Security Analysis of Behavioural Biometrics for Continuous Authentication

Simon Eberz

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Acknowledgements

Personal

First and foremost, I would like to thank my parents for their unwavering support throughout my whole life. You gave me the chance to pursue my dreams no matter what and without your support I would not be where I am today.

I thank Allison for reminding me that life consists of more than papers, rowing and archery. It has been an incredible journey so far and I can't wait to start the next chapter of our lives together.

Thank you to all my fellow archers, rowers, MOEs and all of my friends in Oxford and around the world. You gave me the much needed counterbalance to work and make life so much more worthwhile.

I thank all my collaborators from inside and outside Oxford. All of you made the inevitable late-night paper polishing so much more enjoyable.

Thank you to my assessors Cas Cremers, Ivan Flechais, Andrew Martin and Gene Tsudik. Between my transfer, confirmation and viva you helped to make this thesis what it is today.

Last, but most definitely not least, I would like to thank my supervisors Ivan and Kasper. Throughout the years, you supported me in my good ideas and stopped me from pursuing the terrible ones. It is your guidance and motivation that kept me going through these four years and I'm looking forward to (hopefully) many future projects together.

Institutional

I thank the EPSRC, the Department of Computer Science, St Cross College and armasuisse for their generous financial support that allowed me to pursue this work.

Abstract

In recent years, behavioural biometrics have become increasingly popular, with many types of behaviour being explored for the purpose of user authentication. Some of the most common examples are keystroke dynamics, mouse movements, touchscreen inputs and human gait.

Unlike physiological biometrics (e.g., fingerprints), behavioural biometrics are often believed to be relatively hard for adversaries to collect, but nevertheless have been subject to active attacks, including presentation, signal injection and imitation attacks.

In this thesis, we take a holistic view on the design, evaluation and security analysis of behavioural biometric recognition systems. First, we underline their usefulness by designing a novel authentication system based on distinctive eye movement behaviour. We evaluate this system under different adversary models and show that eye movements can be used for both user authentication and judging a user's task familiarity. Drawing from insights gained from this project, we go beyond the state of the art to develop metrics and methodologies that more accurately reflect a system's real-world performance and security. This approach is centred around reflecting a biometric's systematic false negatives (i.e., attackers that consistently go undetected) more accurately.

A frequent focus of related work is how to present previously obtained biometric data to a behavioural authentication system (e.g., through imitation or mimicry attacks). However, the challenge of obtaining this data in the first place is far less explored. In this thesis, we perform a series of experiments to judge the usefulness of biometric data collected through a variety of sources. The idea is to measure the security impact of the plethora of biometric data that is involuntarily created through our day-to-day interactions with diverse systems.

Contents

List of Figures

List of Abbreviations

Acc . . . . . Accuracy
AUROC . . . . Area under ROC curve
AWG . . . . . Arbitrary Waveform Generator
BLE . . . . . Bluetooth Low Energy
CM . . . . . Confusion matrix
DR . . . . . Detection rate
ECG . . . . . Electrocardiography
EER . . . . . Equal error rate
FAR . . . . . False accept rate
FPR . . . . . False positive rate
FRR . . . . . False reject rate
GC . . . . . Gini Coefficient
HTER . . . . . Half-target error rate
HVS . . . . . Human Visual System
knn . . . . . k-nearest-neighbours
KS . . . . . Kolmogorov-Smirnov
MI . . . . . Mutual information
NCA . . . . . Nymi Companion App
NEA . . . . . Nymi Enabled Application
NN . . . . . Neural network
rbf . . . . . radial basis function
RF . . . . . Random forest
RMI . . . . . Relative mutual information
ROC . . . . . Receiver operating characteristic
SR . . . . . Success rate
SVM . . . . . Support vector machine
TPR . . . . . True positive rate

И вот оказалось, что только жизнь, похожая на жизнь окружающих и среди нее бесследно тонущая, есть жизнь настоящая, что счастье обособленное не есть счастье...

And so it turned out that only a life similar to the life of those around us, merging with it without a ripple, is genuine life, and that an unshared happiness is not happiness.

— Boris Leonidovich Pasternak, Doctor Zhivago

1 Introduction

Contents

1.1 Motivation

Passwords are currently the most common authentication mechanism, both to secure access to local systems (such as desktop computers or laptops) and web-based services. Passwords scale extremely well, are conceptually simple, have minimal requirements on the server side (both in terms of storage and computational power) and users are familiar with them. Despite their widespread use, passwords suffer from a number of security and usability problems. First and foremost, only strong (i.e., sufficiently long and random) passwords are secure against bruteforcing attacks. In practice, this is rarely the case as users tend to choose weak passwords even for high-value accounts [1]. Users also often have difficulty remembering passwords, especially when platforms enforce the use of requirements, such as special characters. This leads to users often writing down

passwords, which makes them vulnerable to local adversaries. Password reuse across different platforms and accounts is also a common problem [2], as the compromise of one account leads to the compromise of many others. Overall, most of the problems associated with passwords are user-related (i.e., the result of bad habits and security practices). Password managers have been proposed as a mitigation technique and allow the use of individual, strong passwords for all accounts. However, they result in a single point of failure if either the master password is compromised or cloud-based password managers turn out to be malicious. As a result, a fundamental solution to the shortcomings of passwords remains elusive.

Biometric recognition is a promising approach to solve some of these limitations and mitigate security concerns of passwords in many environments. In recent years, the push towards biometrics has been particularly driven by the increasing number of smartphones that support unlocking using biometric modalities. This is most commonly achieved through fingerprint scanning or face recognition. Both of these modalities are examples of physiological biometrics, i.e., biometrics that use distinctive physical features to distinguish users. The main concern with these types of biometrics is that they are easy to observe and once they are compromised they can not be changed or revoked. For fingerprints, it is possible to lift them off smooth surfaces (such as a coffee cup or smartphone screen). However, with the increasing availability of high-resolution cameras, photographing them from a distance has become an additional threat vector. In 2016, hackers successfully obtained the fingerprints of German minister of defence Ursula von der Leyen, using only a few high-definition photographs [3]. Since it is relatively easy to create fake fingers out of materials such as latex or wood, this compromises their security.

While the best-known and most widely used biometrics are physiological, their behavioural counterparts recently experienced significant attention. Examples of behavioural biometrics are keystroke dynamics (distinctive typing patterns), touch dynamics (characteristic touchscreen inputs), gait, eye movements and others. One key advantage is that they are significantly harder to observe as they do not generally leave physical traces that can be lifted or photographed. In addition, their time-varying nature means that any samples an adversary may have been able to obtain will constitute a diminishing threat

over time. In addition, behavioural information can often be collected continuously during system operation (e.g., mouse movements or keystroke dynamics) without requiring any explicit actions on the user's part. This unobtrusiveness is what enables *continuous authentication*. Continuous authentication is an approach to establish the user's identity not just once (e.g., during login time), but continuously[1] while the user is operating the system. This technique has the benefit of also detecting a change in user identity after the initial login which can occur when the user leaves the system unlocked and unsupervised. In addition, continuous authentication prevents a user from unlocking the system for someone else, as all actions following initial authentication still have to be carried out by the user. This is a particularly useful property in the context of insider threats.

The goal of this thesis is to provide a systematic analysis of the security provided by behavioural biometrics. As such, we focus on the design, evaluation and security analysis of behavioural biometrics. We have identified three main components for this thesis: (a) design of a biometrics-based continuous authentication system, (b) evaluation methodologies and (c) security analysis against active attacks. Our research goals are as follows:

- **Highlight the challenges of behavioural biometrics through the design of an authentication system based on eye movement patterns**
- **Investigate common limitations in the evaluation of biometric recognition systems**
- **Evaluate the vulnerability of deployed biometric recognition systems against realistic adversaries**
- **Develop and evaluate approaches to measure and improve the resilience of biometric features to active attacks**

[1] In practice, it is only possible to periodically, rather than continuously, confirm the user's identity.

1.2 Contributions of our Research

This section explains the contributions of our published work, which provides the foundation of this thesis. While all publications are joint work with others, I have only included published work where I was the first author. As such, I was the main contributor to the projects, although I had help from my co-authors in conducting the experiments in [7, 8] and preparing the publications.

- Drawing from insights gained from the medical and neuroscience domains, we developed an authentication system based on eye movement biometrics. The results of this study are published at the *2015 Network and Distributed Systems Symposium (NDSS)* [4]. An extension of this work exploring the influence of task selection, classifier and data quality was published in the *ACM Transactions on Privacy and Security (TOPS)* [5]. We elaborate on the foundation of this biometric and resulting security implications in Chapter 3.

- Chapter 4 discusses limitations of evaluation methodologies frequently used in related work. These limitations mainly relate to the use of error metrics and the approach used to simulate system operation on static datasets. Going beyond the state of the art, we investigate the effects of skewed error distributions and the influence of training data selection and attacker modelling on error rates. The results of the study were published at the *2017 ACM ASIA Symposium on Computer and Communications Security (ASIACCS)* [6].

- In Chapter 5 we provide an in-depth security analysis of ECG biometrics and use our insights to develop a cross-device attack against the Nymi Band. The attack relies on capturing the victim's ECG data through a variety of sources and creating a forged signal with a low-cost audio player. The results of this study were published at the *2017 Network and Distributed Systems Symposium (NDSS)* [7].

- Chapter 6 generalizes our previous cross-device attack to a cross-context attack against a variety of biometric modalities. This work was published at the *2018 IEEE Symposium on Security and Privacy (S&P)* [8].

1.3 Scope of this Thesis

The main focus of this thesis is the design and security evaluation of behavioural biometrics in the context of continuous authentication. While the line between physiological and behavioural biometrics is often blurred (as outlined in Chapters 3 and 5), we will focus on time-varying signals. As such, we do not consider authentication mechanisms more commonly used for one-off authentication, such as fingerprint scanning. The reasoning is that continued sensing of these biometrics (such as repeatedly asking the user to scan a fingerprint) are hardly practical.

Usability improvements are often cited as an advantage of biometric recognition. While we collect user feedback throughout the numerous experiments that form this thesis, rigorous usability-focused user studies are not the focus of this work.

1.4 Ethical Considerations

We acknowledge that this work raises two major ethical concerns. Firstly, the collection of biometric data through user studies is of potential concern, as the data could in principle be used to impersonate the user at any system that uses this particular modality. We mainly address this through anonymisation of the data, which makes it much more difficult to match the biometric data to individual users.

The second concern is that some of the data we collect has uses outside of authentication. This is mainly the case for eye movement behaviour (Chapter 3) and Electrocardiography (ECG, Chapter 5). Specific eye movement patterns have been linked to disorders such as Alzheimer's and schizophrenia. However, diagnosis of these disorders requires specific controlled stimuli which are not used in our experimental design. ECG is widely used to diagnose heart conditions, which raises significant privacy concerns. Collecting ECG data allows, in principle, to diagnose conditions that the experiment participants may not even be aware of. The type of measurement devices, coupled with the fact that none of the researchers involved are medical professionals, led us to the decision of not attempting any medical diagnosis during the collection process. In line with this policy we disabled any diagnostic capability of the devices used. In addition, we

anonymise the data before publishing to make it impossible to link any possible future diagnosis to individual users. All study participants are informed of these concerns and practices during the informed consent collected before enrolment in the study.

As required for all experiments involving human participants, we have sought approval from Oxford's Central University Research Ethics Committee (CUREC). Approval has been granted under reference numbers SSD/CUREC1/13-064, SSH_C1A_15_118, R50977/RE001, R50977/RE002, R42894/RE001.

1.5 Outline

The thesis is structured as follows:

- **Chapter 2** places the thesis in the context of related work. We will discuss the state of the art of both biometric system design and published attacks on these systems. This chapter focuses specifically on biometric modalities relevant for the following chapters.

- **Chapter 3** describes the design and evaluation of an authentication system based on eye movement patterns. We design and evaluate distinctive biometric features based on neuroscientific research. Eye movement patterns enable us to both perform transparent continuous user authentication and judge a user's task familiarity.

- **Chapter 4** highlights several limitations of state-of-the art practices used to evaluate biometric recognition systems. Particularly, we outline how widely used metrics are insufficient to quantify the problem of systematic false negatives. We also discuss the implications and pitfalls of evaluating the performance of biometric recognition on a static dataset.

- **Chapter 5** evaluates the security of biometrics against sophisticated adversaries using the example of Electrocardiography (ECG). We obtain data for a cross-device signal injection attack through a variety of sources. The attack is instantiated against the Nymi Band, a wristband marketed to end users that serves as an ECG-based multi-factor authenticator.

- **Chapter 6** generalises the cross-device attack presented in Chapter 5. This generalisation allows us to judge the resilience of individual features against active attacks and enables the design of more secure authentication hardware and features.
- **Chapter 7** summarizes the results, discusses the future work required to more comprehensively secure biometric recognition, and concludes this thesis.

If I have seen further it is by standing on the shoulders of Giants.

— Isaac Newton

2 Related Work

Contents

In this chapter, we will place the thesis in the wider context of related work. The chapter will be split between work describing the design and evaluation of biometric systems and work discussing active attacks.

2.1 Behavioural Biometrics

Over the years, the suitability of many types of behaviour has been investigated for continuous authentication. In this section, we will focus on six biometrics that are of particular interest for the remainder of this thesis: keystroke dynamics, mouse movements, touch dynamics, gait, eye movements and ECG.

2.1.1 Keystroke Dynamics

Keystroke dynamics, the use of distinctive typing patterns (rather than *what* is typed), is possibly the oldest instance of behavioural biometrics. The earliest work by Gaines et al. dates back to 1980 [9]. During this experiment seven professional typists were asked to type an identical paragraph while the timings between consecutive keystrokes were recorded. This experiment was repeated four months later, with the same participants and identical texts. The basis for the analysis are the times taken by the typist to type a specific digraph (two specific consecutive letters, such as ”er”). This metric was computed for all typists and for all digraphs that commonly appeared in the text. Out of all timings for each typist the authors computed the mean, variance, skewness and kurtosis. The authentication decision was then made using the t-test, testing the hypothesis that both samples for the same digraph possess identical variance but different means. Initially all digraphs were used, in order to reduce the system's error rates the entire set was then narrowed down to five digraphs. Why those digraphs were particularly distinctive or if this property is limited to that set of subjects remained unclear.

These initial promising results sparked a large interest in the exploration of keystroke dynamics. Gaines et al. used the mean digraph latency as the only feature. In future work this remained the dominant mode of distinction, however multiple additional features have been used, such as key hold times [10], key pressure [11] and n-graphs [12, 13]. Most proposed systems still use fairly simple statistical hypothesis tests to perform the authentication decision. In recent years, machine learning techniques such as neural networks and support vector machines have become more common [14]. Another important distinction between different papers lies in the experimental design, specifically the use of restricted or free typing. In restricted typing the subject copies a text determined by the researchers, in free typing they are usually given a writing prompt but are free in what they choose to write. Typically, classification on free text is more challenging as the prevalence of specific digraphs and n-graphs can not be controlled and many unpredictable pauses are introduced into the typing process (such as subjects pondering over spelling or what to write). Gunetti et al. investigated the impact of different languages on keystroke

dynamics [13]. Surprisingly, their results showed that a subject's keystroke dynamics are largely independent from the language the text is written in.

The error rates of keystroke dynamics vary wildly depending on features, classifiers and (possibly most importantly) the experimental design. Most systems achieve average EERs of around 5% while EERs as low as 0.5% have been reported [12].

2.1.2 Mouse Movements

Mouse movement biometrics capture the distinctive properties of mouse events, such as movements and clicks. Similar to keystroke dynamics, they are an interesting candidate biometric as mouse movements are usually part of interacting with any desktop computer. In this context, we consider both conventional mice and mouse-like input devices such as track-pads.

The most commonly used features make use of the shape of mouse strokes (strokes can be identified as the movements between two consecutive clicks or through a predetermined duration) and the properties of clicks. Stroke shape features capture the length, curvature, speed and acceleration of strokes while the most commonly used property of clicks is the click duration (i.e., the time between a click-down and a click-up event). Depending on the hardware and driver, the click duration is not always reported for track-pads.

Gamboa et al. propose a web-based authentication system that makes use of mouse movement biometrics [15]. During her first visit a user would authenticate using conventional credentials (e.g., username and password) and is then asked to spend some time on the website. This stage acts as an enrolment phase, during which the user's mouse movement behaviour is learned. After her first visit, the user's behaviour is continuously validated against the stored template. In order to evaluate their system the authors ask 50 volunteers to play a memory game for about 10-15 minutes while their mouse movements are recorded. The reasoning behind this experimental design is that it generates a high number of strokes in a short time (as clicks are required to turn individual tiles), thus simplifying the authentication procedure. However, the frequency of strokes might be much lower in a real-world environment where clicks are less frequent (e.g., during web browsing a considerable amount of time might be spent reading or scrolling). Depending

on the number of strokes used for the decision EERs between 48% (for a single stroke) and 0.2% (for 200 strokes) are achieved.

Zheng et al. argue that many movement-based features might be dependent on different hardware and system parameters (such as screen resolution) [16]. While this is certainly true, the authors don't give reasons why automatically scaling distance-based features to different resolutions would not be practical. They propose to exclusively use angle-based features such as direction angle, curvature distance and angle of curvature. The advantage of angle-based features is that they are independent from the screen resolution or screen size. Using this limited feature set a FAR of between 4.57% and 0.86% is achieved while the FRR varies between 18.79% and 2.96%.

In order to address the somewhat artificial setting of previous studies Nakkabi et al. use data collected on the subjects' machines [17]. They provide the lowest error rates published so far, at a FAR of 0.36% and a FAR of 0%. However, at least a part of the distinctiveness of features may be due to differences in hardware and systems, rather than users [18]. Additionally the uncontrolled data collection raises concerns regarding the distinctiveness of the biometric. As users are not observed during data collection there is no guarantee the system is always used by the same user (thus potentially underestimating distinctiveness), however there might also be positive influences due to user-chosen tasks that would not be present under a real-world threat model.

2.1.3 Touch Dynamics

Touchscreen-enabled devices, such as smartphones and tablets, have become increasingly common in recent years. At the same time many users choose not to enable security mechanisms such as PINs and unlock patterns for convenience reasons. Even when unlock patterns are used they might be susceptible to so-called smudge attacks, which attempt to recover the pattern based on oily residues left on the screen [19].

In order to provide additional security without creating inconvenience for the users, the use of touchscreen input as a behavioural biometric has been proposed. Initial work has focussed on using touch-gestures to perform authentication [20, 21]. These approaches augment (rather than replace) traditional pattern-based authentication by not

only checking which pattern is drawn by the user but also *how* it is drawn. As such they provide a second layer of defence once the attacker has managed to obtain the unlock pattern (e.g., through shouldersurfing, coercion or the smudge attack described above). However, they don't provide further protection once the device is unlocked and they don't solve the usability problem created by the unlock pattern itself.

Unlike the previous approaches Frank et al. propose the use of a user-transparent continuous authentication system that does not require specific user actions (or even user knowledge) [22]. Some of the 41 features are conceptually similar to those used in mouse movement biometrics and capture velocity, acceleration and curvature of swipes. A swipe in this case can be either horizontal (e.g., when swiping through photos) or vertical (e.g., when reading a long text or e-mail). In terms of system operation it does not matter where on the screen a swipe is performed (with the potential exception of some games). Consequently the start and stop coordinates of a swipe might form a behavioural feature (although it is likely more a changeable habit), and it proves to be quite distinctive. In addition a touchscreen allows to harvest the touch pressure and the area covered (although it has been shown that on many low-cost phones these properties are derived from the same sensor [23], suggesting highly correlated information). The experiment is designed to collect a large number of both horizontal and vertical swipes. Horizontal swipes are collected in an image comparison game that requires the user to repeatedly switch between two images in order to determine their differences. The other task involves reading a long wikipedia article, thus prompting frequent scrolling (i.e., vertical swipes). In order to evaluate the time stability of features multiple repetitions were performed. Depending on the time distance between training and testing phase an EER between 0% (intra-session) and 4% (over two weeks) was achieved.

Frank et al.'s work was based on a controlled lab experiment carefully designed to induce many swipes over a short time. This behaviour is not necessarily common in the real world and day-to-day use of a device might introduce a number of swipes not observed in Frank et al.'s experiments. In order to address these concerns Feng et al. implemented a similar system while collecting data on users' own phones [24]. Their results show that touchscreen-based authentication is more challenging in uncontrolled

environments, yielding a FAR of 9% and a FRR of 7%. In addition, the system also consumes an average of 88mV, amounting to up to 6.3% of the device's entire energy consumption. As with mouse movement biometrics the uncontrolled experiment design raises concerns with regard to the influence of user-chosen tasks.

The arguably biggest accuracy improvement to touch dynamics was proposed by Bo et al., who combine touch dynamics with accelerometer data to capture the reaction of the device (e.g., vibration) to the touch event [25]. The authors consider two scenarios: a stable scenario while the user is stationary and a mobile scenario in which the device is used while walking. In both cases, the touch features are similar to previous work. In the stationary setting, the touch features are augmented with accelerometer and gyroscope data. The former consists of the amplitude of vibration measured through the accelerometer and represented by the summation of acceleration vector. The latter measures the angular velocity obtained from the device's gyroscope. As such, it gives a measure of the device's variation in space when touched by the user. In the mobile scenario, these features are likely suppressed due to the much larger scale disturbance introduced by movement of the entire body. To still provide reliable authentication, the authors use accelerometer data to identify the user's gait. After segmenting the accelerometer's time series into individual steps, the authors extract (1) Vertical displacement of each step, (2) Current step frequency and (3) Mean horizontal acceleration for each step. In the static scenario, a 20% EER is achieved with a single action, although this drops to $<$1% after 12 actions.

