an environment in which having scenarios with increased heart rates is not likely to happen.

9.3.2. BioECG: design, optimization and recognition

Same scenario: variations in the same day

The different results in the same day are summarized in Table 9.6; it shows one-day and two-days data in enrollment, respectively. Results were slightly better in the one-day enrollment, as all the data in the verification process belonged to the same day and visit, D1V1. The best results required a lower value of d than in the best results for the two-days enrollment. Even when requiring the same number of samples, the two-days enrollment had half of data related to D1V1, which results in requiring higher enrollment lengths to achieve equally good verification rates when verifying with D1V1. However, the number of available samples were split in half between D1 and D2 in a two-days enrollment, resulting in worse performances for visits in D1.

In the case of S2 for Table 9.6, d = 0.9 provided a very different result, as it was the only non-zero value. This could be a result of the individual tuning for every value of d, which could be not as accurate as in the rest of the cases. On the contrary, performances for D1V1 were better in S2 than in S1 in Table 9.6, both in identification and verification. In this case, the best metrics were achieved with d = 0.9, although results slightly differed with the remaining values.

Table 9.6: Identification and verification results for same day in S1 and S2 for scenario R. The best considered options for one-day and two-days enrollment are in bold font with d = d_{enr}.

	Dataset	S1|				S2	
	Visit	D1V1		D1V2		D1V1	
	Metric / **d**	**Accuracy (%)**	**EER (%)**	**Accuracy (%)**	**EER (%)**	**Accuracy (%)**	**EER (%)**
One-day enrollment	0.3	92.81	0	93.25	0	96.30	0
	0.5	**98.19**	**0**	**97.96**	**0**	**95.86**	**0**
	0.7	96.90	0	96.42	0	95.80	0
	0.9	96.37	0	97.12	0	74.58	1.58
Two-days enrollment	0.3	94.42	0.04	94.09	0.04	97.24	0
	0.5	90.17	0.45	89.58	0.57	91.64	0.03
	0.7	93.09	0.08	93.12	0.08	97.62	0
	0.9	**94.55**	**0.04**	**95.20**	**0.04**	**96.39**	**0**

Same scenario: variations between days

Results for R in different days were collected in Table 9.7. One-day data results are represented in Table 9.7. The decrease of accuracies and increment in EER show how different ECGs can be between days, even under the same scenario. Results in Table 9.7 belong to those with two-days development data. In this case, the improvement from one-day to two-days enrollment was relevant. From accuracies around 66–76% in Table

9.7 to values in Table 9.7 for d = 0.3. Maximum achieved accuracy reached 98.91% in d = 0.7, comparable to those with same-day scenarios, as seen in Table 9.6. In terms of verification, the EER decreased from 2–6.54% in one-day enrollment to values that ranged between 0% and 0.24% in two-days enrollment.

Table 9.7: Identification and verification results for different days with R scenario in S1 and S2. The best considered options for one-day and two-days enrollment are in bold font with d = d_{enr}.

	Dataset	S1				S2	
	Visit	D2V1		D2V2		D2V1	
	Metric / **d**	**Accuracy (%)**	**EER (%)**	**Accuracy (%)**	**EER (%)**	**Accuracy (%)**	**EER (%)**
One-day enrollment	0.3	66.43	4	67.54	4	67.99	6.3
	0.5	**76.86**	**2**	**76.50**	**2.53**	**66.14**	**4.59**
	0.7	71.58	3.71	71.62	3.06	67.78	5.45
	0.9	70.39	4	71.53	4	48.72	6.54
Two-days enrollment	0.3	92.54	0.04	92.27	0.04	98.81	0
	0.5	87.81	0.24	88.31	0.24	93.75	0.17
	0.7	94.83	0.12	94.95	0.12	98.91	0
	0.9	**97.68**	**0**	**97.17**	**0**	**98.12**	**0**

The different results between Tables 9.6 and 9.7 can be a product of the intra-individual variability between days, where doubling data related to D1V1 did not give enough information about the ECG variability in the long term. Moreover, comparing results in Table 9.7, results were noticeably better in all the values of d and in all the visits in D2 when applying a two-days enrollment. When d = 0.9, accuracies went from 48.72–71.53% to 97.17–98.12%.

Different scenario: different positions

The data that also considered the change of position was acquired in the same day as the enrollment but in a different visit. This implies that observed differences could only be related to the position and short time variation. Table 9.8 summarizes those results. Even though test data was acquired on the same day as in Table 9.6, the change of position made the results decrease noticeably: almost a third in accuracy and going from almost ideal values up to 5.45% in EER in the worst-case scenario. This information confirms that changing the position affects the verification results even considering experiments in the same day.

Table 9.8 collected performances when changing into two-days enrollment. The results improved noticeably in comparison to those in Table 9.8. Even though the second day of data used in training did not provide information about the change of position, it has also helped to generalize in this case. It resulted in almost doubling the accuracy in the worst previous result, while decreasing the different EERs up to 50%.

Table 9.8: Identification and verification results for scenario S. The best considered options for one-day and two-days enrollment are in bold font with d = d_{enr}.

	Dataset	S2	
	Visit	D1V2	
	Metric / **d**	Accuracy (%)	EER (%)
One-day enrollment	0.3	37.59	3.64
	0.5	67.32	5.45
	0.7	**68.80**	**3.64**
	0.9	63.78	4.54
Two-days enrollment	0.3	80.42	1.82
	0.5	73.88	3.64
	0.7	79.45	1.99
	0.9	**85.14**	**1.28**

Different scenario: different heart rate

Performances for the change of heart rate are collected in Table 9.9. Both identification and verification presented a huge decrease in performance for one-day enrollment in Table 9.9. Opposed to those results in Table 9.8, this verification data belongs to a different acquisition day, adding extra variations that may not be related to the heart rate. However, comparing results in the same scenario in Table 9.7 with those in Table 9.9 help in the assumption that the performance decrease is due to the heart rate variation.

Table 9.9: Identification and verification results for scenario Ex. The best considered options for one-day and two-days enrollment are in bold font with d = d_{enr}.

	Dataset	S2	
	Visit	D2V2	
	Metric / **d**	Accuracy (%)	EER (%)
One-day enrollment	0.3	42.91	11.43
	0.5	44.36	9.83
	0.7	**50.62**	**7.27**
	0.9	29.26	14.16
Two-days enrollment	0.3	**59.91**	**3.64**
	0.5	58.58	3.64
	0.7	56.28	5.45
	0.9	58.04	3.90

On the other hand, Table 9.9 presents results with two-days enrollment. Results were still not good in terms of identification, remaining below 60%. However, in verification, EERs decreased noticeably. It was proved that adding an extra day of information for

training improves the system noticeably in terms of verification, once it was compared to Table 9.9. In d = 0.9 EER dropped almost 10% and close to 7% in the case of d = 0.3.

9.3.3. Final configuration and extended verification

Considering the different results in the previous section, the solution to the final system's configuration was not unique. Depending on its purpose, some factors needed to be taken into account.

The enrollment process needed to be long enough to provide good information for the development set. However, if the enrollments were too long (i.e., a greater value of d), the user might get tired. Adding an extra day of acquisition has been proven to provide better results. Unfortunately, users are usually reluctant to extend the enrollment process to several sessions. However, if the system was required to have higher performance, it may be worth the effort.

If the purpose of the system focused on fast recognition more than high performances, the enrollment process could get shorter and easier. It would also depend on the probability that users come up with different positions or heart rate throughout recognition, e.g., members working out regularly may have different heart rate as when they go out, because they may have not been fully recovered yet.

Once these issues have been addressed, this work suggests one specific configuration as a trade-off choice, which was considered to provide good general verification results in all different scenarios. Doing a two-days enrollment is key for increasing the verification performance and even more if there are heart rate variations in the recognition process. The chosen enrollment size is d_{enr} = 0.5 as it was a frequent value when obtaining the best discussed results. However, when it was not the best of all the proportions, it still performed properly while allowing to have a shorter enrollment process. Considering the 250 samples per user, using that ratio implies 125 samples per user between two-days. That means around 63 QRS samples per visit, which requires two ECG signal acquisition as 50 complexes are extracted from each one. It summarized in an enrollment process of a maximum duration of 140 s, as every 50 peaks requires 70 s of acquisition. Setting the number of detected peaks to a greater value would allow for the enrollment of people with one signal acquisition, depending on the user's resting heart rate.

The hyperparameter tuning provided as the best configuration b = 20, L = 2, and n = 64. The tuning reached 151 epochs following the early stopping criteria. The mean time taken for training each fold in cross-validation was 147.7 s for the whole S2 dataset.

The system was trained according to these hyperparameters for S1 and S2 independently. The same process was achieved using S1 and S2 as a whole dataset too, S1+S2, providing heterogeneous samples as D1V2 and D2V2 are different scenarios. Verification performance results under these conditions are collected in Table 9.10. All the FNMR and FMR curves are plotted in Figure 9.4 for the different type of enrollments

and data. Testing for D1V1 resulted in values close to 0 in all possible combinations, so they are not represented for simplicity. These graphs also provide a representation of the different EERs, which are marked in black and summarized in Table 9.10. All subfigures provide the same axes for easier comparison.

Table 9.10: EER (%) results for every database, visit and type of enrollment with d_{enr} = 0.5.The values in parenthesis are the percentage of improvement with respect to the one-day enrollment.

Visit	**D1V2**		**D2V1**		**D2V2**	
Set \ **Enroll**	**One-Day**	**Two-Days**	**One-Day**	**Two-Days**	**One-Day**	**Two-Days**
S1	0	0.571	2.305	0.240 (−89)	2.700	0.240 (−91)
S2	5.432	3.928 (−25)	5.268	0.168 (−97)	10.124	3.636 (−64)
S1 + S2	1.905	0.700 (−63)	8.013	0.009 (−99)	14.309	1.352 (−91)

In relation to the different enrollments, for the three sets there is a clear improvement in two-days enrollment with respect to one-day enrollment. The S1 verification has lower EER than S2, as a result of having the same scenario in all the collected visits. Adding a second day of enrollment decreases the performance for D1V2, but in exchange there is noticeable improvement for experiments in the second day, which improve up to 91%. In the case of S2, there is also an improvement from one-day and two-days even in the case of D1V2, which means that including a second day in enrollment adds extra information even for experiments in different days and scenarios.

Comparing S1 and S2, there is a clear decrease of performance when adding extra scenarios in S2. D1V2 and D2V1 do not show great differences in the one-day enrollment, but D2V1 gets really affected by a second day of enrollment. As it has been observed throughout the work with this database, D2V2 still provides the worst performance in S2. However, its improvement with a two-days enrollment makes this verification go from not good to acceptable.

Finally, it is easy to observe that S1+S2 results in EERs are decreased compared to those in S1 or S2 alone. The number of users almost doubles with the entire database, making the verification more complex. For one-day enrollment, results are worse for the second day than the observed in S1 and S2 individually. As observed in Figure 9.4e, the thresholds for these days are very low, which implies that some mated data does not score high and the distributions get overlapped. For the two-day enrollment, D2V1 gets mated scores with higher thresholds, as this visit contributes in enrollment. Thresholds are slightly incremented for D1V2 and D2V2, but their EER gets noticeably decreased with a maximum of 1.352%.

For the sake of brevity, this subsection is only focused on those results for S1+S2, i.e.: the entire BMSIL database, as it helps for further comparison with the previous algorithms experimented in this thesis. The average EER result for one attempt verification based on the number of samples in each attempt (H_A) is plotted in Figure 9.5, with the corresponding standard deviation. Similarly, results for all the attempts are represented in Figure 9.6. Observing two-days enrollments, Figure 9.5b has EER values that are always

lower than 0.005% and reach 0% in the majority of the cases. In addition, Figure 9.6b all the results dropped to 0. As a result, the numeric results for the two-days enrollment were not collected in Table 9.11.

The graph for one-day enrollment is in Figure 9.5a, where the value of H_A does not have dramatic impact in the general average. H_A = 30 gets the lowest EER for most of the scenarios, including D2V2. The two-days enrollment reduces all the EER to 0% when H_A = 30 or 20, when reaching close to 0% values in the rest.

For one-day enrollment when considering all attempts, the results are noticeably bad for the second day, specifically in D2V2. This results in EERs that are superior to those in the initial verification of the one-day enrollment observed in Figure 9.4e. The maximum in the initial verification was 14.309%, but in this case it reaches values that range between 13.964% to 22.409%. On the contrary, the data provided in Figure 9.6b uses the same information for verification than those to obtain the FNMR and FMR curves in Figure 9.4f. However, the way the scores are calculated impacts the result positively. In the former solution, which is equivalent to using H_A = 1, the two-days enrollment gave a range of 0.240%-3.636%, whereas increasing the H_A results in 0% of EER, benefiting from averaging the scores in groups instead of taking each score individually.

Table 9.11: EER (%) results for one-day enrollment with d_{enr} = 0.5 using BioECG and different values of H_A.

Visit	D1V1		D1V2		D2V1		D2V2	
Type / H_A	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts	One attempt	All attempts
5	0	0	0.009±0.010	0	1.515±2.760	4.142	5.746±10.780	13.964
10	0	0	0.008±0.005	0.009	2.441±2.838	3.503	5.412±10.140	15.683
15	0	0	0.008±0.004	0.018	1.422±2.766	4.133	4.765±8.990	18.135
20	0	0	0.009±0.004	0.009	1.180±2.493	4.343	4.102±8.603	17.206
25	0	0	0.006±0.005	0.009	1.516±2.871	4.708	5.219±9.751	17.358
30	0	0	0.004±0.004	0.009	1.236±3.000	7.360	3.773±9.130	22.409

9.4. Conclusion

The present chapter has included different approaches for biometric recognition using Deep Learning algorithm. The initial focus was into data with different morphological conditions, provided by S2 subset of the BMSIL database. CNNs were tested as a potential approach on their own, but the best performances were obtained jointly with LSTM networks, forming the BioECG architecture. After tuning the different hyperparameters using the detected QRS complexes considered in chapter 6, we have provided an in-depth analysis of the recognition, assessing several factors such as enrollment length, days in enrollment and type of verification. These results and observations were extended to the entire BMSIL database, to allow further comparisons with previous chapters.

A summary under the S2 database is detailed in Table 9.12, only collecting those

results for the initial verification, as it is the common approach for MLP, CNN and BioECG. The CNN algorithm does not improve results from the MLP algorithm, considering they have the same enrollment proportion. Using BioECG improves MLP when adding a second day of enrollment improves all results using the same number of samples. However, it is not possible to tell if the results was caused by the classifier or the fact of adding an extra day in enrollment. These results reinforce MLP as a good classifier for one-day enrollment when using the first differentiation.

In the case of using the entire BMSIL database, the results are collected in 9.13. This time, CNN is not considered, so the results can include the extended verification. Again, as the database contains S2 the tendencies are similar, where BioECG performs worse than MLP for the initial verification using $H_A = 1$ and all the attempts. Using one attempt, the average EERs are not drastically distant for one-day enrollment, but reaching better results for MLP. Using all the attempts, MLP performs noticeably better than BioECG, providing results that are close to 0%. In addition, adding extra enrollment has a huge impact in BioECG, going from bad results to a maximum of 1.352% in EER.

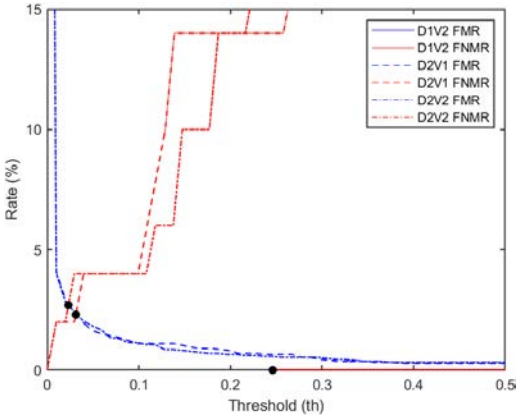
As a conclusion, this chapter has shown that the complexity of Depp Neural Networks is not as successful as simpler algorithms for the ECG user verification. In addition, we have proven the viability of not using any modification to the QRS complex when adding an extra day of enrollment to the process. In addition, we have observed that there are great differences depending on the type of verification, where taking too many attempts does not always result in a good verification. However, the selection of the verification samples is key in the case of not providing homogeneous data.

Table 9.12: Comparison between algorithms with the initial verification. The value d_{enr} is divided by two so the number of total samples is the same as one-day enrollment.

