A Key to Your Heart: Biometric Authentication Based on ECG Signals

Nikita Samarin

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2018

Abstract

Over the recent years, biometric authentication has become ubiquitous. We make use of biometric authentication when we unlock our smartphones using fingerprint scanners, access bank services using voice recognition and enter other countries using automated border control gates with facial recognition technology.

In this project, I investigated a biometric based on the electrical activity of the human heart as recorded by the electrocardiogram (ECG) signals. In order to design an ECG-based biometric system, I started by creating a dataset containing records of 55 users. The data collection was performed over a period of four months, which allowed me to investigate the stability of ECG-based biometrics. After collecting the ECG data, I designed and fully implemented the biometric authentication system, which included signal processing, feature extraction and ECG trace classification. The obtained biometric system was evaluated using two different approaches, and compared with the results from existing studies.

The results from this work support the claim that ECG-based biometrics can be successfully used for personal authentication. Nevertheless, more research needs to be done in order to improve the long-term performance of ECG-based biometric systems.

Acknowledgements

Most of all, I would like to thank my supervisor, Don Sannella, for his guidance and support throughout the work on this project. I appreciated his responsiveness and great attention to details, especially during the writing of this report.

I would also like to thank everyone who took their time to help me with this project by taking part in the data collection experiment; this project would not be possible if not for your help.

Table of Contents

Chapter 1

Introduction

In today's world, we observe an ever-increasing digitization of most areas of our lives. Every day we make use of online services such as mobile banking or email communication and we do not hesitate to keep our personal information on our devices or cloud storage. Unfortunately, the digital era has also paved the way for a series of new attacks and exploits, including unauthorized access to our personal data and devices by adversaries. It is remarkable how, by and large, users are still burdened by numerous passwords, which have been used for access control since the earliest days of computing.

There has been a recent shift of interest towards the field of biometric authentication, which proves the identity of the user using their biological characteristics. The most common one involves using a fingerprint scanner, such as seen in modern smartphones and laptops. While this is a big step forward, there are still problems related to fingerprint usability and reliability. In this project, I investigate a biometric based on the electrical activity of the human heart in the form of electrocardiogram (ECG) signals. Past research has shown that ECG is unique to each individual [8] and thus could potentially be used for authentication. Nevertheless, current challenges include extracting relevant features from ECG signals, designing accurate models for pattern recognition, proving long-term stability of the biometric and protecting against presentation attacks.

1.1 Contributions

This work addresses some of the limitations of existing research related to the use of ECG signals for biometric authentication. I started by performing an extensive literature review on this area. While there are papers describing how to create a personal authentication system using ECG signals, most of them rely on existing datasets and only few investigate stability and usability of ECG as a biometric.

In contrast, data collection was a central component of this research for two main reasons. First, it removed the dependency on public ECG datasets, which are usually designed for medical analysis and, therefore, contain more features and less noise than

data collected by sensors used in actual biometric authentication systems. In particular, ECG signals for this study were recorded using a consumer-grade ECG monitor, which is more affordable and less intrusive than medical-grade devices. Second, data was collected from the same participants over two sessions separated in time by four months, which allowed me to examine not only uniqueness, but also stability of human ECG.

Another challenge of using ECG data for biometric authentication is the continuous nature of the signal. It becomes important, therefore, to extract distinct features from the signal that could be used to match users with their records. Most existing work focuses on locating reference points within the ECG trace to are used to create the input features. While these features provide good authentication performance, correctly identifying the reference points is a non-trivial task in itself and may constitute a single point of failure if performed incorrectly. As a compromise, the approach that I used in this project relies only on locating the R-wave peaks, which is the most prominent feature of the trace. These peaks were used to segment the continuous signal into smaller heartbeat waveforms, which formed the basis for the biometric templates.

In order to perform authentication, the system has to match the stored template of the user with biometrics recorded during the operation of the system. For this research, I developed several classifiers to perform template matching, including logistic regression, k-nearest neighbours and support vector machines. The hyperparameters of these models were optimised using a grid search performed with k-fold validation. This approach allowed me to compare the influence of different parameters and make a decision based on the overall best performance.

Finally, many existing studies do not provide a comprehensive evaluation of their proposed biometric system. For this project, I decided to focus on the issues that occur when evaluating the performance of ECG-based biometrics. This include improper training set selection and lack of explicit attacker modelling. I performed two evaluations of my proposed system, one of which is a standard approach found in existing literature and another one, which provides a more in-depth analysis of the mistakes made by the system.

1.2 Report Organization

Chapter 2 starts by presenting background information about biometrics and their use in authentication systems. It also discusses the nature of ECG signals and briefly overviews existing research in ECG-based authentication. Chapter 3 describes the data collection procedure, experimental setup and participant demographics. Afterwards, Chapter 4 introduces authentication system design and presents techniques for feature extraction and template matching. Chapter 5 presents the best practices for biometric system evaluation and uses two approaches to assess the performance of the proposed system. The report concludes with a summary and future work directions in Chapter 6.

Chapter 2

Background

In this chapter, I start by presenting a brief overview of biometrics, their characteristics and applications. I also highlight the difference between authentication and identification. This is followed by an introduction to the physiology behind an electrocardiogram and how it presents a suitable candidate for a biometric modality. I conclude this chapter with a literature review and a presentation of existing results from past work.

2.1 Biometrics

The term 'biometrics' is used to describe measurable and distinctive characteristics that can be used to perform recognition of individuals. These characteristics are often divided into two categories: physiological and behavioural [41]. Physiological biometrics relate to human physiology; these include fingerprints, facial features, iris patterns or DNA. Behavioural biometrics are based on human behaviour, such as keystroke dynamics, voice or gait.

Biometrics are becoming increasingly used in access control and user authentication. Most existing security applications require using something that you know (e.g. passwords) or something that you have (e.g. secure tokens). With biometrics, it is possible to use something that you *are*, which improves system usability, as users are no longer required to remember any secrets or always carry a physical token. Access control applications can also combine multiple modalities, which improves security even further.

In order for a biometric to be applicable for access control, it must have the following characteristics [41]:

- **Universality.** The biometric is present and measurable in every person.
- **Uniqueness.** The biometric is sufficiently different from person to person.
- **Stability.** The biometric properties remain invariant over the person's lifetime.

Additionally, another four characteristics should be considered:

- **Collectability.** The biometric can be easily measured within a person.

- **Performance.** The biometric is robust, reliable and easily analysed when used for personal identification.
- **Acceptability.** Biometric collection and usage is socially accepted.
- **Circumvention.** The biometric cannot be easily imitated and "spoofed" by a substitute.

Biometrics vary considerably among these dimensions, and deciding which one to use depends on specific problem at hand. Table 2.1 presents a comparison of commonly used biometrics.

Biometrics	Universality	Uniqueness	Stability	Collectability	Performance	Acceptability	Circumvention
Fingerprint	Medium	High	High	Medium	High	Medium	Low
Face	High	Low	Medium	High	Low	High	High
Keystrokes	Low	Low	Low	Medium	Low	Medium	Medium
Iris	High	High	High	Medium	High	Low	Low
Retinal scan	High	High	Medium	Low	High	Low	Low
Signature	Low	Low	Low	High	Low	High	High
Voice	Medium	Low	Low	Medium	Low	High	High
Odor	High	High	High	Low	Low	Medium	High
DNA	High	High	High	Low	High	Low	High
Gait	Medium	Low	Low	High	Low	High	Medium

Table 2.1: A comparison of several biometric modalities. Table adapted from [41].

While biometrics offer many advantages over traditional means of access control, such as passwords and secure tokens, there are still serious concerns over security and privacy of stored biometrics. Once compromised, biometrics cannot easily be changed, as they depend on physiological or behavioural characteristics of an individual. Even if the adversary fails to extract sensitive information from the security application itself, there are other means of collecting the required biometrics to launch an attack. For instance, facial features can be easily collected from photos available on social media websites, and voice can be obtained by recording phone calls. It can be beneficial therefore to assume a strong attacker, in which the attacker can imitate or collect samples of the victim's biometrics. The goal of the security application then is to implement an efficient presentation attack detection algorithm, which would mitigate the consequences of compromised biometrics. The most common presentation attack detection consists of liveness detection, which aims to detect whether the biometric is presented by a living individual.

2.2 Authentication and Identification

Biometrics can be used to achieve two important access control goals, user authentication and identification. Biometric *authentication* involves the user presenting an identity claim and a biometric sample. The system then decides whether this claim is

valid based on the recorded biometric for this identity. For instance, presenting your passport at the border can be seen as user authentication, where you claim the identity of the person to whom the passport belongs. In contrast, user *identification* involves finding the closest match to presented biometrics among the stored records. In this case, there are no identity claims. For instance, identifying wanted criminals from CCTV footage is an example of user identification. Identification is further divided into closed-set and open-set. In the former case, it is assumed that there is a biometric sample of the user that is already stored in the system. Otherwise, the problem is considered open-set.

Biometric authentication or identification is often performed by *template matching*. More specifically, a *biometric template* is a recorded instance of the biometric characteristic collected during the *enrolment* of the subject in the system. This template is stored in a *gallery* of templates. A *query* consists of recording a sample of the biometric during the operation of the system. This query is then matched against templates specific only to the claimed identity (in authentication) or against all templates in the gallery (in identification) [38].

Throughout this report, I will focus my attention on biometric authentication, as opposed to identification. Nevertheless, the ideas presented in this report can be also readily applied when designing an identification system.

2.3 Physiology of the Heart

Before discussing the utility of an electrocardiogram as a biometric, this report presents an overview of the electrical conduction system of the heart and how it relates to electrocardiograms.

The heart is the muscle that pumps blood filled with oxygen and nutrients through the blood vessels to the body tissues [43]. The heart contains four chambers: the upper two chambers (left and right atria) are entry-points into the heart, while the lower two chambers (left and right ventricles) are contraction chambers sending blood through the circulation. The cardiac cycle refers to a complete heartbeat from its generation to the beginning of the next beat, comprising several stages of filling and emptying of the chambers. The frequency of the cardiac cycle is known as the *heart rate* (measured in beats per minute, bpm) [20].

In order to pump blood, the heart muscle must contract, which requires an electrical impulse. This impulse comes from the sinus node (located in the right atrium), which is transmitted via specific pathways throughout the heart, enabling regular contraction and relaxation [7]. The electrical impulse generated by the heart can be detected on the surface of the body using electrodes placed on the skin, which is done during an electrocardiogram (ECG or EKG) test. An ECG trace captures the process of depolarisation and repolarisation of the heart chambers, which causes them to contract and relax. The connection between an ECG and the electrical activity of the heart can be seen from Figure 2.1.

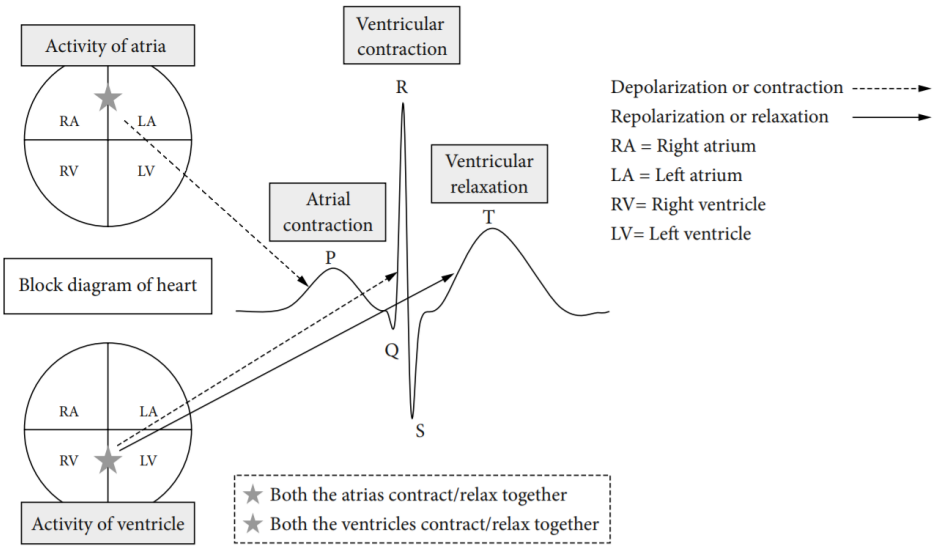


Figure 2.1: Generation of an ECG trace from electrical activity of the heart. Image adapted from [42].

ECG monitors are used to record the electrical activity of the heart using pairs of electrodes placed on the skin. Each pair of electrodes is known as a *lead* and provides an electrical view of the heart from a different angle. There are 12 leads that are used in cardiology, obtained from a combination of 10 electrodes. Different ECG monitors are distinguished by the number of leads that they can record [3].

An ECG trace (for some specific lead) for a single cardiac cycle consists of several parts [3]:

- **PR interval.** The time between the beginning of the P wave and the beginning of the Q wave.
- **P wave.** Corresponds to atrial depolarisation.
- **PR Segment.** The time between the end of the P wave and the beginning of the Q wave.
- **QRS complex.** Corresponds to ventricular depolarisation.
- **ST segment.** The time between the end of the S wave and at the beginning of the T wave.
- **T wave.** Corresponds to ventricular repolarisation.
- **QT interval.** The time between the beginning of the QRS complex and the end of the T wave.

A visualisation of an ECG for a single cardiac cycle is presented in Figure 2.2.

Additionally, we can also measure the RR interval, which starts at the peak of one R wave and ends at the peak of the next R wave. RR intervals can be used to compute the heart rate from a recorded ECG signal.