2.1.4 Gait

Human gait (i.e., the way people walk) has become a popular biometric in recent years, fuelled by the increasing availability of accelerometers in smartphones and smartwatches.

There are two main approaches to capture a person's gait: video recordings or through an accelerometer. Video-based gait tracking [26, 27] is most commonly used to authenticate or identify people in buildings or some public spaces using stationary cameras. With the increased prevalence of accelerometers contained in smartphones the biometric has been applied as a theft detection mechanism by locking the device when fresh gait samples do not match the owner's template [28–30]. Many users typically carry

their smartphone at all times, making gait detection based on the phone's accelerometer an interesting approach for continuous authentication. Conversely, continuous (rather than periodic) video-based monitoring of a person's gait does not seem practical.

The main difference in experiments evaluating accelerometer-based gait tracking lies in the type of the sensor and, more importantly, its position. Early work by Gafurov et al. used an AVR Butterfly evaluation board equipped with two accelerometers [29]. The device samples at a rate of 16Hz and was firmly strapped to the subjects' legs. The authors argue that this position is likely to yield the least noisy samples. As a basis for authentication the authors compute the *combined acceleration signal* which measures the alignment of the resultant acceleration to the sideways-axis. During the training (or enrolment) phase a normalized histogram of the combined acceleration signal is computed. The same is done during the testing phase, the (bin-wise) distance between the histograms yields a score that is compared against a threshold (which defines the trade-off between FAR and FRR). While Gafurov et al. only collect a single sample during the testing phase this approach can generally be extended to continuous authentication by using a single step or a fixed time-window as a sample.

2.1.5 Eye Movements

Eye movements have previously been studied as an input channel that is resistant to shouldersurfing attacks. These systems still rely on a conventional PIN, a password or a passphrase. The authors of [31] developed a system using a Tobii 1750 gazetracker and report a password entry time of 9 to 12 seconds with error rates between 3 and 15%. Similar work used eye gestures instead of passwords and reduced the fraction of successful shouldersurfing attacks to 55% with an average input time of 5.3 seconds [32].

One of the earliest works using eye movements as a biometric was published in 2005 [33]. The authors use gaze velocity and the distance between pupils as features and achieve identification rates of up to 92%. However, the error rates rapidly increase without relying on the pupil distance, a feature more commonly associated with face recognition than with eye tracking. Kinnunen et al. use a Tobii X120 gazetracker with a sampling rate of 120 Hz to capture a subject's eye movements while she is watching

a movie and use short-term eye gaze direction to construct feature vectors which are modeled using Gaussian mixtures [34]. Depending on the amount of training data, an EER of 28.7 to 47.1% is reported. The authors do not state whether the type of video affects the templates (e.g., whether training and testing with different videos is possible). A different approach by Cantoni et al. attempts to distinguish individuals by the way they look at different images [35]. Unlike our work, this approach requires display of controlled stimuli, as such it is more suited for one-time authentication (e.g., to replace a password to login) rather than continuous authentication. Using density and duration of fixations as their main features they report an EER of 27.06%. Similarly, Liang et al. measure the eye's tracking behaviour when a moving stimulus is displayed [36]. They use the acceleration of eye movements while the subjects are pursuing a moving shape as input to both Support Vector Machines (SVM) and a Back-Propagation neural network. In an experiment with five subjects, they achieve an identification accuracy of 82%.

More recently, Sluganovic et al. presented an eye movement-based login mechanism that is resilient to replay attacks [37]. For the login, the user is asked to look at a dot that moves to a different part of the screen once the user's gaze gets sufficiently close. The replay attack protection is achieved by comparing the number of successfully "gazed" points to a predefined threshold. Since the position of points is randomised for each login, the positions of replayed gaze points is unlikely to match those of the presented stimuli. The system achieved a 6.3% EER with an authentication time of 5 seconds.

2.1.6 ECG

Electrocardiography (ECG) measures the electrical activity of the heart over time through electrodes placed on the subject's body. Among other things, an ECG can be used to measure the rate and rhythm of heartbeats, the size and position of the heart chambers, the presence of any damage to the heart's muscle cells or conduction system, the effects of cardiac drugs, and the function of implanted pacemakers [38]. Differences in ECG are a result of the activity of the heart. Therefore, ECG is not generally considered a behavioural biometric. Nevertheless, it provides a time-varying series of measurements and is therefore still of interest for this thesis.

Zhang et al. propose to use ECG to securely generate shared secrets between implanted medical devices in the context of Body Area Networks (BANs) [39]. This approach is interesting as its energy consumption is lower than traditional approaches (such as the Diffie-Hellman key exchange) and the devices may have to measure ECG for medical purposes anyway.

ECG has also been used for continuous authentication and research on this application has recently resulted in a commercial product[1], the Nymi Band. The Nymi band does not perform continuous authentication but verifies the wearer's identity the first time it is put on. Once authenticated it allows to authenticate the user towards cars or gates (replacing keys) or adjust settings in a smart home. Most earlier work focusses on the time domain of the ECG signal by detecting the PQRST signature [40, 41]. The four components of this signature (P, QRS and T) signify voltage changes caused by atrial depolarization, ventricular depolarization and ventricular repolarization. Based on these components the duration of each individual wave, as well as the amplitudes, have been used as features. As the absolute amplitude of the waves represents a voltage reading at the surface of a person's skin it is inherently unstable as water, sweat or skin oils can change the skin's conductivity and cause dramatic changes in the reading. To mitigate this issue the relative amplitude between different waves is often used instead.

Besides the above time-domain features, Wavelet Decomposition [42] and Fourier Transform have been used to generate frequency-domain features for the purpose of authentication. Most ECG-based systems are independent of the subject's heart rate (which frequently changes due to a number of factors, including excitement, stress or physical exercise). There has also been work on anonymising ECG data [43] such that it can be distributed over the public internet. However, research on practical attacks has been limited.

2.2 Active Attacks

The vast majority of papers presented in Section 2.1 assume a zero-effort threat model. Typically, this means that for each user in the evaluation dataset, the data of each remaining

[1] https://www.nymi.com/

user represents the ”attacker“. Beyond this model, there are two types of active attacks: imitation attacks and presentation attacks. Presentation attacks relate to physiological biometrics (such as fingerprints) and involve crafting a physical artefact representing the biometric trait. For fingerprints, this is usually a fake finger made from materials such as wood or latex. Imitation (or mimicry) attacks relate to behavioural biometrics and involve an attacker that attempts to mimic the victim's behaviour. As behavioural biometrics are the main theme of this thesis, we will focus on imitation attacks for the remainder of this section.

Keystroke dynamics. Serwadda et al. demonstrate that the typical zero-effort model underestimates the success rate of attackers even when no information about the victim is available [44]. The authors analyse the statistical distribution of and dependenceis between features to synthesise a feature vector that is more likely to be accepted than that of any random user (as would be the case for a zero-effort attack). Overall, the attack increases the EER by between 28.6% and 84.4% depending on the classifier used.

Conversely, Tey et al. develop an imitation attack under the assumption that the victim's feature vectors are known [45]. The authors develop a system that provides positive and negative feedback to the attacker regarding their closeness to the correct timings. While the attack could also be implemented through a USB dongle registering as a keyboard (as in [46]), the proposed manual approach may be more applicable in some environments (e.g., when video surveillance is used or the hardware is tamper-proof). The results are very promising and show that the attack can even be carried out by novice users with high reliability.

Touch dynamics. Authentication based on touch dynamics relies on a high-dimensional featureset. Frank et al. argue that it would be difficult for a human attacker to imitate all features simultaneously [22]. In addition, features such as acceleration may be difficult for a human to translate to an imitation attack. This intuition initially seemed to be supported by other researchers. Zhang et al. perform an experiment where three different attackers observe users during their authentication procedure [47]. Their results show that all attacker struggle to improve significantly from their baseline FAR. However, the analysis is limited both in terms of the number of attackers (3) and the information

available to them. More recently, Khan et al. have carried out a successful manual imitation attack using either shouldersurfing or an offline training phase [48]. The authors use a similar concept to Tey et al. [45] to train attackers in the offline attack. Instead of providing feedback regarding individual features, they show a target swipe or scroll signature and ask the attacker to retrace it. They achieve an attack success rate of 84% for the shouldersurfing attack and 86% for the offline attack. Overall, this work questions the security of touch dynamics against human imitation attacks. It is particularly troubling that shouldersurfing attacks (a relatively weak attacker model for local authentication) show such high success rates.

Serwadda et al. take a different approach by using a purpose-built Lego Mindstorm robot to automatically carry out swipes [49, 50]. Most features (e.g., speed and coordinates) can be adjusted individually while a single value is used for pressure and area covered. In the first version of the attack, all victims are attacked using a common imitated feature vector. While this attack shows a remarkable success rate, this might be a result of the baseline EER of the attacked system being much higher than those reported in related work. Intuitively, any single feature vector should not be accepted for a large number of users in a low-EER system. The authors extend their attack to a targeted version which assumes that the attacker knows the victim's template or has otherwise obtained a victim-specific feature vector. A similar approach using a more adaptable humanoid robot has been presented in [51].

Gait. There has been some limited success in imitating gait recognition. Gafurov et al. selected imposters with similar physical characteristics to the victim (e.g., height and weight) and tasked them to learn and imitate the victim [52]. While the system's EER increased from 6.7% to 16%, this is a relatively modest success. Mjaaland et al. conduct a limited-size study with a single victim and seven attackers. Despite training imposters over the course of six weeks, none of them could impersonate the victim with any reliability. The authors argue that "gait mimicking is a very difficult task, and that our physiological characteristics work against us when we try to change something as fundamental as the way we walk". Since the authors only used a single victim, it is

unclear if some inherent property (such as erratic gait) of the victim may have artificially made the attack more difficult.

Kumar et al. expanded on the previous efforts by giving the imposters access to a treadmill [53]. Use of the treadmill allows the imposters to more easily control fundamental gait characteristics such as step length, step width, speed and thigh lift. Their attack increases the baseline FAR of 5.8% to 43.66% while the FAR of eleven out of eighteen users increased to 70% or more.

Eye movements. Due to the involuntary nature of eye movements, attacks have focused on altering the pupil diameter. This is aided by the high distinctiveness of pupil-based features and their resulting high impact on the classification decision. Griswold-Steiner et al. demonstrate an attack on eye movement biometrics that uses changes in ambient light to cause a change in the attacker's pupil diameter [54]. The attack leads to a 50% increase in EER on average and a 500% increase in individual users. While the system design and features are modelled after [4], the authors use a Tobii X120 eye tracker. The X120 has a lower sampling rate than the tracker used in [4] (120Hz compared to 500Hz), which leads to less distinctive temporal features. This shift makes the pupil-based features more important and therefore the attack more likely to succeed. As this attack uses changes in ambient light, it could be easily detected by the authentication system. In fact, tracking ambient light would be required to prevent the distorting effect of (benign) changes in ambient light from affecting the authentication decision. However, using more focused light (as suggested by the authors) would defeat this countermeasure and require more sophisticated detection techniques.

You can't depend on your eyes when your imagination is out of focus

— Marc Twain

3

Eye Movements as a Biometric

Contents

3.1 Introduction

The goal of this chapter is to evaluate the effectiveness of using eye movement biometrics as a novel defence against the ”lunchtime attack“ by an insider threat. An insider threat in this context refers to a person with physical access to a workstation that she is not authorised to use (e.g., using a coworker's workstation while she is at lunch). As such, our system serves as a second line of defence after the workstation's primary mode of authentication (such as a password) has already been compromised. This means the attacker either obtains the authentication credentials (such as passwords or hardware tokens) or accesses the already unlocked system. In this work we consider both careless and colluding users. The first class of users merely fails to adequately secure their system (e.g., by failing to lock it when they are away) or falls victim to social engineering attacks. Conversely, colluding users actively seek to facilitate the attacker's access to the system. This effort is required as, unlike passwords, biometric features are not as easily shared with others. This scenario makes the attack notoriously difficult to defend against. Based on these scenarios, we propose a set of features that can be extracted from human eye movements and analyse their distinctiveness and robustness using a systematic experimental design.

The human eyes offer a rich feature space based on voluntary, involuntary, and reflexive eye movements. Traditionally, the analysis of eye movements has been used in the medical domain to facilitate diagnosis of different ocular and neuronal disorders. Eye tracking devices have become much cheaper within the last years and low-cost open-source hardware and software are available. Recent advances in video-based eye tracking technology makes eye tracking applicable to a conventional workplace as it does not require physical contact with the users (more detail on eye tracking is given in Section 3.2).

Our experimental design captures the unique characteristics of each user's eye movements as measured by the eye tracker. We also consider ways in which the attacker could use her position to gain inside information about the user and the system through observation or social engineering. We define metrics to measure this advance knowledge through eye movement data and determine whether it affects the authentication decision.

We consider three scenarios in particular: (i) no prior knowledge, i.e., no information advantage; (ii) knowledge gained through a description, e.g., the adversary is provided with a textual description by a colluding legitimate user; and (iii) knowledge gain through observation, e.g., by looking over the shoulder of a legitimate user performing a task (shoulder-surfing). In addition, we perform a set of experiments to investigate the effect of four different computer-based tasks on eye movement behaviour: (a) reading, (b) two different videos, (c) web browsing and (d) typing. We perform these experiments with 30 participants recruited from the general public. Each set of experiments is repeated twice: Once the same day to test inter-session time stability and once after two weeks to measure potential long-term feature degradation.

The core contributions of this chapter are as follows: We define a set of 20 biometric features and perform measurements that confirm that these features are suitable to perform transparent continuous user authentication. We use different metrics to measure the quality of these features and quantify the effects of increasing time distance on both feature groups and individual features. In order to evaluate whether the eye movement biometric can be used in conjunction with cheap consumer-level hardware, we also determine the impact that the reduced sampling rate of these devices has on both feature quality and the performance of a continuous authentication system. We also propose a novel approach to correct the user's pupil size measurements for distortions caused by different screen brightness.

3.2 Visual System Background

This section provides a brief introduction to the specifics of the human visual system (HVS) required to understand the rationale behind this work and discusses the eye-tracking and gaze tracking technologies and their respective applicability to the security domain. For a systematic overview of the HVS and eye-tracking related research see, e.g., [55].

The HVS has been part of neurophysiological research for many decades. The current understanding of the human brain includes considerable knowledge about the connections between the retina and the brain regions which are responsible for generating eye movements. The experimental design and feature definitions used in this work are

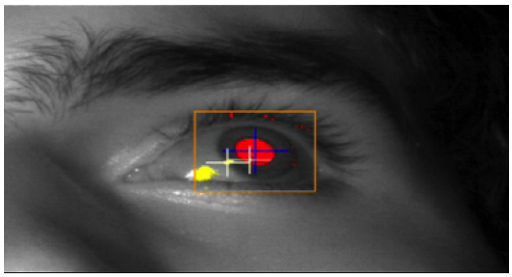


Figure 3.1: Video-based eye tracking: The gaze position is calculated using the distance between pupil position and the corneal reflections (shown as two white crosses).

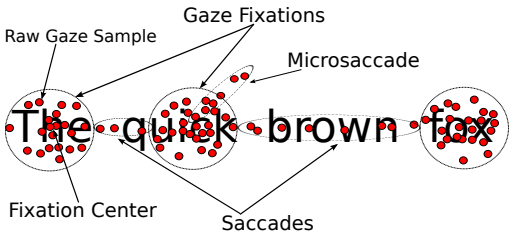


Figure 3.2: A simplified example of gaze tracking: the raw gaze samples are collected by the eye tracking device and subsequently clustered into fixations, saccades, and microsaccades.

inspired by neuroscientific insights related to fixational eye movements. These movements are a particular type of eye movement behaviour involved in processing static scenes which are typical for working with a desktop machine. Besides the neuroscientific foundations, eye tracking technology is of particular importance for this work. This technology has to enable capturing the eye movements of a person performing everyday tasks without posing significant usability disadvantages.

3.2.1 Characteristics of Eye Movements

The human eye moves within six degrees of freedom with six muscles responsible for the movement of the eyeball. The main types of eye movements used in perceiving a stationary object or scene, or reading a document can be categorized into *saccades* and *fixations*. The neural signals controlling these eye movements can be categorized as voluntary, involuntary and reflexive.

Saccades are rapid stepwise movements of both eyes in the same direction that typically last 10-100 ms, depending on the distance covered [55]. They are used to move the fovea[1] to another location. Once the saccadic movement has been signalized by the related neurons, the movement must be completed, i.e., neither the saccade's position nor its velocity can be consciously altered, even if the target has changed its position [56].

In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100-400 ms. They are used to stabilize the retina over a stationary object of interest. Yet, the eyes are never perfectly still, they make involuntary

[1] The fovea is a part of the retina that allows the central, high-resolution vision.

movements even during visual fixations. The main reason for such movements is to counteract retinal fatigue and to prevent visual fading, i.e., if a person attempts to artificially fixate eyes on an image by strongly focusing on a single fixation point, the image would start to fade away and the scene would become blank. One type of such movements are microsaccades, characterized by high velocity and acceleration often away from the fixation centre [57]. Related to microsaccades are movements called Saccadic Intrusions (SI), which consist of involuntary movements away from the previous eye position, followed by a return to that position after a short duration [58]. SIs are characterized through a high velocity and significantly higher amplitude compared to microsaccades. This terminology is visualized in Figure 3.2.

Conventionally, the studies of fixational eye movements have been concerned with medical diagnosis, such as Alzheimer's [59] and schizophrenia [60]. Yet, with an advance in eye-tracking technologies, analysing eye movements has proven to be valuable in many other areas, such as marketing (e.g., for analysing visual attention as a measure of effective advertising) [61, 62], human-computer-interface design [63], pilot training [64], or detecting fatigue and drowsiness in drivers [65–67].

Besides the eye movements, the pupil diameter is also an interesting feature which can be included in the analysis of eye behaviour. The *range* for this feature in a single subject is largely determined by eye physiology, gender and ethnicity and is relatively constant during adulthood [68]. Nevertheless, multiple causes that affect the pupil diameter have been found, including memory and cognitive workload [69], lighting conditions [70] and drug consumption [71]. The pupil size also shrinks as a person ages, an effect which is particularly pronounced in low lighting conditions [72].

3.2.2 Eye and Gaze Tracking Techniques

Eye tracking is the process of capturing a person's eye movements and measuring their positions. If the eye positions are calibrated with respect to an external display then the process is called gaze tracking. There are many types of eye tracking techniques, with the main trade-off between temporal/spacial accuracy vs. intrusiveness and usability. Traditional eye tracking techniques require either a head-mounted device or electrodes

attached to the subject's face. One such example is electrooculography (EOG), which is a technique for recording eye movements by measuring electric potential at the electrodes placed around the eyes. While this technique can be used to capture the eye movements even during sleep (e.g., to monitor REM sleep), its main disadvantage is the high intrusiveness since the electrodes must be attached to a person's face. As such, it is unsuitable for the office scenario assumed in this chapter.

The most widely used eye tracking technology today is video-based. Video-based eye tracking uses a camera which focuses on the pupils and records their movements and size. To improve the tracking accuracy, these devices typically use a source of controlled infrared or near-infrared light to create distinctive reflection patterns (see Figure 3.1). Importantly, current video-based eye tracking is non-invasive and remote, operating without any contact with the subject. The required hardware is only a standard webcam capable of recording infrared light as well as a source of infrared light. For example, the ITU Gaze Tracker is an open source project which offers eye tracking software that can be used by many low-cost webcams. Some smartphone manufacturers such as Samsung have also recently started to include basic eye tracking capabilities in their phones. While these low-cost solutions enable eye tracking at a reasonable accuracy, their sampling rate is still limited by the camera's frame rate. In order to capture microsaccades (which occur over a few milliseconds), a high-speed camera is required.

Given the increasing availability and simplicity of eye tracking, it is likely that the trend of using eye tracking outside of the medical and research domain will continue. The current non-invasive eye tracking technology already enables an easy access to a rich and distinctive feature space of fixational eye movements. Their distinctive capabilities and involuntary nature makes them a potentially valuable biometric.

3.3 Threat Model

The adversary model considered in this paper focuses on insider threats. A well known example of an insider threat is the so called "lunchtime attack" where an adversary temporarily gains access to a co-worker's workstation while the co-worker is away for lunch. Other examples include cleaning staff getting access to workstations after

hours, or the trivial case where one employee simply allows another employee to use her workstation or access credentials. In all these scenarios, an adversary might gain access to a fully operational system, already logged into a privileged account, and with access to everything that the legitimate user of the workstation would normally have access to. Any subsequent attack mounted from such a compromised workstation can be very hard to trace back to the real attacker. A 2011 study has shown that 33% of electronic crimes in companies are committed by insiders [73]. 60% of these attacks use a compromised account [74]. Account compromise is particularly difficult to detect as the account used to carry out the attack typically was not associated with suspicious activity before. Furthermore, it is more difficult to trace back the attack (and investigation may even put false blame on the victim). Most organisations allow their employees remote access (e.g., via SSH or a VPN connection), nevertheless 43% of attacks are performed locally using *physical access* to the workstation [74]. As such, defending against local attacks is the focus of this chapter.

In our model the adversary is aware of the gaze tracking system and will do her best to imitate the behaviour of the legitimate user. Due to the involuntary nature of fixational eye movements (see Section 3.2), we assume the attacker will not be able to directly modify the nature of her fixations and saccades. However, she can control where she looks at the screen on a larger scale. To this end, we consider an attacker that will attempt to copy the victim's "high-level" eye movement behaviour. To achieve this goal, she attempts to familiarise herself with the legitimate user's behaviour when using the system. This also allows her to complete tasks in a faster and more direct manner, thereby helping her to evade detection.