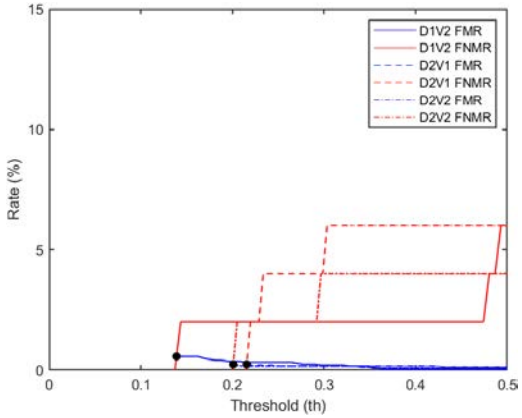
Classifier	Database	Features	Enroll	Verification	EER (%)
MLP	S2	FD	D1V1 ($d_{enr} = 0.9$)	$H_A = 1$. All attempts.	0–5.454
CNN		ND	D1V1 ($d_{enr} = 0.9$)		1.730–10.190
BioECG		ND	D1V1 ($d_{enr} = 0.5$)		0–10.124
			D1V1 ($d_{enr} = 0.25$) D2V1 ($d_{enr} = 0.25$)		0–3.928

Table 9.13: Comparison between algorithms for all types of verification with S1+S2 data. The value d_{enr} is divided by two in the case of two-days enrollment, as it represents the data proportion in the proper visit.

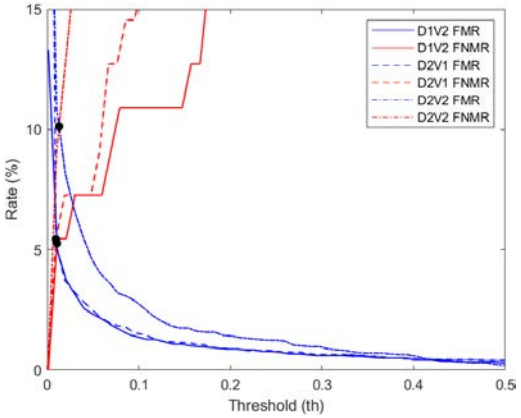
Classifier	Database	Features	Enroll	Verification	EER (%)
MLP	S1+S2	FD	D1V1 (d_{enr} = 0.7)	H_A = 1 All attempts.	0–6.324
				H_A = 30 One attempt.	0–2.711
				H_A = 30 All attempts.	0–0.247
BioECG		ND	D1V1 (d_{enr} = 0.5)	H_A = 1 All attempts.	1.905–14.309
				H_A = 30 One attempt.	0–3.773
				H_A = 5 All attempts.	0–13.964
			D1V1 (d_{enr} = 0.25) D2V1 (d_{enr} = 0.25)	H_A = 1 All attempts.	0.009–1.352
				H_A = 20 One attempt	0
				H_A = 5 All attempts.	0



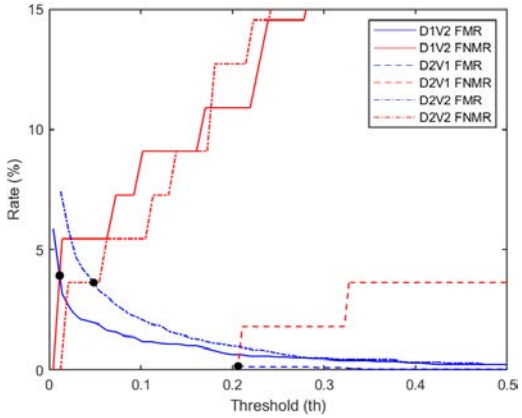
(a) One-day enrollment with S1.



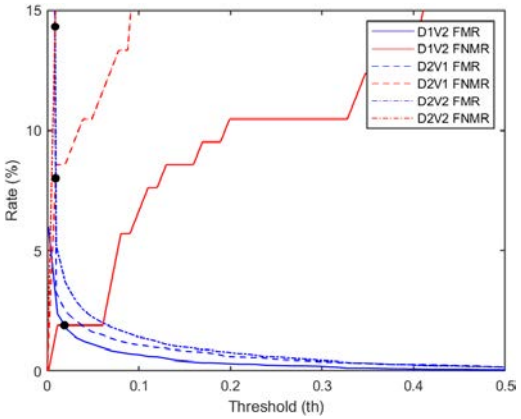
(b) Two-days enrollment with S1.



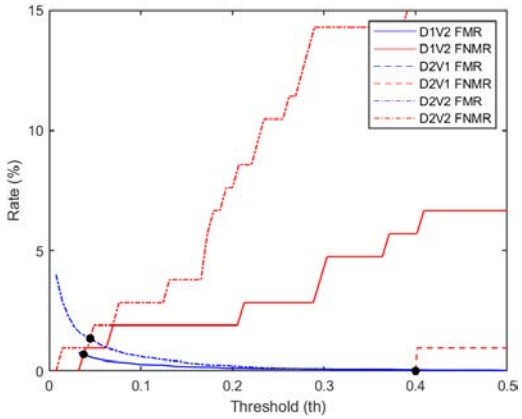
(c) One-day enrollment with S2.



(d) Two-days enrollment with S2.

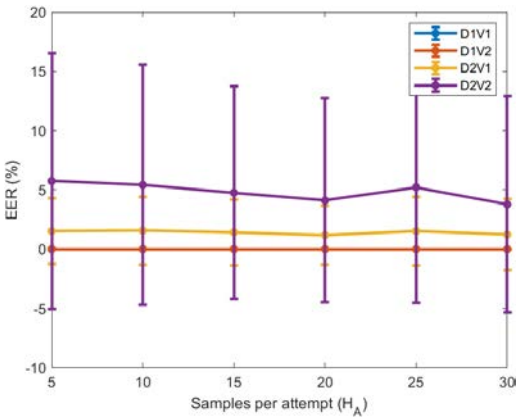


(e) One-day enrollment with S1+S2.

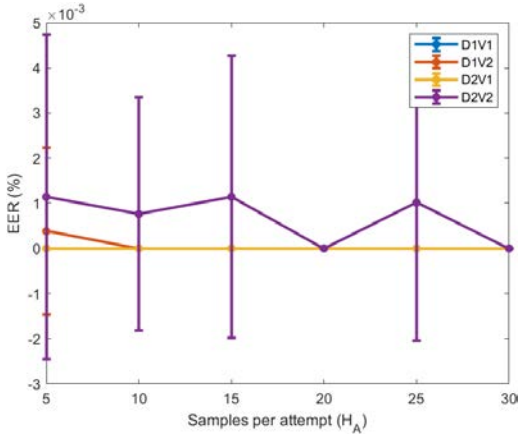


(f) Two-days enrollment with S1+S2.

Figure 9.4: FNMR and FMR curves for the different subsets of data S1, S2 and S1+S2. The EERs are marked with black dots.

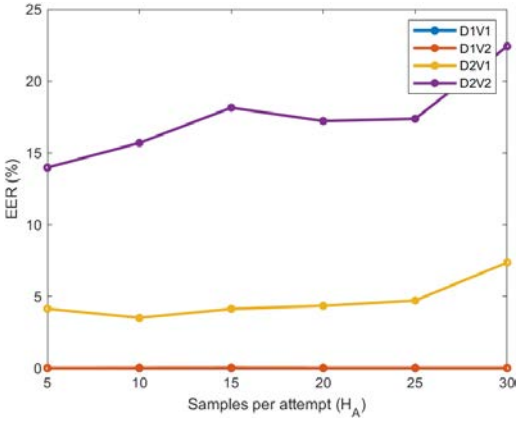


(a) One-day enrollment.

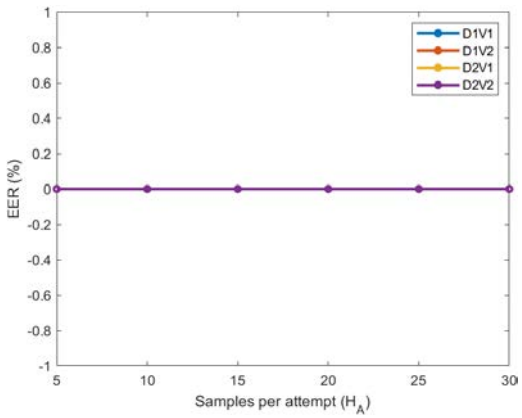


(b) Two-days enrollment.

Figure 9.5: Performance results using data from the entire BMSIL database, S1+S2, using one attempt with different samples.



(a) One-day enrollment.



(b) Two-days enrollment.

Figure 9.6: Performance results using data from the entire BMSIL database, S1+S2, using all attempts with different samples.

10. VIABILITY OF HUMAN VERIFICATION WITH A SMARTBAND PROTOTYPE

The experimentation in previous chapters has been carried out using the BMSIL database. This database has helped setting a baseline, observing how effective every algorithm and approach is for ECG recognition. However, this quality is provided in exchange of convenience. The capture device in the BMSIL database is not user-friendly for a biometric environment, as it involves sensor placement and experience with the acquisition process. For this reason, the knowledge collected in the chapter 8 is applied to the smartband databases specified in chapter 4.

10.1. Peak detection algorithms

The peak detection algorithm of choice has been the one specified in section 5.2.1 throughout all the experiments with the BMSIL algorithm. However, the smartband databases, BMSIL-SB and GUTI, do not have suitable data for this algorithm and it performs poorly. This issue lead to the development of a custom R peak detection algorithm for smartband data, but also considering extra alternatives.

10.1.1. Custom algorithm for smartband

Both databases were collected with the same smartband prototype. However, the GUTI database presented more challenges as the acquisition protocol aimed to be closer to a realistic environment by reducing the supervision in the process. For this purpose, the person in charge of recording the user's data was not required to observe the retrieved data and check its correct collection in any of the different visits. Instead, the user was told to be as still as possible, but avoiding to take long periods of time to increase acceptability.

These mentioned characteristics resulted in lower quality signals in both smartband databases, where signals with abrupt fluctuations and noise are more frequent in the GUTI database. These events affected the performance of the initial peak detection algorithm, creating peaks with high amplitudes that interfere the detection of valid peaks. The Pan-Tompkins algorithm did not perform properly either. The initial approach to solve this problem was to set a quality criteria to discard signals with fluctuations. As fluctuations reduced the number of detected peaks or mislead the detection, the criteria was based on determining the number of peaks and discard those signals with less peaks that the ones given by a threshold.

Despite of devising this approach as a quality criteria, it required a complete peak detection, which finally became a peak detection algorithm itself, instead of being used for

discarding signals. The deployed algorithm was achieved with simple signal operations, involving thresholds for some of the stages. The scheme of the algorithm is represented in Figure 10.1 and are further discussed below. The following figures represent the outcomes of each step for two different users in GUTI database in resting. Parameters wSize and minPeakDist are fixed to 0.2 s or 200 data points. In the case of BMSIL-SB, 0.2 s correspond to 100 data points.

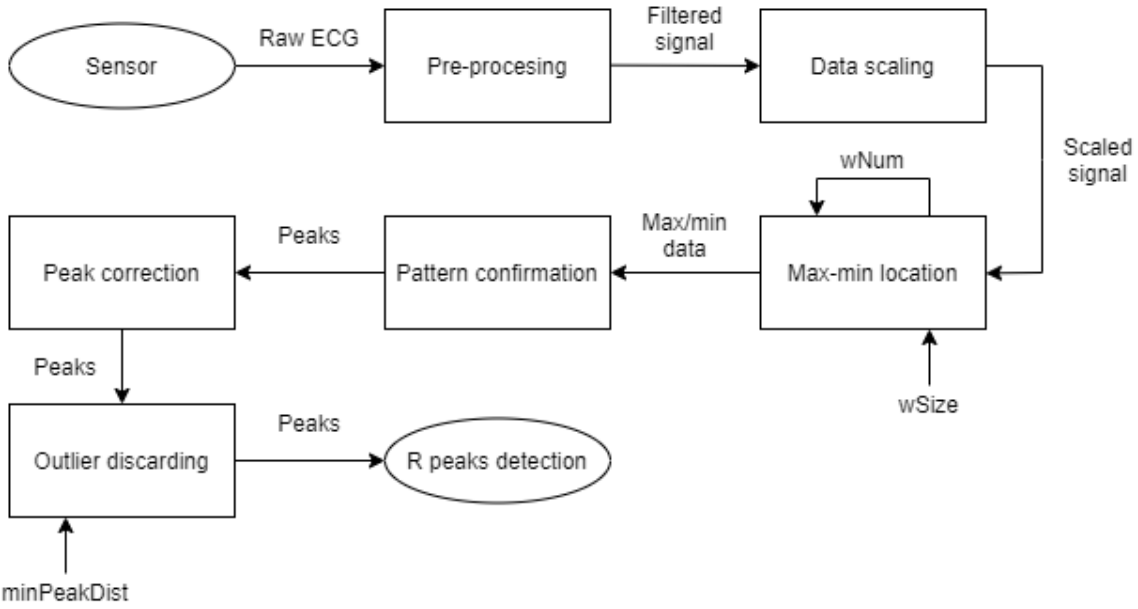


Figure 10.1: Scheme for the peak detection algorithm for low fidelity signals. Parameters wSize and minPeakDist are fixed by observation, and wNum is derived from wSize.

Signal scaling

The filtered signal returned by the pre-processing stage got scaled. The scaling is done based on the maximum absolute value of the signal. The scaling is in Equation (10.1), where F represents the ECG after pre-processing and F_{scale} is the resulting scaled signal. This method allowed to keep the baseline in 0 and amplitude-related thresholds between 0 and 1.

$$F_{scale} = \frac{F}{max(|F|)} \tag{10.1}$$

Max-min location

This block of the algorithm applied fixed length overlapping windows and obtained their local maximum and minimum for each one. The number of final windows, m of length k considering an overlap of r from an original signal of length n is determined by the Equation (10.2). Particularizing for a window length of 0.2 s, its value based on the

sampling frequency resumes as wSize = $0.2 \cdot f_s$. Determining a 50% overlap, the final number of windows is only defined by the original signal duration, as seen in Equation (10.3). As the windows overlap, the max-min location would sometimes get repeated. If the case, the repeated point would be discarded. As a consequence, some final pairs would be very close to each other as observed in Figure 10.2.

$$m = \frac{n - r}{k - r} \tag{10.2}$$

$$wNum = 10(t_s - 0.1) \tag{10.3}$$

Pattern confirmation

Based on the shape of a QRS complex, R peaks are immediately followed by a minimum, which is the S point of the complex, as previously detailed in Figure 3.1. In an ideal case scenario where all the maximum and minimum were correctly found, their location in time would be alternate, i.e.: after a maximum there would always be a minimum, and the other way around. To ensure that the resulted data was fulfilling this condition, there was a detection of points that were two consequent maximum or minimum, deleting the first one that was observed. This step allows to discard redundant points resulting from the previous windowing process, as observed in Figure 10.3.

Peak correction

The previous detected peaks were sometimes placed in their corresponding T waves or found inside an abrupt fluctuation which would produce incorrect data. This part of the algorithm encapsulated two types of correction, one related to abrupt fluctuations and another one regarding to the T wave location corrections.

- Abrupt fluctuations: the signal value in the different retrieved time points was obtained as a vector. The variance got calculated for every point in the vector in order to observe how they independently fluctuate. Normal fluctuations were considered to have a variance value lower than 0.05. Therefore, every greater value was considered an abrupt change. User A did not present obvious fluctuation so it remained unaffected after this step, in Figure 10.4a. Nonetheless, the user B presented an initial distortion in the middle, so the peaks detected in that area were discarded, as seen in Figure 10.4b.
- T wave location: According to section 3.1, the QRS has a duration of 0.12 s. The given time point was applied for the center of a windowing process with 0.2 s of duration, to ensure the encapsulation of the QRS complex. If the point location belonged to a T wave, there has to be a maximum in the window with higher

amplitude. Otherwise, the obtained point corresponded to a correct R peak. In the case of user A, represented in Figure 10.5a, there were several instances in which the R peak was detected around the T wave location. In some cases they did not even correspond to the said T wave, but represented a similar shape. This method discarded said points and corrected them into the proper location. However, in the case of user B in Figure 10.5b there were a few incorrect identifications of the T wave error, as the R present a slightly lower amplitude in some cases. However, most of them remain correctly identified.

Outlier discarding

In some cases, after the previous stages, some peaks remain incorrectly detected. Some peaks remain too close to those correct R peaks, and stated before, there cannot be two R peaks in a window with higher duration than the QRS. When two peaks were detected with a proximity of less than 0.2 s, the lowest peak was discarded. User A did not result in any changes after this process, as observed in Figure 10.6a. On the contrary, user B did present some mistakes that were fixed with this criterion, represented in Figure 10.6b. However, there is a mistake that keeps appearing since the T wave correction and belongs to a fluctuation issue.

The two user examples provided in the previous pages are a representation of different types of signals present in both smartband database, considering that data in BMSIL-SB was more stable in general. Even though some R peaks could not be properly detected, and/or some of them were missed, in Figure 10.7 there is a representation of the average of all the detected peaks for both users A and B. Every window is scaled with the same criteria from Equation (10.1). Even considering the different performance and nature of these signals, for both users the most constant part for every window is the QRS complex. As observed, the mean QRS fits accurately with the joint plot of the individual windows. Figure 10.7a represents less noise than Figure 10.7b, but this noise compensates to form a clear ECG waveform after calculating the mean.

The same algorithm was also tested in the BMSIL-SB database, considering the f_s = 500Hz. Figure 10.8 collects examples for resting in two different users, C and D. Figure 10.8a shows more stable QRS complexes in comparison to those in the GUTI database, resulting in an average signal with less fluctuations. In Figure 10.8b the algorithm shows some detection errors related to a high amplitude in the T-wave. This may have happened due to the T-wave having an absolute higher amplitude than the R peaks or the corresponding window not being big enough to encapsulate the corresponding R peak, which is a result of a low heart rate. This issue leads to a normalization with respect to the highest point, which would correspond to the T-wave. However, the algorithm still retrieves valuable information, and the average shows a good QRS are.