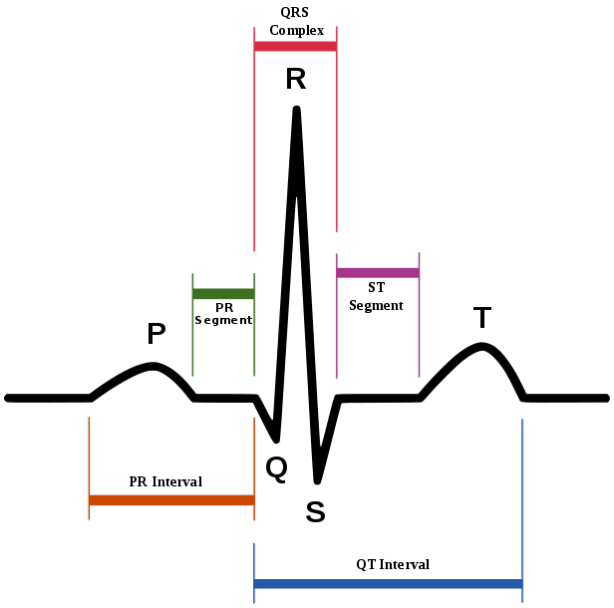


Figure 2.2: Single heartbeat waveform. Image taken from [10].

2.4 Electrocardiogram as a Biometric

Having discussed the physiology behind an electrocardiogram, we can now consider whether it could be used as a viable biometric. Recall that a biometric is applicable for access control if it is *universal*, *unique* and *stable*. Clearly, ECG is universal, as it is conditional on the electrical activity of the heart, which occurs in every living individual. Most existing work, therefore, focuses on establishing uniqueness and stability of ECG.

2.4.1 Uniqueness

Some authors claim that the composition and activity of the human heart is unique, as it inherits uniqueness from the individuality of DNA [20]. The argument that they provide is that by the "central dogma" of molecular biology, genetic information flows from the DNA to RNA (ribonucleic acid) to proteins, which are responsible for the structure and regulation of internal organs, including the heart. Using this argument, we can conclude that ECG of each individual is caused by a unique set of factors. However, the inverse, that each ECG is unique because it is produced by unique set of factors, does not necessarily follow. Therefore, uniqueness of ECG for practical applications needs to be established with empirical evidence.

Most works that explore ECG for personal identification do not assess the performance of their ECG authentication systems on very large datasets, as was done for other biometric modalities. A notable exception is a study by Carreiras et al., which focuses on the uniqueness of ECG signals [8]. The authors of the paper evaluated the performance of their biometric system on a database of ECG recordings collected from 618 subjects using a 12-lead ECG and obtained high recognition rates. The results from this work

provide a positive outlook on the issue of ECG uniqueness.

2.4.2 Stability

Considerably fewer studies investigate the stability of ECG signals. While proving uniqueness can be achieved using data from a single point of time, proving stability requires data to be collected from the same individual over a sufficiently long period of time. Creating large databases of such longitudinal data is expensive and involves significant time investment, which explains the small number of studies that examine ECG stability. A study by Silva et al. collected ECG data from 63 subjects, with two data acquisition sessions separated by a 4-month interval [12]. Their results indicate that biometric authentication performs worse for longitudinal ECG data, but is still viable for real-world applications.

Additionally, we would like to consider *collectability*, *performance*, *acceptability* and *circumvention* characteristics of ECG signals. A discussion of these four characteristics is presented below.

2.4.3 Collectability

Traditional 12-lead ECG machines require 10 self-adhesive electrodes to be placed on the subject chest and limbs. While such machines do not require any effort from the subject to perform an ECG recording, they are often stationary, expensive and take time to set up. Recording ECG using medical-grade monitors is also invasive, requiring the subject to expose their chest and limbs. With the rise of personalised healthcare, however, consumer-oriented ECG monitors are becoming more widespread. These monitors are portable and can be used to record a single-lead ECG trace using electrodes that make contact with the wrists (e.g. smartwatch bands) or fingers (e.g. sensors installed on surface). While consumer-grade ECG monitors provide less data than 12-lead ECG machines, they can be used to record ECG in a non-invasive manner, applicable for biometric systems.

2.4.4 Performance

The performance of a biometric system also depends on the quality of signal preprocessing and feature extraction. Some biometrics have established methods for transforming the raw signal to features that are used for recognition of individuals. Fingerprint scanners, for instance, detect very specific fingerprint features called minutiae, which are used to establish the similarity between a biometric template in the system and a query [45]. ECG as a biometric is much less researched and, thus, there is less consensus over which features should be used. Furthermore, some existing ECG preprocessing techniques are computationally expensive to perform, which might prevent the deployment of ECG-based biometric systems on a large scale.

2.4.5 Acceptability

With the introduction of reliable consumer-grade ECG sensors, there has been more opportunities to create ECG-based biometric systems that are non-invasive and socially accepted. Particularly appealing are "off-the-person" approaches for signal acquisition, in which biometric sensors are embedded into existing systems, such as keyboards, ATM panels and vehicle steering wheels [8]. Nevertheless, there have been no known studies that investigate the attitude of users towards using ECG as a biometric. As with other biometric modalities, using ECG data for personal identification poses considerable security and privacy concerns. For instance, compromised ECG signal can be used to learn about certain health conditions of enrolled users. Therefore, secure storage and usage of biometrics is required to ensure that the biometric system is trusted by its users.

2.4.6 Circumvention

All biometric systems are subject to presentation attacks, which attempt to subvert the system with an artefact or contraption. Nevertheless, biometrics differ in the amount of time and resources that is required to design a suitable artefact. In order to compromise an ECG recording, the attacker has to steal the records from a medical institution or perform a social engineering attack to manipulate the victim into giving their ECG. Once that is achieved, the adversary has to digitalise the recording (if it is on paper) and fake the voltage levels at the electrodes of an ECG sensor using a device that outputs electrical waveforms (e.g. an arbitrary waveform generator). The second part of this attack was demonstrated by Eberz et al., who shows that technological barriers for the attacker are extremely low [14]. A common way to counter presentation attacks is by using liveness detection, which aims to detect whether the biometric is presented by a living individual. While some authors claim [24] that ECG offers an inherent liveness detection (being only present in a living subject), this still does not resolve the problem of presentation attacks via signal injection. More work in this area is required to establish a viable defence against ECG data compromise.

To summarise, ECG remains a strong candidate to be used as a biometric for personal recognition. Several studies have demonstrated uniqueness and stability of ECG, albeit on a small scale. The introduction of low-cost ECG sensors also provides an opportunity for system designers to embed these sensors into existing access control systems. At the same time, there is still insufficient research into extracting features from ECG signals, preventing spoofing attacks and guaranteeing that ECG-based biometric systems are accepted by the general public.

2.5 Related Work

Several literature reviews on the topic of ECG-based biometrics have been performed [1] [20] [30], which the reader should refer to for an in-depth coverage of existing

efforts in this area. This section will provide a brief classification of approaches used in previous studies and an overview of achieved results.

Previous work on ECG-based biometric systems can be differentiated according to the design choices made with respect to *data acquisition*, *feature selection* and *template matching* techniques.

2.5.1 Data Acquisition

Data acquisition techniques can be further subdivided into three dimensions: "on/off-the-person" sensor location, "one-shot" versus continuous authentication and period of data collection.

Most of the studies examined the "on-the-person" approach for signal acquisition, such that electrodes are located directly on the individual. These are done using either a complete 12-lead ECG [5, 8, 22, 16] or only a subset of leads (including single lead) [21, 29]. Studies that followed the "off-the-person" approach are few and far between, but they illustrate a more realistic use-case scenario for ECG-based recognition systems. These include works by Falconi et al., who installed ECG sensors into a smartphone case [2], Silva et al., who embedded the sensor into keyboard wrist rest [12], and Lourenco et al., who installed an ECG monitor into the steering wheel of a car [26].

Biometric systems proposed in previous work mostly perform "one-shot" authentication, verifying the identity presented by the user only at a single point of time when a resource is requested (e.g. access to a building or a banking service). At the same time, some applications might benefit from using a continuous authentication process, such that the identity is verified as long as the resource is being used (e.g. being logged into a computer system or driving a car). Examples of continuous ECG-based authentication system are implemented in the works of Pinto et al. [33] and Coutinho et al. [11].

Finally, existing studies can be differentiated depending on the period of time over which ECG data was collected. Performing longitudinal studies is expensive and time-consuming, therefore most of the existing works use data collected during a single ECG acquisition session, as seen in [5, 2, 21, 25, 11, 16, 24]. These studies, however, cannot draw any conclusions about the stability of the biometric. While several authors used longitudinal ECG data in their studies [29, 22], only Silva et al. explicitly provided a side-by-side comparison of results achieved using both single-session and multiple-session data collected over a period of four months [12]. They concluded that ECG-based biometrics exhibit promising recognition rates using both short-term and long-term data.

2.5.2 Feature Selection

With respect to feature selection, existing approaches can be broadly classified as fiducial, partially fiducial and non-fiducial [12]. Fiducial methods use waves onset/offset, peaks, angle and other measurements derived from the reference points within the signal (e.g. P-QRS-T complexes) to create the feature vectors that form the biometric templates. Fiducial methods have been successfully used in several studies, including [5, 21, 2, 25, 22], but require specialised algorithms to locate the reference points within the ECG trace. Partially-fiducial methods locate only the R-peaks (in the QRS complex), which are used to segment the ECG signal into single heartbeat waveforms. After preprocessing, these waveforms are then adopted by the biometric system as the templates. Partially-fiducial approach was used by several authors, for instance [12, 29, 8]. Other studies use non-fiducial methods that do not require any reference points, as seen in [11, 33]. Instead, these methods extract features from the signal by representing it in the frequency domain or by applying statistical approaches. Several papers also combine fiducial and non-fiducial features to create the biometric templates [40, 24].

2.5.3 Template Matching

Template matching refers to the process of verifying the biometric query against the stored templates. Existing studies differ with respect to recognition methods used to perform the verification. Some authors have explored discriminative models, such as the k-nearest neighbours algorithm [12, 16, 8], support vector machines [12, 40, 24, 33] and neural networks [33]. Generative models were also used, including linear discriminant analysis [29, 25, 21] and Gaussian mixture models [33]. Furthermore, many authors have experimented with different classifiers, choosing the one that offers the best performance on the validation data.

2.5.4 Results

Comparing results reported in the literature proves to be difficult in practice, as no standardised dataset exists for ECG-based biometric research. Furthermore, there is a significant discrepancy between the evaluation metrics used by different authors. Most studies report the accuracy of their biometric systems, expressed either as the classification accuracy (CA) or the equal error rate (EER). Some papers also present other metrics, such as false acceptance rate (FAR), false rejection rate (FRR), true acceptance rate (TAR), true rejection rate (TRR), error of identification (EID), receiver operating characteristics (ROC) curve and area under the ROC curve (AUROC). These metrics are discussed in Chapter 5 of this report.

Table 2.2. summarises some of the existing studies on biometric authentication. The papers presented in the table use ECG data obtained using the “off-the-person” approach with a single-lead ECG monitor. EER is used as the performance metric, with

lower values indicating better performance. Table 2.2. also presents the results obtained by Silva et al. using both short-term (minutes apart) and long-term (4 months apart) ECG signal.

Study	Subjects	Features	Matching	Results
Coutinho et al [11]	19	Non-Fiducial	Custom	**0.4%**
Silva et al. [12]	63	Partially-Fiducial	SVM	**1.0%** (Short-Term)
Singh et al. [40]	126	Mixed	SVM	**3.4%**
Silva et al. [12]	63	Partially-Fiducial	SVM	**9.1%** (Long-Term)
Falconi et al.† [2]	10	Fiducial	Custom	**9.8%**
Komeili et al. [24]	70	Mixed	SVM	**11.0%**
Carreiras et al.‡ [8]	63	Partially-Fiducial	KNN	**13.3%**

Table 2.2: Results from existing studies on biometric authentication. Third and fourth columns refer to feature selection and template matching classifier, respectively. EER is used as the performance metric (lower score indicates better performance).

† Paper does not provide EER results, thus a similar HTER metric is presented instead

‡ Results for the baseline model (main model used 12-lead ECG)

Chapter 3

Data Collection

One of the aims of this project was to investigate whether ECG signals collected using an affordable consumer-grade ECG monitor would provide sufficient data resolution for security applications. There are several publicly available datasets that can be used for evaluating ECG-based biometric systems. These datasets, however, comprise of ECG signals recorded using medical-grade 12-lead ECG monitors that are neither affordable nor usable for deployment in practical access control systems.

One such dataset, the MIT-BIH Arrhythmia Database [28] [18], is sometimes used as a benchmark for comparing results of ECG-based applications. While this dataset is useful for medical analysis, it does not provide a balanced distribution of ECG signals among the general population. This would make it hard to evaluate the proposed biometric authentication system, as irregularities in the signal recordings could facilitate personal identification among a closed user set. Furthermore, the MIT-BIH Arrhythmia Database does not contain any ECG signals collected periodically from the same individual, which makes it impossible to validate the stability of ECG biometrics.

A recent initiative by Silva et al. [34] aimed to create a standardised database to promote research in ECG biometrics. As a result of the work by the Check Your Biosignals Here initiative (CYBHi) two public datasets were released, for short-term and for long-term assessment, with ECG data collected at palms and fingers. A total of 63 subjects were enrolled to create the long-term dataset, which involved two data acquisition sessions with a four-month interval in between, which also made it possible to investigate ECG biometrics stability.

Generally speaking, the CYBHi long-term dataset presents an adequate benchmark for evaluating ECG-based biometric systems. Nevertheless, there are several limitations of the CYBHi database. The device used in the experiments performed 130 Hz bandpass filtering, possibly distorting the signals of other frequencies. Each session recorded the ECG trace of a participant for 2 minutes only, which might not provide enough data to train and test the classifier used for template matching. Finally, the demographics of the study are not representative of a wider population, as the average age of the subjects is 20 years. These issues prompted me to perform my own data collection, which is described in the next section.