We consider two models of knowledge transfer to help the adversary familiarize herself with a system: (1) The adversary has gained knowledge about the system by reading (or being told) how the system works; and (2) the adversary has seen (e.g., by shouldersurfing) how a legitimate user operates the system.

We assume the adversary cannot disable the gaze tracking system, nor can she interfere with its operation in any way (such as covering up the eyetracker's camera), as doing so would quickly lead to the detection of her attack. We don't consider insider threats

which involve the attacker using her own workstation and credentials as, for all intents and purposes, the authentication system is working as intended in this case. These attacks can be traced back to the actual attacker much more easily and are better dealt with through behavioural monitoring [75]. The aim here is to show that gaze tracking is a viable way of identifying users, as well as gauge a user's level of knowledge and familiarity with a particular task.

3.4 Experimental Design and Data Collection

In this section, we give an overview of the two sets of experiments designed to test the feasibility of eye movement biometrics. The first experiment reflects the two classes of knowledge transfer outlined in the previous section (textual descriptions and shouldersurfing). The second set of experiments aims to extend eye movement authentication to a variety of everyday tasks.

3.4.1 Design Goals

The experiments described in this section are designed to test the feasibility of building an authentication system based on the distinctiveness of human eye movement patterns. Such a system should continuously monitor the user's eye movement behaviour in the background without requiring any modifications in the user's behaviour or even her knowledge or consent. In addition, the experiments should allow us to test whether eye movements reveal information about a user's task familiarity, i.e., whether eye movement behaviour changes significantly between familiar and unfamiliar users. This distinction could be used to detect outside attackers as they can be assumed to be significantly less familiar with the system they attempt to access than legitimate users (or inside attackers).

In order to design experiments that show whether or not gaze tracking is suitable as an authentication mechanism, we have to determine which tasks the test subjects should perform while they are being monitored. One option is to give them an entirely free environment in which the subjects can choose what to do for the duration of the experiment. This is probably the experiment that best captures actual user behaviour, but since it is likely that each subject will choose a different set of tasks, it is very hard

to guarantee that the distinguishing power of the resulting classifier is really capturing differences in users, rather than differences in behaviour or tasks. While we choose features such that their computation does not depend on specifics of the task, it is difficult to rule out that some differences in their *distributions* are due to the user-chosen task. As even the variations within a single task (such as different types of websites for a web browsing task) might already cause features to change the number of sub tasks to test would be enormous. If each user were to choose a different task, which possibly results in specific feature characteristics, this would lead to an overestimation of classification accuracy, as the classifier performs task distinction instead of user distinction. Conversely, a fixed task for all users means that any differences between the datasets are due to differences between users.

In order to overcome these sources of error, we define a specific set of tasks that all users must complete. Our goal is to determine whether the users' eye movements are distinguishable, even if they are completing the same task the same way with the same knowledge. If this is indeed the case, this means that there are *inherent* differences between users that can not be attributed to different ways of completing a single task. Nevertheless, we choose our features such that their *computation* does not make any assumptions about the task.

3.4.2 Knowledge Transfer Experiments

We first introduce a terminology to make it easier to refer to different parts of our interaction with test subjects, see Figure 3.4 for a visualization. We refer to one sitting of a test subject as a *session*. Two weeks after the first session, the test subject comes back for a second session. This is done to make sure our results are consistent over time. To verify that our results are not only consistent over longer periods but also across two subsequent sessions on the same day, our test subjects do a third session about an hour after completing session 2. All three sessions are identical, and each consists of three different *experiments*.

Each experiment has a similar structure. The test subject is initially presented an empty screen with a grey background. Once the experiment begins, a red dot with a

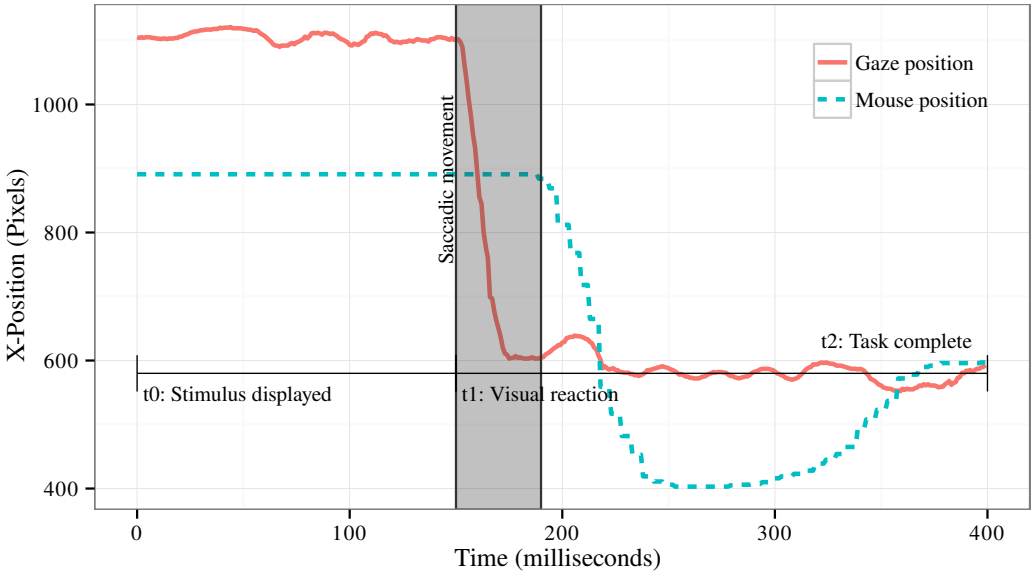


Figure 3.3: At time t_0 a new dot appears followed by a period of inactivity (reaction time) in which neither the gaze, nor the cursor move significantly. After about 150 ms, at t_1, a visual reaction in the form of a large saccade occurs (the grey area) and the gaze and cursor converges to the position of the stimulus.).

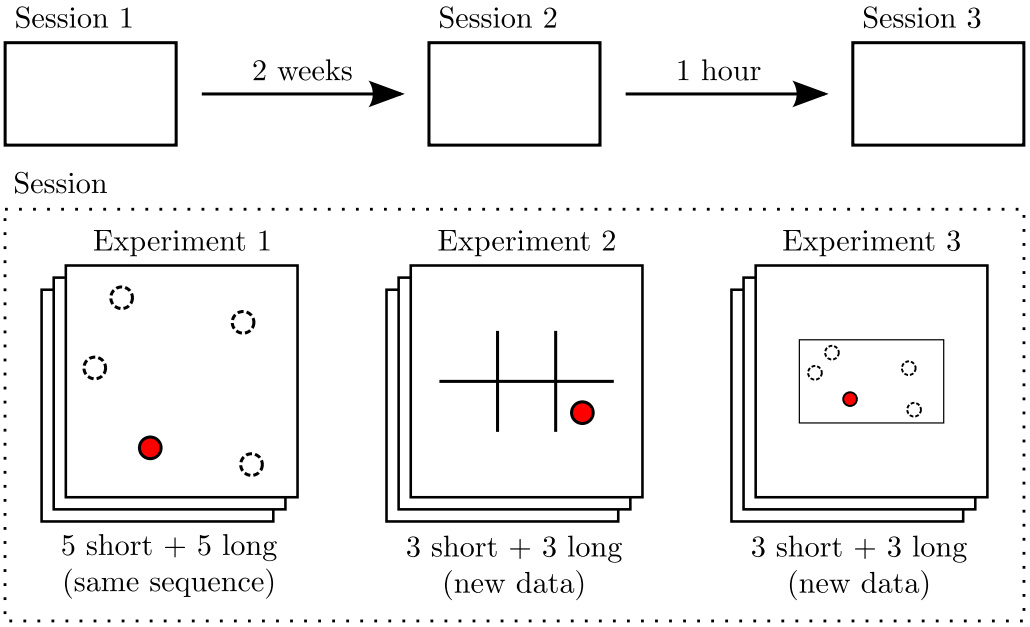


Figure 3.4: Experiment structure. Each session is divided into three experiments, each of which is repeated a number of times. The entire session is repeated after two weeks, and again an hour after the second repetition.

white centre appears at a random location on the background. The user is then asked to click on the dot as fast as possible. Once the dot is clicked the next one appears after a short delay, during which the screen is reset to the grey background. All instructions are displayed on-screen before the experiment begins, and the experiments differ in the nature of the instructions given to the subject. Additionally, each experiment comes in a short and a long version. This allows us to capture potential effects of training and memory for both simple and more complex tasks.

Experiment 1 (no prior knowledge) provides no instructions to the test subjects beyond asking them to click the dots as fast as possible. The short version has five dots and the long version has 7 dots. The idea behind Experiment 1 is to model a scenario in which an adversary sits down at a workstation without prior knowledge of the task she is facing. We assume that the subject's performance is affected by increasing task familiarity as well as memory-based learning effects when she completes the *same* sequence of dots multiple times. These effects reflect those observed in real environments when users become accustomed to their typical working environment. During the experiment the test subject learns the position of the dots over time, but in addition gains a general familiarity with the nature of the experiments. This experiment can also be transferred to an attacker that performs tasks she is accustomed to on a victim's workstation to cover her own tracks.

At each repetition the test subject is informed that the sequence will remain the same for the next iterations. We would expect the learning effects in short sequences to be bigger compared to long sequences. In order to test this each user performs five repetitions of the short sequence and five repetitions of the long sequence. The random seed used to generate the position list was kept identical for all subjects in order to eliminate distortion effects caused by the dot positions. The 5-dot and 7-dot sequences are chosen independently, as such the long sequence is not merely an extension of the short sequence. This design ensures that the user does not benefit from sequence-specific knowledge gained during the previous sequence.

Experiment 2 (Knowledge through description) provides the test subject with textual information about the dot positions before the dot sequence is shown. The screen is divided into six areas, numbered 1 through 6, and the positions of the dots in the

sequence is give in terms of a sequence of numbers that correspond to an area. This experiment models a scenario where a trusting (or even actively collaborating) user provides the adversary with knowledge about her workstation, as outlined in our threat model. This knowledge could relate to the location of icons, buttons or similar areas of interest, thereby allowing the attacker to more easily find and interact with these items. Such information transfer is rarely perfect so we model the transferred knowledge by giving the test subject the rough location of the dots, i.e., one of the six areas, before they appear on the screen. The test subject has no time limit when looking at and trying to remember the dot positions. We repeat the experiment 3 times with different 5-dot sequences and 3 times with different 7-dot sequences, to capture both simple and more complex tasks. Unlike the previous experiment we consider knowledge transfer rather than natural learning, consequently each sequence is only used once. We make this choice as repetitions of identical sequences would combine the effects of knowledge transfer (i.e., giving external information to the user) and natural learning (i.e., the user learning from completing a sequence more than once). This combination would then make it hard to isolate the individual contributions of each source of information.

Experiment 3 (Knowledge through observation) provides the test subject with a visual representation of the exact dot positions before the dot sequence is shown. This models the case where the adversary is able to observe the legitimate user while he performs her tasks, also known as "shouldersurfing". While a legitimate user's gaze position is not visible through observation in an office environment, things like the cursor position are still likely to reveal some information. This experiment represents the maximum amount of information an adversary is able to obtain before attempting the task herself.

We collected the data for the knowledge transfer experiment from 30 participants, 20 male and 10 female. Participants were recruited through public advertisements, mailing lists and social media. Aside from a minimum age of 18, there were no further exclusion criteria. The age distribution, as well as whether the subjects are wearing glasses or contact lenses, is given in Figure 3.5. The experiments are conducted with the approval of the ethics committee of the University of Oxford, reference SSD/CUREC1/13-064.

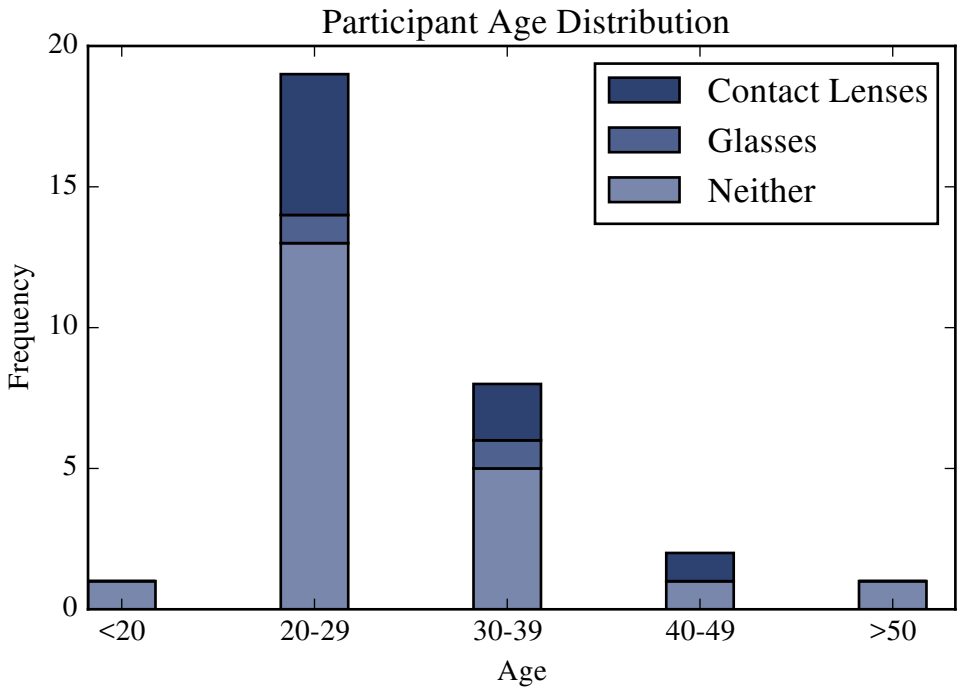


Figure 3.5: Participant age distribution in decades. Out of 30 participants 2 are wearing glasses and 9 are wearing contact lenses.).

3.4.3 Task Selection

The set of tasks described in the previous Section is suitable to both detect the effect of learning on eye movement behaviour as well as to measure the amount of identifying information that can be obtained through eye movements. However, it does not reflect tasks that would typically be performed in an office environment. As such, it is hard to draw reliable conclusions with regard to the performance of a continuous authentication system in a real-world environment. In order to amend this, we define a second set of experiments that more closely resembles such an environment. In line with our initial reasoning we keep the tasks identical for all participants in order to avoid classification of tasks rather than individuals. Our task set consist of reading, writing, two videos and web browsing.

Reading: As part of this task the participants are presented an excerpt from Daniel Defoe's *Robinson Crusoe*. The black text is displayed on a white background in a single column centred horizontally on the screen, as it is common with many types of e-book software.

Writing: During this task the users are asked to copy part of the text they read before. To this end they are presented with the original text on the top half of the screen and an empty text box below the text. One might argue that restricting users in what they are typing might potentially influence features, however it is difficult to truly capture daily typing behaviour in a lab experiment. Even when displaying a writing prompt (such as asking participants to recapitulate their day) many participants would likely be at a loss of what to write, thus greatly limiting the fraction of time spent typing. While it would be possible to ask participants to perform tasks they were intending to perform regardless of the experiment, the unfamiliar environment would likely affect behaviour and constitute a confounding factor. Besides the test of task dependence, we hope to gain another insight from this task: Due to the nature of optical eyetracking, samples might be lost when the user is looking at the keyboard (which is likely to be frequent, especially for inexperienced typists). Based on this assumption we will also quantify the fixation rate, as it directly impacts the speed with which authentication decisions can be made (see Section 3.6.2 for details).

Browsing: An obvious choice for this task would be to give users a fixed time limit and don't restrict the websites they visit. Naturally this might lead to users choosing wildly different sites, such as streaming sites (e.g., YouTube), news sites or online games. Compared to the number of possible groups of websites, the number of users is relatively small. This would likely lead to (partially) profiling website types rather than users. To address the trade-off between a real-world environment and the need to fix the task, we used a Wikipedia browsing game. As part of this game the user is initially shown a random Wikipedia article and asked to exclusively use links within that page to reach the article "University of Oxford". Once this goal is accomplished, the user is asked to use Wikipedia's "Random Article" function to start over until the task's time limit is reached.

Videos: With the increasing popularity of online streaming sites such as YouTube and Netflix, we considered it important to include watching a video as one of the tasks. While it is infeasible to include all varieties and genres of videos in a single lab study we selected two different videos as representatives. The first video shown to participants is "Big Buck Bunny", a popular short computer-animated comedy film produced by the

Blender Foundation[2]. The rationale behind this choice is that the film is released under an open-source licence (and can be shown as part of the experiment without further legal restrictions) and is likely to keep participants engaged in the experiment. The film itself features both slow fading as well as rapid cuts, together with frequently changing colour schemes. Our second choice is an educational video titled "The Problems with First Past the Post Voting Explained", which is also freely available on YouTube[3]. Unlike Big Buck Bunny, the video contains very limited movement. Quick scene changes are mostly absent and the colour scheme is bright and lacks frequent changes.

We collected the task dependence data from 10 participants, 6 male and 4 female. The recruitment process was identical to that used for the previous experiment. The experiments are conducted with the approval of the ethics committee of the University of Oxford, reference SSH_C1A_15_139.

3.4.4 Feature Stability Over Time

For eyetracking to be a useful defence against insider threats, the features measured from our test subjects must be relatively stable over time, otherwise false rejects would occur frequently as the template becomes outdated. While this can be countered by sporadically retraining the classifier this constitutes a serious challenge, as the user identity has to be established reliably during this time. We present a full list of features in Section 3.6.1 (Table 3.1). In this Section, we present the main reasons why time stability is a challenging problem:

Changes in the environment Features such as the pupil diameter may change depending on lighting conditions. While the screen brightness is kept constant across all subjects and all sessions, the level of daylight may change. It is important that the classifier accounts for these changes.

[2]https://peach.blender.org/, last visited 01/25/2016
[3]https://www.youtube.com/watch?v=s7tWHJfhiyo, last visited 01/25/2016

Changes in the user's physical and mental state Neuroscientific research shows that a person's eye movement behaviour can change depending on states like drowsiness, exhaustion, stress or euphoria (see Section 3.2 for details).

Technical Artefacts A recent study shows that the duration and number of fixations and saccades can depend on the gazetracker precision and the fraction of missing samples [76]. As these values rely on the calibration of the gazetracker, they may change slightly across different sessions.

The changes described above can manifest themselves both within the same session and across multiple days or weeks. Technical artefacts may be particularly prevalent when using data collected in different sessions due to the fact that a separate calibration has to be performed before each session. Despite these difficulties we show in Section 3.6.2 that we are able to collect a classifier training dataset that is rich enough to reduce the influence of these error sources. By including training data from several session we are able to capture, and adjust for, both long-term and short-term feature decay.

3.4.5 Experimental Setup

Figure 3.6 shows our experimental setup. We use an SMI RED500 eyetracking device with a sampling rate of 500 Hz to collect the raw gaze data. The stimuli are displayed on a 24 inch Dell U2412M monitor with a resolution of 1920 x 1200 pixels. The viewing distance between the subjects and the screen is approximately 60 cm. In order to reduce distractions and to minimize the influence of the experimenter on the subjects, all instructions were displayed on-screen during the session. Although the eyetracker compensates for minor head movements during the data collection, we asked the participants to move as little as possible.

Before the session the eyetracker has to be calibrated for each test subject. This stage consists of a calibration phase and a verification phase in which the error between actual and estimated viewing angle in degrees is determined. In order to ensure as high a data quality as possible, we reject calibrations with a viewing angle error of more than 1°, either horizontally or vertically. If the error is too high the calibration has to be repeated.

Figure 3.6: Our experimental setup consists of an SMI RED500 gazetracker that determines the user's gaze position on a 24 inch screen with a 1920x1200 resolution

At the end of the session we repeat the verification phase in order to test whether the initial calibration is still valid. A large verification error at this stage indicates low quality data, most likely due to excessive movements during the experiments. During testing we observed an average error of $0.49°$ in the X-direction and $0.52°$ in the Y-direction immediately after calibration. These errors increased to $0.74°$ and $0.72°$ respectively over the course of the experiment. Given that the error rates are lower than our threshold even at the end of the experiment we are confident in the quality of our data.