10.1.2. Other alternatives

None of the available databases for this thesis provided R peak labeling. As a consequence, it is not possible to achieve a numerical performance evaluation of the algorithm. The given parameters and steps taken in the algorithm have been visually assessed using samples from all the different experiments in the GUTI database, relying initially on the D1V1 visit in the sitting scenario. As this database contains more complex data with lower quality, its performance deals properly with the BMSIL-SB database.

In order to check a common and feasible state-of-the-art algorithm, the Pan-Tompkins algorithm has been included in this stage, and the experiments are carried out using this approach, too. Visually, the performance is adequate when used in BMSIL-SB, for not suitable in the case of GUTI database. This may be a result of the first database having more quality, as the Pan-Tompkins algorithm was designed for professionally collected ECG signals.

The lack of algorithm validation for the available data could give uncertainty when carrying out experiments that rely on said algorithms. To be sure that the verification results are only a product of the classifier or the data quality, an extra R peak detection has been achieved manually for the BMSIL-SB. However, it was not used for validating the peak detection algorithm, as the indexes may slightly vary between the automatic and manual detection. This would require extra criteria to determine the evaluation, and it is not easy to determine.

The final number of detected peaks for BMSIL-SB and GUTI database are summarized in Table 10.1 and Table 10.2. We must consider that the first database was formed by heterogeneous numbers of samples of 206 users, whereas the latter was formed up to 72 users, hence the big difference between their detected peaks.

Table 10.1: Number of detected peaks for the BMSIL-SB database and both algorithms. The average number of peaks per user is in parenthesis.

Algorithm / **Visit**	Custom	Pan-Tompkins	Manual
Rest	10507 (51.00)	12883 (62.5)	12792 (62.1)
Exercise	14665 (71.2)	17404 (84.5)	18002 (87.4)

In both databases, the Pan-Tompkins algorithm is the one with most detected peaks. This could imply more errors, or detecting peaks that were not considered in the manual detecton. In addition, there is a clear increment in the exercise experiments, as a result of the increased heart reate. However, in the custom detection in the GUTI database that does not happen, as it could be a result of avoiding fluctuations, which may be more present in that scenario.

It is important to highlight the differences between the manual detection and the algorithms in Table 10.1. The Pan-Tompkins algorithm detects more peaks, which could

Table 10.2: Number of detected peaks for the GUTI database and both algorithms. The number of peaks per user is in parenthesis.

Visit	Scenario	Custom	Pan-Tompkins
D1V1	Sit	4053 (56.3)	5471 (76)
	Walking	3717 (51.6)	5734 (79.6)
	Exercise	3368 (46.8)	6460 (89.7)
D1V2	Sit	3894 (54.1)	5387 (74.8)
	Walking	3621 (50.3)	5475 (76)
	Exercise	3445 (47.8)	5917 (82.2)
D2V1	Sit	3721 (54.7)	4729 (69.5)
	Walking	3451 (50.7)	4797 (70.5)
	Exercise	3263 (48)	5815 (85.5)
D2V2	Sit	3852 (57.5)	4815 (71.9)
	Walking	3630 (54.2)	4670 (69.7)
	Exercise	3242 (48.4)	5422 (80.9)

probably be related to detection mistakes. The custom algorithm always detects less than the other two approaches. This is a result of the algorithm being conservative, as it is designed to preferably detect less but correct data, instead of detecting wrong points in the signal.

10.2. Data preparation and tuning

Once the R peak algorithms are specified, the segmentation, differentiation and classifier are selected based on the results obtained throughout this Thesis. However, extra feature transformations are included to the differentiation as it is expected to have worse results with smartband data.

The taken procedures are based on those in [90], which applied them to the BMSIL database. There are some differences, as the referenced work applies an initial longer segmentation, considering 0.4 s before and after the R peak detection. In addition, the applied algorithm for R peak detection is different, and there is more data available with less users. The different parts of the experiment are applied to both databases as follows:

- Segmentation: centered R peaks with 0.1 s before and after. The peak detection algorithms are the following:
  - BMSIL-SB: manual, Pan-Tompkins and custom peak detection with the previously specified and 500 Hz sampling frequency.
  - GUTI: Pan-Tompkins and custom peak detection with 1000 Hz sampling frequency.

- User discarding: users with a total number of peaks less than 5 is discarded.
- QRS features transformations:
  1. FD: First differentiation of the QRS complex.
  2. FD + SWT: SWT with Daubechies 9 wavelet for the FD, selecting level 4 coefficients.
  3. FD + SWT + IFS: IFS to the SWT coefficients, choosing the selected features to use.

The tuning process is similar to the one in MLP, but with extra specifications:

- Each tuning process is independently achieved depending on the type of features for both databases.
- BMSIL-SB only allows one-day enrollment, using the sitting experiment with d = 0.8 development set.
- GUTI database allows one and two-days enrollments:
  - One-day: using D1V1 sit in tuning and development set proportion d = 0.8.
  - Two-days: using D1V1 sit and D2V1 sit in tuning. The d_{enr} value refers to the data proportion considering both days altogether. In this case, the d gets values of 0.4 and 0.8: the first uses a similar number of samples to one-day tuning, and the second one takes advantage of the larger number of samples, doubling it.
- Each training, including those in tuning, implement a 5-fold cross-validation with 80-20 proportions.
- Three final different models are trained with these hyperparameters specifying different enrollment proportions, where d_{enr} can be 0.5, 0.7 and 0.9 including the two-days enrollments.

10.3. Verification experiments

10.3.1. BMSIL-SB

Even though the main focus is put in verification, identification results are also considered to observe how different these two approaches can be in the process. The BMSIL data only collects two types of scenarios, which simplifies the analysis.

10.3.2. GUTI

In the case of the GUTI database, the results are also achieved as in the BMSIL-SB up to the exhaustive search. Once this search is done for all the three feature transformations and two peak detection algorithms, the viability of every approach is observed through training with the different enrollments, without needing to test the remaining scenarios.

Based on the training results, the best approach is selected with its according enrollment size. Then, they are analyzed based on the different type of scenarios: sitting, walking and exercise. The final performances are obtained individually for one-day and two-days enrollments, considering the best enrollment proportions for each case.

10.4. Results

10.4.1. Effects of SWT and IFS

As a preliminary observation for the SWT and IFS features, an extra section is included providing the behavior of both smartband databases. This procedure also determines the fixed features that are selected from the IFS algorithm for each of the databases, so they are crucial for verification. For representation purposes, the SWT complexes are averaged and they are divided by the maximum. This division is also done with the weights from the IFS, to represent the feature relevance.

BMSIL-SB

Figure 10.9 represents the results for both experiments in the BMSIL-SB, with the three peak detection alternatives. The manual detection presents wider coefficients, followed by Pan-Tompkins algorithm with the custom algorithm being last. As the manual detection was conservative, we can assume the averaged coefficients as only correct ones. As the x-axis separates from the enter, in Pan-Tompkins and custom algorithms, the average is more constant. We can infer this as a result of different small amplitudes canceling each other on the average calculation.

Considering the same peak detection algorithm, there are clear similarities in shape between scenarios. Even though outliers could vary this mean result, there is constancy in the SWT coefficients. The observed weights in the exercise scenario peak in the extremes of the representation, which implies the starting of a new QRS complex. This does not happen in rest scenarios as the complexes are wider due to the lower heart rate.

The weight shape on the manual peak detection is smoother, considering the surroundings of the R peak the most relevant features, and decreasing clearing the further it gets. However, in the case of the custom algorithm and Pan-Tompkins algorithms, the weights are not as clear, but they are also in the same range. This is a result of detected

outliers, which did not happen in the manual peak detection. In addition, the custom algorithm presents more fluctuations in the extremes than Pan-Tompkins, which implies more mistakes detecting the R peak, which is probably shifted towards the highest weight is.

GUTI

The GUTI database only had the possibility of using Pan-Tompkins and the custom peak detection algorithms. However, the number of scenarios and experiments are larger. The observations are based on the three scenarios given in D1V1 with both algorithms. The results of the normalized mean SWT and weights are in Figure 10.10. The axis is doubled with respect to those in BMSIL to represent the same lapse of time, as the sample frequency is doubled.

The left column with Figures 10.10a, 10.10c, 10.10e refer to the custom peak detection algorithm. In contrast to those in the right column, Figures 10.10b, 10.10d, 10.10f, they show more noise. However, the weights from the Pan-Tompkins algorithm are more chaotic, implying possible mistakes in detection. More specifically in the case of exercise, where it is clear that some of the complexes are shifted, as most of the high weights are represented in the left side of the graph.

As opposed to the BMSIL database, the GUTI database has narrower complexes in general. This proves that the BMSIL database had more supervision in the data collection, as the users had a lower heart rate as they were more calmed.

Regarding the weights in the IFS database, results tend towards those in the BMSIL. The most relevant features are around the R-peak. In this case, the number of involved features could appear to be lower, but we have to consider the increased number of samples. The highest weighted features are close to the R-peak, collecting around 50 features. The selection for these experiments will be a window segment centered in the R peak, representing 50 ms.

10.4.2. Verification with BMSIL-SB

Hyperparameter tuning

The results using a development set of 80% of the rest scenario in the BMSIL-SB database are collected in Table 10.3. These results are summarized in Figure 10.11 to easily observe how each factor affects to the mean EER.

The graphical representation of the mean EER shows that the Pan-Tompkins achieves the best results of all the algorithms in the three cases, being the FD + SWT transformations the most successful. This transformation is also the one with the best results in manual peak detection and the second best in the custom algorithm.

Regarding the manual detection with FD it is remarkable that the achieved results are the worst of the three. It can be a result of having a conservative manual detection by a naked eye with no medical expertise. However, it benefits from using the SWT coefficients, as the EER drops drastically. This might be a result of the noise deletion in the SWT.

In the custom algorithm, the SWT does not improve the results. It its probably a consequence of wrong R peak detection. If the complexes are misplaced, the noise reduction would not impact the results.

For the three algorithms, reducing the data with IFS does not result in improvements, reaching the worst results in two of the three cases.

Table 10.3: Best models with Exhaustive Grid with d = 0.8 for the three peak detection algorithms and possible feature transformations.

Transformations	Peak detection	Hidden layers	Activation	Alpha	Tolerance	Mean EER (%)
FD	Manual	500	ReLU	0.0005	0.01	1.979
	Custom	350		0.005		1.289
	Pan-Tompkins	300		0.0001		0.538
FD + SWT	Manual	700	ReLU	0.0001	0.01	0.782
	Custom	700	Tanh	0.0001		1.750
	Pan-Tompkins	400	ReLU	0.001		0.392
FD + SWT + IFS	Manual	700	ReLU	0.0001	0.01	1.395
	Custom	400		0.0001		2.198
	Pan-Tompkins	500		0.001		0.952

Recognition with different enrollment sizes

Even though the tuning has been achieved using the 80% of the data, the three enrollment sizes are applied independently with the hyperparameters specified in Table 10.3.

In this case, the graphic representation is left out as results show a clear trend and drastic differences. Therefore, both EER and accuracy are represented in Table 10.4 considering the FD + SWT transformations after the Pan-Tompkins detection.

The rest scenario results in desirable EER values, but reaching the best one with d = 0.7. However, when looking at the identification accuracy, these results do not map to the good results in verification. Using this metric as the evaluation metric in the tuning process would impact performances in verification. The accuracy evaluates sample per sample and classifies based on the maximum score. However, the implemented EER considers more than one sample.

Regarding the recognition process with exercise scenarios, the results change dramatically. The performances for both verification and identification drop dramatically. From almost ideal EER results up to almost 14%. The change in the accuracy is more significant, reaching a little over 20% of correct classification in the best case scenario.

Table 10.4: Results for accuracy and EER with the BMSIL-SB database and different enrollment sizes. The complexes have the FD + SWT feature transformation and Pan-Tompkins peak detection.

Scenario	Rest			Exercise		
$\mathbf{d_{enr}}$	0.5	0.7	0.9	0.5	0.7	0.9
EER (%)	0.142	0.078	0.485	12.035	13.530	13.445
Accuracy (%)	72.930	78.243	78.139	18.455	21.734	22.244

10.4.3. Verification with GUTI

Hyperparameter tuning

Following the scheme for the BMSIL-SB database, the best model in the Exhaustive Grid and the corresponding mean EER are summarized in Table 10.5 for the two available peak detection algorithms.

When considering the custom algorithm, 255 complexes are deleted from the initial detected ones, as a result of not considering signals with less than 5 samples. In the elimination of these samples, the final number of users goes from 72 to 67 in the D1V1 sitting experiment. In the two-days tuning, the D2V1 sitting scenario also gets 121 samples deleted under this criteria.

Table 10.5: Best models with Exhaustive Grid with d = 0.8 for the possible peak detection algorithms and feature transformations.

Transformations	**Peak detection**	**Hidden layers**	**Activation**	**Alpha**	**Tolerance**	**Mean EER (%)**
FD	Custom	400	ReLU	0.01	0.01	4.590
	Pan-Tompkins	700		0.01	0.05	9.695
FD + SWT	Custom	700	ReLU	0.01	0.01	4.512
	Pan-Tompkins	350		0.05		8.522
FD + SWT + IFS	Custom	700	ReLU	0.01	0.01	6.189
	Pan-Tompkins	450		0.001		13.347

For easier observation the results are plotted graphically in Figure 10.12. In this database, the Pan-Tompkins database clearly performs worse than the custom algorithm. This might be a result of wrong R peak detection given the data present in the GUTI database. The BMSIL-SB database could be more similar to professional collected ECG present in the data used to develop the Pan-Tompkins algorithm.

As it happened in the BMSIL-SB database, the best mean EER comes after the FD + SWT transformation, as reducing data with IFS could be getting rid of more valuable than misleading information.

Considering these transformations and peak detection criteria, another tuning process is achieved using two days, and applying two types of proportion. The final values for the different enrollments and proportions are referred in Table 10.6. Doubling the number of

samples in two-days enrollment clearly lower the mean EER in training, therefore this is the chosen proportion in two-days enrollment.

Table 10.6: Final hyperparameters and the mean EER for the different types of development sets and proportions for GUTI database. 255 and 121 samples are deleted from D1V1 and D2V1 sit scenarios, respectively.

Type	d	Hidden layers	Activation	Alpha	Tolerance	Mean EER (%)
One-day	0.8	700	ReLU	0.01	0.01	4.512
Two-days	0.4	500	Tanh	0.0005	0.01	10.935
	0.8	450		0.001		7.737

Recognition with different enrollment sizes

Considering the best performing peak detection approach, the custom algorithm, three models are trained with different enrollment sizes for one-day and two-days enrollments. The verification for the remaining data in both sets is summarized in Table 10.7. The higher the value of d, the lower number of samples used for recognition.

These results show that the GUTI database is clearly not suitable for identification purposes. We can observe that the accuracy results even for the same scenario as in training, are just coincidences. This observation exposes that the classifier is incapable of distinguishing the user with only one sample. Regarding the EER results, even though they are higher than those in BMSIL-SB, are still suitable for verification purposes.

Considering that the two-days experiment implies using the double amount of information, the results do not improve as it could be expected. This could be a consequence of overfitting or distortion when including more samples in the verification process. In addition, it could be also a product related to the lower data available in the same scenario experiment, as we con only use those samples that were not included in the enrollment set.

In the case of one-day experiments, the best EER is obtained with 0.7, as this happens using 0.9 in two-days enrollment.

Table 10.7: Results for accuracy and EER with the GUTI database and different enrollment sizes. The complexes have the FD + SWT feature transformation and custom peak detection.

Type	One-day			Two-days		
$\mathbf{d_{enr}}$	0.5	0.7	0.9	0.5	0.7	0.9
EER (%)	3.586	2.510	4.568	7.465	5.970	5.925
Accuracy (%)	1.106	0.877	0.526	1.211	0.789	1.316

As the selection of scenarios and different collection dates in the GUTI database is larger, we need to observe the results based on the criteria specified in section 10.3.2.

Given the initial results in accuracy, this metric is not consider in the following verification experiments.

Verification in sitting scenario

In Figure 10.13 the different EER results are represented with their corresponding numerical values. Two things need to be considering when interpreting these results: D2V1 has representation in enrollment when using two-days, so it should result in better generalization, and the proportion is also higher than the one in one-day enrollment.

In general, two-days enrollment result in better EER. In general, the more different the visits are from the enrollment, the higher the EER results.

For one-day enrollment, the results are clearly impacted when the visit is not the same in enrollment. However, we do not observe that those observations are worse when changing the day of acquisition, which could imply that it does not really matter how much time passes by between the enrollment and the verification process.

It is a noticeable result for D2V1 in the case of two-days enrollment. This could be a consequence of overfitting or using too many samples from that visit in training, i.e.: the proportion was 90% for enrollment in this case. We can assume is a bad generalization, as in D2V2 the results skyrocket to 30%.