3.1 Experimental Setup

The main motivation of performing ECG signal acquisition experiments as opposed to using an existing dataset is the ability to control the conditions of the experiments. In order to explore the stability of ECG biometrics, data collection was performed over two sessions with a period of four months in between. All experiments were conducted indoors, in the Mini Forum 1 area of the Informatics Forum building.

Please note that the experimental methodology described in this section adheres to the ethics regulations of the University of Edinburgh and the setup was reviewed and authorised by the School of Informatics Ethics Panel.

3.1.1 Electrocardiogram Monitor

For the experiments, I used an AliveCor Kardia Mobile ECG monitor containing two electrodes. This device was chosen as the best compromise between ease of data acquisition and approximation to a biometric sensor that could be deployed in a real authentication system. It is also affordable, with a price of just under £100. The technical specifications of the device are covered in Table 3.1.

Characteristic	Specification
Dimensions	8.2 cm x 3.2 cm x 0.35 cm
Weight	18 grams
Lead Count	Single Lead
Input Dynamic Range	10 mV peak-to-peak
Recording Duration	30 seconds to 5 minutes
Sampling Rate	300 samples per second
Resolution	16 bit

Table 3.1: Technical characteristics of the ECG monitor.

In order to record an ECG, the user has to place two fingers from each hand onto each of the two electrodes, as shown in Figure 3.1. During operation, the monitor is connected to a smartphone application, which stores the data as a single-lead ECG recording. The application provides an interface that allows the user to access and share the ECG in the form of a PDF report.

3.1.2 Participants

Most of the subjects enrolled in this experiment were affiliated with the university, either as students, support staff or faculty members. A total of 55 subjects were enrolled

in the experiment, with 53 participants taking part in both sessions. According to the demographics survey, 30 males and 25 females participated in the study with reported ages ranging from 18 to 60. The median age was 22. In terms of education, all subjects have obtained at least a high school diploma and the majority of them were enrolled in undergraduate or graduate studies. Furthermore, none of the participants reported any serious health issues, though several were feeling exhausted or sleep deprived at the moment of the experiment. You can find more information on participant demographics in Appendix A of this report. There were no restrictions on eligibility, as long as the subject was at least 18 years old.

3.1.3 Experimental Procedure

The experiment was designed to take around 10-15 minutes. The participants came to the Mini Forum 1 area of the Informatics Forum, where the purpose and conditions of the experiment were explained. The experimental setup is demonstrated in Figure 3.2. All subjects signed a consent form, which confirmed their voluntary participation in the data collection procedure. The experimental procedure differed depending on the session.

In the first (October) session, the subject recorded their ECG trace using the monitor for 4 minutes. The recording was performed twice for a total of 8 minutes, with a break in between. During the recording, participants were asked not to move their hands in order not to introduce noise or interrupt the measurement. If the connection with the smartphone was lost, the participants were asked to repeat the 4-minute recording, unless less than one minute remained. Otherwise, subjects were not restricted in their actions, and were allowed to talk and to perform movements. During the break, the participants were asked to fill in a demographics questionnaire (a sample of which, along with the other documents, is available in Appendix B). As part of the survey, subjects were asked to self-report their physical and emotional states, although this did not have any impact on the course of the experiment. After both signal measurements, the participants were given the option to receive a copy of their ECG trace via email.

In the second (March) session, the first 4-minute recording was performed in the same way as in the first session. Two additional participants who did not take part in the October session were also asked to complete the demographics questionnaire after the first measurement. The second ECG recording aimed to elicit a small heart rate change by introducing an acoustic stimulus, as part of the research on presentation attack detection. During this recording, the participant had wireless headphones on and was expected to stay in a resting position. The first minute of the recording was silent, followed by a minute of music; both conditions were repeated twice for a total of 4 minutes. Before this recording, the subjects were given the opportunity to familiarise themselves with the presented music and adjust the volume, if needed. After this session, all participants received a £5 Starbucks gift card as a compensation for their participation in the study and were also given the option to receive a copy of their ECG trace.

Figure 3.1: ECG monitor connected to the smartphone application.

Figure 3.2: Experimental Setup.

3.1.4 Data Extraction

As I mentioned previously, the ECG monitor used for data collection produced PDF reports as an output. In order to perform any meaningful analysis, it was necessary to convert the PDF files into a dataset containing vectors of amplitudes, corresponding to the ECG signal.

As a first step, PDF reports were converted into SVG (Scalable Vector Graphics) image files. The advantage of using SVG is that files in this format use XML to describe how images should be rendered on the screen. Hence, by parsing XML in these files I could extract the information about how ECG graphs are displayed in the SVG images and use it to construct arrays of amplitudes.

PDF to SVG conversion was performed using Inkscape, a free and open-source vector graphics editor tool. XML parsing was done using a script originally written by Simon Eberz as part of his work on ECG-based biometrics [14]. The script was improved and modified to produce a single JSON file containing all the records together with metadata.

3.1.5 Discussion

The creation of a custom dataset for this project achieved several goals. Even though the ECG monitor did not provide a straightforward interface to access the raw ECG trace, it was still sufficient to mimic an "off-the-person" sensor that could be used in a realistic access control system. Furthermore, performing two data acquisition sessions with 4 months in between allowed me to investigate the stability of ECG-based biometrics. For each subject in the dataset, an ECG trace of a total of 16 minutes was recorded, which provides enough training and validation data to evaluate the template classifiers.

While not the focus of this project, blocks of 4-minute ECG measurements also provide enough contiguous signal to explore a continuous authentication process, rather than "one-shot" authentication. Finally, ECG data that was obtained from the subjects presented with an acoustic stimulus can be used to design presentation attack detection algorithms based on heart rate variability.

Chapter 4

System Design

This chapter describes the design of the proposed biometric authentication system based on ECG signals. According to the classification given in the Related Work section, the proposed system uses the "on-the-person" approach for signal acquisition and performs "one-shot" authentication. Features were extracted from the ECG trace using a partially-fiducial approach, specifically, the signal was segmented into distinct heartbeat waveforms by locating the R peaks. Finally, several discriminative models were tested for template matching, including logistic regression and support vector machines.

In this project, I adopted the conceptual design of the system described by Silva et al. [12]. The pipeline for the proposed biometric system includes the following steps:

1. Signal Filtering
2. Signal Segmentation
3. Outlier Removal
4. Feature Selection
5. Template Matching

The raw ECG signal used in the design of the system has a sampling frequency of 300 Hz and included data from 53 participants of the October session. 52 subjects had over 125,000 sampling points (approximately 7 minutes of recording). 1 subject had less than 7 minutes of ECG measurement, due to technical problems with the monitor. I decided to exclude the data from this subject and use the first 125,000 sampling points of the remaining users.

Furthermore, 20% of the signal for each subject was set aside as test data to ensure an accurate evaluation of the system at the end. Consequently, 80% of the data was used to develop the prototype of the system.

4.1 Signal Filtering

In general, every raw signal has a noise component, the magnitude of which varies depending on the quality of the sensor and the measurement procedure. An example of common interference is power line noise, which has a frequency of 50 or 60 Hz [44]. Additionally, raw ECG signal can suffer from a baseline drift (i.e. low-frequency noise) and high-frequency noise.

Depending on the type of interference, an appropriate filter can be used to improve the quality of the signal. The ECG monitor used for data acquisition, for instance, includes a default Mains filter, which removes any power line interference. In order to remove the baseline wander noise, a high-pass filter can be applied. In a similar way, a low-pass filter can be used to reduce high-frequency noise.

A common filter used to reduce overall noise is a band-pass filter, which allows frequencies within a specific range (passband) and attenuates frequencies outside of that range (stopband) [4]. One such filter is Butterworth band-pass filter, which is characterized by a very even response to frequencies within the passband [44]. The order of the Butterworth filter can be adjusted to increase the attenuation of the signal outside of the passband, with a higher order suppressing the signal more aggressively.

In order to reduce noise from ECG measurements, I decided to experiment with the Butterworth band-pass filter. Visually, a third-order filter with a passband from 1 Hz to 20 Hz gave the best results. The result of applying this filter on a raw ECG signal is presented in Figure 4.1.

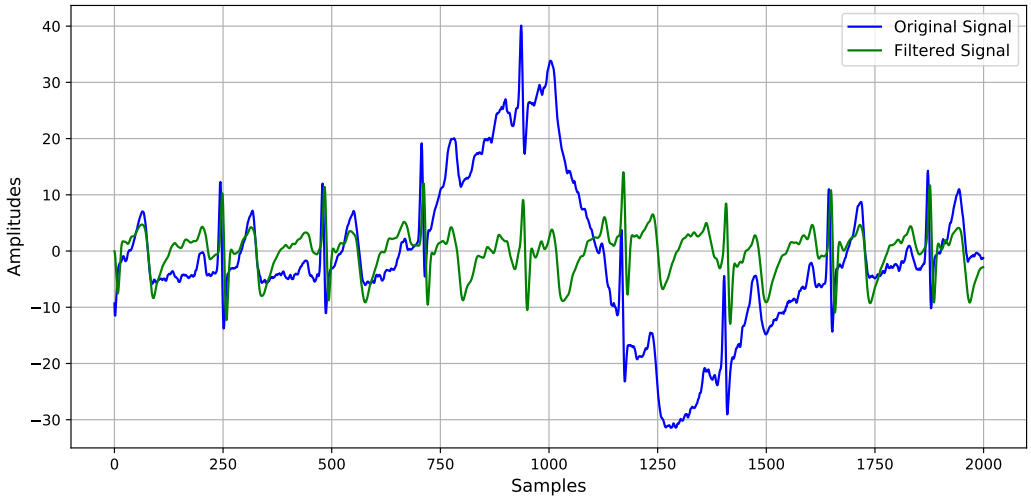


Figure 4.1: Butterworth band-pass filter applied to an ECG trace.

The ECG monitor used for data collection also provided the option to apply a band-pass filter. Unfortunately, the order and the passband of filter are not given in the technical specification of the device.

Using a custom filter is essential for all-in-one platforms, which include both hardware (acquisition sensors) and software components. Nevertheless, this project was mostly

focused on the software side of the biometric system. For this reason, I decided to use signal filtering capabilities provided by AliveCor to preprocess the acquired ECG signal.

4.2 Signal Segmentation

After de-noising the signal, the ECG trace can be used to extract features that comprise a biometric template of the individual. This in itself is a non-trivial task, which is broken up into several steps described in this and the next sections.

For this project, I decided to follow a partially-fiducial approach, which involves extracting features from individual heartbeat waveforms. For this reason, an appropriate segmentation algorithm was required, which would locate the R peaks within the ECG trace and perform appropriate partitioning of the signal, based on those peaks.

4.2.1 QRS Complex Detection

The first step in performing ECG signal segmentation is locating the R peaks within the ECG trace. More specifically, the R peaks are located within the QRS complex, thus a QRS complex detection algorithm could be applied for this purpose.

There has been extensive research into detecting QRS complexes for medical analysis. A real-time algorithm developed by Pan and Tompkins, which became known as the Pan-Tompkins algorithm, is commonly used for QRS complex detection [32]. The Pan-Tompkins algorithm is based on the differentiation and squaring of the signal, in order to amplify the characteristic features of the QRS complexes within the ECG trace.

More recent approaches started using wavelets to perform R-peak detection. In this project, I implemented my own QRS complex detection algorithm by performing a discrete wavelet transform of the signal and then applying a threshold-based R-peak detector. The ideas for this part were adapted from a tutorial on R wave detection from MathWorks [27] and an article on ECG analysis techniques by van Gent [44].

Wavelet transform is similar to Fourier transform, in that it can be used to analyse and process signals. Both transformation can be used to obtain information about the frequency components that exist in the signal. Fourier transform, however, lacks resolution between time and frequency domains, meaning that it only gives information about the frequency (i.e. spectral) components that exist in the signal without positioning them in time. This makes Fourier transform applicable for stationary signals, in which the frequency content does not change over time. However, ECG signals are non-stationary in nature, as the frequencies of the signal change depending on the current heart rate. [35]

In order to obtain a time-frequency representation of the signal, wavelet transform is used instead. While Fourier transform decomposes the signal in the time-domain by

presenting it as a linear combination of sine and cosine waves, wavelet transform uses a linear combination of short waves, called *wavelets*, to express the signal. The wavelets are generated from a function ψ, known as the *mother wavelet*, by scaling and shifting:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t - k2^j}{2^j}\right) \tag{4.1}$$

In the formula, k is the shift parameter and j is the scale parameter. By changing the shift parameter, it is possible to obtain the spectral components of the signal at a specific time, essentially by 'sliding' the wavelet along the time domain. The scale parameter can be used to represent the different frequencies that exist in the signal, by adjusting the width of the wavelet.

There are different types of wavelets, including Haar, Daubechies and Biorthogonal. For R-peak detection, Symlets 4 (Sym4) wavelet is commonly used, as it closely resembles the QRS complex.

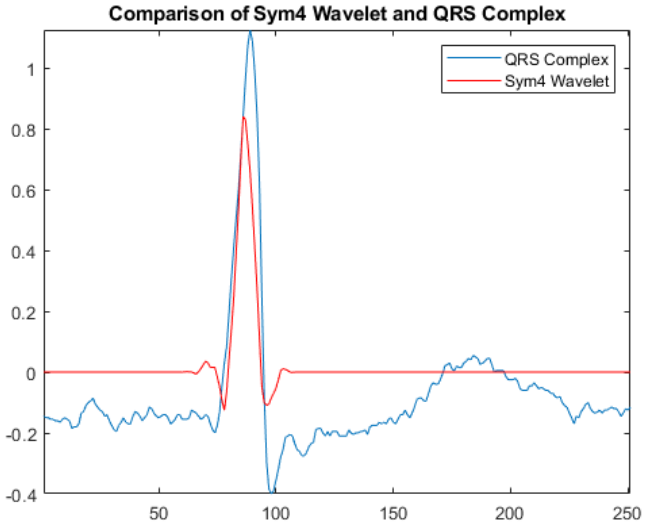


Figure 4.2: Comparison of Sym4 wavelet and the QRS complex. Image taken from [27].