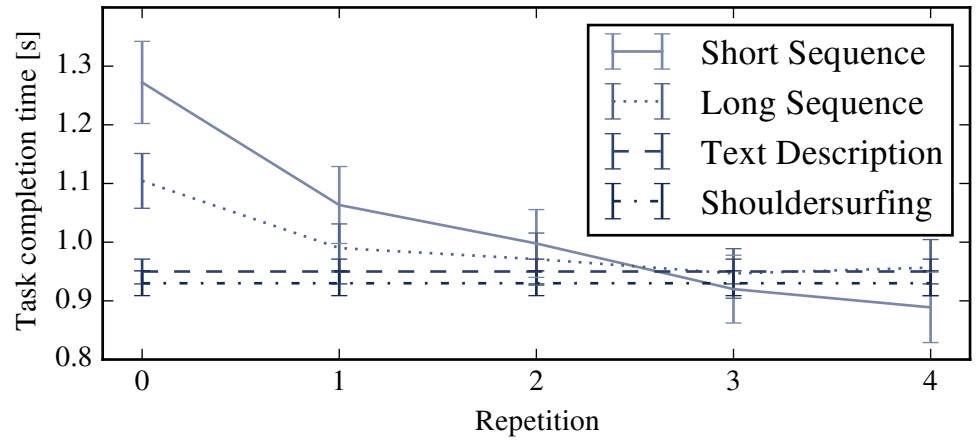
3.4.6 Modifying Sampling Rate

As outlined above, all the data in our study was collected at a sampling rate of 500 Hz. Capturing data at the highest available sampling rate provides the benefit of exploring exactly which distinctive features are contained in human eye movements, even though this sampling rate might not be available in equipment used in many productive environments. While it would be possible to repeat the experiments with different hardware, this would make comparisons more difficult due to external factors (e.g.,

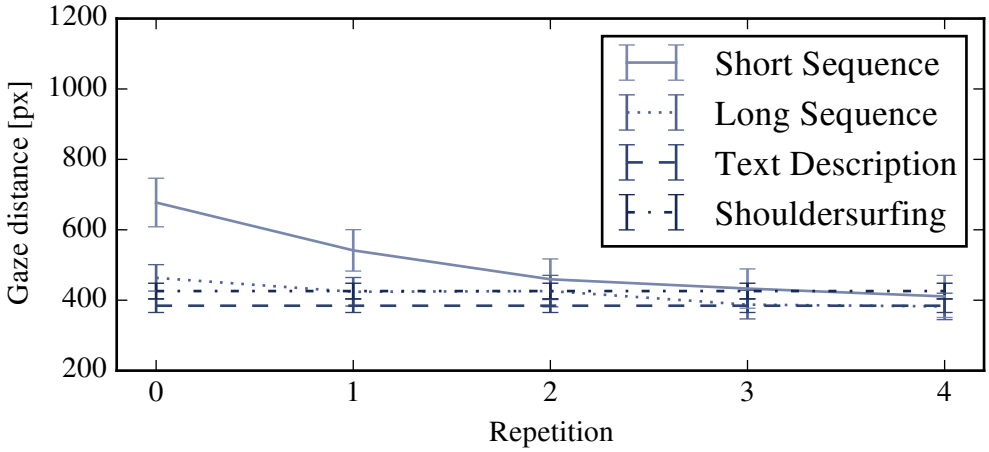
lighting, different individuals) that are virtually impossible to control for. Even when keeping all factors identical by simultaneously collecting data with multiple devices the number of datasets that is collected is inherently limited by space constraints when placing the devices. In order to provide both insights into the maximum distinctiveness of the biometric as well as the performance that can be expected with consumer-level hardware we perform downsampling on our dataset to simulate the sampling rate of these devices. We employ downsampling factors of 1,2,5 and 10. A downsampling factor of *n* means that every n^{th} sample is *kept*. Consequently, based on the initial sampling rate of 500 Hz, we generate individual datasets with 500, 250, 100 and 50 Hz.

3.5 Measuring Task Familiarity

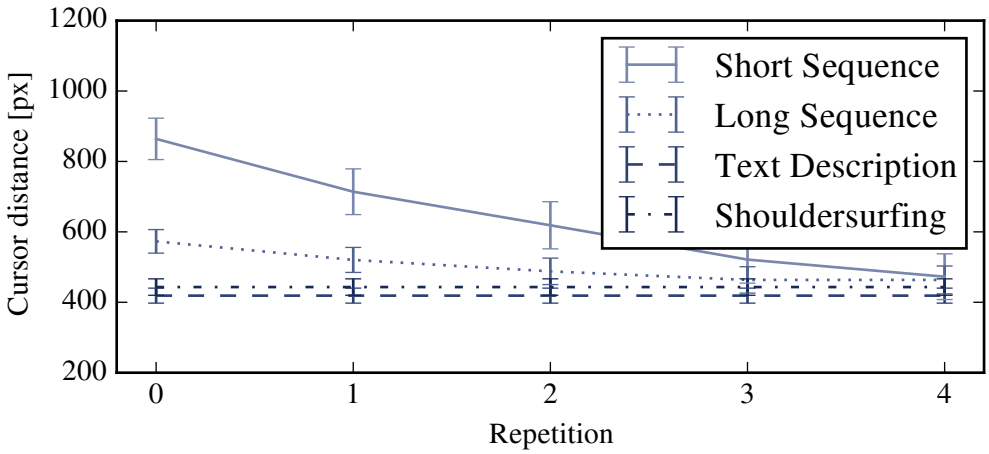
As outlined in our threat model, the attacker will attempt to familiarise herself with the legitimate user's behaviour and the task at hand. As such, we have to confirm that users are actually improving (gaining familiarity) over the course of the experiment. In order to provide ground truth for this assessment, we define two metrics to measure task performance for the knowledge transfer experiment. This allows us to test whether any improvements in the gaze-based metrics correlate with improvements in actual user performance. (1) *Response Time* is the time it takes a test subject to complete one dot of the sequence in the experiment. It is important to note that this metric refers to a *single dot*, rather than a *sequence of dots*. This definition allows us to compare both short and long sequences using the same metric. As the users are asked to complete the sequences as quickly as possible, this measure is the most natural measure of performance. (2) *Cursor Distance* is the distance between the cursor location and the position of the stimulus, right before it is displayed. Reaction time and mouse movement time are the most significant components of the response time (Figure 3.3 illustrates this for a single user). As such, using information about the (predicted) dot location to position the cursor closer to the dot will significantly improve the task completion time. A (significant) decrease in the cursor distance can only be due to a better prediction of the dot location by the user, showing that the user has gained and used information either through natural learning or knowledge transfer.



(a) Response Time



(b) Distance from Gaze to Stimulus



(c) Distance from Cursor to Stimulus

Figure 3.7: Changes of three different performance metrics caused by natural learning, text descriptions and shouldersurfing. The error bars indicate the 95% confidence intervals.

In addition to these ground truth measures, we introduce the gaze distance, defined as the distance between a user's gaze position and the position of the stimulus (the dot) just before it is displayed. If this measure follows the same trend set by the two ground truth metrics, it shows that it is possible to learn about users' task familiarity by observing their eye movement behaviour.

Figure 3.7 shows the results of our experiment. As we do not perform repetitions with identical sequences for Experiment 2 and 3 (text descriptions and shouldersurfing), the figure shows the average over all sequences.

There are two ways in which the users improve over the course of the experiments: (1) by learning the game mechanics, and (2) by predicting the location of the dot before it appears. The second component then allows the participant to use the blank period between dots to reposition the cursor. With an increasing number of repetitions, the users improve with regard to all three metrics. This improvement shows in the response time decreasing from 1.33 seconds to 0.9 seconds for the short sequence, with similar effects on gaze and cursor distance. Generally, this effect could be due to either (1) or (2), as users become more familiar with both the game and the sequence. However, once they proceed from the short sequence to the (different) long sequence, they only benefit from an improved grasp of the game mechanics (as the short and long sequences are independent from each other). This improvement, shown as (1) in Figure 3.7, is present for both of the ground truth metrics (response time and cursor distance), as well as the gaze-based metric. On average, users improve from 1.33 seconds for the first iteration of the short sequence, to 1.11 seconds for the first iteration of the long sequence. In both cases the sequence was unknown, suggesting that this improvement is only due to an improved grasp of the game mechanics. The improvement resulting from learning the sequence (i.e., becoming better at predicting the position of the next dot) is marked as (2) in Figure 3.7, showing the difference in performance between a familiar and an unfamiliar sequence. While users take an average of 0.90 seconds for the last repetition of the short sequence, this increases to 1.11 seconds as the effect of having learned the sequence disappears.

The gaze-based metric follows similar patterns, showing an improvement of 132px for (1) and 30px improvement for (2), although only the former is statistically significant

($p < 0.01$). For all metrics, there were no statistically significant differences ($p > 0.05$) in user performance, regardless of whether information was obtained through natural learning, shouldersurfing or text descriptions. This suggests that gauging task familiarity through gaze patterns is particularly effective against outside attackers without access to these sources of information.

3.6 Continuous Authentication

In this section, we will discuss the design and performance of the eye movement authentication system. We use both the task familiarity and task dependence datasets as basis for the system's evaluation. Due to the larger sample size of the task familiarity dataset, we will use it to quantify overall error rates. The task dependence data serves to quantify how features change across tasks and how this effect can be mitigated most effectively.

3.6.1 Biometric Features

Feature selection criteria

An important consideration when choosing features is what data is required to compute them and whether there are any constraints regarding the environment in which they are collected. In order to make the authentication system usable in a standard working environment, the calculation of the features must only use raw eyetracking data without using application-level data from running processes. This approach allows transparent continuous authentication, rather than providing a one-off login. This assumption distinguishes our approach from related work, which measures the user's reactions to controlled stimuli, and is therefore unsuitable for transparent continuous authentication [35–37].

It is important to know to which degree features are influenced by the task the user performs while the features are collected. As eye movements are always a reaction to a stimulus, perfect task independence can never be guaranteed and some features are more susceptible to such influences than others. Largely task-independent features allow conducting the training phase with a task different to the one performed during the system's actual operation. This is particularly desirable in an office environment, as a

wide variety of tasks are performed on a daily basis. A higher degree of task independence will significantly reduce the error rates exhibited by the system.

We choose our features such that their computation does not depend on any specific experimental design. As such, we don't use features for authentication that depend on the dot-clicking game (e.g., the gaze position relative to the dot position) or any of the real-world tasks. The main advantage of this approach is that the experimental design (i.e., the tasks performed by the subjects) is interchangeable and the authentication can be transparent and occur without the user's cooperation or even knowledge. While the features can be computed regardless of the task, their *distributions* might still be affected. We evaluate the effect of task selection and changing feature distributions in Section 3.6.4

Determining feature quality

Having a measure of feature quality is important for two reasons: (a) to be able to select the best features when the entire set is too high-dimensional and (b) to gain better insights into *why* the biometric works. Additionally, it allows to measure how external factors (such as different hardware or collection time span) affect each feature. Even initially highly distinctive features might be unusable if one or more of these factors severely degrade its performance. In order to ensure the robustness of the ranking of the features in our set, we employ three different measures: The relative mutual information (RMI), the Kolmogorov-Smirnov statistic of a two-sample KS-test and the Bhattacharyya distance. Initially, an amount of uncertainty is associated with the user ID (its entropy). This amount depends on the number of classes (i.e., users) and the distribution of the samples between users. Each feature reveals a certain amount of information about the user ID, this amount can be measured through the mutual information (MI). In order to measure the mutual information relative to the entire amount of uncertainty, we use the relative mutual information (RMI) which measures the percentage of entropy that is removed from the user ID when a feature is known [22]. The RMI is defined as

$$\mathrm{RMI}(uid, F) = \frac{H(uid) - H(uid|F)}{H(uid)}$$

where H(A) is the entropy of A and H(A|B) denotes the entropy of A conditioned on B. The range of this feature is between 0 (indicating that the feature contains no information

about the user) and 1 (meaning that all users can be uniquely identified through this feature). In order to calculate the entropy of a feature, it has to be discrete. As most features are continuous, we perform discretization using an Equal Width Discretization (EWD) algorithm with 20 bins [77]. This algorithm typically produces good results without requiring supervision. In order to limit the drastic effect that outliers can have when using this approach, we use the 1^{st} and 99^{th} percentile instead of the minimal and maximum values to compute the bin boundaries. A high RMI indicates that the feature is distinctive on its own, but it is important to consider the correlation between features as well when choosing a feature set. Additionally, several features that are not particularly distinctive on their own may be more useful when combined.

The relative mutual information relies on discretization of feature values, the number of bins and the algorithm used to filter outliers might not only change the absolute values of the measure, but also the relative ranking of features. In order to gain additional insights, we calculate the Kolmogorov-Smirnov statistic of a two-sample KS test. We consider the two one-dimensional probability distributions for two users with regard to a single feature. The KS-test then tests whether the two samples are drawn from the same distribution (null hypothesis). A feature would only be distinctive if the null hypothesis can be rejected for a high number of user pairs. As the information whether the null hypothesis is rejected at a certain confidence does not provide any information about the *magnitude* of the differences between the samples we use the Kolmogorov-Smirnov statistic as a metric, computed as

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of two different subjects. The measure is computed for all pairs of subjects, Table 3.1 provides the averages and standard deviations. Defined as the difference between two empirical distribution functions (that lie between 0 and 1 at each point), the KS-statistic also takes a value in that interval. A value of 1 indicates that the average difference between two users regarding this feature is maximal, thus suggesting a distinctive feature. None of the features used in the biometric

follow a normal distribution ($p < 0.001$), as such the fact that the test does not assume any specific distribution of the data is critical.

Additionally, for each pair of users (p,q) we compute the Bhattacharyya distance of a feature as

$$D_B(p,q) = -ln\left(\sum_{x \in X} \sqrt{p(x)q(x)}\right)$$

The Bhattacharyya distance measures the similarity of two continuous probability distributions, in this case the probability distributions of the same feature for two users. Higher values indicate bigger differences between the distributions, resulting in higher distinctiveness of this feature. This metric has been shown to correlate well with classification accuracy for a number of classifiers and datasets [78].

Grouping of samples

The eyetracker reports raw samples containing X/Y coordinates and the current pupil diameter. Without strong contextual information, a single raw sample does not contain any distinguishing information. Therefore, it is necessary to combine multiple raw samples and use the relationships between these samples (i.e., movements instead of static positions) as features. Given the nature of the data, we consider fixations to be the most natural level of abstraction. The RED500 groups samples collected over at least 50 ms that lie within a 30-pixel radius into a fixation (see Figure 3.2 for an illustration). In the context of this section, the term sample will refer to one fixation (i.e., a set of raw samples). In our data, we observe one fixation on average every 250 ms, yielding a sampling rate of 4 Hz. This rate may change depending on the user task (e.g., reading will lead to longer fixations and a lower sampling rate), calibration accuracy, overall tracking quality and across different users.

Feature types

A complete list of our features is given in Table 3.1. We consider three different types of features: pupil features, temporal features and spatial features.

Pupil features can be split into static and dynamic features. As outlined in Section 3.2, the *range* of the pupil diameter is largely constant for each person. We capture this static

range using the maximal, minimal and mean pupil diameter that is observed during one fixation. The dynamic component is reflected by the short-term changes of the pupil diameter. These changes can be caused by cognitive load or different stimulation through lighting. While these external stimuli are equal for all participants their *reactions* to them may not be. We model these changes through the standard deviation and the difference between the minimal and maximal pupil diameter observed during a fixation.

Temporal features include both the duration of saccades and fixations as well as speed and acceleration. Both the peak and the average velocity of movements within a fixation have been shown to differ greatly between people in related neuroscientific work (see Section 3.2). These differences are mainly caused through different prevalence of saccadic intrusions and microsaccades, both of which are characterized by high velocity and acceleration. Different studies report similar ranges for these values, even though their experimental designs differ significantly. This suggests that these features show a high degree of task independence, which makes them particularly desirable for classification. We compute the velocity between each pair of consecutive samples and only use the magnitude of acceleration (i.e., we do not use the direction). The reasoning behind this is that the direction of acceleration depends on the location of the target stimulus and is therefore task-dependent [79].

Spatial features are a method to measure the steadiness of a person's gaze. A fixation is a group of samples within a fixed-size circle, which consists of the samples and a centre point (see Figure 3.2 for an illustration). While the total area that can be covered by a fixation is limited by this definition, the spatial distribution of samples within this area can still be different. If a person's gaze is steady, the samples will be clustered closely around the fixation centre, with few samples outside of this group. If a person has trouble focussing, their gaze the samples will be spread more evenly. We compute both the distance between each raw sample and the center point as well as the distance between each pair of raw samples. As some movements may be more pronounced in the vertical or horizontal direction we also make this distinction. The distance between two fixations (as measured by the euclidean distance between their center points) allows

Feature	RMI	K-S Statistic	Bhattacharyya distance
Pupil features			
Pupil Diameter - Max	19.84%	0.61±0.28	0.78±0.88
Pupil Diameter - Mean	20.27%	0.62±0.28	0.84±0.97
Pupil Diameter - Min	20.26%	0.61±0.29	0.82±0.97
Pupil Diameter - Range	1.19%	0.12±0.07	0.02±0.02
Pupil Diameter - Stdev	0.98%	0.11±0.06	0.02±0.01
Temporal features			
Acceleration - Max	2.49%	0.18±0.12	0.05±0.06
Acceleration - Mean	0.35%	0.07±0.03	0.01±0.00
Duration of Saccade	1.09%	0.12±0.05	0.02±0.02
Duration of Fixation	0.9%	0.10±0.06	0.01±0.02
Pairwise Speed - Max	4.95%	0.25±0.16	0.10±0.12
Pairwise Speed - Mean	5.36%	0.26±0.17	0.11±0.14
Pairwise Speed - Stdev	1.77%	0.14±0.09	0.03±0.04
Spatial features			
Distance from Center - Max	1.2%	0.12±0.06	0.02±0.02
Distance from Center - Mean	2.52%	0.20±0.12	0.04±0.05
Distance from Center - Min	0.72%	0.11±0.06	0.01±0.01
Distance from Center - Stdev	1.21%	0.13±0.07	0.02±0.02
Distance from previous fixation	0.66%	0.10±0.05	0.01±0.01
Max Pairwise Distance	1.23%	0.13±0.07	0.02±0.02
Max Pairwise Distance X only	1.06%	0.13±0.07	0.02±0.02
Max Pairwise Distance Y only	0.84%	0.11±0.05	0.02±0.01

Table 3.1: List of pupil, temporal and spatial features that are computed for each fixation. For each feature, we report the relative mutual information (RMI) shared with the user ID. Additionally, we compute the average and standard deviation (as indicated by the ± sign) of the Kolmogorov-Smirnov statistic of the two-sample KS test, and the Bhattacharyya distance for all pairs of users. For all metrics, higher values indicate higher feature quality.

us to measure how many points between two areas of interest (i.e., target stimuli) are actively focused and processed by the subject.

Discussion

Table 3.1 shows how each feature performs with regard to the three metrics discussed in the previous section. Each metric has different value ranges, as such the values are not directly comparable. However, the relative rankings of features in the set are similar between the three metrics, which suggests that the feature discretization performed

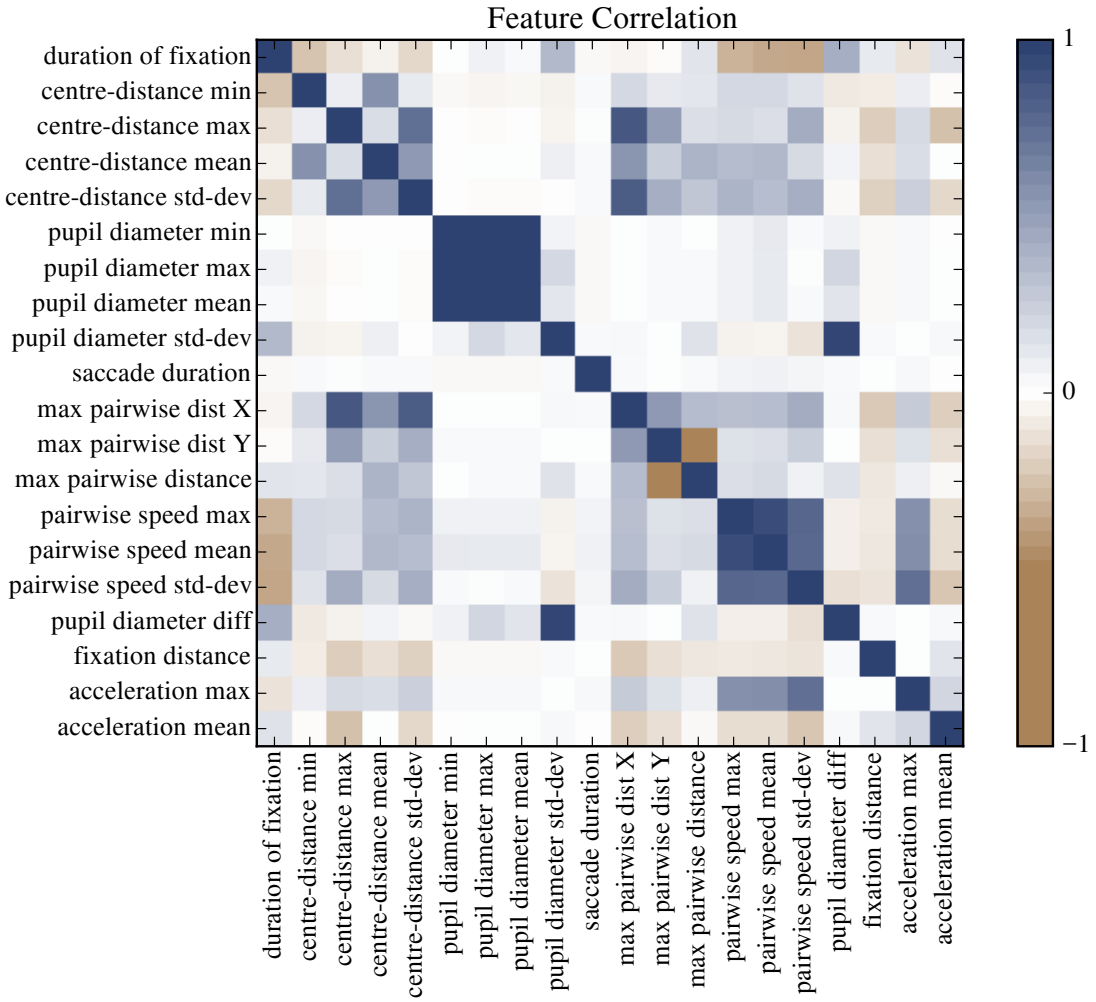
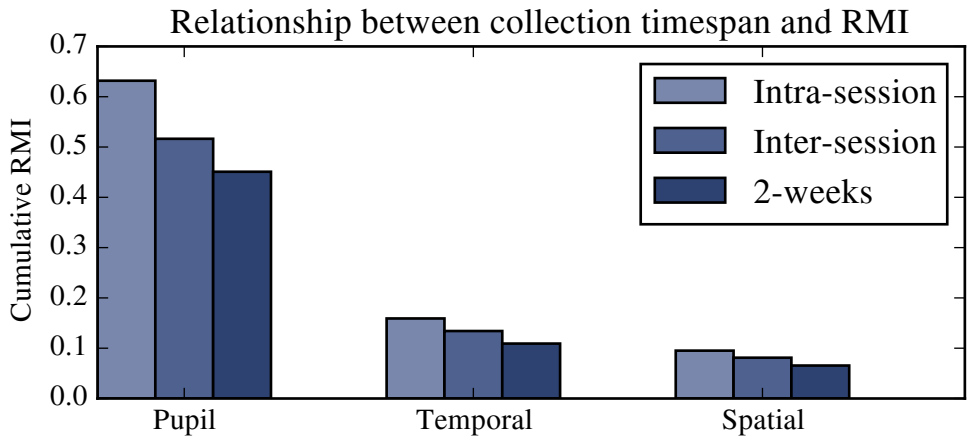


Figure 3.8: Feature correlation measured by the pearson correlation coefficient. A value of 0 indicates no correlation, values of 1 and -1 signify positive and negative correlation, respectively.