In general, we can assume the bad signal quality in D2V2, as none of the enrollments consider this scenario and both models are not capable of separating users in verification.

Verification in walking scenario

Similarly, results for walking scenarios are represented in Figure 10.14. Generally, all the results are worse than those in the sitting experiment, so we can assume the smartband collection and data modeling is negatively affected by this change of scenario.

Again, the two-days enrollment is the best approach in all the visits. D2V1 is noticeably less than the result with one-day due to the representation of that visit in enrollment. In general, it helps with better generalization throughout all the scenarios.

Verification in exercise scenario

Finally, the results under exercise are plotted in Figure 10.15. This scenario is also better generalized using two-days enrollment, but the trend is coherent with all the observations throughout this thesis: both types of enrollment have an increase in their EER when considering exercise.

This scenario also shows noticeable differences in D2V2, which happened for sitting and walking experiments too. It allows to confirm that this specific visit may have had

present additional issues while collecting or detecting the R peaks.

10.4.4. Final system for smartband recognition

BMSIL-SB

Selecting the best enrollment proportion as 0.7, the final system results in the FNMR and FMR curves that correspond to both rest and exercise scenarios are plotted in Figure 10.16. The estimated EER are 0.078% in the rest scenario, and 13.530% in exercise. The graphic representation of the EER is greater than the estimated result, yielding in more than 16%.

The FNMR and FMR curves show the differences between the final thresholds for both scenarios, and their evolution throughout this variable. In the case of exercise, the model struggles to clearly distinguish the user to verify, most of the mated users are contained in lower scores. This is not what happens in the rest scenario, where most of the mated users have a score higher than 0.1, which does not happen in non mated data.

The final system in the BMSIL-SB database uses the Pan-Tompkins algorithm resulted in an average 62.5 samples per user. Only 70% of those samples were required for enrollment, resulting in 43.75 samples for enrollment. Depending on the heart-rate, considering a minimum of 60 bpm in resting state, this would only imply less than a minute for enrollment.

GUTI

This database has not been found suitable for identification purposes, as the accuracy results are around or less than 1%. However, the final system's performance related to verification is represented based on the different scenarios in Figure 10.17. The enrollment is done throughout two days.

In the case of the sitting scenario in Figure 10.17a, results in the highest thresholds and lowest EER among all the possible scenarios. However, there are clear differences in the evolution of the FNMR curves between visits. The two visits that take part in enrollment, D1V1 and D2V1, have the lowest EER being D2V1 almost ideal, which could imply overfitting. Second visits result in higher EER, as they have not taken part in enrollment, which could be interpreted as the most realistic situation.

The EER results, as previously referred, increase when changing the scenario to walking as observed in Figure 10.17b. The pattern of obtaining lower EER with those visits that took part in enrollment is repeated in this case. However, in this case, D1V1 and D2V1 have very similar behaviors as the threshold increases. In this case, the second visits also turn out different, where D2V2 is still the one with worst results.

The observed trends in walking are reproduced similarly in the case of the exercise but

with higher values in terms of EER. As expected, it is a result of lower quality data given after exercise. Again, D1V1 and D2V1 have similar tendencies, with worse performances of D1V2 and D2V2.

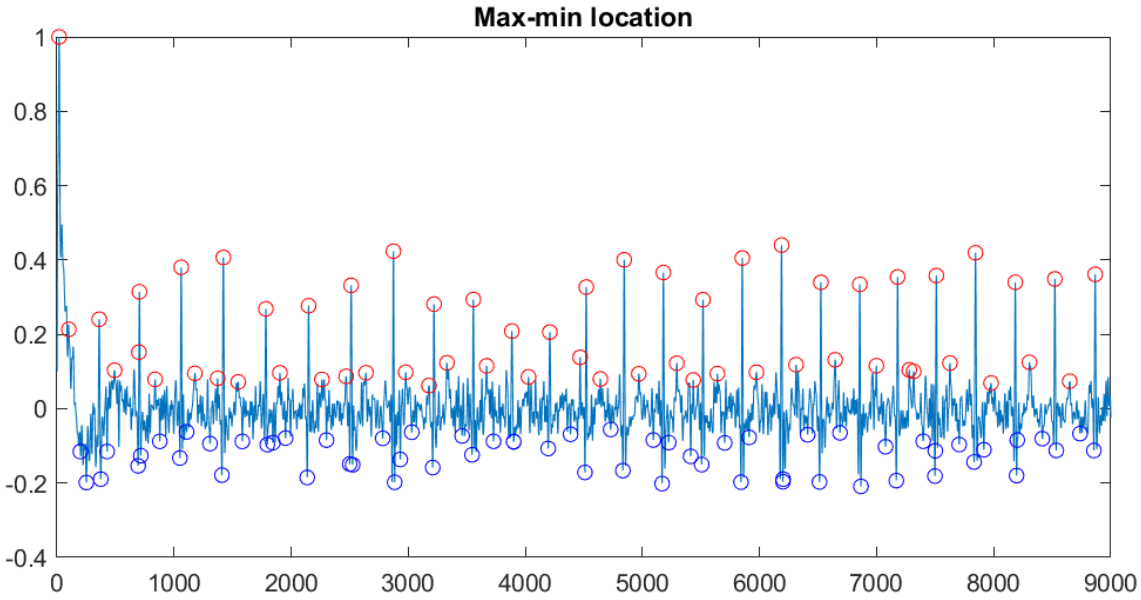
10.5. Conclusions

In this chapter we have studied the potential of implementing ECG recognition using a low cost smartband prototype. The most successful tools and techniques have been applied to this case, but also some adaptation was required given the database characteristics.

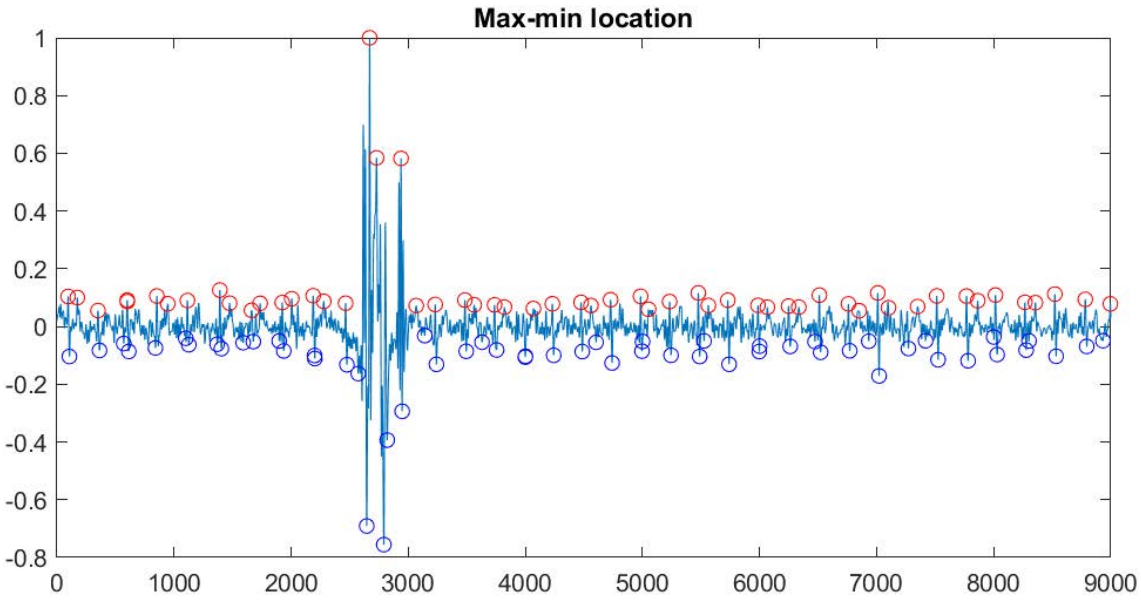
The collected data with this prototype are very susceptible to bad positioning or collection. The BMSIL-SB has showed that the data has chances of being used as an identification tool, which could probably improve when the number of enrolled users is lowered. However, for this purpose, the data needs to be collected carefully and the probability of mistakes in identification is high, reaching 78.2% of accuracy in a database of 206 users. In terms of verification, this approach is viable when the recognition scenario is consistent to the one in enrollment, reaching almost an ideal EER. These results are altered if the user presents an increased heart rate, reaching an EER of 13%.

A more realistic acquisition is achieved in the GUTI database, proving that precision in this process is key considering the used prototype. This database has shown that user-friendly collection is not appropriate if the system's goal is identification. The one-day enrollment resulted in a system with a range of 2.510%–40.678% EER. The addition of a two-days enrollment has been showed to improve the general verification results. Nonetheless, these values are still high for a commercial use, ranging between 0.068% to 31.669% in the worst case scenario.

To achieve a successful ECG verification using this smartband prototype, the user must have some expertise using the device. In addition, the environment in which this system is applied to, must be constraint and ensure of having users with similar conditions as those in a sitting scenario.

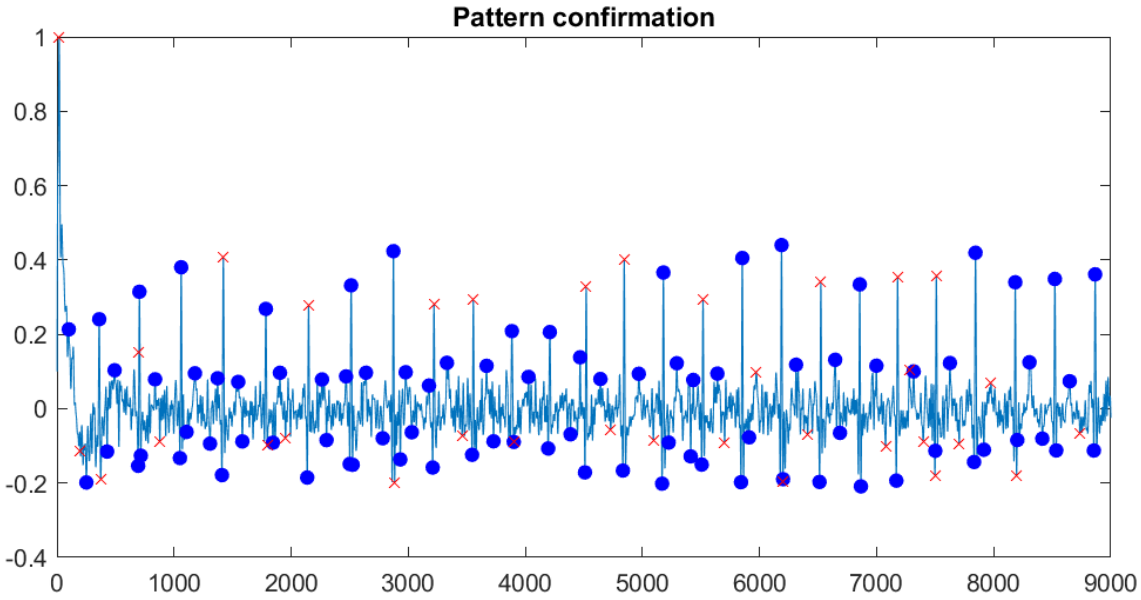


(a) User A.

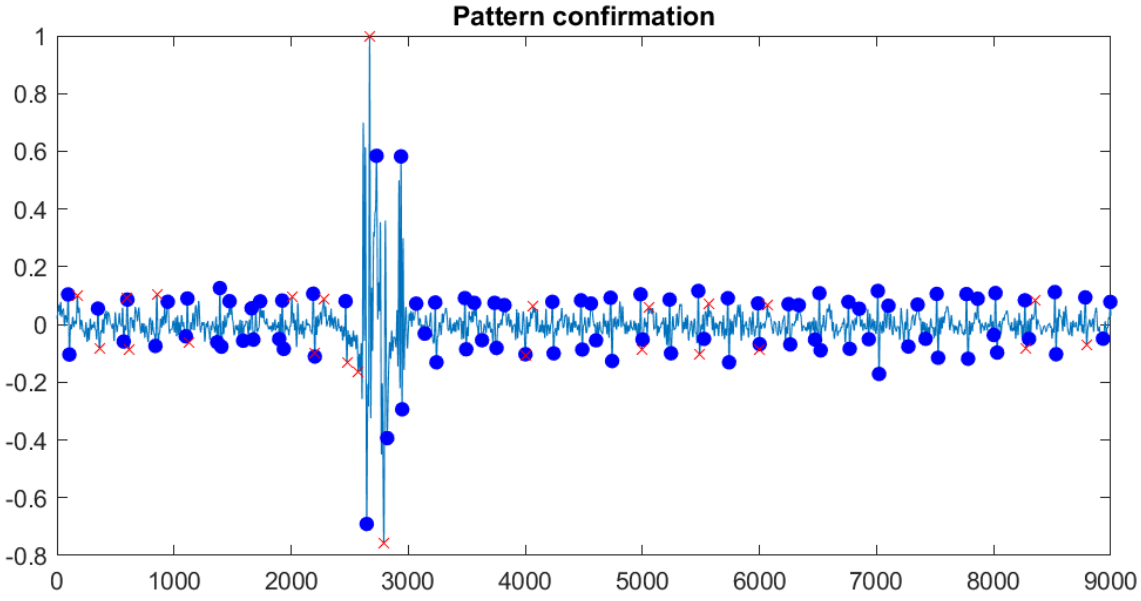


(b) User B.

Figure 10.2: The circles represent the maxima (red) and minima (blue) locations in the scaled signal.

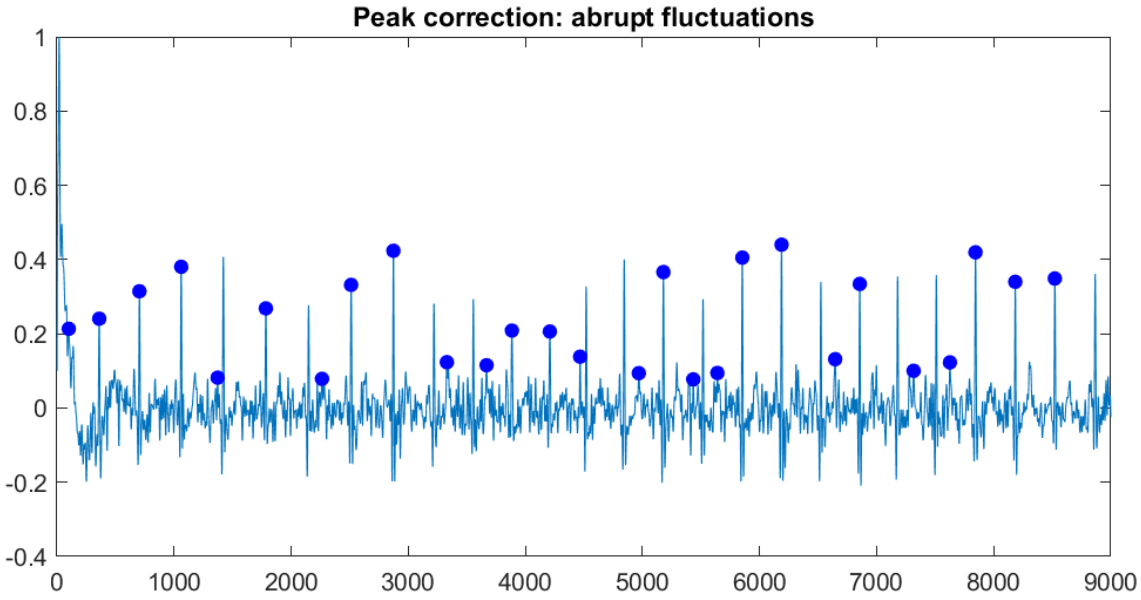


(a) User A.

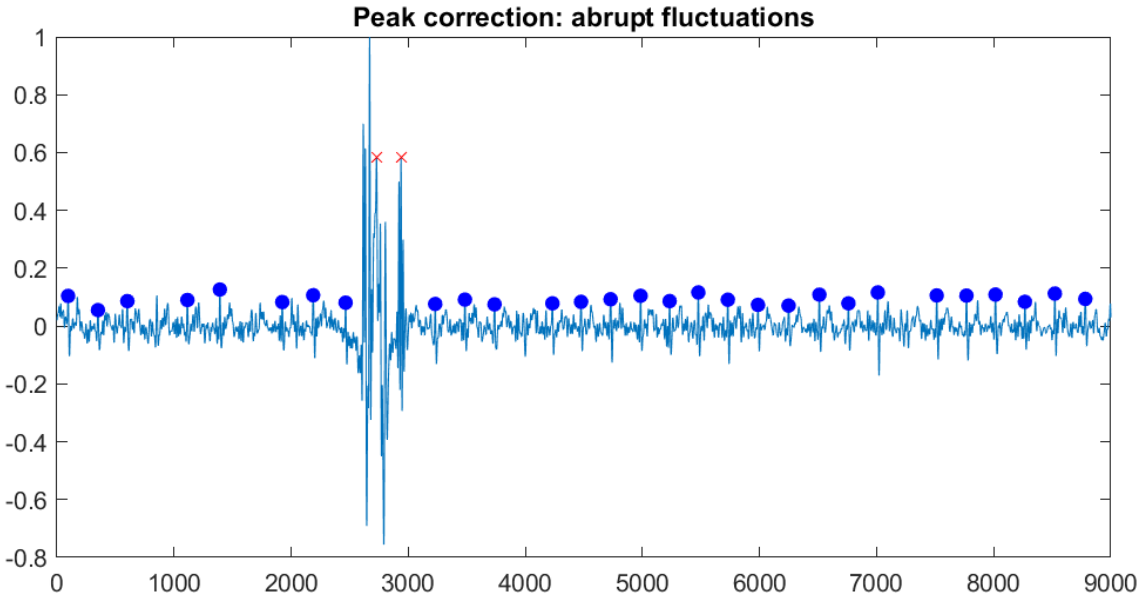


(b) User B.