We can represent the ECG signal as a discrete function $f(t)$, which maps a time point t to a signal amplitude $f(t)$. By applying the discrete wavelet transform, we can obtain *wavelet coefficients* $c_{j,k}$. These coefficients provide a similarity measure between $f(t)$ and $\psi_{j,k}(t)$ and can be used to reconstruct the original signal (using inverse discrete wavelet transform). While it makes sense to shift the wavelet over all values t, for which $f(t)$ is defined (i.e. set $k = t$), the scale j can be chosen arbitrarily, depending on the frequency content that need to be represented. For ECG-based applications, Elgendi et al. showed that the optimal frequency range for detecting QRS complexes is 8-20 Hz [17].

I experimented with several values of j and found that QRS complexes become the most pronounced when the signal is reconstructed using wavelet coefficients at scales

$j = 3, 4, 5$. If the sampling rate of the signal is 300 Hz, then the scales correspond to the following approximate frequency bands, according to the Nyquist-Shannon sampling theorem [31]:

- Scale 1 - $[75,\ 150)$ Hz
- Scale 2 - $[37.5,\ 75)$ Hz
- Scale 3 - $[18.75,\ 37.5)$ Hz
- Scale 4 - $[9.38,\ 18.75)$ Hz
- Scale 5 - $[4.69,\ 9.38]$ Hz

Thus, scales 3, 4 and 5 correspond to the passband [4.69, 37.5) which covers the optimal frequency range for detecting QRS complexes as described above.

The resulting signal was then squared to further accentuate the peaks. Figure 4.3 shows an example of such a reconstructed signal plotted against the original ECG trace.

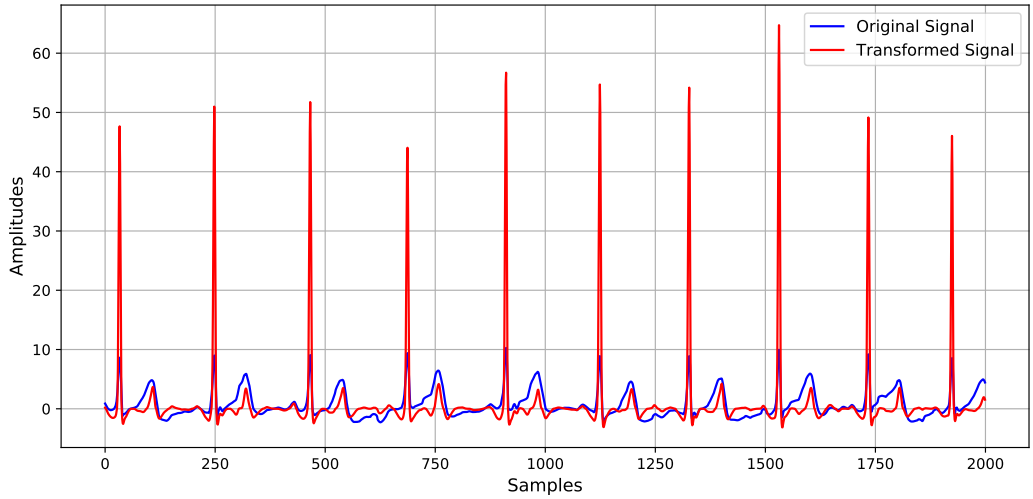


Figure 4.3: Wavelet coefficients (scales 3 to 5) used to reconstruct an ECG trace.

Note that the implementation for maximal overlap discrete wavelet transform, its inverse and other supplementary functions were adapted from the code provided by 'quanly_mc' on Stack Overflow [36].

4.2.2 Peak Detection

After localising the QRS complexes, the next step was to mark the R-wave peaks, which will be used for signal segmentation. While wavelet transform accentuated the R-wave peaks on most ECG traces, some signals also had a strong influence from the T wave and residual noise, which made detection challenging.

As a first step, I segmented ECG traces for all subjects into equal blocks, each containing 5000 samples, corresponding approximately to 17 seconds of recording. This was

done to improve the quality of the peak detector, as some blocks may contain more noise than others.

The algorithm for R-wave peak detection consists of three parts: threshold selection, local maxima detection and signal partitioning.

4.2.2.1 Threshold Selection

A threshold is a line that defines regions of interests, i.e. areas of the graph located above the threshold. These regions of interest contain peaks that are likely to be the R-wave peaks. First, I tried using a simple straight line as a threshold. While it worked well in many cases, there were still some parts of the transformed ECG signal which had noise or T-wave peaks dominating the R-wave peaks. An example of a misclassification that occurred when using straight lines as the threshold is illustrated in Figure 4.4.

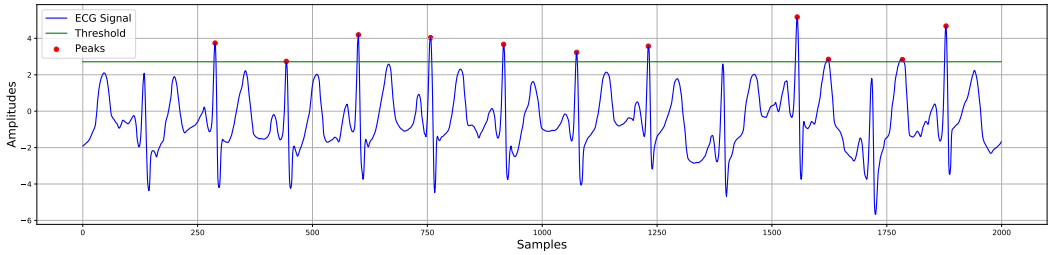


Figure 4.4: Peak detection using a threshold based on a constant line. Notice how the algorithm misses several R-wave peaks and misclassifies two T-wave peaks.

In order to improve this approach, I started using thresholds defined by the running mean of the signal. As the QRS complexes usually involve a rapid increase in the amplitudes that is short in duration, they only had a small effect on the running mean. On the other hand, T waves usually involve a steady increase of the signal that lasts longer, leading to the increase of the running mean. This meant that the R-wave peaks remained above the threshold, while T-wave peaks were located under, as shown in Figure 4.5.

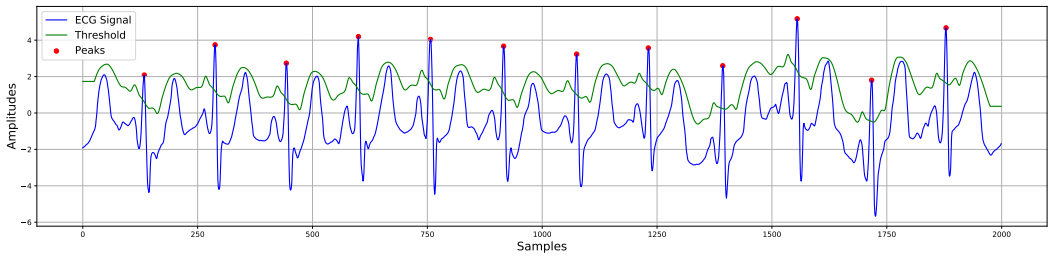


Figure 4.5: Peak detection using a threshold based on the running mean. All R-wave peaks in the trace are detected correctly.

Using the threshold based on the running mean improved R-wave peak detection, but still resulted in some errors. For this reason, the algorithm was modified to return a set

of candidate thresholds. Each candidate threshold is computed by shifting the running mean graph upwards, resulting in more constrained regions of interest.

4.2.2.2 Local Maxima Detection

For each set of regions of interest defined by a candidate threshold, a set of peaks was extracted. As there are several candidate thresholds, this resulted in several sets of candidate peaks. Selection of the most optimal set was based on the heuristic rules described by Paul van Gent [44], which are presented in Procedure 1.

As a reminder, the distance between an adjacent pair of R-wave peaks is known as the RR interval. If the sampling frequency is known, then the RR intervals can be used to compute the average heart rate (measured in beats per minute) and heart rate variability from a set of peaks.

Procedure 1 Establish Optimal Peaks

Input: List of candidate sets of peaks *CList*
Output: Optimal set of peaks *B*

```
B ← {}
for C in CList do
    compute avgHeartRate using C
    compute varHeartRateC using C
    if avgHeartRate > 130 or avgHeartRate < 30 then
        continue to next C
    else if B = {} then
        B ← C
        compute varHeartRateB using B
    else if varHeartRateC < varHeartRateB and varHeartRateC ≠ 0 then
        B ← C
        compute varHeartRateB using B
    end if
end for
return B
```

Essentially, if the average heart rate derived from a set of candidate points is less than 30 or higher than 130 beats per minute, it is always rejected, as it does not correspond to the range of human heart rates. Otherwise, the best points are the ones that have the least variance, as high variance can signify that the peak detector has misclassified R-wave peaks.

Peak detection was performed on the wavelet-transformed signal. The location of the most optimal set of peaks on the transformed ECG trace was assumed to be the location of the R-wave peaks on the original ECG signal.

4.2.2.3 Signal Partitioning

Finally, the locations of the R-wave peaks were used to perform the signal partitioning into individual heartbeat waveforms. The location of the boundaries of individual heartbeats was computed using formula (4.2).

$$E_i = P_i + (P_{i+1} - P_i) \times \frac{2}{3} \tag{4.2}$$

where E_i is the i-th boundary and P_i is the i-th peak

This meant that 1/3 of the RR interval before and 2/3 of the RR interval after a given peak was used as a single heartbeat, as it captures the short PR interval, the QRS complex and the long ST segment. Each heartbeat waveform was resampled to contain 250 samples, in order to mitigate the fact that subjects had a variable heart rate during ECG data collection.

Individual heartbeat waveforms obtained by the signal segmentation procedure for a single subject are illustrated in Figure 4.6 and for 8 different subjects in Figure 4.7.

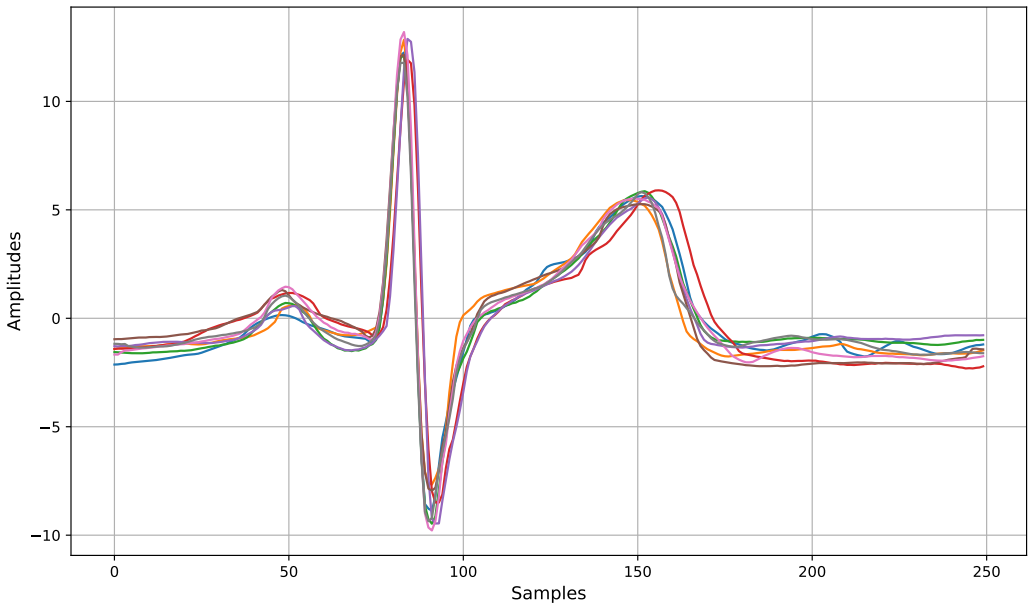


Figure 4.6: ECG variation within a single individual (8 waveforms).

4.3 Outlier Removal

No matter how well the signal partitioning is performed, it is almost guaranteed that some segments will be corrupted. In most cases, this is caused by noise dominating some parts of the ECG signal, for instance, if the subject performs an abrupt movement

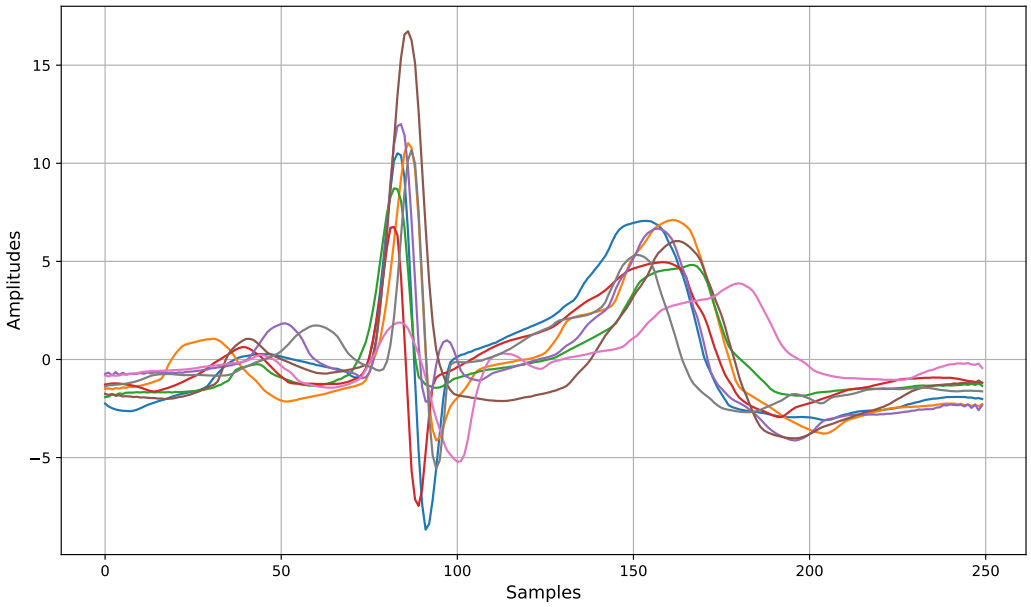


Figure 4.7: ECG variation among 8 individuals (1 waveform each).

during the recording. In order to mitigate the effect of such interference, I decided to develop a simple outlier removal procedure, which would drop the corrupt segments.