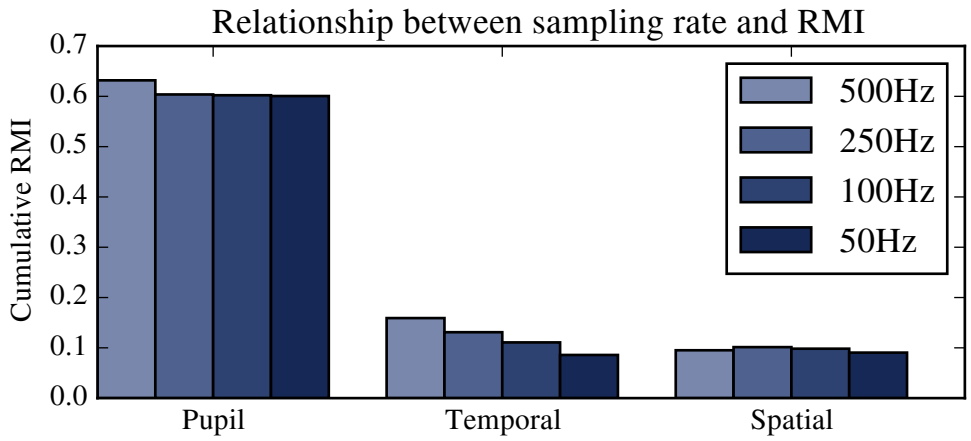
before computing the RMI does not distort the overall ranking and all three metrics are indicative of the features' quality.

The static pupil diameter features (i.e., min, mean and max) share the most information with the user ID. The dynamic pupil diameter features (i.e., the standard deviation and the min-max difference) are less distinctive, which suggests that the pupil diameter is more a result of different genders, ethnicities, ages and eye shapes, rather than a behavioural feature.

While the behavioural features, both temporal and spatial ones, show a lower distinctiveness than the pupil diameter, they still contribute significant amounts of information. The fact that both peak speed and acceleration exhibit comparatively high scores with



(a) Collection Timespan



(b) Sampling Rate

Figure 3.9: Effects of sampling rate and collection timespan on cumulative RMI for different feature groups.

regard to all metrics shows that we accurately model the distinctive capabilities of saccadic intrusions and microsaccades.

When selecting which feature candidates should form the final feature set there are several aspects that have to be considered: Each of the features should be hard to imitate in a given threat model. As we focus on insider threats this rules out features that can be easily observed and copied. Given the insights from Section 3.2, we suspect that it may be possible for a sophisticated attacker to modify her own pupil diameter to a certain degree. Specifically, it has been demonstrated that the pupil reacts almost instantaneously to external light stimulation, while the reversion to the baseline occurs slowly [80]. Consequently, an attacker could decrease their own pupil diameter by

constantly shining a bright light in their own eyes, even though the reverse might be harder to achieve. While it might be possible to recognize such a stimulation (e.g., by monitoring ambient light intensity), we still consider this a valid threat. In order to address this issue, we also investigate the performance of a feature set that does not make use of the pupil diameter features. When putting the system into operation, it can then be decided which feature set should be used, depending on the threat model and the capabilities of potential attackers. We will discuss the impact of not using the pupil diameter as a feature (and thereby raising the bar for an attacker trying to perform an imitation attack) in Section 3.6.2. Figure 3.8 shows that the correlation between features belonging to the same group (i.e., pupil diameter, temporal or spatial) is relatively high, while the inter-group correlation is considerably lower. This suggests that all three groups contribute to the distinctiveness of the biometric and no group can be replaced entirely by another. This also makes sophisticated imitation attacks more difficult, as a number of very distinct features have to be emulated simultaneously.

Feature degradation

As outlined in Section 3.4.4, we consider two main factors that could negatively impact feature quality: (a) increased data collection timespan and (b) reduced sampling rate. Figure 3.9 shows the effects of increased data collection time and reduced sampling rate on the cumulative RMI of the three feature groups. Not surprisingly, the intra-session dataset (which only includes the first session) results in the highest total information. After that, the features' information content declines with increasing time distance between the sessions, with the effect being more pronounced for the features involving the pupil diameter. This effect is most likely due to changes in lighting, fatigue and cognitive effort as outlined in Section 3.2.

In line with our expectations, reducing the sampling rate also reduces the information content of features. Pupil diameter features suffer slightly from halving the sampling rate, any further reduction has no measurable effect. Conversely, temporal features decrease almost linearly with every further increase of the downsampling factor. This effect can be explained through the short-lived nature of the main physiological processes causing

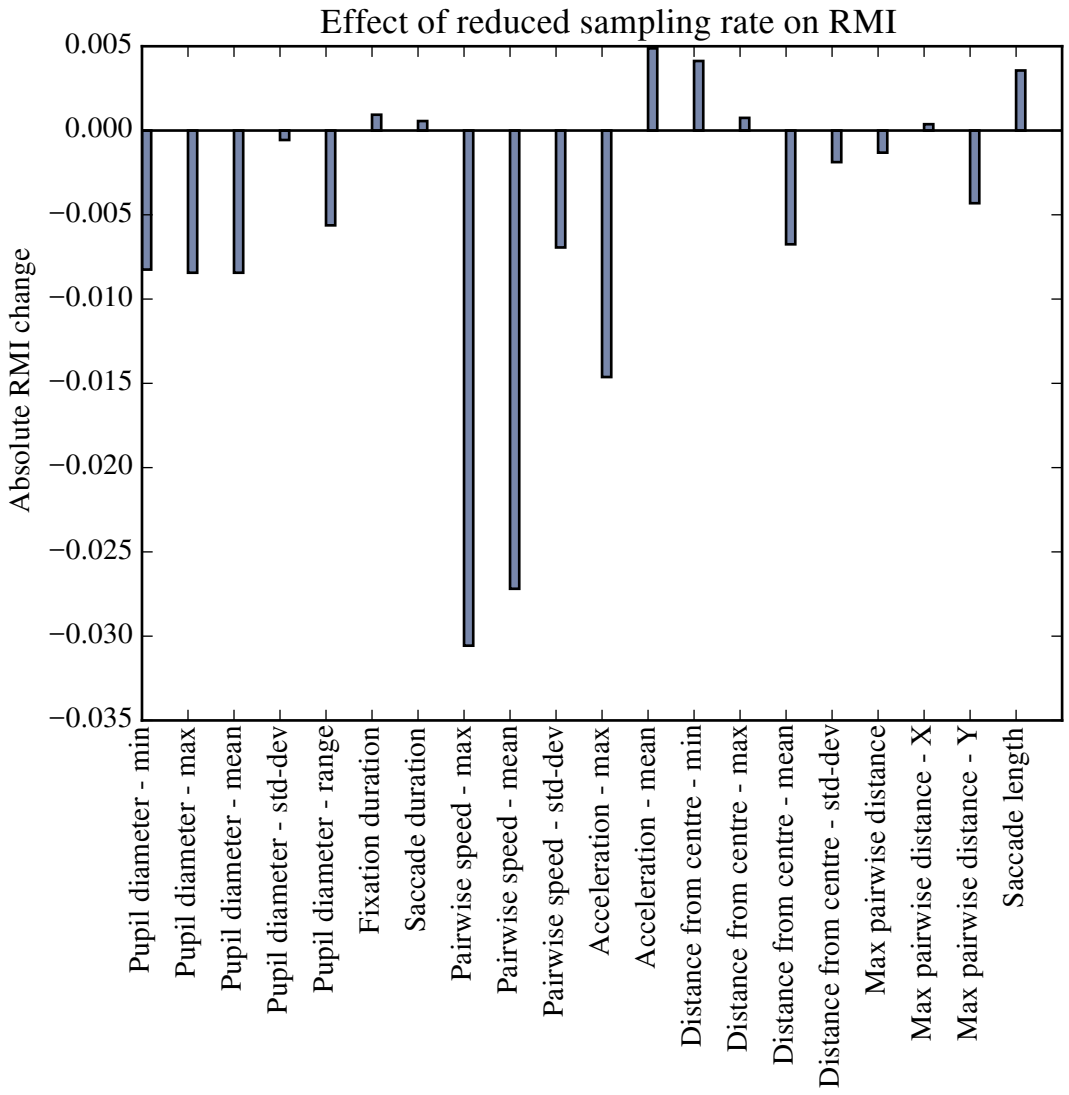


Figure 3.10: Changes to the RMI of individual features when reducing the sampling rate to 50Hz. The temporal features reflecting the properties of microsaccades (speed and acceleration) are most strongly affected.

the distinctiveness of this feature group. As described in Section 3.2, microsaccades and saccadic intrusions are distinctive, but only last a few milliseconds. As such, it is not surprising that reducing the sampling rate below a certain threshold prevents their distinctive capabilities from being harvested. Figure 3.10 breaks down the degradation of features when reducing the sampling rate to 50 Hz. Maximal and average speed as well as peak acceleration, the features most strongly associated with microsaccades, suffer from the biggest degradation of all features in the set. The degradation is statistically significant for all features in this group ($p < 0.05$). Conversely, the spatial features, with

the exception of the mean distance to centre, show no statistically significant changes ($p > 0.05$) even for the lowest sampling rate setting.

3.6.2 Classifiers and Metrics

In this section, we will describe a number of classifiers for continuous authentication. We will distinguish between open-set and closed-set classifiers and analyse their respective advantages when used with the eye movement biometric. We also discuss the impact that feature selection, sampling rate and time distance have on the classifier performance and present metrics that make it possible to gauge the real-world performance of the system. Finally, we will give insights on how different parameters of our system can be chosen to reflect different security requirements.

Closed-set Classifiers

Training of closed-set classifiers requires samples for each potential user of the system. The output of the classifier for a new sample is then the predicted class, chosen out of the set of classes it was initially trained with. This type of classifier is useful in an insider threat scenario, as an employer using a biometric system likely has collected templates for all employees. The advantage of such a system is that it not only detects an unauthorized user (by virtue of the claimed identity, as established though a user name, not matching the user recognized by the classifier), but can also reveal the attacker's identity. The major disadvantage is that once an external attacker (i.e., somebody not enrolled in the system) attempts to access a workstation, the classifier will recognize her as the user with the closest template. This might lead to incorrectly granting access or framing an innocent user for the failed attack.

We consider two closed-set classifiers, the k-nearest-neighbors (knn) algorithm and Support Vector Machines (SVM). In order to determine the optimal parameters for these classifiers, we perform a grid search through a defined subset of the parameter space. For the knn classifier, we tested values of k between 1 and 20 and weighting samples uniformly or by euclidean distance. For the SVM, we tested a linear, a polynomial and a radial kernel (rbf) function. For all three kernels, we varied the soft margin constant C

in powers of ten between 1 and 10000. The polynomial kernel was used with degrees between 2 and 5 and for the radial kernel function, we tested values of γ between 0.00001 and 10. The best results were achieved with k=5 and weights based on euclidean distance for knn and an rbf-kernel with C=10000 and γ=0.001 for the SVM.

After completing the training phase, the classifier continuously assigns labels (i.e., user IDs) to fresh samples. If the system is used for authentication (which is the focus of this work) rather than identification, this decision can be transformed to either an accept (i.e., the predicted user matches the claimed user) or a reject (i.e., the predicted user is different from claimed user). The decision's robustness can be increased by combining multiple samples before making a final decision. Combining multiple samples will increase the accuracy of the decision but also introduces a delay before an imposter can be detected. As eyetracking provides a stream of new samples at a constant and high rate we choose to combine several samples for each authentication decision. Our authentication system is parametrized through the size n of a sliding window and the threshold t which defines how many of the samples in this window must be classified as benign. As such, a fresh sample is accepted only if at least t out of the last n labels output by the classifier match the claimed user ID. Therefore, a higher value of n increases the system's robustness but delays the detection of an attacker while the value of t controls the tradeoff between the false accept rate and false reject rate.

Open-set Classifiers

When used for authentication, one-class classifiers are only trained with data belonging to a single legitimate user. For each new sample, the classifier determines whether it is sufficiently close to the data observed during training. As such, the output of the classifier is only a yes/no decision, rather than a user ID (the output of a closed-set classifier). Therefore, it is unable to reveal the identity of an attacker. However, identification of attackers is impossible even for a closed-set classifier unless reference data is available for all potential attackers. While this may well be the case for an insider threat scenario, it is unrealistic for a more general setting. Additionally there are usually other means of identifying an attacker who is physically present, such as video surveillance. The major

advantage of one-class classifiers is that training does not require any knowledge about either the set of possible attackers or their individual biometric templates. As such, it is well-suited to detect both inside and outside attackers.

In this section, we analyse the performance of the one-class Support Vector Machine. This classifier is parametrized by the kernel coefficient γ and the regularization parameter ν. The value of ν is an upper bound on the fraction of *training* errors. Higher values of ν will increase false rejects while reducing false accepts. While it would be possible to tailor the value of ν to achieve a certain split between the two error rates, this is generally undesirable as changing the parameter requires retraining the classifier. Therefore, we choose both values to minimize the total number of errors within the development data set. Using a grid search on a development set different from the data used for training and testing, we identify $\gamma = 0.001$ and $\nu = 0.18$ as optimal parameters. The actual authentication decision is made using the same sliding window technique we use for the closed-set classifier. A sample is accepted only if at least t out of the last n classifier decisions are accepts.

Metrics

Most papers proposing new biometric systems provide both the false accept rate (FAR) and false reject rate (FRR) of their chosen biometric. Most of the time some sensitivity parameters can be adjusted to trade lower FAR for a higher FRR, and vice-versa. In an attempt to make different systems more comparable the equal error rate (EER) is often given as well. The EER is the error rate of the system when its parameters are adjusted such that both the FAR and FRR are equal.

In order to ensure comparability of our biometric with previous work we also provide the EER (see Section 3.6.3). However, this metric suffers from a number of practical problems that makes it difficult to draw valid conclusions from it: In a continuous authentication scenario (which is the environment in which behavioural biometrics often compare most favourably with hard biometrics), it is crucial to know how the errors are distributed between both legitimate users and user-attacker pairs. A FRR of 5% could signify that all users are rejected exactly once every 20 samples, or that 5% of the

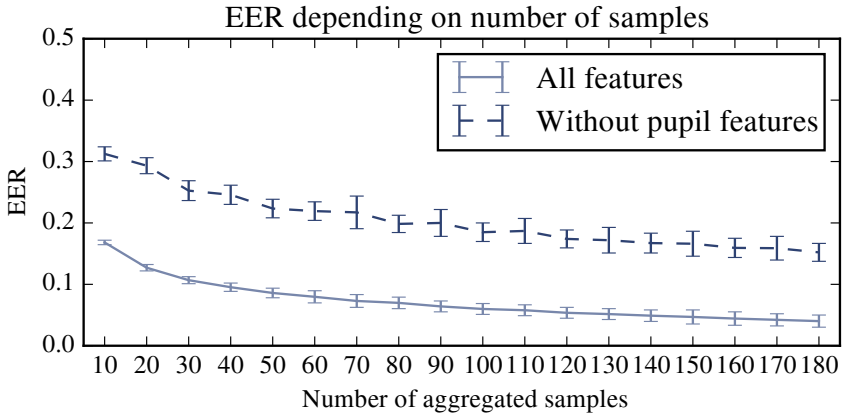
users are rejected consistently (or, most likely, something in between). Obviously, these cases pose very distinct challenges. The latter case could be addressed by authenticating the users that are consistently rejected using a different mechanism (such as another biometric), while the first case renders the entire biometric largely useless. The same can be said for the FAR, questioning the usefulness of the EER as a measure of security. Additionally, it is impossible to derive a biometric's typical attack detection speed without knowing its sampling rate.

To address these issues, we provide two more metrics: The systematic false negative rate (sys-fn) and the median time until an attacker is detected (med-ttd). The systematic false negative rate is the fraction of attackers that are never detected (within the scope of our data). This is usually due to their biometric template being close to that of a legitimate user. As with the conventional measures (FAR and FRR), these values depend on the system's sensitivity settings. In order to provide results that are easy to compare, we report them at a setting that never rejects legitimate users. As such, any additional security provided by the system comes at no cost in terms of user inconvenience or needing a mechanism to handle false alarms.

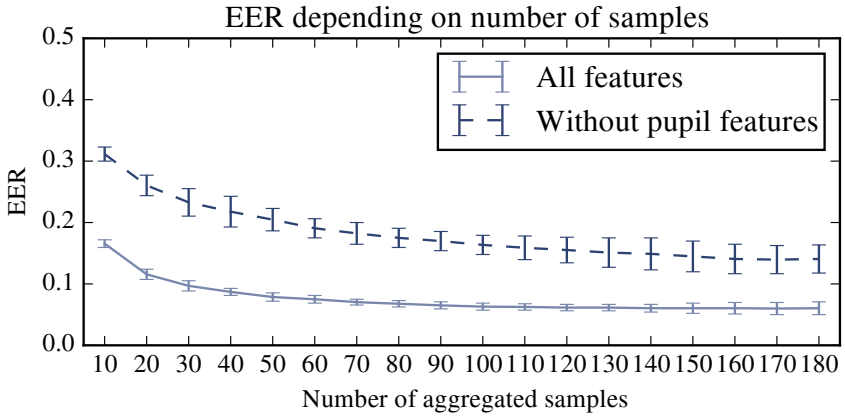
3.6.3 Task Familiarity Experiment

The performance of the closed-set classifier depending on the number of samples is shown in Figure 3.11. Depending on the dataset, the biggest reduction in the EER is achieved by combining 60 samples, with little benefit beyond 100 samples. As would be expected, the error rates increase with the collection timespan of the data, with the 2-weeks dataset showing the highest EER.

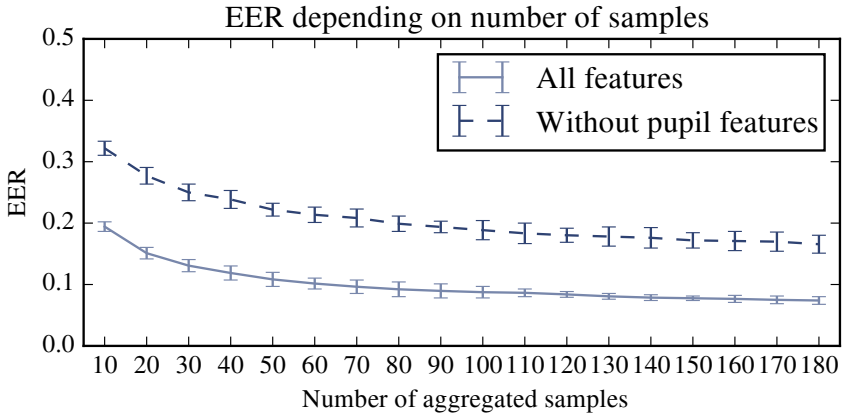
The performance of the open-set classifier for all combinations of dataset, feature set and sampling rate for the task familiarity experiment (see Section 3.4) is given in Table 3.2. Most notably, the equal error rate decreases greatly when compared to the closed-set classifier (Figure 3.11). The relative relationships between the error rates for different datasets (i.e., intra-session, inter-session, over two weeks) is preserved, increasing time distance also increases the equal error rate. This effect is present for all sampling rates, although it is most pronounced for the 500Hz setting as the intra-session



(a) Intra-Session dataset



(b) Inter-Session dataset



(c) 2-weeks dataset

Figure 3.11: Average Equal Error Rates obtained through 5-fold stratified cross validation on three different datasets using the closed-set SVM classifier. The error bars indicate 95% confidence intervals.

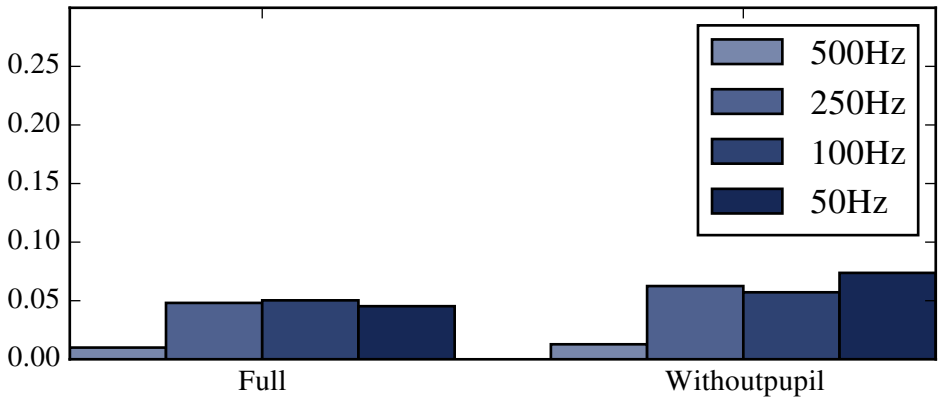
			Full			Without Pupil Diameter		
Dataset	**Users**	**SR**	**EER**	**sys-fn**	**med-ttd**	**EER**	**sys-fn**	**med-ttd**
Intra	30	500 Hz	1.00%	15.44%	30.50s	1.29%	51.80%	∞
Intra	30	250 Hz	4.83%	24.88%	37.50s	6.26%	60.04%	∞
Intra	30	100 Hz	5.05%	25.35%	37.50s	5.71%	63.63%	∞
Intra	30	50 Hz	4.52%	13.63%	22.25s	7.36%	50.28%	∞
Inter	20	500 Hz	2.06%	8.24%	27.50s	2.44%	42.88%	370.00s
Inter	20	250 Hz	7.53%	16.14%	35.50s	7.65%	42.81%	378.25s
Inter	20	100 Hz	7.14%	17.08%	37.50s	7.53%	44.50%	384.25s
Inter	20	50 Hz	8.19%	10.74%	27.00s	7.77%	35.37%	263.50s
2-weeks	20	500 Hz	3.92%	2.69%	27.50s	3.67%	31.70%	323.75s
2-weeks	20	250 Hz	10.00%	25.14%	84.25s	8.96%	68.38%	∞
2-weeks	20	100 Hz	8.99%	25.52%	105.50s	8.81%	66.38%	∞
2-weeks	20	50 Hz	7.82%	9.90%	31.50s	6.64%	54.57%	∞

Table 3.2: Performance of the one class SVM classifier. sys-fn is the fraction of attackers that are not detected within the scope of the data, med-ttd is the median time until an attacker is detected. Note that all metrics are computed at a parameter setting that never rejects a legitimate user (i.e., not at the parameter setting that achieves the equal error rate).