Figure 10.3: The red crosses represent those peaks that were discarded after pattern confirmation. The blue filled circles are the valid maximum-minimum.

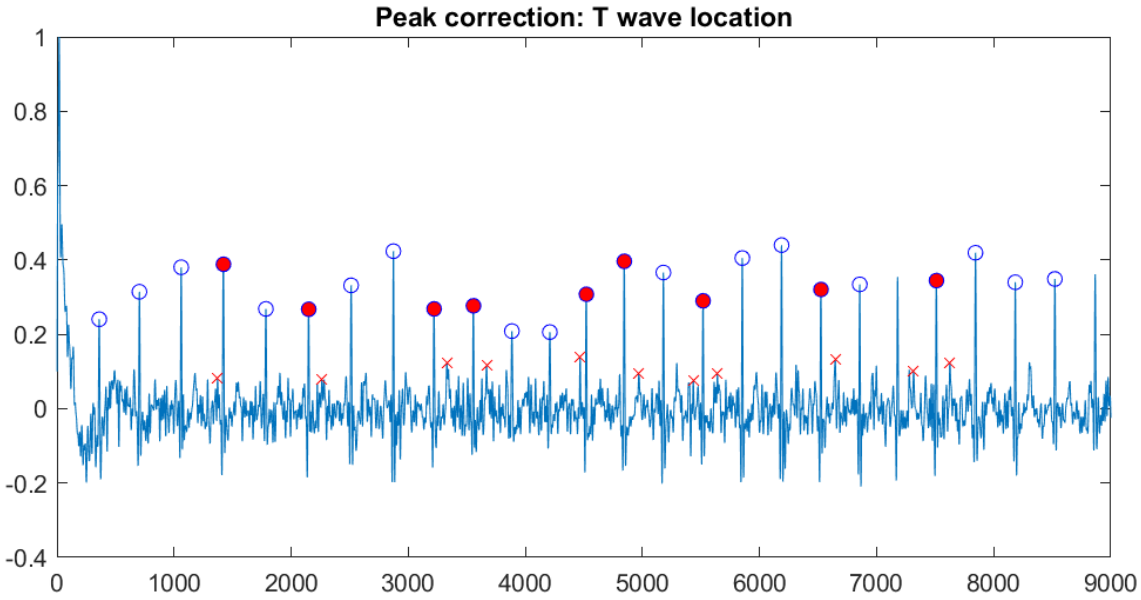


(a) User A.

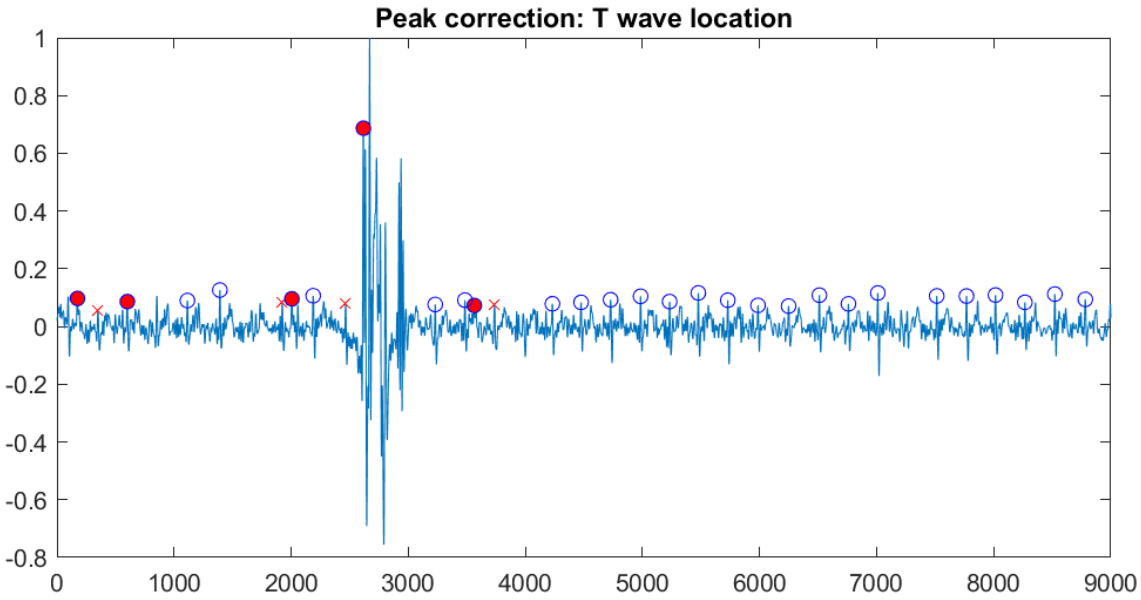


(b) User B.

Figure 10.4: The red crosses represent those peaks that were discarded as they were part of an abrupt change in the signal. The blue filled points belong to those that remain valid for the next block.

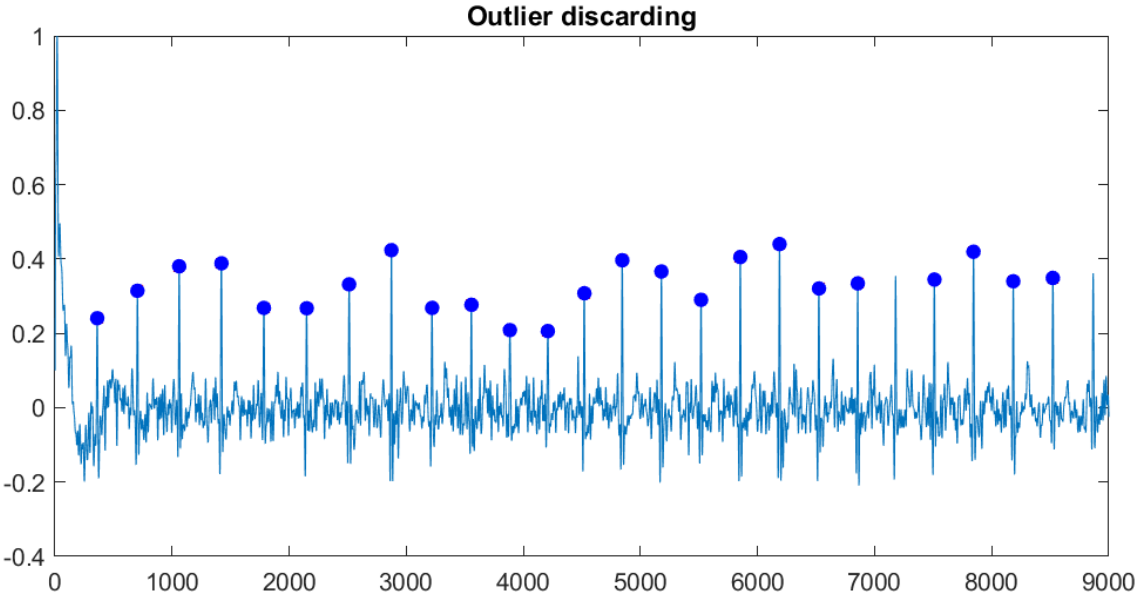


(a) User A.

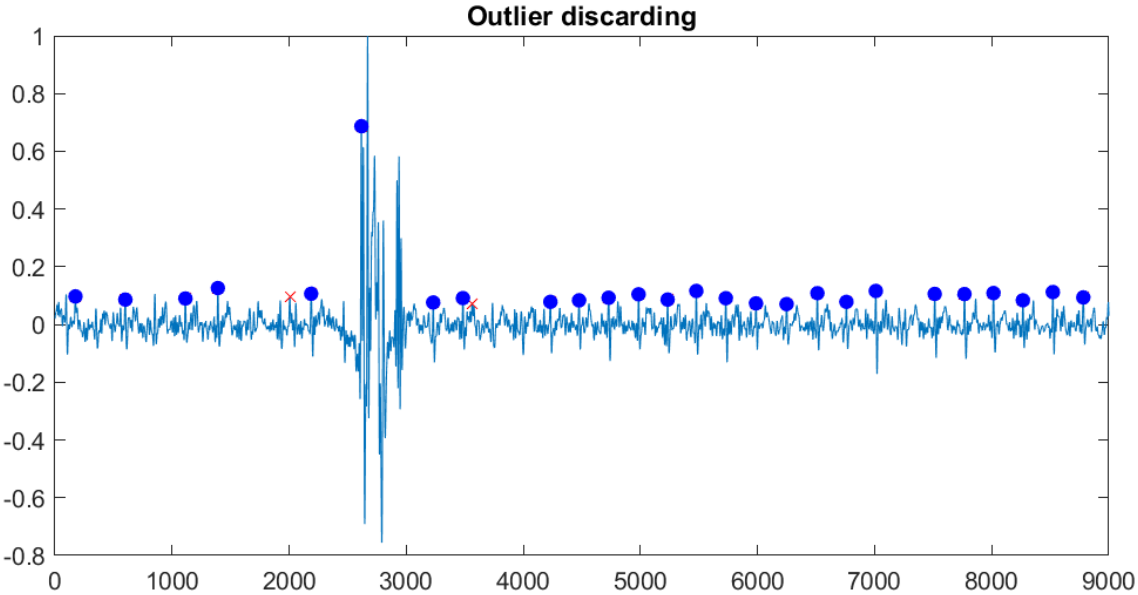


(b) User B.

Figure 10.5: Red crosses represent the discarded peaks. Red filled dots are the corresponding new peak assignation. Remaining blue circles belong to the peaks that remained the same in this stage.

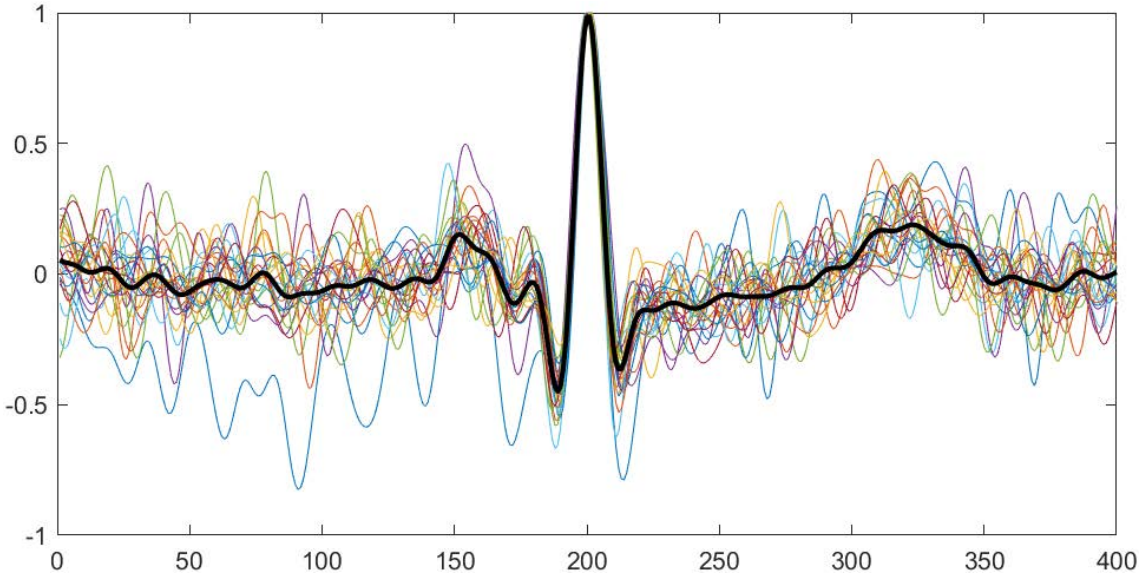


(a) User A.

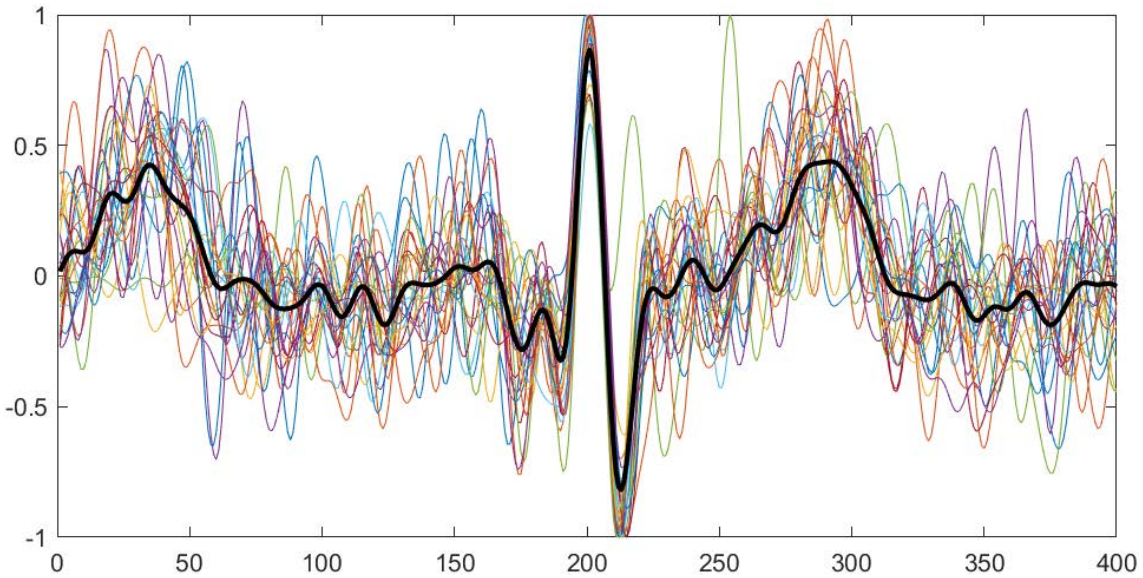


(b) User B.

Figure 10.6: R peaks in their correct position.

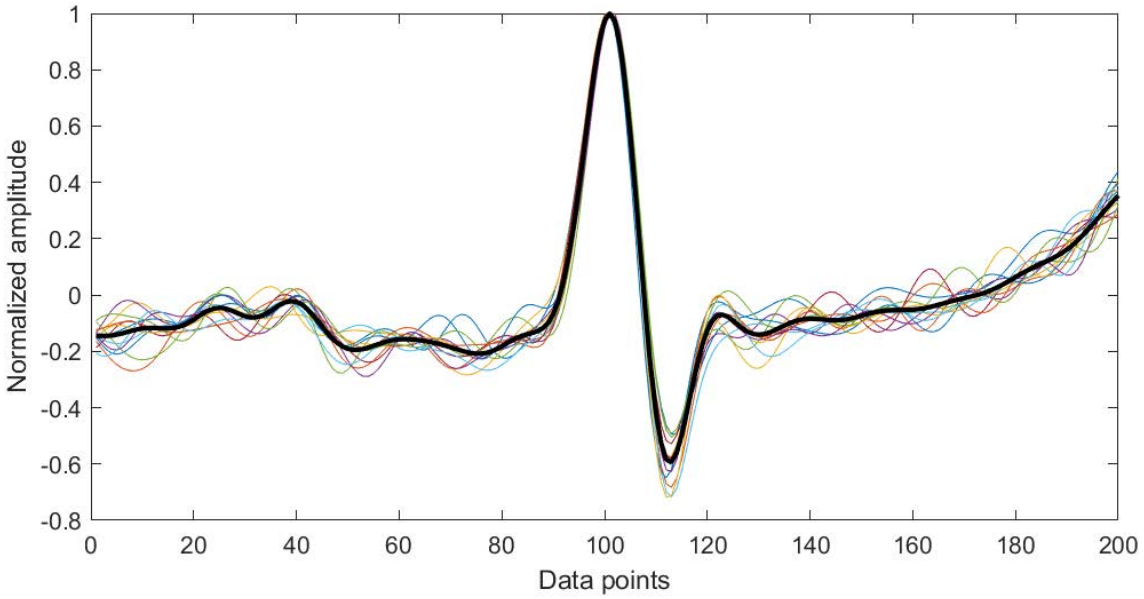


(a) User A.