We can present all the heartbeat waveforms for a particular subject in the form of a matrix, where each row is a 250-dimensional vector representing a single heartbeat. As a first step, the algorithm creates a median waveform by taking median values for each of the 250 dimensions. The median was chosen, as it is not so affected by the outliers, as opposed to the mean.

Afterwards, the algorithm computes the distance between each heartbeat vector and the median. For this purpose, I used the Euclidian distance, though other metrics could be applied as well. Finally, the algorithm drops 20% of the most distant vectors.

In practice, this outlier removal procedure performed well even with ECG signals severely corrupted by noise. However, it has a significant drawback of setting a hard cut-off point of 20%. For some ECG traces of high quality, more heartbeat waveforms were dropped than necessary. Similarly, ECG traces that contained a lot of noise may have required even more pre-filtering. An improvement would be to use more advanced statistical methods to detect outliers, for instance, the Modified Z-scores or the IQR method [19].

4.4 Feature Selection

After removing the outliers, the heartbeat waveforms can be essentially used as the templates. However, there are several steps that are essential to improve the performance of the classifier that will be used for template matching. These include feature standardisation and dimensionality reduction.

4.4.1 Feature Standardisation

Standardisation (or Z-score normalisation) is a common preprocessing technique used in many machine learning applications [37]. Feature standardisation rescales the values of each feature, such that they have the properties of a standard normal distribution with a zero mean and a unit variance. Z-score normalisation can be performed by finding the standard scores (z-scores) of each feature, as given in (4.3).

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{4.3}$$

where x_i is the i-th input feature, μ_i is the mean i-th feature and σ_i is the standard deviation

Generally speaking, feature standardisation ensures that all features are on the same scale and, thus, contribute equally to the prediction made by the machine learning algorithm. Some algorithms benefit in particular from z-score normalisation [37]:

- **K-Nearest Neighbours with Euclidian distance.** Standardisation ensures that all features contribute equally to the result of classification.
- **Algorithms using gradient descent for cost optimisation.** Without standardisation some weights may update faster than others.
- **Dimensionality reduction algorithms (e.g. Principal Component Analysis).** Without standardisation the algorithm will be influenced by features that have a higher variance caused by the differences in scale.

For these reasons, feature standardisation was performed as the first step before applying any further transformations.

4.4.2 Dimensionality Reduction

At the current stage, each heartbeat waveform is represented by a 250-dimensional feature vector. If the dimension of the vector is too high, training the classifier becomes computationally expensive. It may also lead to sub-optimal performance caused by the sparsity of the feature space at high dimensions, which results in the classifier overfitting to the training data and generalising poorly to unseen data. This problem is referred to as the curse of dimensionality [9].

In order to alleviate this problem, I decided to perform principal component analysis (PCA) to reduce the dimension of feature vectors. PCA is a statistical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components* [13]. The first principal component accounts for the most variability in the original data, with each succeeding component accounting for as much of the remaining variability as possible. Therefore, PCA can be applied to reduce the dimensionality of feature vectors while retaining as much of information about feature variability as possible.

I experimented with different number of principal components by applying PCA to the matrix of standardised features representing heartbeat waveforms and discovered that the first 25 principal components account for over 99% of total variability in the training data. The relationship between the explained variance and the first 50 principal components is displayed in Figure 4.8.

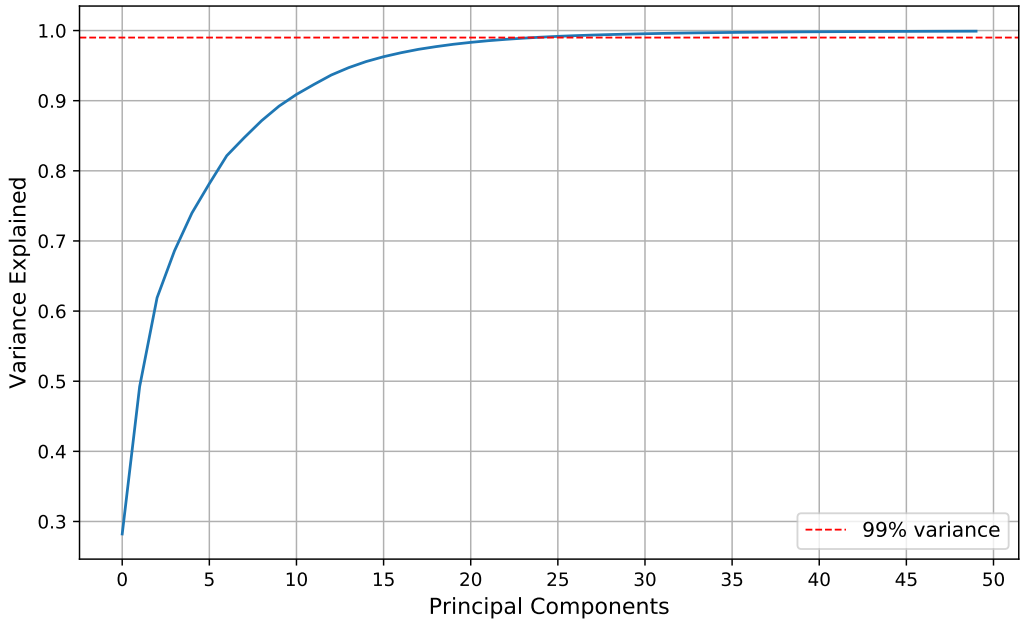


Figure 4.8: Relationship between the explained variance and the first 50 principal components.

In the end, I decided to use the first 25 principal components as the input features, which were used to train the biometric system classifier.

4.5 Template Matching

After preprocessing the ECG signal and creating the biometric templates, I focused on choosing a machine learning model that would yield the most optimal classification performance. As a reminder, this project aims to tackle the problem of authentication, as opposed to identification.

Formally, we can represent the set of users as $C = \{1, 2, ..., K-1, K\}$ where K is the number of users enrolled in the system. Furthermore, we can represent the set of recorded templates for user k as $B_k = \{b_k^{(1)}, b_k^{(2)}, ..., b_k^{(n-1)}, b_k^{(n)}\}$ where n is the total number of templates recorded, with n potentially differing for each user. The set of all biometric templates is therefore $B = \{B_1, B_2, ..., B_{K-1}, B_K\}$.

During the training stage, the system is trained to predict whether a set of biometrics B_k belongs to the identity $k, \forall k \in C$. Thus, a separate model M_k is created for each user k. For each model M_k, we assign the recorded templates B_k to the positive class, by

labelling them as '1'. All other biometric templates $B_p, \forall p \in C \wedge p \neq k$ are assigned to the negative class '0'. We can say that each model M_k is trained to solve a binary classification task.

A side-effect of solving the binary classification problem is that, for each user k, there are far more examples of the negative class B_p than there are examples of the positive class B_k. This can be partially mitigated by assigning more importance to correct classification of positive class observations during the training process.

During the inference stage, a user presents their identity k and the query q_k, representing a biometric template recorded during the authentication process. The system then uses M_k to assign a confidence score that q_k belongs to k. Depending on the policy, the decision threshold can be adjusted to improve usability (reduce number of false negatives) or security (reduce number of false positives) of the system.

I experimented with three discriminative models: logistic regression, k-nearest neighbours and support vector machines. In order to choose the best hyperparameters for the model, I performed a grid search with k-fold validation, which is explained in the next section.

As a side note, most of the machine learning tools used in this project were provided by the scikit-learn Python library. Nevertheless, some features, such as the performance metrics, were not available and had to be implemented manually.

4.5.1 Hyperparameter Optimisation

Hyperparameters differ from model parameters in that they cannot be explicitly learned from the data during the training phase. In the discussion below, the term "hyperparameter" is used interchangeably with "parameter", while the trainable model settings will be referred to as "weights".

There are several ways to optimise the hyperparameters of a machine learning classifier. The most straightforward approach is to perform a *grid search*, which is an exhaustive search through a specified subset of the parameters of a model. A classifier is trained with every setting of the parameters, and the performance of each model is recorded. The configuration that is chosen at the end corresponds to the model that achieved the best performance [23].

Traditional grid search is very expensive to perform when the number of parameters is high, as the cost of an exhaustive search increases exponentially with the number of tuned parameters. At the same time, each configuration of the model can be trained in parallel, making grid search feasible for a small number of parameters.

Once a classifier is trained, it needs to be evaluated on unseen data, in order to ensure good *generalisation* of the model to future data. When performing hyperparameter optimisation, evaluation of each setting of the parameters cannot be performed on the test set, as that "tunes" the parameters of model to a specific instance of the test set that was available during the production of the system.

One way to mitigate this problem is to further divide the training data into a training and a validation set, which is used to assess the performance of the model. This evaluation method is known as *cross validation*. One way to perform cross validation is by using the *holdout method*, such that the data is split into three sets from the start: training, validation and test. The model is then always trained on the training set, hyperparameters are optimised using the validation set and the final performance is assessed on the test set [39]. A downside to this approach is that it makes less data available for training the model, and the evaluation results (including the optimised hyperparameters) are significantly affected by the data points that are chosen to be in the validation set.

A better way to perform evaluation is to use *k-fold cross validation* [39]. In this approach, the training set is divided into k subsets (also known as folds), and the holdout method is repeated k times. During each iteration, one of the k subsets is used as the validation set and the other $k-1$ subsets are used together as the training set. The final performance metric is computed as the average performance of the k models obtained during each trial. An illustration of k-fold cross validation is shown in Figure 4.9.

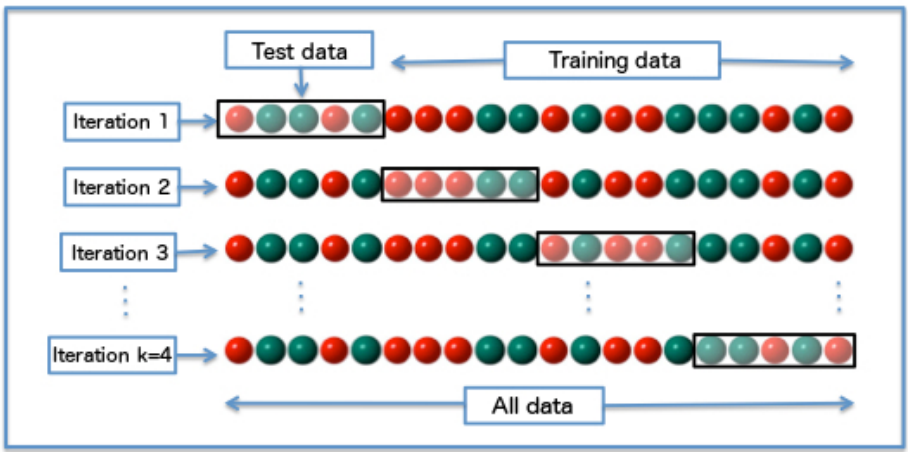


Figure 4.9: Example of k-fold cross validation with k = 4.

For this project, I performed hyperparameter optimisation by running a grid search over specified subsets of parameters. For each model configuration, I performed 5-fold cross validation to ensure the robustness of achieved results. The performance is measured using equal error rate (EER), a common metric used in biometric system evaluation. An in-depth discussion of EER and other performance metrics is presented in Section 5.1.

In the next sections, I will present information about the classifiers that were used to run the experiments. Explaining every detail of the machine learning models used in this project is beyond the scope of this report, and the reader should refer to [6], which provides an in-depth treatment of the subject.

4.5.2 Logistic Regression

Logistic regression is a discriminative model that can be used to perform classification. At its core, logistic regression uses the sigmoid function (4.4) to assign observations to a discrete set of classes. The range of the sigmoid function is [0, 1], thus the output of the logistic regression can be viewed as the probability that the observation belongs to the positive class.

$$\sigma(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \tag{4.4}$$

The input to the sigmoid function z is a linear combination of features. The constant coefficients are known as weights, and the logistic regression learns the best set of weights that maximises the performance (or, equivalently, minimises the error) of the classifier. In our example, a biometric template is represented by a set of 25 features, thus 25 weights need to be learnt plus a constant term called the bias w_0:

$$z = w_0 + w_1 x_1 + w_2 x_2 + \ \ldots \ + w_{25} x_{25} = w_0 + \sum_{i=1}^{25} w_i x_i \tag{4.5}$$

A common problem in machine learning is *overfitting*, which occurs when the model learns to represent the training set too closely. When this happens, the model becomes capable of near-perfect predictions on the training set, but performs poorly on the validation and test sets. Overfitting is mitigated by adding a *regularisation* term to the error function. Regularisation forces logistic regression to learn a simpler representation of the data, by reducing the weights of the model towards zero. The effect of the regularisation term is controlled by the hyperparameter α. The derivation of the error function of logistic regression and more information about regularisation is presented in [6].

I experimented with several parameters of the model by using a grid search with 5-fold cross validation. In particular, I investigated the performance of L1 and L2 regularisation terms, which are commonly used in practice. For each term, I experimented with different values of α, including 0.00001, 0.0001, 0.001. Finally, in one case I assigned more importance to correct classification of data points belonging to the positive class, while in the other case both classes were treated equally.

The results for the best 5 configurations are shown in Table 4.1. EER is used as the evaluation metric (measured in percentages), and a lower score indicates better performance of the model.