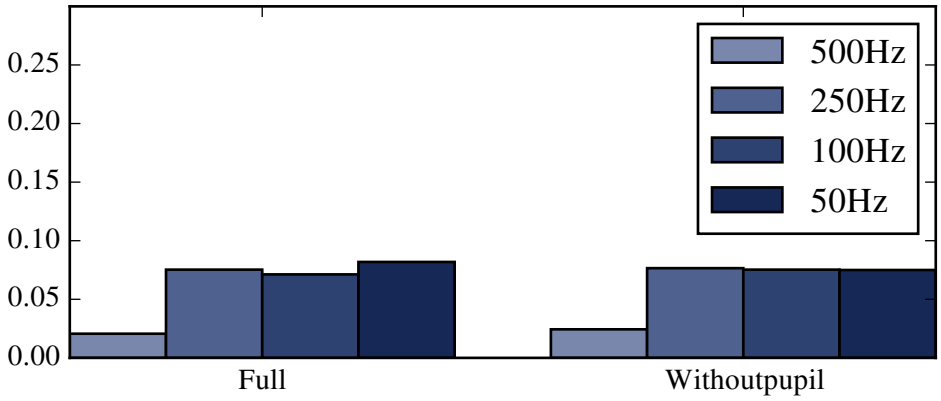
error rates are much lower (see Figure 3.12). Interestingly, the same can not be said when switching from the full feature set to a set without the pupil diameter. While the error rates increase for two out of the three datasets they even *decrease* slightly when using data gathered over two weeks. This suggests that the pupil diameter is not only a potential security concern, but an actual source of error when the biometric is used over longer time spans. This is most likely due to the fact that the effects of external stimuli (such as lighting) overshadow the initial distinctiveness provided by physical characteristics. Consequently, this feature group should not be used for long-term operation without regular classifier retraining (e.g., at the start of a session).

As would be expected, reducing the sampling rate reduces classification accuracy and consequently increases the equal error rate. The sharpest increase can be observed when reducing the initial sampling rate of 500 Hz to 250 Hz. Any subsequent changes don't result in a statistically significant increase in error rate. Surprisingly, the reduction from 100 Hz to 50 Hz *lowers* the equal error rate for some datasets.

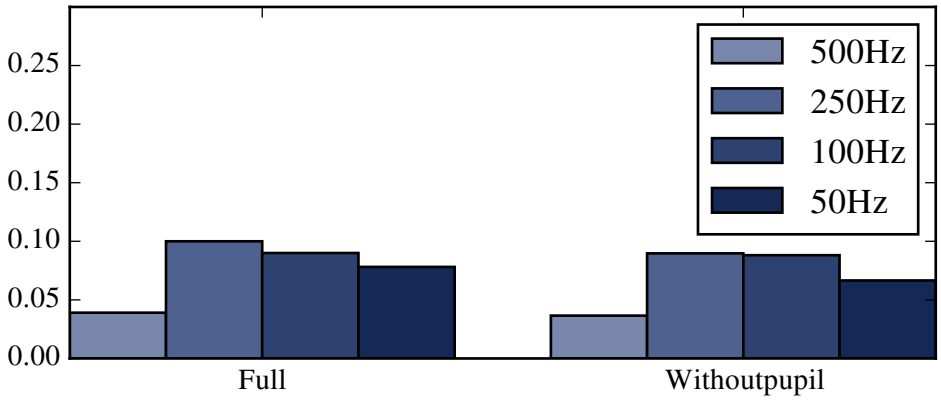
Regarding the newly introduced metrics "median time to detection" and "systematic



(a) Intra-Session dataset



(b) Inter-Session dataset



(c) 2-weeks dataset

Figure 3.12: Average equal error rates obtained with the oneclass-SVM classifier for different sampling rates. The error bars indicate 95% confidence intervals. Reducing the sampling rate to 250Hz has a large effect on the error rates, any subsequent reduction does not produce a statistically significant increase in the EER.

false negative rate" (see Section 3.6.2 for details), it is interesting to note that they don't correlate well with the equal error rate. While sys-fn increases when reducing the sampling rate, the effect is not nearly as strong as would be expected given the enormous increase in EER. The effect is even more pronounced when not using the pupil diameter for classification. While the error rates increase only marginally (or even decrease in the 2-weeks dataset) the fraction of undetected attackers increases immensely. In some datasets this fraction increases to over 50%, resulting in a median time to detection of ∞. Note that these metrics are computed for a parameter setting that never results in legitimate users being rejected within the scope of our data. As such, the detection of attackers comes at no additional costs with regard to user inconvenience or handling of false alarms.

3.6.4 Task Dependence Experiment

The previous subsection has shown the feasibility of eye movement authentication based on data from the task familiarity experiment. We designed the feature set to be independent of any specific task. As such, artificial tasks, as used in this experiment, are still well-suited to test the feasibility of continuous authentication. However, as outlined in Section 3.2, the distribution of biometric features may still depend on a specific task, thereby affecting their respective discriminative power. In this section, we will use the data from the task dependence experiment (Section 3.4.3) to explore this concern.

Pupil Diameter Correction

Medical research has shown that a person's pupil diameter is greatly affected by light stimulation (see Section 3.2 for details). Naturally, the tasks described in Section 3.4.3 will result in different screen brightness (with the videos being generally darker than text on white background). Consequently, it is likely that classification accuracy would suffer when the image or screen brightness differs between enrolment and operation. However, a software-based continuous authentication system can monitor the brightness of the currently displayed image and correct the raw pupil diameter reported by the eye tracker accordingly. In order to perform this correction, two pieces of information

are needed: (a) the brightness of an image and (b) the way a person's pupil diameter depends on this brightness.

During each of the main tasks, we continuously record the image brightness along with the eye tracking data. This is necessary, as the screen brightness during a task might depend on user actions (e.g., the set of web pages visited during the browsing task). To compute the brightness of the image, we compute the average brightness of all RGB pixels. As different components of an RGB colour contribute differently to the overall brightness (green being perceived as brighter than blue, for instance), we use the following formula, as proposed by the W3C[4]:

$$Br = (R \times 0.299) + (G \times 0.587) + (B \times 0.114)$$

Given a maximum value of 255 for each of the colour components, the brightness is a value between 0 and 255, with 0 being black and 255 being white. When examining the correlation between screen brightness and pupil diameter for the combination of all tasks, an average Pearson correlation coefficient of -0.68 was observed. This correlation is statistically significant ($p < 0.001$) and suggests a correction for this relationship would be beneficial.

In order to perform the actual correction, it is crucial not to use information derived from the tasks used for classification. Before starting the main tasks described in Section 3.4.3, we display a black screen, followed by a white screen for 20 seconds. We choose this order as the pupil's adaptation to bright light is almost instantaneous, whereas adaptation to darkness is gradual [80]. Based on these tasks, we observe that on average the pupil diameter decreases by 0.005 when brightness is increased by one. The difference in this slope between users is minimal, as such we use a single value for all users. This approach also has the advantage of not requiring a per-user calibration of this adjustment function. The corrected pupil diameter is then obtained according to the following formula:

$$d_{new} = d_{raw} + (Br \times 0.005)$$

[4]https://www.w3.org/TR/2000/WD-AERT-20000426#color-contrast, last visited 01/25/2016

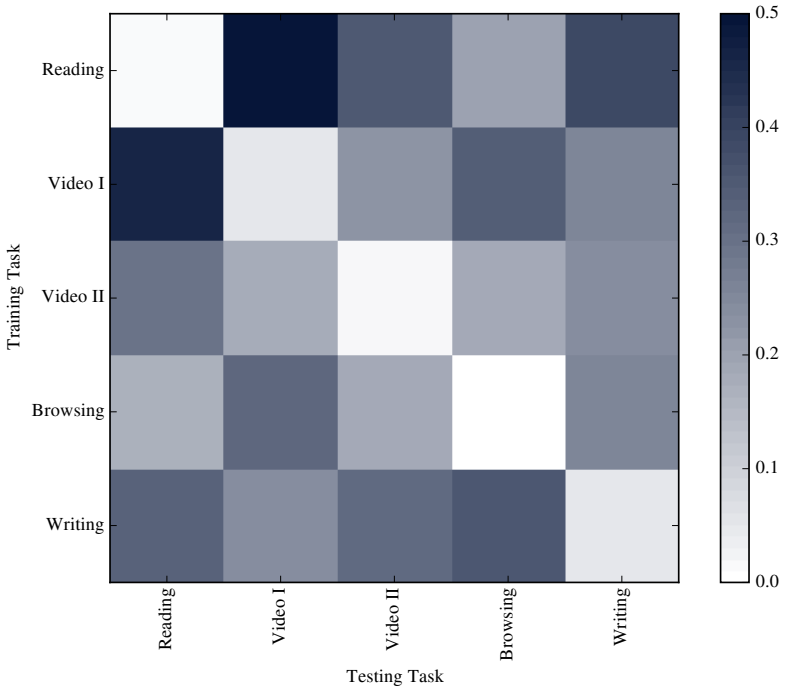
When using the pupil diameter correction, this formula is applied to all samples, both in the training and testing sets. After applying the correction, we observe no statistically significant correlation between corrected pupil diameter and screen brightness ($p > 0.05$) for any of the users in the dataset.

Evaluation

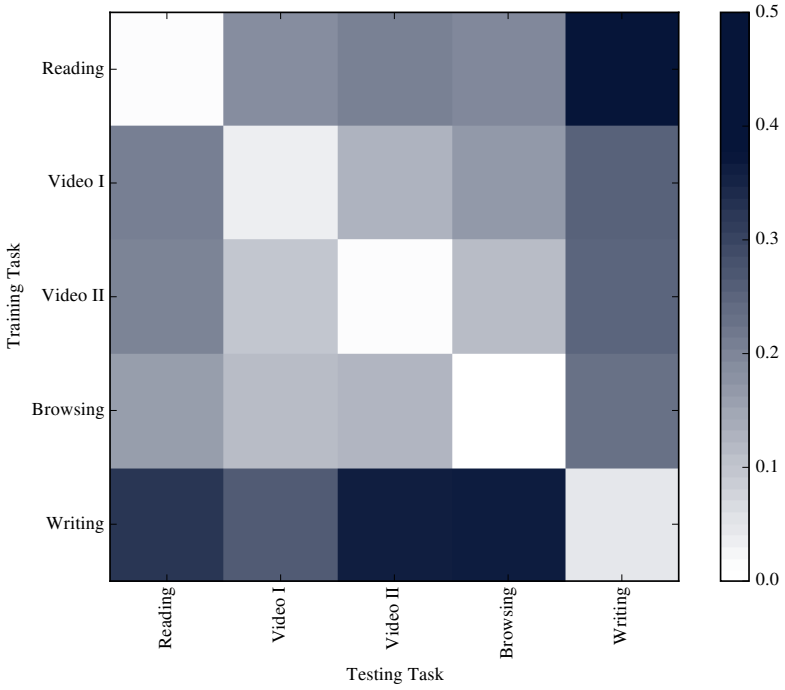
Figure 3.13 shows the results of our analysis. It is apparent that the EER is low when using data from the same task for training and testing, with values between 0.04 % for browsing and 4.9% for the first video. These error rates are comparable to those exhibited with the task familiarity experiment described in Section 3.4. Within a single task, the pupil diameter correction only provides a meaningful change for the first video, decreasing the EER from 4.9 % to 3.3 %. The remaining tasks exhibit significantly smaller variation in screen content brightness, which explains the limited benefit of correcting for this variation.

Classifier performance changes significantly when using one task for enrolment and another for operation. The most apparent increase in EER results in the combination of the first video and the reading task, regardless of which of these two tasks is used for training. In this scenario, the EER approaches 50%, thereby only providing a marginal over random guessing. This is likely the result of the pupil diameter being greatly affected by the bright background of the reading task compared to the relatively dark video. Consequently, the classification accuracy is actually worse than not using the pupil diameter at all, as the pupil diameter of Alice reading a text might match that of Bob watching a video. This EER is drastically lowered when applying the pupil diameter correction, reducing the error rate from 49 % to 18.7 %. A similar reduction can be observed for the other task pairs, with the one exception being the writing task.

When using writing data for either training or testing, performance is poor, regardless of whether or not the pupil diameter correction is used. The most likely cause of this lies in the imperfect nature of eye tracking when the user is not looking directly at the screen. During the experiment, we noticed that many users were either inexperienced typists or not familiar with the keyboard layout, resulting in them frequently looking



(a) Without pupil diameter correction



(b) With pupil diameter correction

Figure 3.13: Equal error rates for different combinations of training and testing task. The bottom figure uses the pupil diameter correction described in Section 3.6.4.

away from the screen. This apparent behaviour is backed by a drop in the sampling rate (when only counting valid samples) from an average of 470 Hz to 360 Hz for the writing task. However, once the device loses track of a person's eye there is a brief recovery period which results in low-quality tracking. In addition, movements of the user's head and even entire body cause an overall loss of accuracy. These factors lead to a fixation rate of 0.9 Hz, as opposed to the average of 4 Hz over the other tasks. Both the sampling rate and the EER suggest that the current eye tracking technology is not suitable for authentication during tasks centred largely on writing or other tasks that frequently draw the user's gaze away from the screen. While the gaps in the data created by these distractions will ultimately remain a problem, we believe that improvements in gaze tracking technology (especially given the rise of eye tracking in the entertainment sector) will likely reduce the length of these gaps and the loss of tracking accuracy around them, thereby mitigating this additional source of error.

The results show that the error rates are comparable to, or even lower than those observed for the task familiarity experiment. While the error rates increase significantly when performing enrolment on a different task this issue can be somewhat mitigated through pupil diameter correction. Additionally, the results suggest that grouping similar tasks (e.g., text-based tasks) for the purpose of enrolment might yield sufficiently low error rates, thereby balancing acceptable error rates with simplified enrolment. The generation of templates for different task groups could be coupled with automated task detection to further improve accuracy without giving up the benefits of transparent authentication.

3.7 Discussion

There are a number of possible limitations to consider when evaluating this work. All data has been collected in a lab study, while the features can be computed in any environment and error rates are low for all tested real-world tasks, the participants were still restricted in their actions. While this may seem like an obvious limitation, we argue that it is a necessary first step to draw meaningful conclusions about the biometric. Regardless of the number of subjects, there is always the danger of each subject choosing an individual (variation of a) task, which could lead to classification accuracy being overestimated as

the classifier distinguishes tasks instead of users. The same reasoning applies for the need to conduct the study in a controlled environment (i.e., in a lab study). Under the insider threat model the attacker would always use the workstation in the *same environment* as the victim, as the biometric is meant to secure local access. Consequently, the environment (and resulting factors such as lighting) have to remain the same for all users and this level of control can only be reliably established in a lab study. While a field study would certainly give interesting additional insights, for example by instructing users to attempt to impersonate their co-workers, we consider a lab study a necessary first step into investigating the suitability of eye movements as a biometric.

We performed the experiment with 30 users for the task familiarity experiment and 10 users for the analysis of task dependence. Naturally, a higher number of participants would give greater confidence in the robustness of our results. Nevertheless, our participants cover a wide variety of age groups, both genders and a number of participants with glasses or contact lenses (see Figure 3.5. Together with the narrow confidence intervals shown in Figure 3.11 this suggests that similar results could be expected in a larger study. Our recruitment process is based on social media and mailing lists, which, along with the natural selection of people willing to participate in experiments in general, might introduce a bias that influences our results. Additionally our sample might be subject to additional unknown sampling bias (as some subsets of the entire population may be particularly hard or particularly easy to distinguish), although this can not necessarily be avoided even with higher sample sizes if the source of the bias is unknown. So far there has been, to the best of our knowledge, no research exploring how the distribution of our features change across different subsets of the population, with the exception of the pupil diameter. Without establishing these effects first, it is hard to draw a sample from the entire population that is representative with regard to the biometric.

The threat model in this work assumes a zero-effort attack, with participants not actively trying to modify their own eye movements with regard to our feature set. If an attacker is able to record a victim's eyes, she might attempt to match her victim's eye movement patterns when attempting to access the victim's workstation. While it is, by the nature of the problem, impossible to show that such an attack is infeasible, we

consider it unlikely to succeed in practice. Medical research shows that subjects have not been able to permanently suppress microsaccades (the type of eye movement likely responsible for the distinctiveness of temporal features) and that temporary suppression leads to a higher rate of microsaccades shortly after [81]. Modifying the exact duration and magnitude of acceleration would arguably be even more difficult. To the best of our knowledge the only feature which has been shown to be susceptible to influence through stimulation is the pupil diameter. Besides manual imitation, another attack vector would be the creation of an artificial eye that moves according to the attacker's specification. However, due to the millisecond-scale of fixations the control would have to be extremely precise and the attack could be countered by implementing liveness detection (i.e., distinguishing between human and artificial eyes).

There might be factors influencing a person's eye movements we have not accounted for (such as fatigue, or the effects of medication or alcohol). While we have likely captured many different confounding factors by recording data across three sessions, more research is needed to measure potential long-term changes in eye movement patterns.

The transparent nature of the proposed authentication system allows the establishment or confirmation of a user's identity without her active cooperation or even knowledge. This property, which is shared by other behavioural biometrics (such as mouse movement behaviour), poses a privacy concern. However, the work presented here performs authentication, as such it is necessary for the user to make an initial identity claim (e.g., by entering a user name). This necessity mitigates the privacy concern, as the system does not allow identifying an anonymous user. The use of eye movement technology in general still raises concerns due to their diagnostic capabilities (see Section 3.2), although these are not the focus of this thesis.

3.8 Future Work

Based on these results, we have identified several future research directions.

Ambient light correction. In this chapter, we measure the brightness of the screen content and correct its effect on the pupil diameter. In future work we plan to extend the same approach to ambient light. This will allow us to account for varying levels

of daylight and artificial lighting. This can be achieved by incorporating a brightness sensor in the eytracker itself or make use of existing sensors (e.g., those used in smart homes). This correction would also defeat active attacks based on changing the ambient light intensity to modify the attacker's pupil diameter (although attacks with very targeted light would still be possible).

Binocular tracking. We operated the eyetracker in a mode that only reports the average of the gaze position of both eyes. Binocular tracking (i.e., independent tracking of each eye) might enable us to discover additional features. Promising candidates will be the difference in gaze position as well as pupil diameter between both eyes.

Uncalibrated data collection. In this experiment, the eye tracker has always been calibrated to achieve maximum accuracy (i.e., to ensure that the reported and actual gaze position are as similar as possible). However, most of the features rely more on precision than accuracy. We never use the actual gaze position but instead base features on the relationship of samples within one fixation. As such, we hypothesise that the effect of skipping the calibration and the resulting lower accuracy on features will be relatively minor.

Challenges of practical deployment. As this experiment was conducted in a lab study, it is difficult to predict whether similar performance would be achieved in an uncontrolled setting. Such a setting raises challenges both relating to the user and the operation of the system itself. User-centric challenges involve distractions and changes in posture that move the user's eyes out of the tracker's field of vision. Even in a controlled setting, we have observed frequent gaps in the data while users are looking at the keyboard.

In order to account for the task dependence of features, it would be necessary for the system to reliably determine the current task and select the most appropriate template. This also requires the selection of a task set for the initial training phase in order to directly or indirectly cover the majority of day-to-day tasks. Following the initial training phase, the templates for one or more tasks have to be updated to account for behavioural drift. The main challenge with template updates is that the user's identity has to be firmly established in order to prevent poisoning or replacement of templates.

Il est dangereux d'avoir raison dans des choses où des hommes accrédités ont tort.

It is dangerous to be right in matters where established men are wrong.