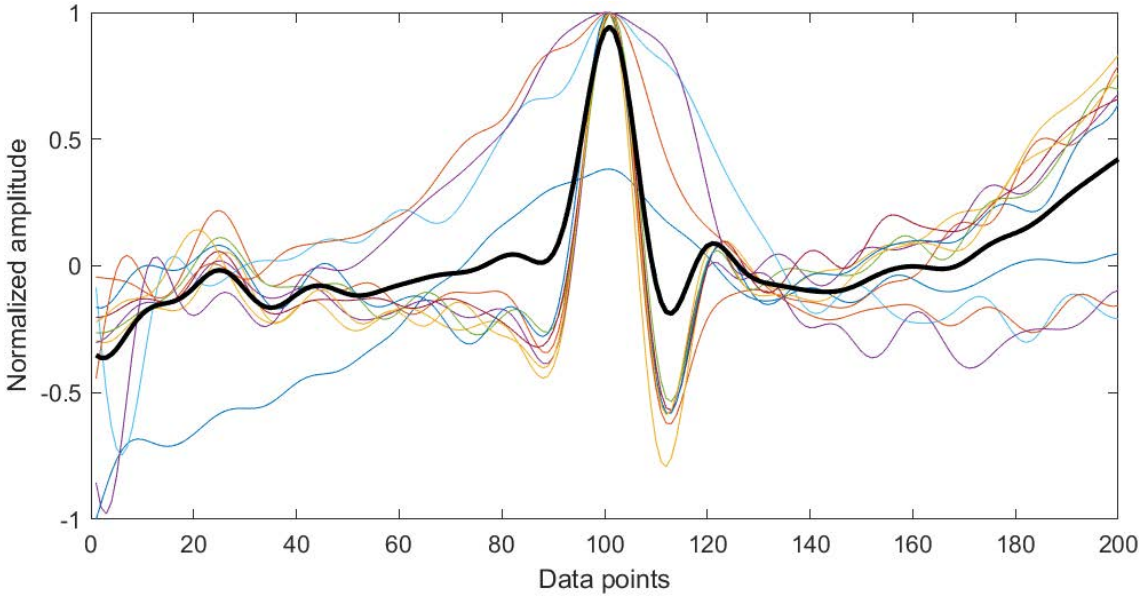


(b) User B.

Figure 10.7: Windows of 0.4 s for GUTI database in sitting scenario. The signals centers correspond to the detected R peaks and their mean signal in black.

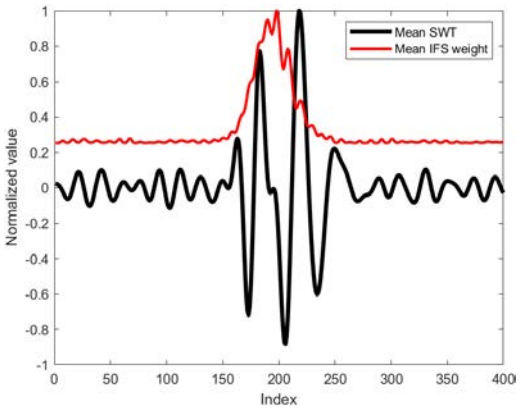


(a) User C.

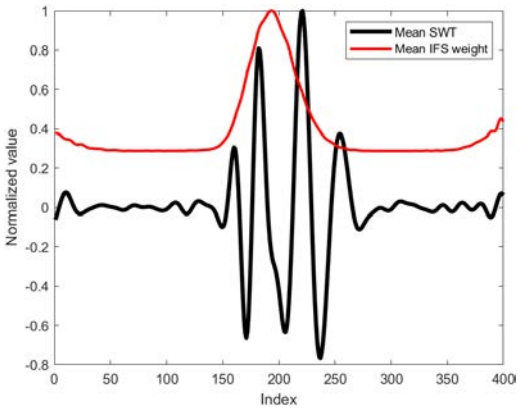


(b) User D.

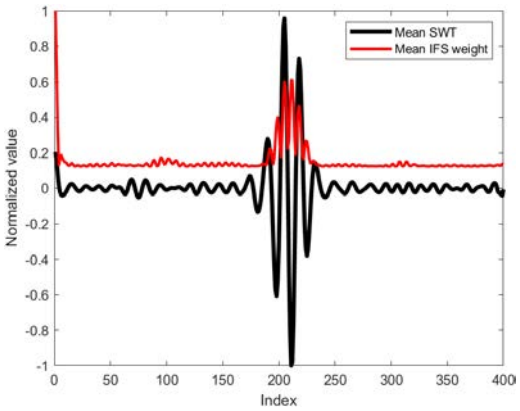
Figure 10.8: Windows of 0.4 s for BMSIL-SB database in resting scenario. The signals centers correspond to the detected R peaks and their mean signal in black.



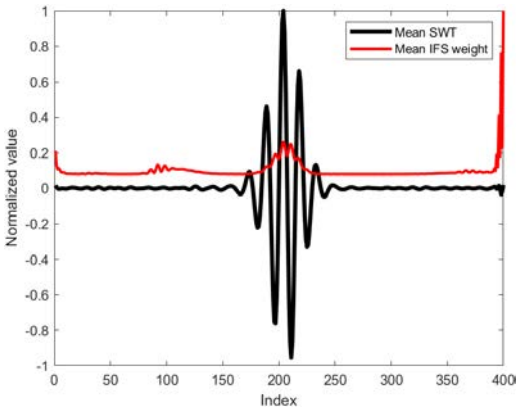
(a) Manual peak detection: rest.



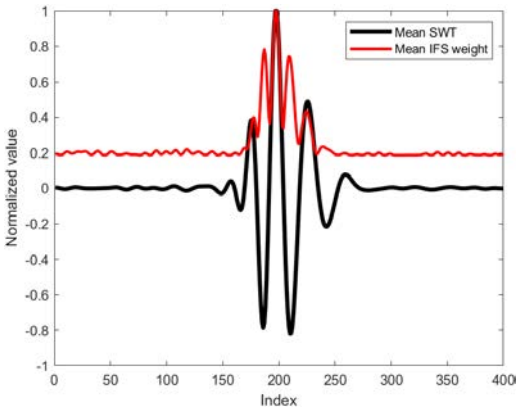
(b) Manual peak detection: exercise.



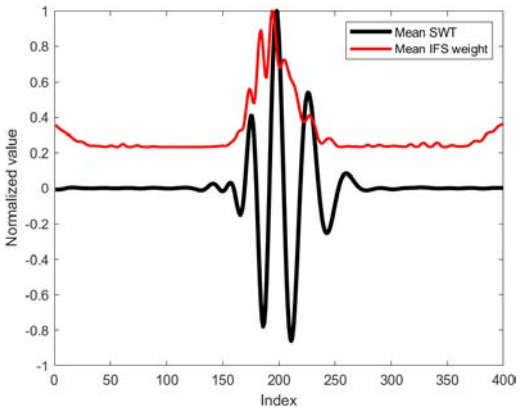
(c) Custom peak detection: rest.



(d) Custom peak detection: exercise.

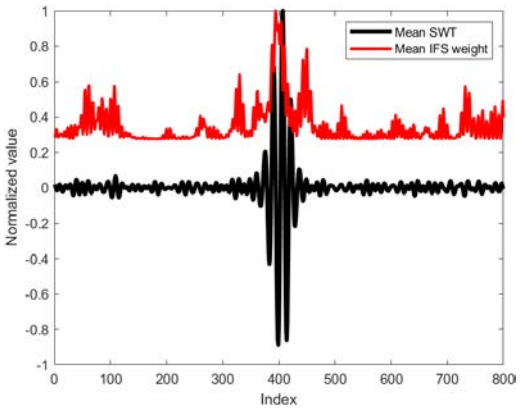


(e) Pan-Tompkins peak detection: rest.

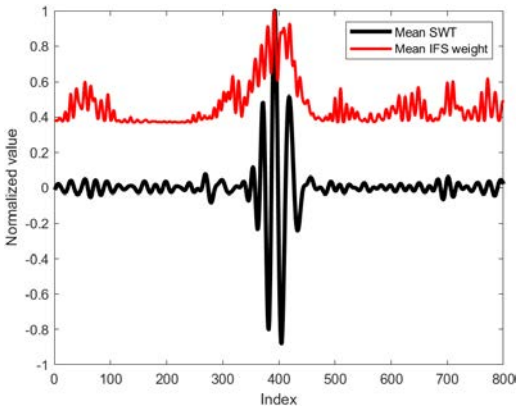


(f) Pan-Tompkins peak detection: exercise.

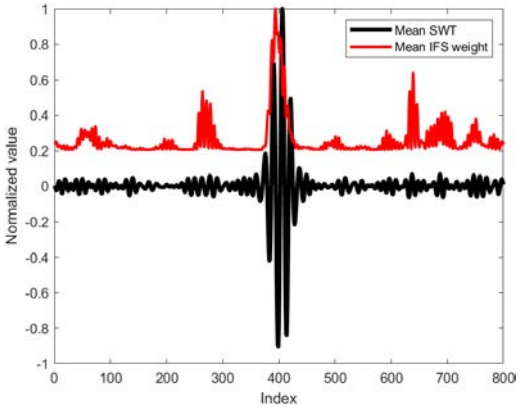
Figure 10.9: Mean normalized SWT for all detected complexes for the different peak detection algorithms and experiments in BMSIL-SB database in black. The normalized weight of each feature from the IFS algorithm is represented in red.



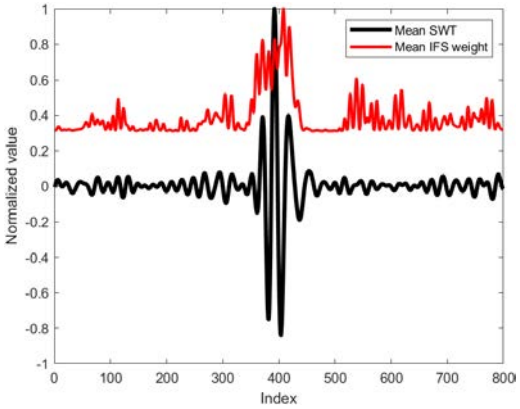
(a) Custom peak detection: sit.



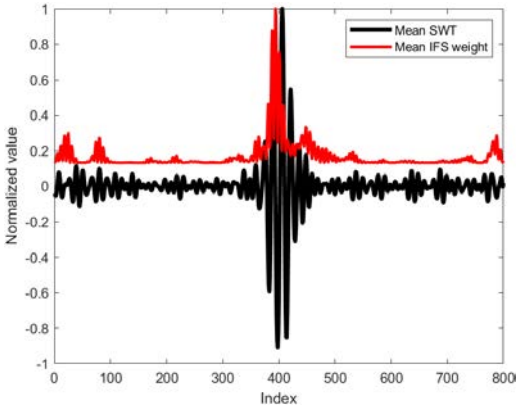
(b) Pan-Tompkins peak detection: sit.



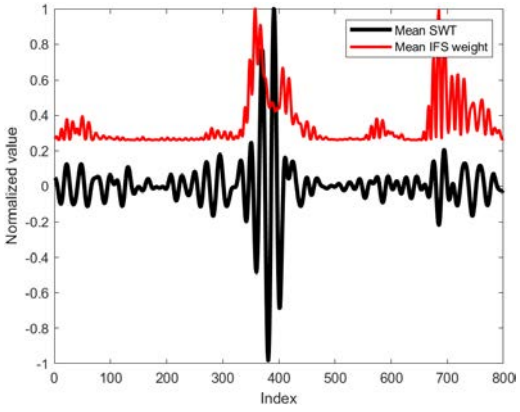
(c) Custom peak detection: walking.



(d) Pan-Tompkins peak detection: walking.



(e) Custom peak detection: exercise.



(f) Pan-Tompkins peak detection: exercise.

Figure 10.10: Mean normalized SWT for all detected complexes for the different peak detection algorithms and experiments in GUTI database in black. The normalized weight of each feature from the IFS algorithm is represented in red.

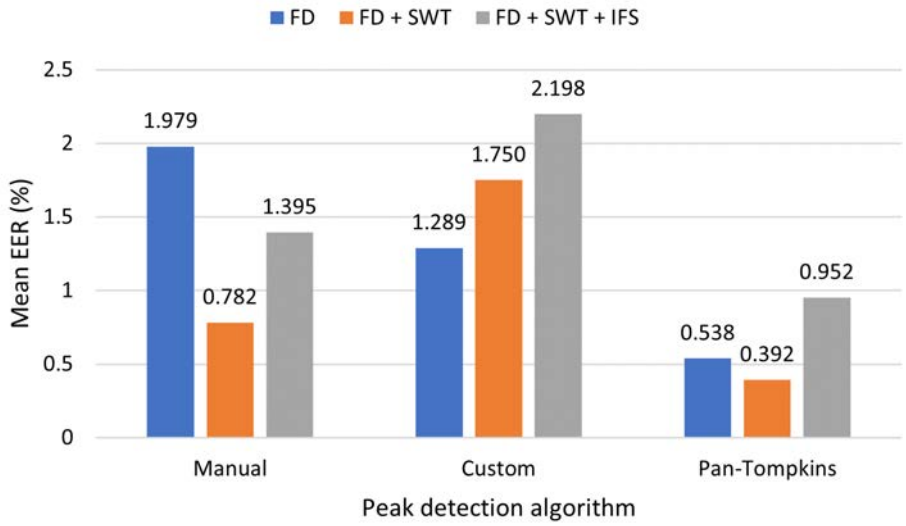


Figure 10.11: Mean EER for the best result in the Exhaustive Grid for all peak detection algorithms and feature transformations in BMSIL-SB database.

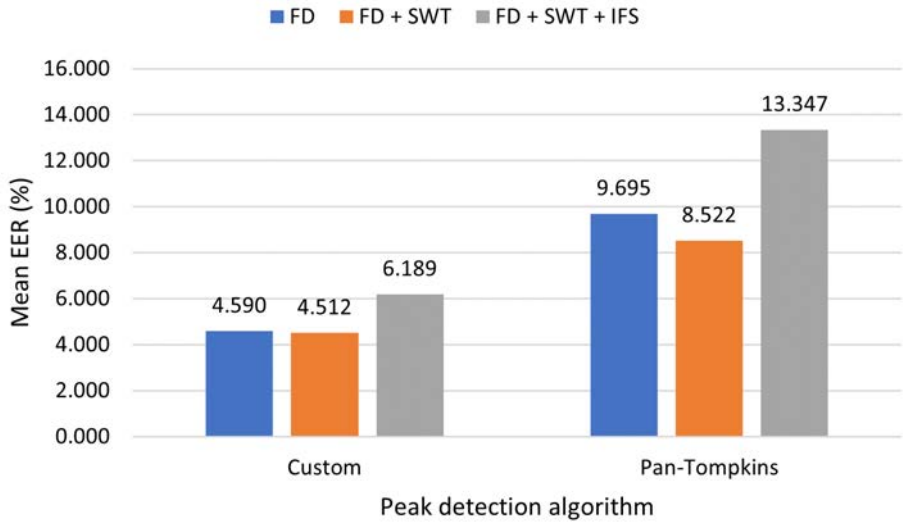


Figure 10.12: Mean EER for the best result in the exhaustive grid for all peak detection algorithms and feature transformations in GUTI database.

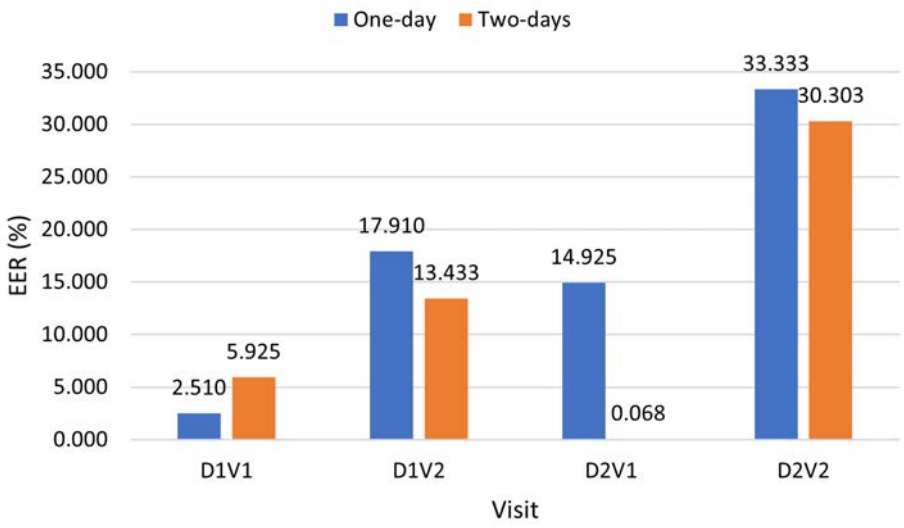


Figure 10.13: EER results for the sitting scenario with GUTI database, considering one-day and two-days enrollment.