4.5.3 K-Nearest Neighbours

K-nearest neighbours (KNN) is another algorithm that can be used for classification problems. It is based on a simple principle: during the prediction phase, KNN com-

Regularisation	α	Equal Classes	EER
L2	0.0001	False	**13.92%**
L2	0.0001	True	**13.92%**
L1	0.0001	False	14.01%
L1	0.0001	True	14.01%
None	N/A	False	14.08%

Table 4.1: Top 5 results obtained by running grid search on the logistic regression model. EER is used as the performance metric (lower score indicates better performance).

putes the distance between each training point and the new observation. The algorithm then assigns a label to the new observation by choosing the class shared by the majority of the k nearest points. As such, there is no explicit "training" phase, which simplifies the design of the algorithm. On the other hand, in order to predict a new label, KNN has to loop over the entire training set, which makes it impractical for large datasets.

The performance of the model is affected by the choice of the parameter k. Smaller values of k make the model prone to overfitting, as the prediction is made based only on a few neighbours, which is sensitive to distortions (e.g. noise and outliers) specific to the chosen training set. On the contrary, higher values of k put less emphasis on the location of the new observation and more on the general frequency of the samples. At the extreme end, setting k equal to the number of training points will force KNN to always predict the label corresponding to the most frequent class.

Another factor that affects predictions is the metric used to compute the distance between pairs of points. A common choice is Euclidian distance between pairs of high-dimensional points given by the input features. Other options include Manhattan distance, Mahalanobis distance and cosine vector similarity.

In this part, I experimented with k set to 3, 5 and 7. For each setting of k, I computed the Euclidian and Manhattan distance between pairs of observations. In one case, once the k nearest neighbours were located, a simple majority vote was performed to decide the label for the new point. In the other case, I assigned higher importance to votes of neighbours that were closer to the new observation than those which were further away. The results of running the grid search for KNN are presented in Table 4.2.

4.5.4 Support Vector Machines

The last model that I experimented with was the Support Vector Machine (SVM) algorithm. SVM performs classification by locating the hyperplane that separates the positive and negative class observations in the feature space. The parameters of the hyperplane are then optimised to provide the highest margin between the nearest data point and the separating hyperplane. The resulting hyperplane is then used to perform

k	Distance	Vote System	EER
7	Manhattan	Distance	**8.82%**
7	Manhattan	Majority	8.83%
7	Euclidian	Majority	9.58%
7	Euclidian	Distance	9.64%
5	Manhattan	Majority	9.67%

Table 4.2: Top 5 results obtained by running grid search on the k-nearest neighbour model. EER is used as the performance metric (lower score indicates better performance).

classification of new data points. This is illustrated in Figure 4.11.

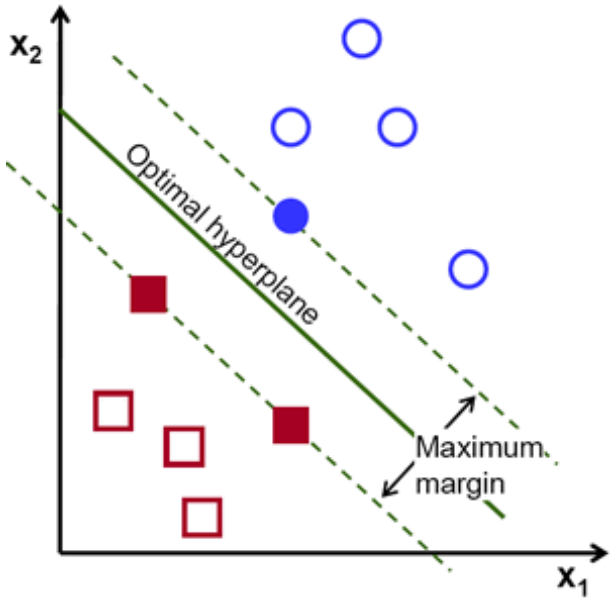


Figure 4.10: Illustration of SVM

In practice, classification problems often involve data points that are not linearly separable. This means that no hyperplane exists that can perfectly separate the positive and negative class observations. SVMs can mitigate this problem by using a technique called the kernel trick. Kernel functions can be used to transform the input space to a higher dimensional space, effectively allowing the observations to be linearly separable. On the other hand, the increase in the number of dimensions also introduces the risk of overfitting, as was described previously in Section 4.4.2. Kernel functions include polynomials, the sigmoid function and the radial basis function, among others.

I experimented with different configurations of SVMs using the radial basis function (RBF) as the kernel function, which has the parameter γ as a tuneable hyperparameter. Additionally, regularisation can be controlled by adjusting the parameter C, such that a lower value of C trades off performance for a simpler model and vice-versa. Finally, in

one case I assigned more importance to correct classification of data points belonging to the positive class, while in the other case both classes were treated equally. The best five model configurations are presented in Table 4.3.

C	γ	Equal Classes	EER
0.1	0.01	False	**5.74%**
1	0.01	False	5.84%
10	0.01	False	6.53%
10	0.01	True	6.85%
1	0.01	True	6.89%

Table 4.3: Top 5 results obtained by running grid search on the support vector machine model. EER is used as the performance metric (lower score indicates better performance).

4.5.5 Discussion

The best results were achieved by running the model with the SVM classifier with a radial basis function kernel. Assigning more importance to correct classification of observations belonging to the positive class makes a difference, because it ensures that the model does not simply predict the majority class (the negative class has significantly more samples than the positive class). Among the top five model configurations, γ was always set to 0.01. The best SVM used C value of 0.1, meaning that a simple model generalises better to unseen data, unlike models trained with C equal to 1 and 10.

The SVM classifier also outperforms both the KNN algorithm and the logistic regression model. Conceptually, SVM extends the capabilities of the logistic regression model. Unlike the KNN classifier, SVM provides more advanced techniques for training the model, for instance, by transforming the input space into a higher-dimensional space using kernel functions. It also includes more parameters for preventing overfitting, such as the C and γ hyperparameters.

The performance results obtained in this chapter should be taken with a grain of salt, as they were obtained in the process of tuning the hyperparameters of the model. The next chapter presents several approaches to biometric system evaluation that are used to validate the proposed system design.

Chapter 5

System Evaluation

Proper evaluation is an integral part of designing a viable biometric system. However, existing papers on ECG-based biometrics rarely include a comprehensive section focusing on the evaluation of the proposed systems. One of the aims of this projects is to address this shortcoming, by providing an overview and applying the best practices for system evaluation. The ideas presented in this chapter are described in more depth by Eberz et al. [15].

This chapter starts by introducing the most common metrics used to assess the performance of a biometric system. I then consider how incorrect training data selection can affect the values of these metrics. I finish by applying the concepts discussed in this chapter to assess the performance of the biometric system proposed in this report.

Throughout this chapter, we assume that the biometric data is split into the training and test sets. The training set is used to optimise the parameters of the biometric system and the test set is used only to obtain the performance measure of the final system design. In this project, 80% of the data was used as the training set and 20% was left aside as the test set.

5.1 Evaluation Metrics

In general, a biometric system can perform two types of errors. A *false accept* happens whenever a system incorrectly accepts an intruder and a *false reject* happens whenever a system incorrectly rejects a genuine user. The probability of these errors occurring is presented by two metrics: *false acceptance rate* (FAR) and *false rejection rate* (FRR) [38]. These are defined as follows:

$$FAR = \frac{False\ Accepts}{All\ Accepts} = \frac{False\ Accepts}{True\ Accepts + False\ Accepts} \tag{5.1}$$

$$FRR = \frac{False\ Rejects}{All\ Rejects} = \frac{False\ Rejects}{True\ Rejects + False\ Rejects} \tag{5.2}$$

Formally speaking, we can imagine a function f that takes a biometric query q_i and returns a confidence score of how likely it matches the recorded template B_i for the claimed identity i. We can represent the threshold for accepting an identity claim as t. In this case, the system accepts the user if and only if:

$$f(q_i, B_i) > t \tag{5.3}$$

Tuning threshold t determines the convenience-security trade-off of the biometric system [38]. Setting a high t will increase the security of the system, as more unauthorised accesses will be prevented. The system then will achieve a lower FAR score at the expense of a higher FRR. On the contrary, setting t to a lower value will improve the usability of the system, as more genuine users will be accepted. The system then will achieve a lower FRR score at the expense of a higher FAR.

In order to improve the overall biometric system performance, the system designer has to minimise both FAR and FRR error rates. Eberz et al. defines additional metrics that incorporate both FAR and FRR [15]:

- **Equal Error Rate (EER).** Error rate that is achieved by tuning the detection threshold of the system, such that FAR and FRR are equal.

- **Half Target Error Rate (HTER).** The average between FAR and FRR at some arbitrary threshold.

- **Receiver Operating Characteristics (ROC) Curve.** A graph that shows the convenience-security trade-off achieved by tuning the detection threshold of the system. The ROC curve allows to derive a set of pairs (FAR, FRR) at which the system can be run by changing the threshold.

- **Area under the ROC Curve (AUROC).** Ranges from 0.5 (random guessing) to 1 (perfect classification) and aggregates the performance of the system at all threshold settings.

- **Confusion Matrix (CM).** Plots the fraction of accepted samples for each user pair. Can be used to derive the FRR for each user and the FAR for each user-attacker pair.

Many papers also include the accuracy score of the classifier, which shows the fraction of samples that are correctly classified. This metric, however, can be misleading if class distribution in training data is unbalanced. For instance, if there is only 1 positive class sample and 99 instances of the negative class, then a constant model $M(x) = 0$ achieves an accuracy of 99%. In practice, however, we cannot assume that samples from the negative class occur more frequently than observations from the positive class.

5.2 Training Data Selection

There are two examples of how inappropriate selection of training data can affect the metrics used to evaluate the biometric system. The first example relates to the continuous nature of ECG signal, while the second one occurs when the attacker class is incorrectly modelled within the system classifier.

5.2.1 Continuous Biometrics

Regardless of whether the biometric system performs "one-shot" or continuous authentication, an ECG trace is represented as a *time series*, i.e. a sequence of amplitudes sampled at regular time intervals. As a consequence, a single sample is correlated with other points located close to it in time. The heartbeat waveforms that are located closer in time also exhibit a higher correlation, as the heartrate, in most cases, changes gradually over time. Therefore, the closer in time two heartbeat waveforms are, the more information is gained about one waveform by observing the other one.

If we consider a dataset consisting of heartbeat waveforms that were collected from a single ECG trace, then the training/test split affects the accuracy of evaluation. For instance, if a random shuffle is performed before splitting the dataset, the test observations will be intermixed with training samples, as demonstrated in Figure 5.1.

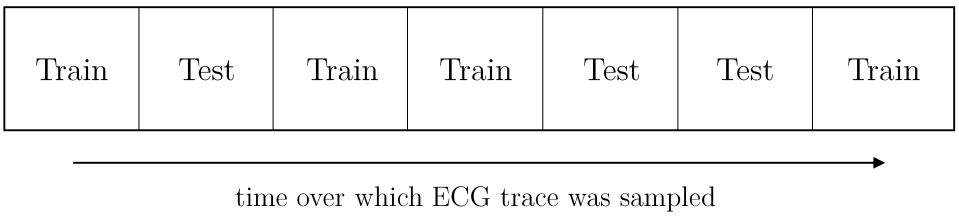


Figure 5.1: Training and test samples are intermixed if a random shuffle is performed.

If the biometric system is developed using such training set, it will be able to make inferences about the test data. For this reason, evaluation cannot be performed on a random data split, as it will demonstrate a better performance of the system than can actually be achieved. A more robust approach is to select the training samples from the start or end of the ECG signal without any shuffling.

5.2.2 Attacker Class Modelling

The second problem occurs during the training phase of the classifiers. In order to train the system to perform authentication, we create a separate model for each user, such that the biometric templates of this user are marked as the positive class, while samples from every other user compose the negative class.

A naïve way to perform evaluation of the system is to assess the performance of each classifier on the test data that contains the biometric templates from the same users as in the training set. An example of such data split is shown in Figure 5.2.

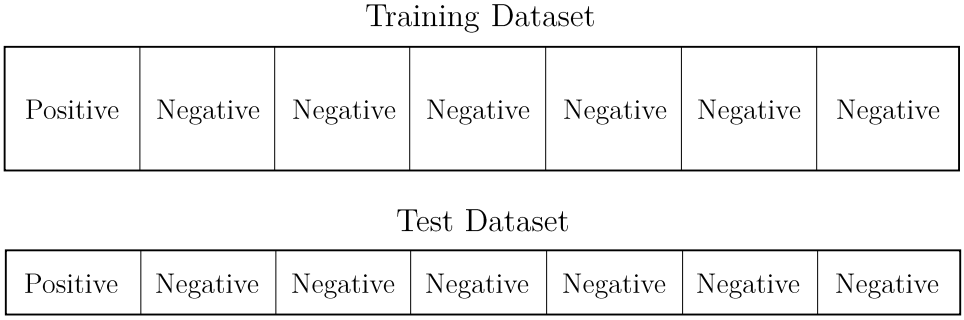


Figure 5.2: Training and testing is performed on the same users. Each block represents a distinct user enrolled in the system.

If we are trying to optimise both for FAR and FRR error rates, then FAR rates will be lower on this test data, as the classifier was already trained to recognise all of the negative class instances during the training phase. As such, if some templates in the system belong to an attacker, the classifier will already know how to distinguish them, as the training and test sets contain information about the same users.

In order to mitigate this problem, data samples from one enrolled user belonging to the negative class should be left out during the training phase. Then the classifier can be tested on the samples belonging to this user, who is the designated attacker. The model is then retrained with another user being left out, providing ECG variability results for every pair of enrolled users. One such split is shown in Figure 5.3.