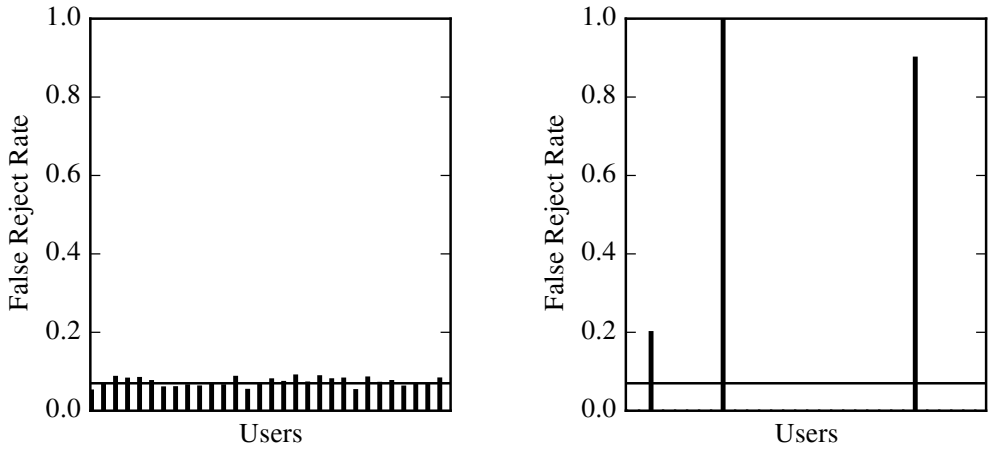
— Voltaire

4 Metrics for Continuous Authentication

Contents

4.1 Motivation

The extensive body of work on behavioural biometrics calls for reliable ways to compare different systems when faced with the choice of which one to implement. In addition, developers will want to have realistic ideas of what security gains can be expected from using biometric recognition systems. As basis of their analysis, most papers (including



(a) Random Errors (b) Systematic Errors

Figure 4.1: Different distributions of the FAR lead to different security challenges, random errors and eventual detection (left) and systematic false negatives (right). The grey line denotes the (identical) 9% FAR of both samples. Note that this figure shows the success of a single attacker in impersonating multiple victims.

the work presented in the previous chapter) collect a static biometric dataset from a number of users. This data is then used to calculate biometric features, which are ultimately classified through statistical or machine learning approaches. This approach is used because it separates data collection from processing and is therefore particularly convenient for research. However, it also means that the operation of a (hypothetical) continuous authentication system has to be simulated based on the previously collected data. As we will show in this chapter, assumptions made when performing this simulation do not always reflect how the system would perform in the real world. As such, the reported theoretical performance measures may be at times misleading, which also leads to increasing difficulties in comparing work conducted by different researchers. We will show that this difficulty stems from three major components: (a) insufficient metrics, (b) different approaches in temporal selection of training data and (c) modelling of the attacker in training data.

The previous chapter has briefly outlined the danger of systematic errors, with a particular focus on systematic false negatives (i.e., perpetually undetected attackers in a continuous authentication system). The great disparity in error rates between users has

first been noted in 1998 by Doddington et al. in the context of speaker recognition [82]. The researchers categorise users into lambs, sheep, goats and wolves, a terminology later nicknamed the ”biometric menagerie“. Despite these early insights, systematic errors are not adequately characterised in the average error rates typically reported in most authors' analyses. Our work in Chapter 3 further confirms that error distributions are often skewed, which means that the average error rates are hardly indicative of “typical” system behaviour.

The second concern is the temporal selection of training data. In real world system operation, the training phase has to take place before any new samples can be classified. While re-training and template updates are possible during the system's operation, this is often costly due to the requirement of strong authentication to prevent template poisoning. This approach of conducting the entire (monolithic) training phase before classifying new samples leads to a problem for biometrics with poor time stability. As the time distance between new samples and the completion of the training phase grows, the accuracy of classifiers is likely to drop. While this is unavoidable in practice, it can be “avoided” in the analysis of a static dataset (albeit without any real improvement of the system). As the entire temporal scope of the dataset is available for the training phase, training samples can be sampled randomly (rather than in a single consecutive chunk). This can be achieved through stratified cross-validation, which ensures each sample is used for testing exactly once. This approach significantly decreases the time distance between a previously unseen testing sample and the “closest” training sample.

The third component impeding the comparability of different systems lies in the modelling of the attacker. Most systems assume a zero-effort threat model in which the attacker does not make a conscious effort to circumvent or trick the authentication system. In the context of a static dataset, this is achieved by selecting one user as legitimate and having all other users act as zero-effort attackers. As such, the researchers possess data for all potential (zero-effort) attackers when performing the analysis. However, this is not true in the real world, outside of specific scenarios such as those exclusively dealing with insider threats. One approach to deal with this discrepancy is to exclude all of the eventual attacker's samples from the training phase (thereby treating her as an outside attacker).

Any analysis that includes some of the attacker's samples in the training phase is likely to report lower error rates without any meaningful improvement of the system itself.

The goal of this chapter is three-fold: First, we will summarise the state of the practice. This allows us to learn which metrics are commonly used in academic papers and what decisions authors make with regard to training data selection and attacker modelling. Secondly, we will collect biometric datasets for four biometrics to observe how they perform in regard to previous and newly introduced metrics. Lastly, we will quantify the impact that the effect of training data selection and attacker modelling have on error rates. Ultimately, the combination of these steps will allow us to determine to what degree different papers can be compared today, and which steps can be taken to improve comparability in the future.

4.2 State of the Practice

In this section, we present a rigorous analysis of the state of the art, both with regard to metrics reported and the machine learning methodology used to obtain the results. In order to cover a cross-section of the field, we have analysed 25 systems based on five different biometrics with a focus on recently published work. While these systems differ in experimental design and underlying features, they all provide continuous authentication. This reflects the focus of this thesis in general and the unique challenges of continuous authentication raised in this chapter's motivation. As such, we do not consider systems that provide enhanced biometric-based login time authentication (such as password hardening or fingerprint scanning).

4.2.1 Metrics

The goal of a continuous authentication system is to quickly identify imposters without incorrectly rejecting a legitimate user. In order to determine which metrics are typically used to quantify these characteristics, we have analysed 25 systems based on five different biometrics. The results of this survey are shown in Table 1, see Figure 4.2 for a summary. In Figure 4.2, we consider explicitly reported metrics (i.e., those reported directly in

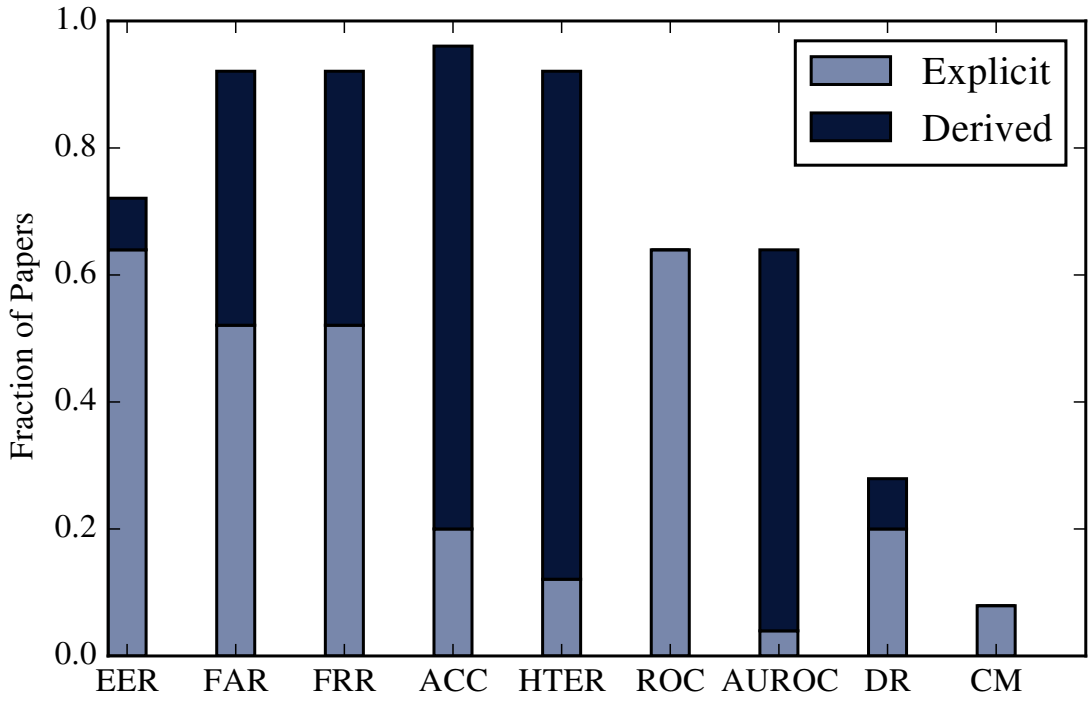


Figure 4.2: Metrics reported in literature

a table or figure) and derived metrics which can be calculated from others. The total set of metrics reported across these papers is follows:

False Accept Rate (FAR) is typically measured as the fraction of intruder *samples* (rather than intruders) that are incorrectly accepted. As such, it is distinct from detection rate and the sys-fn metric reported in the previous chapter. Note that ”intruder“ typically still refers to zero-effort attackers (i.e., other users in the same sample) rather than active attackers.

False Reject Rate (FRR), also known as the False Match (FM) or False Positive (FP) rate, is the fraction of benign samples that are incorrectly rejected.

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

Accuracy is the fraction of samples that is accurately classified, without distinction between the two error types.

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

The *Receiver operating characteristics (ROC) curve*, while not a metric, is a plot that shows the dependency between the FAR, FRR and the system's detection threshold. The ROC curve allows to derive a set of pairs (FAR,FRR) at which the system can be run by changing the threshold settings. As such, it also allows derivation of the EER.

The *Area under the ROC Curve (AUROC)* ranges from 0.5 (random guessing) to 1 (perfect classification) and aggregates the system's performance at all threshold settings.

Detection Rate is a measure of the fraction of attackers that are detected by the system. Unlike the FAR it operates on individual users, rather than samples. As such, it is related to the previous chapter's sys-fn metric (with $DR = 1 - sys - fn$).

The *Confusion Matrix (CM)* plots the fraction of accepted samples for each user pair. As such, it is a representation of raw data, rather than a single numerical metric. The CM shows the FRR for each user on the diagonal and the FAR for each user-attacker pair on the remaining fields. However, as the number of user pairs scales quadratically with the number of users, the space requirements are high for large number of users. In addition the CM is usually given as a plot, which somewhat reduces the space requirement but makes it difficult to obtain more than estimates of the actual numerical results.

Table 1 shows that the EER, as well as derived metrics, are reported by the vast majority of papers, regardless of the biometric. In addition, a plot of the ROC curve is given in 16 out of the 25 reviewed papers, although the AUROC is rarely given as a number (and could only theoretically be extracted from the plot). Reporting of the detection rate is extremely rare, and due to the unknown distribution of errors between attackers it can not be derived from the FAR either. A confusion matrix, which allows the derivation of all other metrics, is only given in two papers, most likely due to the high space requirements.

Limitations

As shown above, EER, FAR and FRR are the most prevalent metrics by a wide margin. However, most works merely present the average over all users or user-attacker pairs. Nevertheless, an intuitive assumption is that a lower EER results in attackers being detected more quickly (and more attackers being detected overall) and users being rejected less frequently. Overall, an EER is assumed to be indicating a better system

and/or feature set. In the context of one-time (i.e., not continuous) authentication this is a sensible and widely accepted metric. However, continuous authentication provides a unique challenge as errors accumulate over the runtime of the system. Without knowing the exact distribution, an FAR of 10% could signify all attackers being detected 90% of the time (resulting in eventual detection), or 10% of the attackers never being detected while all others are exposed immediately. The second scenario exhibits so-called *systematic false-negatives*. These different scenarios are illustrated in Figure 4.1. Unlike regular false negatives, which might be randomly distributed across victim-attacker pairs as well as across the time of a session, systematic false negatives are tied to a combination of attacker and victim and are usually more persistent or even permanent as a result of the behaviour of two users being very similar. These types of errors are more problematic from a security perspective, as the undetected attackers can then access the compromised system for a virtually unlimited time. Part of this property is captured through the detection rate, which measures the fraction of attackers with a non-zero FAR. However, the metric does not account for the difference between undetected attackers and those with simply a very high FAR. In practice, this might even be determined by a single sample being classified differently. The confusion matrix paints a complete picture, but it is neither compact enough to report for large datasets, nor does it enable readers to easily compare two systems. Most likely, these limitations are the reason it is rarely reported in the literature. The authors of [83] propose to report the number of undetected attackers along with the average number of imposter actions (ANIA), a metric related to the false accept rate. However, they recommend reporting only the ANIA (which is, by definition, an average value), with no regards for its distribution between attackers.

While systematic errors are problematic for the FAR, this type of distribution might be desirable for false rejects. A seemingly low, but non-zero false reject rate for all users might still lead to frequent false alarms due to the base rate fallacy [84] if the system is run continuously throughout the day with a moderate sampling rate. If the false rejects were concentrated on few users they could be authenticated through other means (such as a different biometric) instead, without compromising security for the remaining users.

In addition, such a scenario allows the developer of a biometric recognition system to analyse why the system performs poorly for precisely these users.

Ref	Biometric	EER	FAR	FRR	Acc	HTER	ROC	AUROC	DR	CM
[22]	Touch	✓[1,3,4]	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[24]		✗	✓	✓	(✓)	(✓)	✗	✗	✗	✗
[85]		✓	(✓)	(✓)	✓	✓	✓	(✓)	✗	✗
[86]		✗	✗	✗	✓	✗	✗	✗	✗	✗
[87]		(✓)	(✓)	(✓)	(✓)	(✓)	✓	✓[2]	✗	✗
[88]		✓	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[89]		✗	✓	✓	(✓)	(✓)	✓[2]	(✓)	✓	✗
[90]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[91]		✓	✓	✓	(✓)	✓	✓	(✓)	✗	✗
[92]		✗	✓	✓	✓[3]	(✓)	✗	✗	✗	✗
[93]		✓[4]	✓[4]	✓[4]	(✓)	(✓)	✗	✗	✗	✗
[94]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[4]	Gaze	✓[4]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✓	✗
[34]		✓	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[5]		✓[4]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✓	✗
[95]	Pulse Response	✓	✓	✓	✓	(✓)	✓	(✓)	✓	✗
[96]	Gait	✓	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[97]		✓	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[30]		(✓)	✓	✓	✓	✓	✓	(✓)	✗	✗
[98]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[99]	Mouse	✗	✓	✓	(✓)	(✓)	✗	✗	(✓)	✓
[100]		✓	✓	✓	(✓)	(✓)	✓	(✓)	(✓)	✓
[101]		✓[4]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[16]		✗	✓[4]	✓[4]	(✓)	(✓)	✓	(✓)	✗	✗
[102]		✗	✗	✗	✗	✗	✗	✗	✓	✗

✓Explicitly reported (✓) Derived from other metric ✗Not reported Unless indicated otherwise, only means are reported
[1] min, max, median [2] individually for each user [3] as a function of number of users
[4] as a function of number of samples

Table 4.1

4.2.2 Evaluation Methodology

The error rates resulting from the analysis are determined by two key factors: the quality of the dataset itself and the evaluation methodology. The former depends on a variety of

factors, including the level of control (with more controlled environments often leading to lower error rates), selection of users and involvement of the experimenter. The latter is determined by the methodology used when simulating the authentication system's operation on a static, previously collected dataset. As outlined in the motivation, we will focus on this aspect. We perform a literature survey with regard to the following methodological components that have the potential to impact error rates:

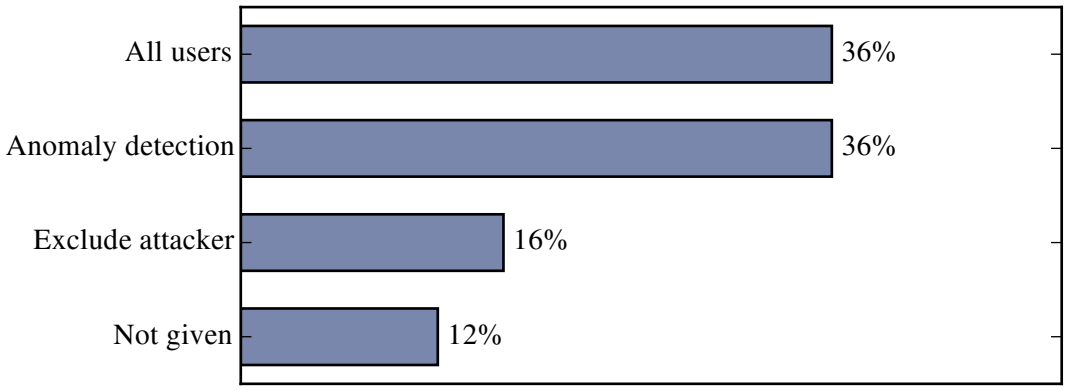
Hyperparameter Tuning. Following the feature extraction and normalization, a suitable classifier has to be chosen. Depending on the classifier, a number of hyperparameters have to be instantiated. Such parameters include the number of data points (the value of k) in the k-nearest-neighbours algorithm and the kernel type and soft margin constant C of a Support Vector Machine. These values are normally determined by testing the performance of different values via a grid search. However, this search has to be performed on a development set different from that used for training and testing.

Attacker Model. Most biometrics are evaluated without a committed attacker in mind, this is commonly referred to as the zero-effort threat model. As such, the "attacker" is another user that attempts to access the victim's system without taking action to either circumvent the authentication system or impersonate the legitimate user. Even in this simplified threat model, it is still necessary to test the system's performance in detecting intruders. This is commonly achieved by comparing a user's template against the samples of all other users (i.e., the "attackers"). An important concern is the building of the user model itself. A common choice is to train a binary classifier with one user's samples as the positive class and samples from all other users (including the eventual attacker) as a single combined negative class. The system is then "attacked" individually by each of the users that jointly form the negative class. This approach means that reference data of the attacker is included in the negative class, even though it only forms a fraction of the overall negative class. In practice, it is impractical to assume that reference data for each potential attacker is available (aside from specific insider threat scenarios, such as [4]) and including this data may lead to overestimating the classifier's performance. We refer to this approach as the **all-users** model. A different approach trains a generic attacker model from other users (again, combining them into a negative class), but withholding

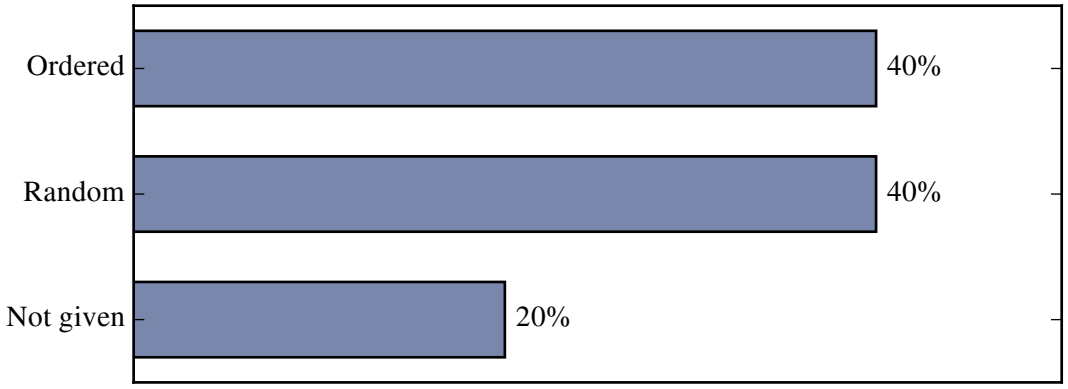
samples from the actual attacker. We refer to this as the **exclude-attacker** model. The authentication system could then be shipped with this (anonymised) reference data. These two scenarios are also considered in [83] and referred to as external and internal scenarios, respectively. A more straight-forward approach is to perform **anomaly detection**, which trains a model from a single user's data without the requirement of providing samples for a negative class. New samples are then classified based on how similar they are to the training examples. This avoids including the attacker in the negative class by virtue of not having a negative class at all.

Selection of Training Data. An operational authentication system always requires reference data for each legitimate user (training data) in order to classify new observations. In practice, the initial training has to occur before any samples can be classified (although the model can be updated based on new observations). Consequently, a common approach to simulate this setting is to use the first part of the recorded data as training data, and the remaining samples as test data. We will call this **ordered training data selection**. Another approach is to randomly sample the training data from the entire dataset, and to use the remaining data for testing. The sampling is often repeated to provide statistical robustness (either by performing several iterations of random sampling or through cross validation). However, this approach violates the requirement that training always has to precede testing (as some training samples may have been recorded after some testing samples). For the rest of the paper, we will call this **random training data selection**, regardless of the specific process of randomisation.

Sample Aggregation. Single feature vectors are often noisy (due to measurement noise, erratic user behaviour or generally low discriminative power). In order to combat this, several samples can be combined to increase robustness. Samples can either be combined before classification (e.g., by computing the component-wise mean of several feature vectors) or afterwards (e.g., by majority votes). In the latter case, instead of simply using the classifier output, it is also possible to use the classifier confidence for each class. Classifier confidence can be measured as the distance to the decision boundary in an SVM or the number of nearby examples of each class for knn.



(a) Attacker Models



(b) Training Data Selection

Figure 4.3: A large fraction of papers use random training data selection and inclusion of imposter data in the training set, both of which are likely to underestimate error rates.