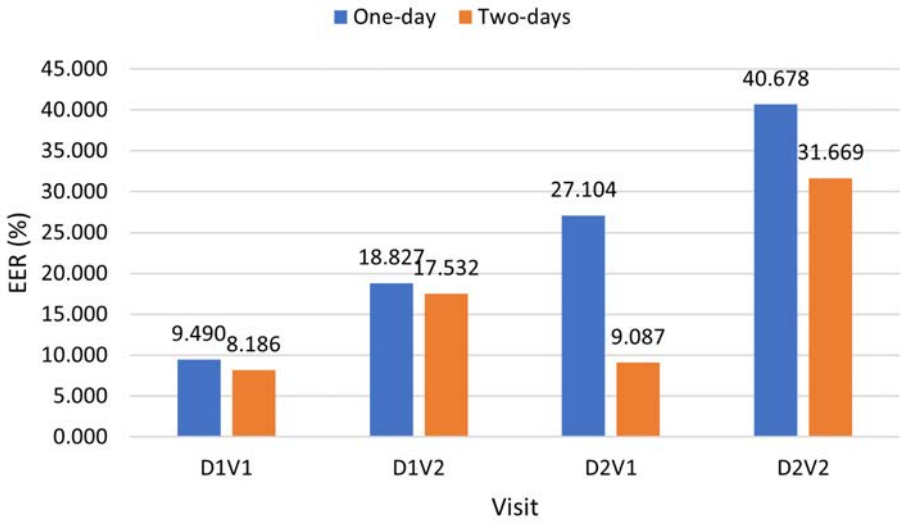


Figure 10.14: EER results for walking with GUTI database, considering one-day and two-days enrollment.

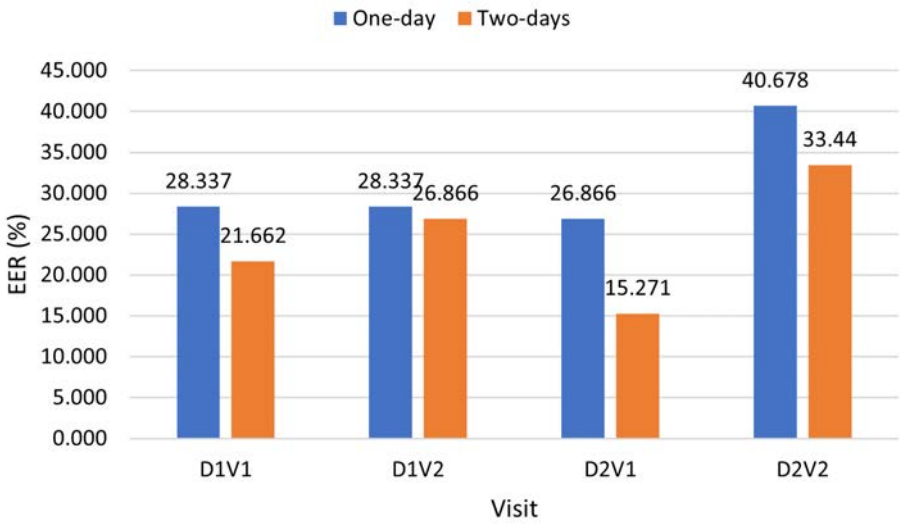


Figure 10.15: EER results for exercise with GUTI database, considering one-day and two-days enrollment.

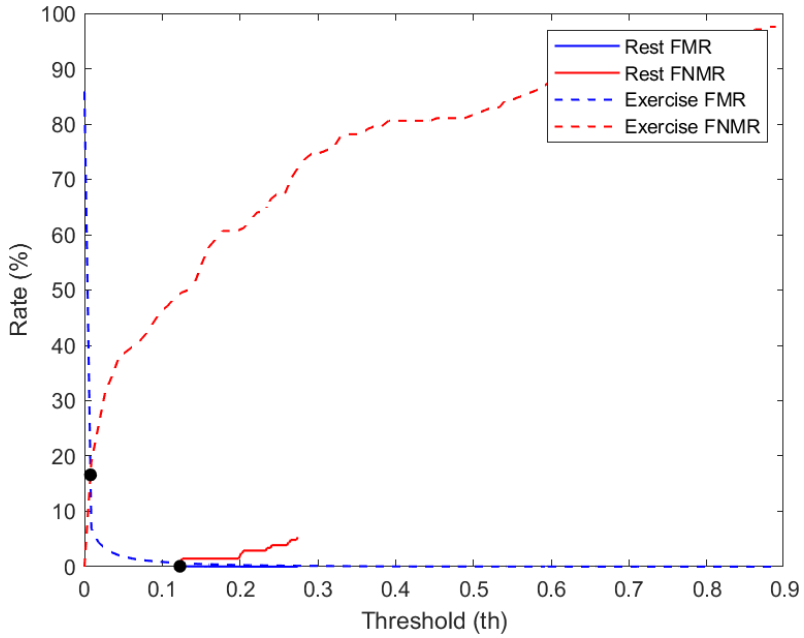
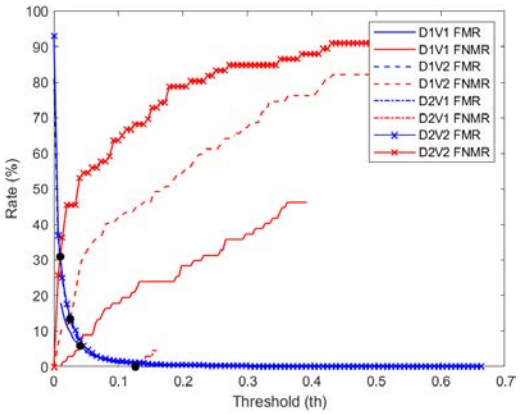
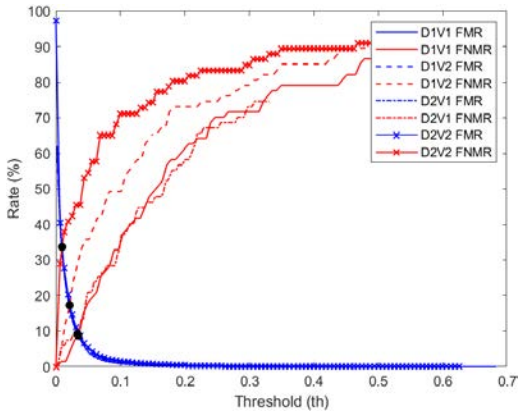


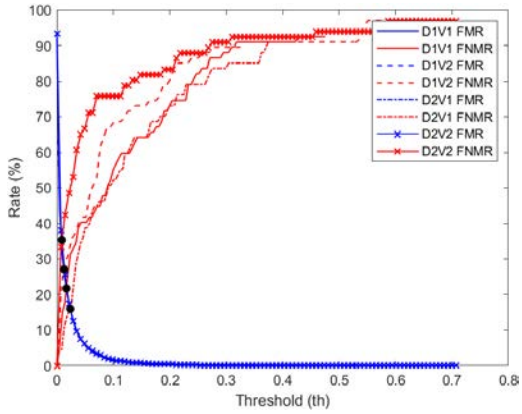
Figure 10.16: FNMR and FMR graphs for both rest and exercise scenarios in BMSIL-SB database.



(a) Sit.



(b) Walking.



(c) Exercise.

Figure 10.17: FNMR and FMR curves for the final configuration of the GUTI database in different scenarios with two-days enrollment.

11. CONCLUSIONS AND FUTURE WORK

11.1. Conclusions

This thesis has observed the viability of using ECG as a biometric signal while getting close to a real case scenario. Through this process, we have obtained the following conclusions:

- ECG can improve the verification performance of fingerprint biometrics, where the best choice is using it in PAD.
- The MLP algorithm can achieve acceptable results under changes of scenario, proving the viability of the modality only using basic transformations.
- Using two days of enrollment clearly enhances the system's performance when considering neural networks, being the best approach.
- Good quality data does not clearly differentiate between changes of position. This situation is the same when considering recognition with data collected a few hours and a few days after the enrollment.
- The way the verification is achieved highly impacts the functionality, which implies a specific design for the number of attempts and samples that are required.
- Low fidelity sensors are very susceptible to position and heart rate changes, requiring very controlled acquisitions and extra noise removal.

The main focus was put into achieving good verification results without the requirements of transforming fiducial data at the expenses of high computational costs. Different classification algorithms have been tested from low to more complexity to achieve this goal. All the different approaches showed the impact of changing the heart rate after exercise, which complicates the fiducial detection and the model generalization.

The consideration of using one model per user was limited and only achieved with DCT features GMM, resulting in 11.26% of EER and problems in convergence, not being suitable for any implementation. In terms of training one model for all the users, results with k-NN and SVM did not perform appropriately, resulting in high processing time and memories or poor performances. The LDA algorithm was successfully tested and was barely affected by the increment of data in enrollment. The obtained verification results of 7.465%–8.906% in EER proved the viability of using ECG as a biometric trait under different scenarios. This improvement in comparison to simpler characteristics proves that increasing complexity in the classification, helps generalizing with ECG data.

Results obtained with LDA showed the potential of this biometric trait even considering various physical scenarios. However, they cannot compete with conventional modalities. Nonetheless we have proven that combining ECG with fingerprint improves the verification in 70.64%. Moreover, when approaching it as a PAD solution, the system fell into an ideal EER.

In the process of improving the performances, we have tested more algorithms, proving them as good solutions. However, they need to be evaluated based on the conditions and requirements of the biometric application. These algorithms have allowed to assess the trade-off between results, hardware and time requirements. We have observed that even the Deep Learning approach, BioECG, is more complex than MLP, the optimization of the latter is a simpler yet successful classification with only one day of enrollment. In addition, through MLP optimization process, we have proven the first differentiation to be the best transformation when trying to enhance the features of a QRS complex.This result is a consequence of enhancing the abrupt changes in the waveform, and making it easier to generalize.This transformation concluded in a range of 0%–6.324% in EER, which improves to 0–0.247% by changing the verification approach.

BioECG, on the contrary, presented hardware and time limitations when the optimization had to be done without cross-validation. This fact interferes in the assessment of the differentiation performance. As a consequence, only the non-differentiated QRS is used, resulting in struggles when generalizing in the most complex scenarios in both verification and identification. However, once the two-days enrollment was introduced, the BioECG system has showed results of 1.352%, getting closer to conventional and commercial biometric traits. This second day of enrollment also impacts the performances with data from the first day, which proves the variation in the second day added extra information that can be extrapolated to other scenarios. Depending on the application and the security demands of the system, this approach is significantly better even considering the complications added in enrollment and training.

In the pursuit of results that relate to a real scenario, we have also implemented extended verification alternatives. These extra approaches in verification have showed how they can change the outcome of the system, based on the number of attempts and the number of samples per attempt: MLP results drop to 0–0.247% and BioECG gets an ideal EER result. This happens with a single attempt with lower samples or grouping the samples in attempts. These improvements are not only numerical, but we can also consider them more realistic, as it is more probable and convenient to be verified when using a single attempt with few samples.

Even considering the various characteristics of all the achieved experiments, there are general trends based on the different scenarios. When the enrollment is done while sitting down and relaxed, the standing position does not significantly affect if we compare them to the same experiment in another day. We can assume that verifying while standing has the same impact as doing the verification another day in the same enrollment scenario.

On the contrary, the heart rate increase clearly impacts the verification, which could be a consequence of a bad data collection or lack of generalization.

The final stage of trying to recreate a real scenario was reached when the device was substituted by a prototype smartband. By comparing two databases collected with different protocols, we have showed how the supervision impacts the system's verification. Even considering the same device, the same R peak detection algorithms, Pan-Tompkins' and a custom algorithm, have opposite behaviors, which shows the different nature of both acquisitions.

The database with more supervised and guided collections, resulted in accuracy up to 78.243%, and EER up to 13.530%, despite of having a greater number of enrolled users. However, the less strict and user-based collection was not even considered for identification due to its poor results, and summarized in an EER ranging from 0.068%–31.669% under exercise. These smartband verification results were the result of improving them with an extra SWT, and considering MLP classifier. with a two-days enrollment. Even considering the similarities with the initial database,there is a clear correlation between the device quality, collection quality and data quantity.

In conclusion, the potential of ECG as a biometric signal has been deeply assessed in this thesis, while avoiding over complicated transformations that could affect the extrapolation to a portable device. We have observed that the heart rate increase is a drawback in this modality but it can be overcome while contemplating alternatives and including extra enrollments. In addition, the expertise of the user in recognition can be key when using a portable device, due to its lower fidelity. Even though this field has more challenges, the research community would probably success in this task.

11.2. Future work

The comparison among works related to ECG biometrics is complicated due to the lack of standardization of this modality. One of the most relevant works that could be achieved in the future should be directed towards this problem, in order to provide some guidelines and feasibility. These requirements should treat issues related to data collection, signal quality and performance assessment.

The lack of public databases that consider different scenarios, users and conditions is also a big initial barrier when researching in this field. Successful experiments in public databases cannot be extrapolated to real case scenarios, and those applied to private databases cannot be evaluated by the research community. The publication of new databases that consider realistic biometric conditions is also a big pending task to improve future works.

The creation of new databases is also limited by the lack of commercial, functional devices that allow to obtain and manipulate raw data. The problem of using prototypes or commercial but non-portable devices always makes difficult the task of trying to replicate

real environments. The data that is collected by prototypes cannot be ensured to have constant quality, and the results in classification are always depending on that. The development of smartband or general portable devices is key for successful research.

Finally, given that a healthy ECG pattern is already a challenge in human recognition, there are not researches that focus on proceeding with data from users with cardiovascular diseases. This fact avoids the research to be universal, and it is heavily related to the lack of public databases and an issue to consider in standardization.

BIBLIOGRAPHY

[1] P. Tirado-Martin and R. Sanchez-Reillo, "BioECG: Improving ECG Biometrics with Deep Learning and Enhanced Datasets," *Applied Sciences*, vol. 11, no. 13, 2021. DOI: 10.3390/app11135880. [Online]. Available: https://www.mdpi.com/2076-3417/11/13/5880.

[2] P. Tirado-Martin, J. Liu-Jimenez, J. Sanchez-Casanova, and R. Sanchez-Reillo, "QRS Differentiation to Improve ECG Biometrics under Different Physical Scenarios Using Multilayer Perceptron," *Applied Sciences*, vol. 10, no. 19, 2020. DOI: 10.3390/app10196896. [Online]. Available: https://www.mdpi.com/2076-3417/10/19/6896.

[3] P. Tirado-Martin, R. Sanchez-Reillo, and J. Park, "Effects of Data Reduction when using Gaussian Mixture Models in Unidimensional Biometric Signals," in *2018 International Carnahan Conference on Security Technology (ICCST)*, IEEE, 2018, pp. 1–5. DOI: 10.1109/CCST.2018.8585514.

[4] P. Tirado-Martin, R. Blanco-Gonzalo, A. Alvarez-Nieto, and A. Romero-Diaz, "Image processing techniques for improving vascular hand biometrics," in *2017 International Carnahan Conference on Security Technology (ICCST)*, 2017, pp. 1–5. DOI: 10.1109/CCST.2017.8167807.

[5] J. Sanchez-Casanova, J. Liu-Jimenez, P. Tirado-Martin, and R. Sanchez-Reillo, "Unsupervised and scalable low train pathology detection system based on neural networks," *Heliyon*, vol. 7, no. 2, e06270, 2021.

[6] P. Fernandez-Lopez, J. Sanchez-Casanova, P. Tirado-Martin, and J. Liu-Jimenez, "Optimizing resources on smartphone gait recognition," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 31–36. DOI: 10.1109/BTAS.2017.8272679.

[7] A. a. R. Jain and S. A. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004. DOI: 10.1109/TCSVT.2003.818349.

[8] ISO/IEC JTC 1/SC 37, *Information technology — Biometric performance testing and reporting — Part 1: Principles and framework (Standard No. 19795)*. 2006, p. 56. [Online]. Available: https://www.iso.org/standard/41447.html?browse=tc.

[9] The University of Nottingham, *Beginners Guide to Normal Heart Function, Sinus Rhythm and Common Cardiac Arrhythmias*, [Accessed 9-9-2021]. [Online]. Available: https://www.nottingham.ac.uk/nursing/practice/resources/cardiology/function/normal_duration.php.

[10] P. W. Foley *et al.*, "Cardiac resynchronisation therapy in patients with heart failure and a normal QRS duration: the RESPOND study," *Heart*, vol. 97, no. 13, pp. 1041–1047, 2011.

[11] W. Einthoven, "The telecardiogram," *American Heart Journal*, vol. 53, no. 4, pp. 602–615, 1957. DOI: https://doi.org/10.1016/0002-8703(57)90367-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0002870357903678.

[12] A. L. Goldberger, *Chapter 3 - ECG Leads*, Seventh Edition, A. L. Goldberger, Ed. Philadelphia: Mosby, 2006, pp. 21–32. DOI: https://doi.org/10.1016/B0-323-04038-1/50004-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B0323040381500044.

[13] Nicholas Patchett, *Spatial orientation of EKG leads. License CC BY-SA 4.0*, [Accessed 30-June-2021], 2015. [Online]. Available: https://commons.wikimedia.org/wiki/File:EKG_leads.png.

[14] S. Maheshwari, A. Acharyya, M. Schiariti, and P. E. Puddu, "Frank vectorcardiographic system from standard 12 lead ECG: An effort to enhance cardiovascular diagnosis," *Journal of Electrocardiology*, vol. 49, no. 2, pp. 231–242, 2016. DOI: https://doi.org/10.1016/j.jelectrocard.2015.12.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002207361500429X.