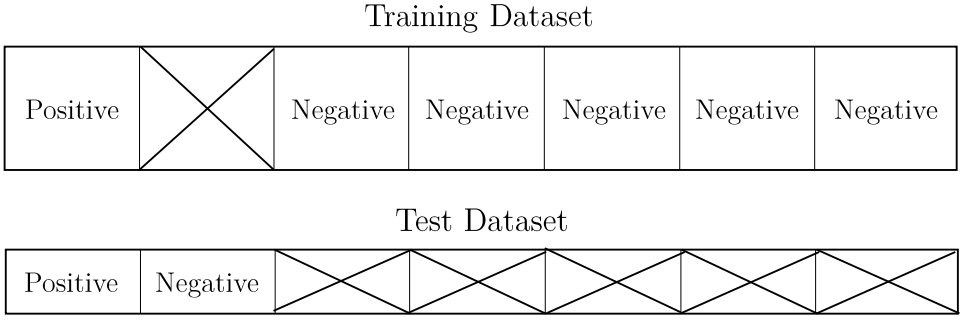


Figure 5.3: Training and testing is performed on different users that belong to the negative class. One user, crossed out in the training set, is the “designated attacker”.

Performing this evaluation is more computationally expensive, as it requires to train $N-1$ models for each user in a collection of N users. However, it provides an unbiased estimate of the performance of the biometric system. Furthermore, it can be used to perform statistical analysis on the error distribution, for instance, examine the class of samples that contribute the most to misclassification.

5.3 Performance Evaluation

The remainder of this chapter will be devoted to the evaluation of the proposed biometric system. I experimented with two evaluation procedures. In the first one, the training and test sets included the same users, such that the attacker was not modelled explicitly, as discussed in the previous section. The result of this evaluation is the EER score for each user, achieved by tuning the threshold such that FAR equals FRR. This metric allowed me to compare my results to those in the existing literature, as most authors publish the EER results of their proposed systems.

The second procedure involved choosing an explicit attacker to test the performance of the system. This was achieved by removing the biometric templates of the attacker from the training set and leaving only the templates of the attacker and the genuine user in the test set. This result of this evaluation was a set of HTER scores, one for every user-attacker pair, which is the average between FAR and FRR at some threshold. The threshold was determined by performing a stratified K-fold validation on the training data. EER was not used directly in this case, as oftentimes it is possible to tune the threshold, such that it separates the templates of the genuine user and the designated attacker, even if the model did not see the samples of the attacker during the training phase.

Evaluation was performed on all subjects that provided their ECG signal during two data acquisition sessions, under the constraint that each subject has over 125,000 sampling points (approximately 7 minutes of recording out of 8). This constraint was enforced to ensure sufficient data to generate the biometric templates (using the training set) and to have the possibility to draw conclusions about evaluation results (using the test set). However, as was mentioned in Chapter 4, one participant had less than 125,000 sampling points in the October session and was excluded from the evaluation set. Three additional subjects did not satisfy this constraint from the March set. For the sake of uniformity, I decided to leave data for 49 subjects across both sessions to use for evaluation. While this solution was not optimal, it still allowed me to obtain results that could be reliably used to compare the proposed biometric system with existing approaches.

For each data session, a training/test split was performed. The first 80% of the ECG trace across 49 subjects was used as the training data and the remaining 20% of the signal was used as the test data. Thus, two training sets and two test sets were established and used for both evaluation procedures.

We assume that the training data corresponds to the stored biometric information, while the test set is obtained during the operation of the system. Both sets were processed by the system separately, for instance, the outlier removal procedure in the training ECG signal did not influence the outlier removal procedure performed on the test signal. The template classifier used for both evaluations is the support vector machine algorithm, with the hyperparameters established in the previous chapter.

5.3.1 Standard Evaluation

Standard evaluation was performed without making any adjustments to the training and test sets, as discussed in the previous section.

Table 5.1 presents the results of standard evaluation performed under four conditions. In the first case, both the training and the test data were taken from the October session, labelled S1. Similarly, in the second case, the training and test data were obtained from the March session, labelled S2. Under the last two conditions, training and test sets were taken from different sessions. The performance of the system is assessed using the EER error rate (measured in percentages), which is averaged over the EERs obtained for each of the 49 enrolled users.

Training	Testing	Average EER	Standard Deviation
S1	S1	3.22%	2.99%
S2	S2	**2.44%**	2.40%
S1	S2	9.65%	11.35%
S2	S1	6.53%	7.87%

Table 5.1: Results obtained by running standard evaluation using the proposed biometric system. S1 refers to the October (first) session, while S2 refers to the March (second) session. The first two columns reflect from which session the corresponding dataset originates. Average EER is obtained by averaging errors for every trained model, and the standard deviation reflects the variance of the performance for different users.

5.3.2 Designated Attacker Evaluation

The second evaluation was performed by leaving out samples of a "designated attacker" from the training set. Recall from the previous discussion that we leave out the biometric templates of the attacker from the training set, such that the model does not learn any information about those observations. Similarly, we only leave the templates of the attacker and the genuine user in the test set, in order to examine the performance of the system when confronted with unknown templates. This is performed for every possible user-attacker pair. Thus, we train 48 models per user (1 genuine user and 48 attackers) for 49 users for a total of 2,352 models.

It is possible to perform evaluation in the same way, only changing the composition of the training and test sets. However, using EER as the error metric shows the performance of the system when the threshold is tuned such that FAR equals FRR. However, a more realistic scenario involves setting the threshold during the training process, and then using the same threshold during the operation of the biometric system.

In order to set the threshold, I performed 3-fold cross validation on the training set and obtained the average threshold such that FAR equals FRR. Using this threshold, I

computed the FAR and FRR error values obtained by testing the model on the test set. By taking the average of these error rates, I obtained the final HTER metric (measured in percentages). This process was repeated for all models trained and the results for each condition are shown in Table 5.2.

Training	Testing	Average HTER	Standard Deviation
S1	S1	5.86%	10.00%
S2	S2	**4.58%**	9.35%
S1	S2	30.02%	17.40%
S2	S1	30.01%	16.66%

Table 5.2: Results obtained by running full evaluation using the proposed biometric system. S1 refers to the October (first) session, while S2 refers to the March (second) session. The first two columns reflect from which session the corresponding dataset originates. Average HTER is obtained by averaging errors for every trained model, and the standard deviation reflects the variance of the performance for different user-attacker pairs.

Chapter 6

Conclusions

In this project, I investigated the performance of an ECG-based biometric authentication system. All parts of the system design were performed with equal importance, starting from data collection and finishing with two different evaluation approaches.

In Chapter 2, I provided a brief overview of the theory behind biometrics and what characteristics they need to possess in order to be applicable for access control systems. I then discussed how the physiology of the heart relates to the ECG signals and whether the electrical activity of human hearts constitutes a suitable biometric modality. This chapter finished with a literature review and a comparison of results from related studies.

I discussed the data collection procedure in Chapter 3. One of the contributions of this project was the creation of a dataset containing four 4-minute ECG traces of 53 subjects collected over a period of four months. As part of my future work on this project, I intend to properly format and make this dataset publicly available for other researchers working in this area. Furthermore, during the March data collection session, subjects were presented with an acoustic stimulus during one of their ECG recordings. This was done in order to elicit a small heart rate change that could be used to research presentation attack detection based on the heart rate variability. This question was not investigated as part of this project, but the availability of the data opens opportunities for future work in this area. Similarly, I did not have the opportunity to compare the performance of my proposed system on the CYBHi dataset [34]. Nevertheless, this can be done in order to promote the use of a standardised dataset for ECG-based biometrics research.

The proposed system design, described in Chapter 4, consists of several parts. Even though I decided to use the signal filtering capabilities provided by the ECG monitor, I implemented the signal segmentation procedure manually. As part of the future work on this project, I will explore established methods for R-wave peak detection, for instance, the Pan-Tompkins algorithm [32]. In addition to that, while this project proposed a basic way to handle outliers, there are more advanced statistical approaches that can yield better results.

Chapter 4 also focused on performing biometric template matching using well-established

methodologies from the machine-learning field. As a first step, I performed feature standardisation using z-scores and transformed the features into a lower-dimensional space using principal component analysis. The newly-created features formed the biometric template that was used for authentication. I ran experiments with several classifiers, including logistic regression, k-nearest neighbours and support vector machines, and optimised the hyperparameters of the models using grid search with 5-fold cross validation. By comparing the results obtained from these experiments, I established that support vector machines provide the best performance for template matching in the proposed system, achieving a cross-validation EER score of 5.74% for the best hyperparameter settings.

I focused my attention on proper evaluation of the system in Chapter 5. First, I discussed how improper training data selection can lead to incorrect performance assessments. After establishing the best practices for biometric system evaluation, I performed two performance measurements.

The standard evaluation discussed in Section 5.3.1 reflects the assessment undertaken by most existing studies. When the training and test sets were taken from the same data acquisition session, the EER of the system was 3.2% and 2.4% for the first and second sessions, respectively. The results suggest that the long-term stability of ECG-based biometrics is worse, obtaining 9.7% and 6.5% error rates when training and testing on data from different sessions. We can now update Table 2.2 originally presented in Section 2.5 with the results obtained in this project. The updated summary is presented in Table 6.1.

Study	Subjects	Features	Matching	Results
Coutinho et al [11]	19	Non-Fiducial	Custom	**0.4%**
Silva et al. [12]	63	Partially-Fiducial	SVM	**1.0%** (Short-Term)
Present Work	49	Partially-Fiducial	SVM	**2.4%** (Short-Term)
Singh et al. [40]	126	Mixed	SVM	**3.4%**
Present Work	49	Partially-Fiducial	SVM	**6.5%** (Long-Term)
Silva et al. [12]	63	Partially-Fiducial	SVM	**9.1%** (Long-Term)
Falconi et al.† [2]	10	Fiducial	Custom	**9.8%**
Komeili et al. [24]	70	Mixed	SVM	**11.0%**
Carreiras et al.‡ [8]	63	Partially-Fiducial	KNN	**13.3%**

Table 6.1: Results from studies on biometric authentication. Third and fourth columns refer to feature selection and template matching classifier, respectively. EER is used as the performance metric (lower score indicates better performance).

† Paper does not provide EER results, thus a similar HTER metric is presented instead

‡ Results for the baseline model (main model used 12-lead ECG)

The results obtained in this project is comparable to previous studies, for instance, by Silva et al. [12] and demonstrates a high potential of using consumer-grade ECG sensors for short-term authentication. Improving the performance of ECG over longer periods of time could be done by synchronising the stored biometric with the new signal after each successful authentication. However, more research needs to be done in order to gain a proper understanding of how to use the ECG signal for long-term authentication.

The designated attacker evaluation performed worse than the standard approach. This was expected, as we do not model the attacker explicitly in the training set, as discussed in Section 5.2.2. We obtain average HTER scores of 6.2% and 3.5% for the first and second sessions, respectively. The error rates reach up to 30.0% when training and testing on data from different sessions. Nevertheless, we can also see that there is high variability in the individual HTER scores, with standard deviation of 17.4% and 16.7%, respectively. This means that, for some specific user-attacker pairs, the system performs far worse than the average, increasing the overall HTER score. As part of the future work on this project, I would focus on analysing the error distribution, in order to understand which user-attacker pairs present difficulties for the system and what changes need to be made in order to improve the overall performance.

The results presented in this report provide a positive perspective on ECG-based biometrics, by showing that individuals can be authenticated by using their ECG trace. This project has also confirmed the results of previous authors showing that the performance of ECG biometrics degrades over time. Nevertheless, it also provides opportunities to explore potential countermeasures. I also presented a unique evaluation of the proposed ECG-based biometric biometric system by collecting the results about pairwise variability of users (in the designated attacker evaluation). These results can be used to assist in our understanding of the errors that the system makes and how to improve the authentication performance in the most challenging cases.

Bibliography

[1] Foteini Agrafioti, Jiexin Gao, and Dimitrios Hatzinakos. Heart biometrics: Theory, methods and applications. In *Biometrics*, chapter 10. InTech, 2011.

[2] J. S. Arteaga-Falconi, H. Al Osman, and A. El Saddik. ECG authentication for mobile devices. *IEEE Transactions on Instrumentation and Measurement*, 65(3):591–600, March 2016.

[3] Euan A Ashley and Josef Niebauer. *Cardiology Explained*. Remedica, 2004.

[4] AspenCore. Passive band pass filter, (Accessed) 2018-07-04.

[5] L. Biel, O. Pettersson, L. Philipson, and P. Wide. ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3):808–812, Jun 2001.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[7] CardioSecur. Electrocardiogram (ECG) - electrocardiography, (Accessed) 2018-07-04.

[8] Carlos Carreiras, André Lourenço, Hugo Silva, Ana Fred, and Rui Ferreira. Evaluating template uniqueness in ECG biometrics. In Joaquim Filipe, Oleg Gusikhin, Kurosh Madani, and Jurek Sasiadek, editors, *Informatics in Control, Automation and Robotics*, pages 111–123, Cham, 2016. Springer International Publishing.

[9] Lei Chen. *Curse of Dimensionality*, pages 545–546. Springer US, Boston, MA, 2009.

[10] Anthony Atkielski (Wikimedia Commons). Electrocardiography, (Accessed) 2018-07-04.

[11] David Pereira Coutinho, Ana L. N. Fred, and Mário A. T. Figueiredo. ECG-based continuous authentication system using adaptive string matching. In *BIOSIGNALS*, 2011.

[12] H. P. da Silva, A. Fred, A. Loureno, and A. K. Jain. Finger ECG signal for user authentication: Usability and performance. In *2013 IEEE Sixth International*

Conference on Biometrics: Theory, Applications and Systems (BTAS), pages 1–8, Sept 2013.

[13] djmw. Principal component analysis, 2016.

[14] Simon Eberz, Nicola Paoletti, Marc Roeschlin, Andrea Patané, Marta Z. Kwiatkowska, and Ivan Martinovic. Broken hearted: How to attack ECG biometrics. In *NDSS*, 2017.

[15] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. Evaluating behavioral biometrics for continuous authentication: Challenges and metrics. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 386–399, New York, NY, USA, 2017. ACM.