[15] B. J. Drew *et al.*, "Practice Standards for Electrocardiographic Monitoring in Hospital Settings," *Circulation*, vol. 110, no. 17, pp. 2721–2746, 2004. DOI: 10.1161/01.CIR.0000145144.56673.59. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.CIR.0000145144.56673.59.

[16] A. Rashkovska, M. Depolli, I. Tomašić, V. Avbelj, and R. Trobec, "Medical-Grade ECG Sensor for Long-Term Monitoring," *Sensors*, vol. 20, no. 6, 2020. DOI: 10.3390/s20061695. [Online]. Available: https://www.mdpi.com/1424-8220/20/6/1695.

[17] M. Bocchiardo and R. Asteggiano, "ECG portable devices: example of e-Health strength and threats," *E-Journal of Cardiology Practice*, vol. 18, p. 25, 2020.

[18] Wikimedia Creators, *A 5-electrode Holter. License CC BY-SA 3.0*, [Accessed 30-June-2021], 2010. [Online]. Available: https://commons.wikimedia.org/wiki/File:EKG_leads.png.

[19] iRhythm, *Zio Monitors*, [Accessed 08-June-2021]. [Online]. Available: https://www.irhythmtech.com/providers/zio-service/zio-monitors.

[20] Imec, *Biomedical sensor systems-on-chip*, [Accessed 08-June-2021]. [Online]. Available: https://www.imec-int.com/en/system-on-chip.

[21] AliveCor, *AliveCor's Kardiamobile*, [Accessed on 08-06-2021. [Online]. Available: https://www.alivecor.es/kardiamobile.

[22] U. Satija, B. Ramkumar, and M. S. Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE reviews in biomedical engineering*, vol. 11, pp. 36–52, 2018.

[23] M. D'Aloia, A. Longo, and M. Rizzi, "Noisy ECG Signal Analysis for Automatic Peak Detection," *Information*, vol. 10, no. 2, 2019. DOI: 10.3390/info10020035.

[24] R. Kher, "Signal processing techniques for removing noise from ECG signals," *J. Biomed. Eng. Res*, vol. 3, pp. 1–9, 2019.

[25] M. Kania, H. Rix, M. Fereniec, D. Janusek, and R. Maniewski, "The effect of precordial lead displacement on P-wave morphology in body surface potential mapping," in *Computing in Cardiology 2013*, 2013, pp. 531–534.

[26] 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, 2001. DOI: 10.1109/19.930458.

[27] H. K. Wolf and P. W. MacFarlane, "Optimization of computer ECG processing," *Journal of Clinical Engineering*, vol. 5, no. 3, p. 264, 1980.

[28] G. Wübbeler, M. Stavridis, D. Kreiseler, R.-D. Bousseljot, and C. Elster, "Verification of humans using the electrocardiogram," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1172–1175, 2007.

[29] S. Pouryayevali, S. Wahabi, S. Hari, and D. Hatzinakos, "On establishing evaluation standards for ECG biometrics," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3774–3778. DOI: 10.1109/ICASSP.2014.6854307.

[30] C.-K. Chen, C.-L. Lin, S.-L. Lin, Y.-M. Chiu, and C.-T. Chiang, "A chaotic theoretical approach to ECG-based identity recognition [application notes]," *IEEE Computational Intelligence Magazine*, vol. 9, no. 1, pp. 53–63, 2014.

[31] J. Sulam, Y. Romano, and R. Talmon, "Dynamical system classification with diffusion embedding for ECG-based person identification," *Signal Processing*, vol. 130, pp. 403–411, 2017.

[32] W. Louis, S. Abdulnour, S. J. Haghighi, and D. Hatzinakos, "On biometric systems: electrocardiogram Gaussianity and data synthesis," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, no. 1, pp. 1–10, 2017.

[33] S. A. Fattah *et al.*, "An approach for human identification based on time and frequency domain features extracted from ECG signals," in *TENCON 2011-2011 IEEE Region 10 Conference*, IEEE, 2011, pp. 259–263.

[34] R. Hoekema, G. J. 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, 2001.

[35] K. N. Plataniotis, D. Hatzinakos, and J. K. Lee, "ECG biometric recognition without fiducial detection," in *2006 Biometrics symposium: Special session on research at the biometric consortium conference*, IEEE, 2006, pp. 1–6.

[36] J. S. Arteaga-Falconi, H. Al Osman, and A. El Saddik, "ECG authentication for mobile devices," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 591–600, 2015.

[37] C. Ye, M. T. Coimbra, and B. V. Kumar, "Investigation of human identification using two-lead electrocardiogram (ECG) signals," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, 2010, pp. 1–8.

[38] S. Šprager, R. Trobec, and M. B. Jurič, "Feasibility of biometric authentication using wearable ECG body sensor based on higher-order statistics," in *2017 40th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, 2017, pp. 264–269.

[39] R. Tan and M. Perkowski, "Toward improving electrocardiogram (ECG) biometric verification using mobile sensors: A two-stage classifier approach," *Sensors*, vol. 17, no. 2, p. 410, 2017.

[40] BIOPAC Systems, Inc., *MP System comparison: MP150 vs. MP100*, [Accessed 11-November-2021]. [Online]. Available: https://www.biopac.com/knowledge-base/mp-system-comparison-mp150-vs-mp100/.

[41] ——, *Remote Monitoring*, [Accessed 11-November-2021]. [Online]. Available: https://https://www.biopac.com/application/remote-monitoring/.

[42] Vernier, *EKG Sensor*, [Accessed 11-November-2021]. [Online]. Available: https://www.vernier.com/product/ekg-sensor/.

[43] Savvy, *Savvy ECG is all-in-one medical device*, [Accessed 11-November-2021]. [Online]. Available: http://www.savvy.si/en/.

[44] Fraunhofer IIS, *FitnessSHIRT*, [Accessed 11-November-2021]. [Online]. Available: https://www.iis.fraunhofer.de/en/ff/sse/health/medical-sensors-and-analytics/prod/fitnessshirt.html/.

[45] Cardiac Designs, *ECG Check*, [Accessed 11-November-2021]. [Online]. Available: https://www.cardiacdesigns.com/.

[46] Nymi, *Nymi's Homepage*, [Accessed 29-June-2021]. [Online]. Available: https://www.nymi.com.

[47] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, e215–e220, 2000.

[48] S. Pouryayevali, S. Wahabi, S. Hari, and D. Hatzinakos, "On establishing evaluation standards for ECG biometrics," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 3774–3778.

[49] T. S. Lugovaya, "Biometric human identification based on electrocardiogram," *Master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University 'LETI', Saint-Petersburg, Russian Federation*, 2005.

[50] A. Taddei *et al.*, "The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *European heart journal*, vol. 13, no. 9, pp. 1164–1172, 1992.

[51] F. Jager *et al.*, "Long-term ST database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia," *Medical and Biological Engineering and Computing*, vol. 41, no. 2, pp. 172–182, 2003.

[52] 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, 2001.

[53] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," 1995.

[54] S. A. Israel, W. T. Scruggs, W. J. Worek, and J. M. Irvine, "Fusing face and ECG for personal identification," in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.*, IEEE, 2003, pp. 226–231.

[55] 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.

[56] K. A. Sidek, V. Mai, and I. Khalil, "Data mining in mobile ECG based biometric identification," *Journal of Network and Computer Applications*, vol. 44, pp. 83–91, 2014.

[57] H. Chen, F. Zeng, K.-K. Tseng, H.-N. Huang, S.-Y. Tu, and J.-S. Panl, "ECG human identification with statistical support vector machines," in *2012 International Conference on Computing, Measurement, Control and Sensor Network*, IEEE, 2012, pp. 237–240.

[58] S. Z. Fatemian and D. Hatzinakos, "A new ECG feature extractor for biometric recognition," in *2009 16th international conference on digital signal processing*, IEEE, 2009, pp. 1–6.

[59] Y. Wang, K. N. Plataniotis, and D. Hatzinakos, "Integrating analytic and appearance attributes for human identification from ECG signals," in *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, IEEE, 2006, pp. 1–6.

[60] C. C. Poon, Y.-T. Zhang, and S.-D. Bao, "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 73–81, 2006.

[61] T.-W. D. Shen, W. J. Tompkins, and Y. H. Hu, "Implementation of a one-lead ECG human identification system on a normal population," *Journal of Engineering and Computer Innovations*, vol. 2, no. 1, pp. 12–21, 2010.

[62] L. Mesin, A. Munera, and E. Pasero, "A low cost ECG biometry system based on an ensemble of support vector machine classifiers," in *International Workshop on Neural Networks*, Springer, 2015, pp. 425–433.

[63] F. Agrafioti *et al.*, *ECG in biometric recognition: Time dependency and application challenges*. University of Toronto, 2011.

[64] F. Porée, A. Gallix, and G. Carrault, "Biometric identification of individuals based on the ECG. Which conditions?" In *2011 Computing in Cardiology*, IEEE, 2011, pp. 761–764.

[65] K. A. Sidek, I. Khalil, and M. Smolen, "ECG biometric recognition in different physiological conditions using robust normalized QRS complexes," in *2012 Computing in Cardiology*, IEEE, 2012, pp. 97–100.

[66] D. Tantinger *et al.*, "Human authentication implemented for mobile applications based on ECG-data acquired from sensorized garments," in *2015 Computing in Cardiology Conference (CinC)*, IEEE, 2015, pp. 417–420.

[67] M. Komeili, N. Armanfard, D. Hatzinakos, and A. Venetsanopoulos, "Feature selection from multisession electrocardiogram signals for identity verification," in *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, IEEE, 2015, pp. 603–608.

[68] Z. Zhang and D. Wei, "A new ECG identification method using Bayes' teorem," in *TENCON 2006-2006 IEEE Region 10 Conference*, IEEE, 2006, pp. 1–4.

[69] J. M. Irvine and S. A. Israel, "A sequential procedure for individual identity verification using ECG," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–13, 2009.

[70] Y. Wang, F. Agrafioti, D. Hatzinakos, and K. N. Plataniotis, "Analysis of human electrocardiogram for biometric recognition," *EURASIP journal on Advances in Signal Processing*, vol. 2008, pp. 1–11, 2007.

[71] T.-W. Shen, W. Tompkins, and Y. 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*, IEEE, vol. 1, 2002, pp. 62–63.

[72] T. Choudhary and M. S. Manikandan, "A novel unified framework for noise-robust ECG-based biometric authentication," in *2015 2nd international conference on signal processing and integrated networks (SPIN)*, IEEE, 2015, pp. 186–191.

[73] M. Derawi, I. Voitenko, and P. E. Endrerud, "Real-time wireless ECG biometrics with mobile devices," in *2014 International Conference on Medical Biometrics*, IEEE, 2014, pp. 151–156.

[74] N. Belgacem, A. Nait-Ali, R. Fournier, and F. Bereksi-Reguig, "ECG based human authentication using wavelets and random forests," *International Journal on Cryptography and Information Security (IJCIS)*, vol. 2, no. 2, pp. 1–11, 2012.

[75] "Signal processing techniques for removing noise from ECG signals,"

[76] C. T. Arsene, R. Hankins, and H. Yin, "Deep learning models for denoising ECG signals," in *2019 27th European Signal Processing Conference (EUSIPCO)*, IEEE, 2019, pp. 1–5.

[77] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[78] C.-C. Chiu, C.-M. Chuang, and C.-Y. Hsu, "A novel personal identity verification approach using a discrete wavelet transform of the ECG signal," in *2008 International Conference on Multimedia and Ubiquitous Engineering (mue 2008)*, IEEE, 2008, pp. 201–206.

[79] A. D. Chan, M. M. Hamdy, A. Badre, and V. Badee, "Wavelet distance measure for person identification using electrocardiograms," *IEEE transactions on instrumentation and measurement*, vol. 57, no. 2, pp. 248–253, 2008.

[80] F. G. S. Teodoro, S. M. Peres, and C. A. Lima, "Feature selection for biometric recognition based on electrocardiogram signals," in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, pp. 2911–2920.

[81] "On evaluating human recognition using electrocardiogram signals: From rest to exercise."

[82] H.-S. Choi, B. Lee, and S. Yoon, "Biometric authentication using noisy electrocardiograms acquired by mobile sensors," *IEEE Access*, vol. 4, pp. 1266–1273, 2016.

[83] V. Mai, I. Khalil, and C. Meli, "ECG biometric using multilayer perceptron and radial basis function neural networks," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, pp. 2745–2748.

[84] 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.

[85] 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.

[86] Q. Zhang, D. Zhou, and X. Zeng, "PulsePrint: Single-arm-ECG biometric human identification using deep learning," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, IEEE, 2017, pp. 452–456.

[87] R. Salloum and C.-C. J. Kuo, "ECG-based biometrics using recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 2062–2066.

[88] M. N. Dar, M. U. Akram, A. Usman, and S. A. Khan, "ECG biometric identification for general population using multiresolution analysis of DWT based features," in *2015 Second International Conference on Information Security and Cyber Forensics (InfoSec)*, IEEE, 2015, pp. 5–10.

[89] E. J. da Silva Luz, G. J. Moreira, L. S. Oliveira, W. R. Schwartz, and D. Menotti, "Learning deep off-the-person heart biometrics representations," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1258–1270, 2017.

[90] J. Kim, D. Sung, M. Koh, J. Kim, and K. S. Park, "Electrocardiogram authentication method robust to dynamic morphological conditions," *IET Biometrics*, vol. 8, no. 6, pp. 401–410, 2019.

[91] Sinikka Halme, *Electrode-skintact. License CC BY-SA 4.0*, [Accessed 30-June-2021], 2020. [Online]. Available: https://commons.wikimedia.org/wiki/File:Electrode-skintact.jpg.

[92] J. Padgette, K. Scarfone, and L. Chen, "Guide to bluetooth security," *NIST Special Publication*, vol. 800, no. 121, p. 25, 2012.

[93] Adafruit, *Adafruit Mini Lipo w/Mini-B USB Jack - USB LiIon/LiPoly charge*, [Accessed 30-June-2021]. [Online]. Available: https://www.adafruit.com/product/1905.

[94] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[95] Y. Zhang, Z. Dong, L. Wu, S. Wang, and Z. Zhou, "Feature Extraction of Brain MRI by Stationary Wavelet Transform," in *2010 International Conference on Biomedical Engineering and Computer Science*, 2010, pp. 1–4. DOI: 10.1109/ICBECS.2010.5462491.

[96] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4202–4210.

[97] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.

[98] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[99] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research (JMLR)*, vol. 12, no. 85, pp. 2825–2830, 2011. DOI: 10.1145/2786984.2786995. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[100] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107 398, 2021. DOI: https://doi.org/10.1016/j.ymssp.2020.107398. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020307846.

[101] A. Miranda-Escalada, "Analysis on the viability of using bio-signals for the recognition of human beings," Unpublished Bachelor's Thesis, 2018.

[102] Ian Nabney, MATLAB Central File Exchange, *Netlab*, [Accessed 12-9-2021]. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/2654-netlab.

[103] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "Security and accuracy of fingerprint-based biometrics: A review," *Symmetry*, vol. 11, no. 2, p. 141, 2019.

[104] Ian Nabney, MATLAB Central File Exchange, *Fingerprint Recognition*, [Accessed 21-9-2021]. [Online]. Available: https://www.innovatrics.com/biometrics-for-oem-solutions/fingerprint-recognition/.

[105] G. Panchal, A. Ganatra, Y. P. Kosta, and D. Panchal, "Behaviour Analysis of Multilayer Perceptrons with Multiple Hidden Neurons and Hidden Layers," *International Journal of Computer Theory and Engineering (IJCTE)*, vol. 3, no. 2, pp. 332–337, 2011. DOI: 10.7763/ijcte.2011.v3.328.

[106] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015, pp. 1–15.

[107] F. Chollet *et al.* "Keras." (2015), [Online]. Available: https://github.com/fchollet/keras.

[108] I. Chamatidis, A. Katsika, and G. Spathoulas, "Using deep learning neural networks for ECG based authentication," in *2017 International Carnahan Conference on Security Technology (ICCST)*, 2017, pp. 1–6. DOI: 10.1109/CCST.2017.8167816.

[109] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[110] N. Pato-Montemayor, "Deep learning application in Electrocardiogram," Unpublished Bachelor's Thesis, 2021.

[111] J. M. Shrein, "Fingerprint classification using convolutional neural networks and ridge orientation images," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2017, pp. 1–8.

[112] X. Sun, L. Liu, C. Li, J. Yin, J. Zhao, and W. Si, "Classification for remote sensing data with improved CNN-SVM method," *IEEE Access*, vol. 7, pp. 164 507–164 516, 2019.

[113] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[114] *Autonomio Talos [ComputerSoftware], 2020.* [Accessed 15-November-2021]. [Online]. Available: http://github.com/autonomio/talos/.

[115] S. Mekruksavanich and A. Jitpattanakul, "Biometric User Identification Based on Human Activity Recognition Using Wearable Sensors: An Experiment Using Deep Learning Models," *Electronics*, vol. 10, no. 3, 2021. DOI: 10.3390/electronics10030308. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/308.

[116] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, PMLR, 2015, pp. 448–456.