[16] Afonso Eduardo, Helena Aidos, and Ana L. N. Fred. ECG-based biometrics using a deep autoencoder for feature learning - an empirical study on transferability. In *ICPRAM*, 2017.

[17] Mohamed Elgendi, Mirjam Jonkman, and Friso De Boer. Frequency bands effects on QRS detection. In *the 3rd International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS2010)*, pages 428–431, 01 2010.

[18] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[19] Boris Iglewicz and 1944 Hoaglin, David C. (David Caster). *How to detect and handle outliers*. Milwaukee, Wis. : ASQC Quality Press, 1993. Includes bibliographical references (p. 73-78) and index.

[20] iMotions. What is ECG and how does it work?, 2017.

[21] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. ECG to identify individuals. *Pattern Recogn.*, 38(1):133–142, January 2005.

[22] Irena Jekova, Vessela Krasteva, and Ramun Schmid. Human identification by cross-correlation and pattern matching of personalized heartbeat: Influence of ecg leads and reference database size. *Sensors*, 18(2):372, Jan 2018.

[23] Jeremy Jordan. Hyper-parameter tuning for machine learning models, 2017.

[24] M. Komeili, W. Louis, N. Armanfard, and D. Hatzinakos. On evaluating human recognition using electrocardiogram signals: From rest to exercise. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4, May 2016.

[25] M. Kyoso and A. Uchiyama. Development of an ECG identification system. In *2001 Conference Proceedings of the 23rd Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society, volume 4, pages 3721–3723 vol.4, 2001.

[26] André Lourenço, Ana Priscila Alves, Carlos Carreiras, Rui Policarpo Duarte, and Ana Fred. Cardiowheel: ECG biometrics on the steering wheel. In Albert Bifet, Michael May, Bianca Zadrozny, Ricard Gavalda, Dino Pedreschi, Francesco Bonchi, Jaime Cardoso, and Myra Spiliopoulou, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 267–270, Cham, 2015. Springer International Publishing.

[27] MathWorks. R wave detection in the ECG, (Accessed) 2018-07-04.

[28] G. B. Moody and R. G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, May 2001.

[29] A.P. Nemirko and T.S Lugovaya. Biometric human identification based on electrocardiogram. In *XII-th Russian Conference on Mathematical Methods of Pattern Recognition*, pages 387–390. MAKS Press, 2005.

[30] I. Odinaka, P. H. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh. ECG biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, Dec 2012.

[31] Bruno A. Olshausen. Aliasing, 2000.

[32] J. Pan and W. J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, March 1985.

[33] João Ribeiro Pinto, Jaime S. Cardoso, André Loureno, and Carlos Carreiras. Towards a continuous biometric system based on ECG signals acquired on the steering wheel. In *Sensors*, 2017.

[34] Hugo Plcido da Silva, Andre Lourenco, Ana Fred, Nuno Raposo, and Marta Aires-de Sousa. Check your biosignals here: A new dataset for off-the-person ECG biometrics. *Computer methods and programs in biomedicine*, 113, 12 2013.

[35] Robi Polikar. Fundamental concepts & an overview of the wavelet theory, (Accessed) 2018-07-04.

[36] quanlymc. Python package for maximal overlap discrete wavelet transform (modwt), 2016.

[37] Sebastian Raschka. About feature scaling and normalization and the effect of standardization for machine learning algorithms, 2014.

[38] Albert Ali. Salah. Machine learning for biometrics. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, chapter 26, pages 539–560. IGI Global, Oxford, 2010.

[39] Jeff Schneider. Cross validation, 1997.

[40] K. Singh, A. Singhvi, and V. Pathangay. Dry contact fingertip ECG-based authentication system using time, frequency domain features and support vector

machine. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 526–529, Aug 2015.

[41] Peter Stavroulakis and Mark Stamp. *Handbook of Information and Communication Security*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[42] Fahim Sufi, Ibrahim Khalil, and Jiankun Hu. ECG-based authentication. In *Handbook of Information and Communication Security*, pages 309–331. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[43] The Johns Hopkins University. Basic anatomy of the heart, (Accessed) 2018-07-04.

[44] Paul van Gent. Analyzing a discrete heart rate signal using python, 2016.

[45] Naser Zaeri. Minutiae-based fingerprint extraction and recognition. In Jucheng Yang, editor, *Biometrics*, chapter 03. InTech, Rijeka, 2011.

Appendix A

Participant Demographics

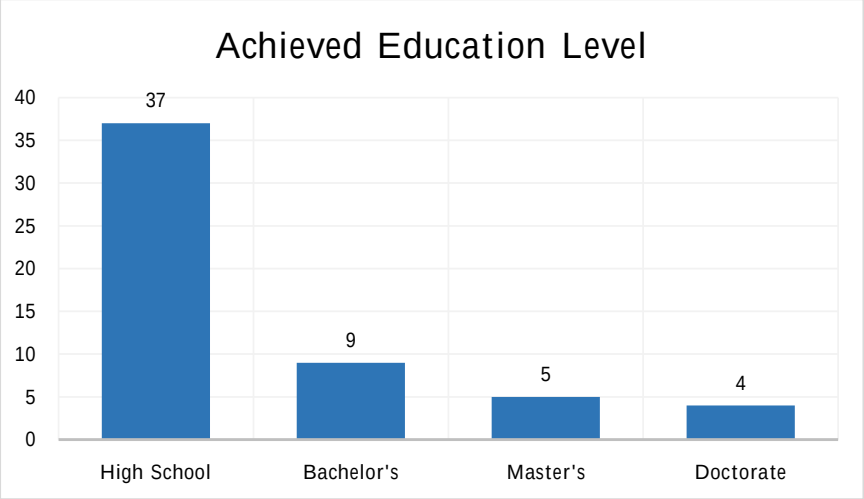


Figure A.1: Education level achieved by the respondents.

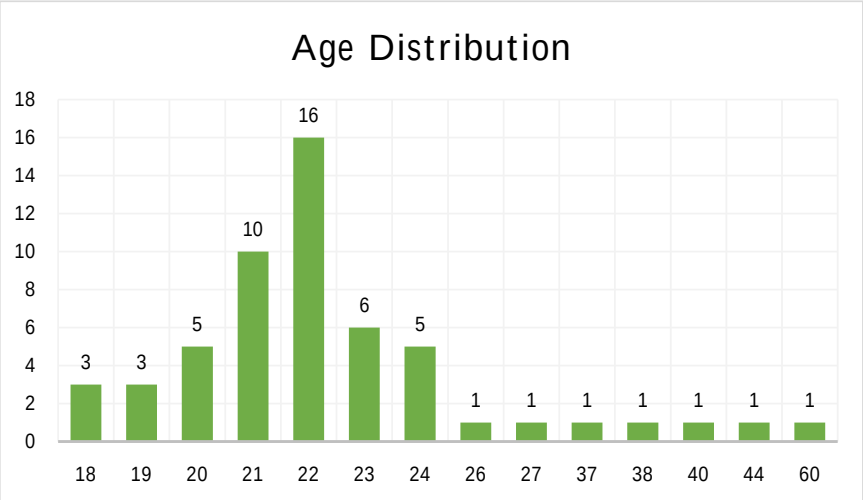


Figure A.2: Age distribution among the respondents.

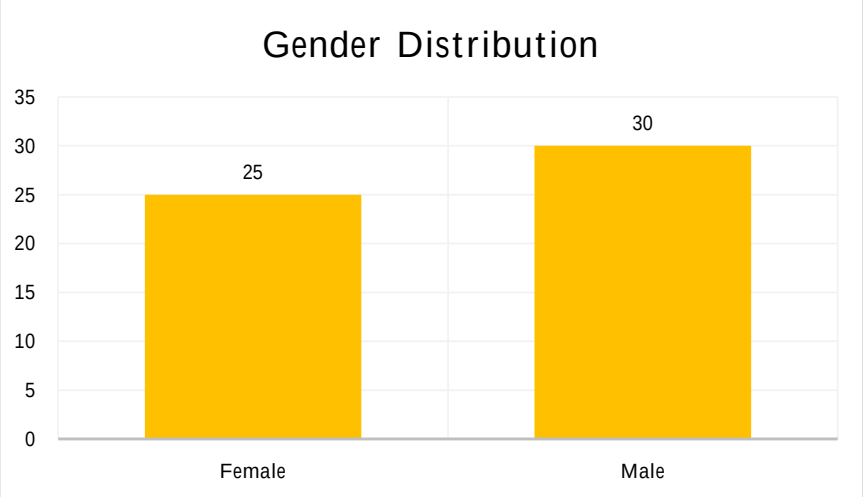
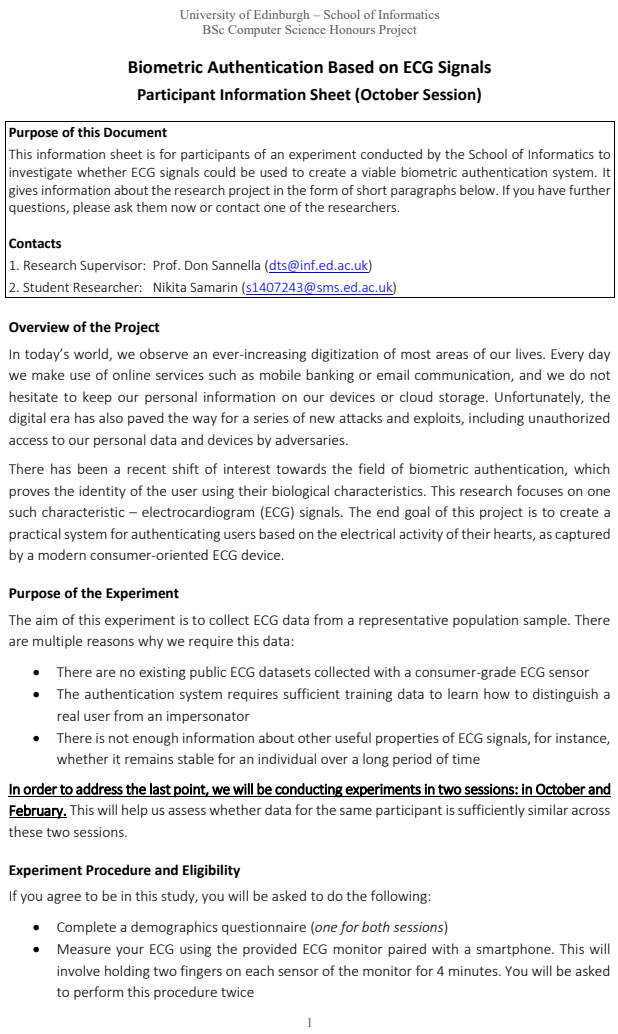


Figure A.3: Gender distribution among the respondents.

Appendix B

Document Samples



University of Edinburgh – School of Informatics
BSc Computer Science Honours Project

Biometric Authentication Based on ECG Signals

Participant Information Sheet (October Session)

Purpose of this Document

This information sheet is for participants of an experiment conducted by the School of Informatics to investigate whether ECG signals could be used to create a viable biometric authentication system. It gives information about the research project in the form of short paragraphs below. If you have further questions, please ask them now or contact one of the researchers.

Contacts

1. Research Supervisor: Prof. Don Sannella (dts@inf.ed.ac.uk)
2. Student Researcher: Nikita Samarin (s1407243@sms.ed.ac.uk)

Overview of the Project

In today's world, we observe an ever-increasing digitization of most areas of our lives. Every day we make use of online services such as mobile banking or email communication, and we do not hesitate to keep our personal information on our devices or cloud storage. Unfortunately, the digital era has also paved the way for a series of new attacks and exploits, including unauthorized access to our personal data and devices by adversaries.

There has been a recent shift of interest towards the field of biometric authentication, which proves the identity of the user using their biological characteristics. This research focuses on one such characteristic – electrocardiogram (ECG) signals. The end goal of this project is to create a practical system for authenticating users based on the electrical activity of their hearts, as captured by a modern consumer-oriented ECG device.

Purpose of the Experiment

The aim of this experiment is to collect ECG data from a representative population sample. There are multiple reasons why we require this data:

- There are no existing public ECG datasets collected with a consumer-grade ECG sensor
- The authentication system requires sufficient training data to learn how to distinguish a real user from an impersonator
- There is not enough information about other useful properties of ECG signals, for instance, whether it remains stable for an individual over a long period of time

In order to address the last point, we will be conducting experiments in two sessions: in October and February. This will help us assess whether data for the same participant is sufficiently similar across these two sessions.

Experiment Procedure and Eligibility

If you agree to be in this study, you will be asked to do the following:

- Complete a demographics questionnaire (*one for both sessions*)
- Measure your ECG using the provided ECG monitor paired with a smartphone. This will involve holding two fingers on each sensor of the monitor for 4 minutes. You will be asked to perform this procedure twice

1

Figure B.1: Participant information sheet for the October session (single side).

University of Edinburgh – School of Informatics
BSc Computer Science Honours Project

Biometric Authentication Based on ECG Signals
Participant Questionnaire (October Session)

1) What is your gender?
   - ☐ Male
   - ☐ Female
   - ☐ Other

2) What is your age?

   ____

3) What is the highest degree or level of school you have completed? (If currently enrolled, please indicate the highest degree you have *received*.)
   - ☐ Less than a high school diploma
   - ☐ High school degree or equivalent
   - ☐ Bachelor's degree
   - ☐ Master's degree
   - ☐ Doctorate (e.g. PhD) or another research or professional doctoral degree

4) Are you feeling well and rested today? Are you currently relaxed and calm?
   - ☐ Yes
   - ☐ No

   If you said 'No', please elaborate:

   _____

5) Assuming there are no unforeseen circumstances, are you available and willing to participate in this experiment again in February?
   - ☐ Yes
   - ☐ No

Participant ID Code: _____
(researcher only)

1

Figure B.2: Participant demographic questionnaire for the October session.