The results of our survey with regard to the above methodological components are shown in Table 4.2, a summary is shown in Figure 4.3. One of the most important observations is the (apparent) reluctance of researchers to make their data freely accessible online. This is a major obstacle when attempting to reproduce individual results. However, it should be noted that our survey only accounts for data that is both available online and referenced in the corresponding paper. We have not contacted individual authors and can not make any statement on their willingness to share data on request. The number of papers using and building on this shared data (most notably, the data published as part of Touchalytics [22]) highlights that this is a valuable contribution to the community. In a similar fashion, the code used to generate the results is not usually published. As

			Hyperparameters					Available Online	
Ref	**Biometric**	**Classifier**	**Values**	**Method**	**Attacker Model**	**Training**	**Sample Aggregation**	**Data**	**Code**
[22]	Touch	SVM,knn	✓	CV	all users	ordered	weighted	✓	✓[1]
[24]		knn	✓	✗	✗	ordered	✗	✗[2]	✗
[85]		SVM	✗	✗	subset	CV-10	majority	✓	✗
[86]		DT	✓	GS+CV	✗	CV-3	✗	✗[2]	✗
[87]		SVM	✗	✗	✗	✗	✗	✗	✗
[88]		sim-score	N/A	N/A	AD	random	N/A	✗	✗
[89]		NN	✓	✗	AD	ordered	N/A	✗	✗
[90]		knn	(✓)	✗	no-attacker	✗	✗	✗	✗
[91]		NN,SVM	✓	✗	all users	random [5]	✗	✗	✗
[92]		SVM,RF	✗	✗	AD	✗	✗	✗	✗
[93]		HMM	✓	CV-5	all users	ordered	mean	✗[2]	✗
[94]		SVM	✓	✓	AD	ordered	N/A	✗	✗
[4]	Gaze	SVM,knn	✓	GS+CV	all users	CV-5	majority	✗	✗
[34]		UBM	✓	✓	all users	✗	N/A	✗	✗
[5]		SVM,knn	✓	GS+CV	AD, all users	CV-5	majority	✗	✗
[95]	Pulse Response	SVM,knn	✓	✓	all users	CV-5	N/A	✗	✗
[96]	Gait	sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[97]		sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[30]		sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[98]		sim-score	N/A	N/A	AD	random	N/A	✗	✗
[99]	Mouse	DT	N/A	N/A	all users	ordered	weighted	✗	✗
[100]		NN	✓	✓	no-attacker	random	N/A	✗	✗
[101]		sim-score	N/A	N/A	AD	✗	N/A	✗	✗
[16]		SVM	✗	✓	no-attacker	ordered	mean	✗	✗
[102]		SVM	✗	✗	all users	random	N/A	✗	✗

✓Reported (✓) Partially reported ✗Not reported
[1] Only feature extraction [2] Uses data from [22] [5] Sampling repeated 10 times

Table 4.2: Simulation Design Choices in Related Work

a number of machine learning steps depend on random numbers (e.g., for randomised training data selection), this might make it particularly difficult to reproduce exact results, even if all decisions are clearly stated and raw data is available.

While the specific values for hyperparameters are often given, the process with which they were obtained is not usually explicitly described. This is problematic, as the selection process is far more interesting (and the values used for an individual datasets are unlikely to be optimal for others). In addition, some processes (such as validating parameters on the entire dataset, instead of just the training or development set) might artificially improve reported results, without resulting from a better system. However, unless the process is documented, it is impossible to tell whether this has been the case.

The vast majority of papers either do not use aggregation of samples, or don't report on the specifics of their mechanism. If samples are aggregated, this is usually done following classification (i.e., not on a feature vector level). 6 out of 25 papers do not report on the use of sample aggregation. This is particularly concerning as sample aggregation can easily reduce the error rates by a factor of 10 (see, e.g., [4]).

Limitations

The previous section has shown that a wide variety of methodologies are used to evaluate the static datasets, which suggests that it might not be possible to directly compare papers even if they use similar metrics. This would not necessarily be a problem if the impact of different methodologies on the reported metrics were to be comparatively small. To the best of our knowledge, this effect has not been quantified in the context of continuous authentication. It is, however, well-studied in malware detection. Specifically, Allix et al. have shown that sampling training data randomly from all available data leads to systematic underestimation of error rates [103, 104]. This is problematic, as reference data for future malware helps in the classification, but would not necessarily be available in the real-world (i.e., to classify newly observed malware). One might assume a similar effect for continuous authentication, as random training data selection would make future samples available to help classifying past ones. This allows the classifier to accurately account for short and long term changes in user behaviour, which would not be possible

when maintaining the temporal integrity of the dataset. The only way to achieve a similar reduction of practical error rates would be through continuous classifier re-training and template update. However, this comes with additional challenges as the user identity has to be established "out-of-band" to prevent template poisoning.

9 out of 25 papers model the attacker by merging all users but the legitimate one into a single negative class, with a further 3 not giving information on their methodology (see Figure 4.3). This approach is somewhat unrealistic, as it assumes reference data for every potential attacker. While this is possible in pure insider threat scenarios (such as a company that wants to detect employees using their co-workers' systems), it is less realistic for other scenarios, such as a stolen phone or any other kind of outside attacker. As the attacker is merged with all other users into a single negative class, the effect might be relatively small, especially for datasets with larger numbers of users. However, the impact of this potential source of additional information for the classifier has to be quantified in order to allow a more informed comparison of papers. 13 papers exclude the specific attacker from the training set, or only perform anomaly detection (i.e., train the model without reference data for any attackers), thereby escaping this problem.

4.3 Evaluation Datasets

In the previous section, we have outlined potential limitations of both commonly used metrics and evaluation methodologies. In order to evaluate the impact of both, we require diverse biometric datasets suitable for continuous authentication. This allows us to gauge if and to what extent these limitations apply to different biometrics. As such, we require datasets covering multiple biometrics and ideally several datasets per biometric (to account for different data collection methodology). For this analysis we use 13 datasets obtained from related work and 3 datasets collected for this study. This section will give a brief overview about the methodology used for each of them.

4.3.1 Gait Biometric

Due to its simplicity and real-world applications, we focus on accelerometer-based gait recognition (as opposed to video-based recordings). We adapt the classification

process of [29] to support continuous authentication. The classification process is identical for both datasets.

We recruited 14 volunteers, 9 male, 5 female. The experiment was carried out with the approval of the University of Oxford's Central Research Ethics Committee, reference number SSD/CUREC1/13- 064. During the experiment, each subject walked an identical 300 meter long route on a footpath in the university parks and returned to the starting point, resulting in two datasets of roughly identical length for each participant. The route was straight and did not involve turns. Data collection was manually stopped before the halfway turn and resumed afterwards. The accelerometer data was collected with an off-the-shelf Samsung Galaxy Note 4 smartphone at a sampling rate of 200Hz. The phone was contained in a standard running armband strapped to the participant's lower leg, just above the calf muscle. On average each dataset contained 190 seconds of accelerometer data, or 38,000 raw samples. Using this data we obtained an average EER of 8.44%. We refer to this dataset as **Gait I**.

The second gait dataset was obtained from the authors of [105]. The set contains data from 27 participants that walked along a footpath at three different paces. While the data was collected for the purpose of evaluating step-counting algorithms, the data format makes it suitable for authentication as well. The data was collected through the accelerometer of a smartphone held in various positions (in a front or back trouser pocket, in a backpack/handbag, or in a hand with or without simultaneous typing). Not all sensor positions are available for each subject. In order to remove potential distinguishing information resulting purely from the sensor position, we only use the subset of traces in which the device was held by the subject without simultaneous typing, limiting the number of subjects to 24. The data was collected at a rate of 100Hz, with an average of 4400 samples (or 44 seconds) per subject. For each subject we extract the portion of the trace during which the subject was walking, using the timestamps provided as part of the dataset. As the first half of the data is used for training it contains mostly slow movements, unlike the testing timeframe during most of which the subjects were moving at a quicker pace.

The system shows an EER of 28.4%. This relatively high value (especially compared to the dataset collected by us) might also be a consequence of a mismatch between

training and test data (which were gathered at different walking speeds). We refer to this data as **Gait II**.

4.3.2 Mouse Movement Biometric

In addition to the gait data, we conduct an experiment to collect volunteers' mouse movements. Our experimental design is conceptually close to that in [106]. During the experiment, each participant was shown 25 rectangles arranged in a 5x5 grid, one of which was red. The user is then asked to click on the red rectangle. This task is repeated 200 times, with the red rectangle appearing in a new, random location for each iteration. The random seed to generate the sequence was kept identical for all users in order to limit the effects of the rectangle's position on our features. The size of the window displaying the rectangles was fixed in order to avoid any distinctiveness created solely by different screen resolutions. In order to control for artificial bias created by different input devices [18], we collect two datasets. The first set was obtained by sending our software to subjects, to be run on their own home or work machine. For the second set we invited a (different) set of volunteers to take part in the experiment on our lab machine. If any features are more distinctive in the first set this would imply that their distinctiveness is at least partially due to the properties of different devices, rather than differences in user behaviour.

We achieve an EER of 9.98% for the lab dataset that decreases to 9.22% when using the data gathered on subjects' machines.

4.3.3 Eye Movement Biometric

We use the two different sets of eye movement data collected in the previous chapter. The first dataset, referred to as **Eye I** represents the data of the task familiarity project. This dataset is further subdivided according to the data collection timespan (yielding the Intra, Inter and 2-weeks divisions) and two featuresets (complete and excluding the pupil diameter). These two divisions result in 6 datasets with 30 users.

The second dataset (labelled **Eye II**) is based on the task dependence experiment and reflects the four tasks of reading, writing, browsing and watching two different videos.

For this dataset, we consider both intra-task (training and testing within a single task) and cross-task classification. This dataset includes samples from 10 users.

4.3.4 Touch Dynamics

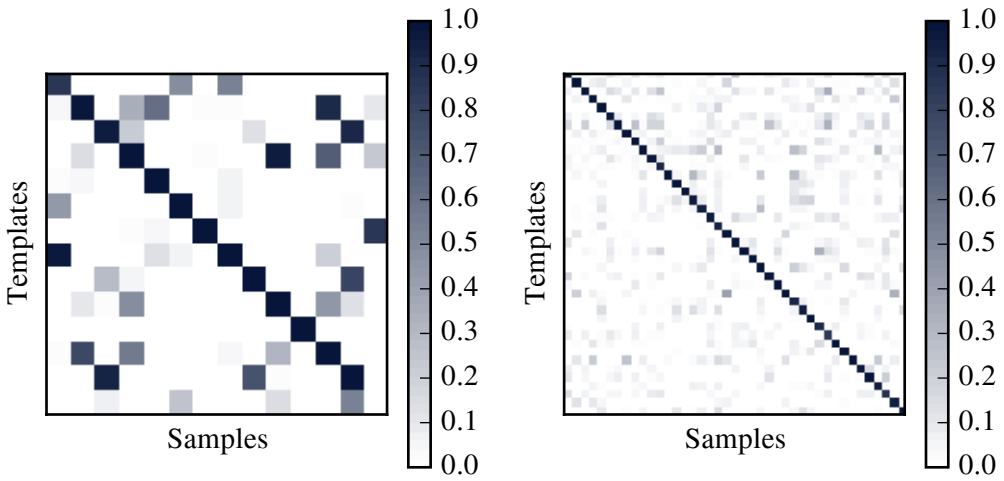
The touch dynamic dataset is based on the data shared in [22] and includes 41 users. The biometric's features describe the properties of swiping motions on touchscreens, including their position, curvature and pressure. Data was collected over two weeks, resulting in an intra-session, inter-session and 1-week dataset. The error rates range from 0% for intra-session authentication to 4% when authentication is performed a week after enrolment. As we are interested in determining the distribution and causes of errors we do not use the intra-session dataset for our comparison.

4.4 Measuring Skewed Feature Distributions

In this section, we investigate the previously described datasets with regard to the distribution of their errors. Based on the insights gained from this analysis, we propose different metrics that reflect these distributions, augmenting the state of the practice outlined in Section 4.2.

4.4.1 Systematic Errors in the Wild

The most complete way to visualize the exact distribution of errors (both FAR and FRR) is a confusion matrix. A confusion matrix shows the fraction of accepted samples for each combination of template and samples (see Figure 4.4 for an example). As such, the TPR (i.e., 1-FRR) is shown on the diagonal and the remaining fields show the FAR for each combination of attacker and victim. The confusion matrix of an ideal system would be 1 on the diagonal and 0 otherwise. As discussed in Section 4.2.1, systematic false negatives (i.e., attackers that consistently remain undetected) are a more severe problem than a moderate, low-variance FAR for all attackers. This is due to the nature of continuous authentication, which requires an attacker to consistently fool the authentication system, rather than only succeed once. In our datasets we actually observe both of these scenarios, leading to a need to accurately distinguish them without the need

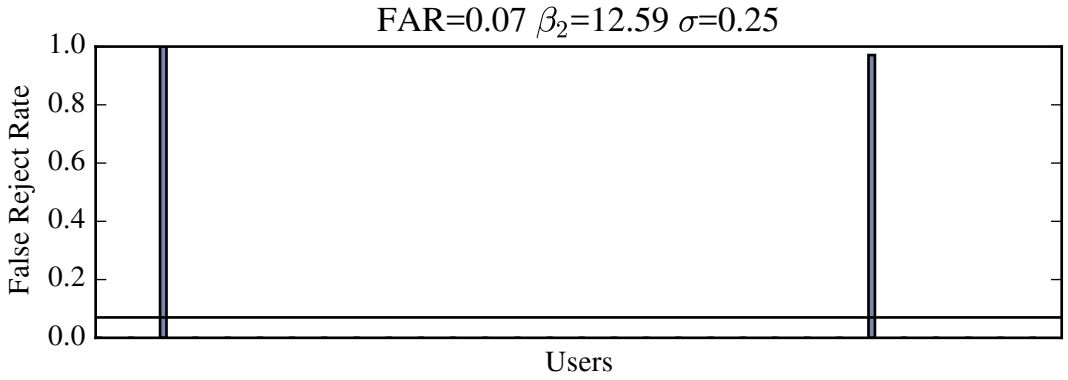


(a) Gait (b) Touch

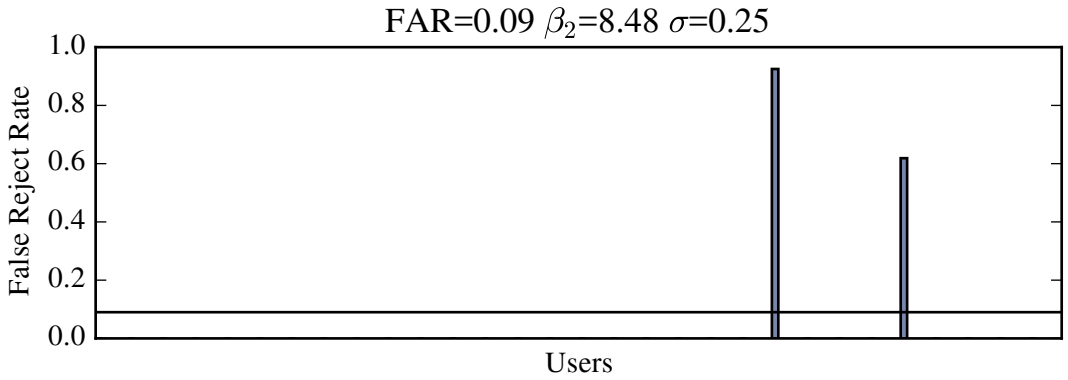
Figure 4.4: Confusion matrices for two biometrics. Each cell shows the fraction of accepted samples for that combination of template ID and samples ID. As such, the diagonal shows legitimate users and the remaining cells attackers.

of manually examining the confusion matrix. Figure 4.4 suggests that the gait biometric show a high number of extreme outliers for the FAR (as indicated by the dark spots off the diagonal). Conversely, the false accepts seem to be more evenly distributed between attackers for the touch input biometric, suggesting it would be better suited for continuous authentication from a pure security perspective.

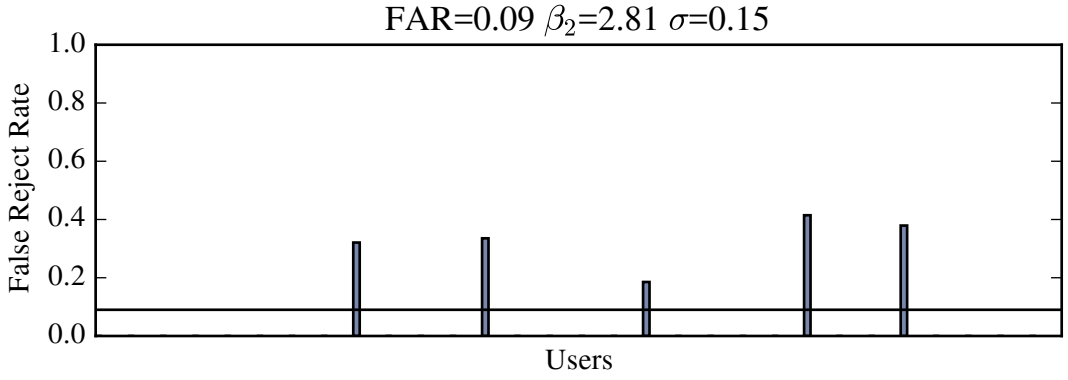
For the FRR we observe similar differences in distributions, although the consequences are different. Systematic rejections of individual users might indicate erratic behaviour (such as excessive head movements or poor calibration for the eye movement biometric), while even distributions of errors suggest a lower distinctiveness of features in general. The former could be mitigated by examining the root cause of error for the affected users and, if these can not be fixed, authenticating users through a different mechanism. Multimodal authentication systems are particularly well-suited for this, as they can dynamically choose biometrics that work well for this specific user. As such, biometrics where the FRR is focused on few users might be easier to use in practice. Figure 4.5 shows the distribution of the FRR for different over-time datasets for the eye movement biometric. Errors are focused on few users given a short time-distance and start to evenly affect more users over two weeks.



(a) Intra-Session



(b) Inter-Session



(c) Over Two Weeks

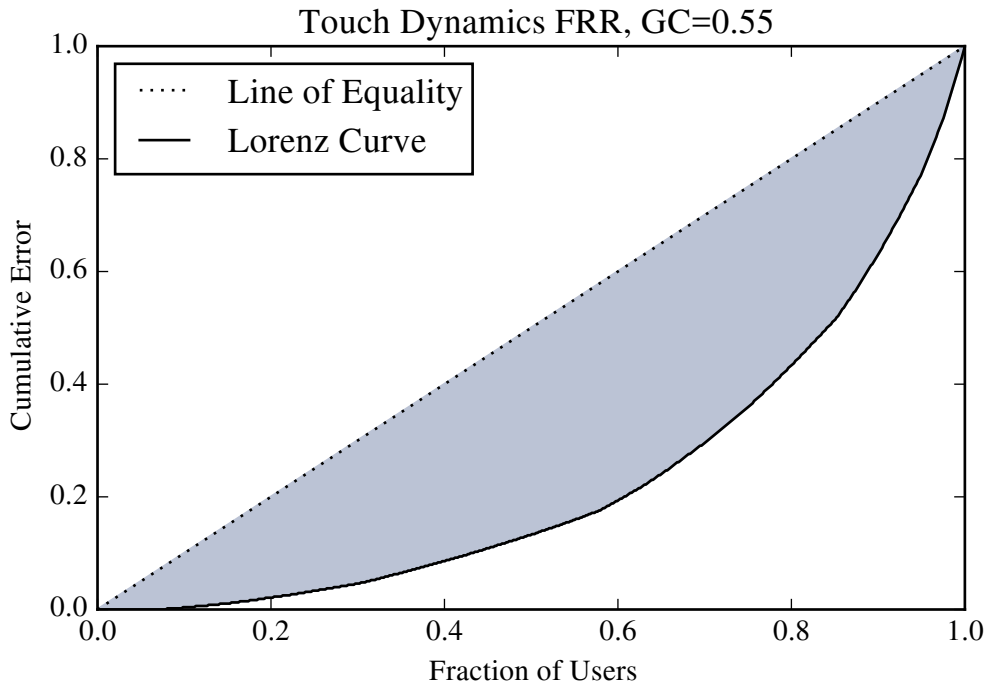
Figure 4.5: Distribution of the FRR between users for three different datasets based on the eye movement biometric using all features. While the average FRR is similar for all datasets, the distributions are not. The two-weeks dataset shows moderate error rates for many users while the errors are concentrated on few users for the other two. This property is modelled by the kurtosis and to a lesser degree by the standard deviation.

4.4.2 Metrics to Quantify Systematic Errors

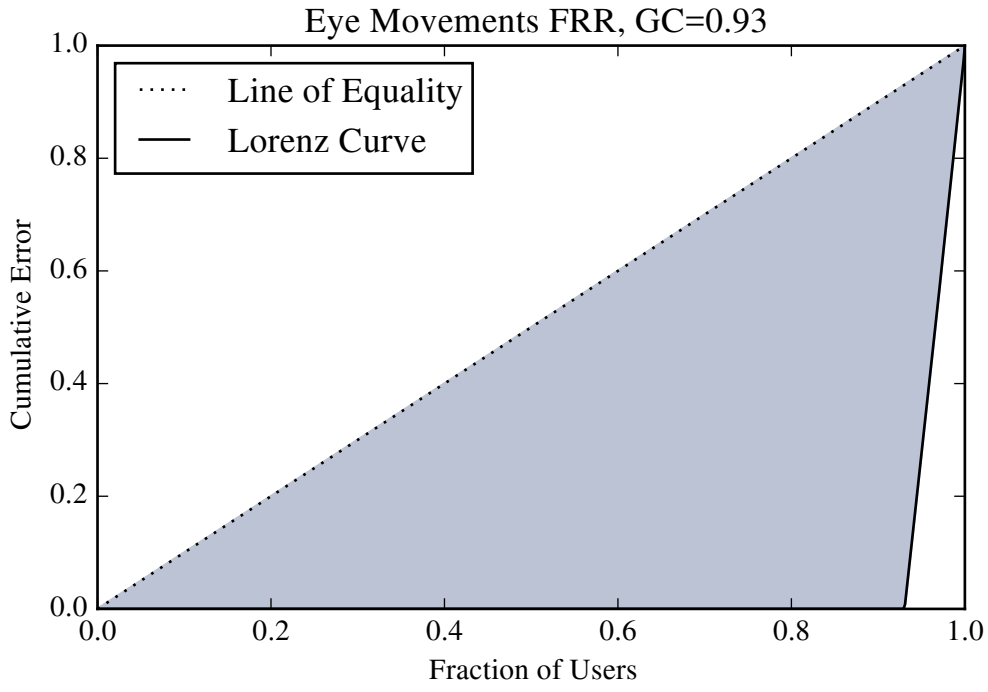
The previous section shows that biometrics exhibit different degrees of systematic errors. In this section we will discuss a number of statistical measures to better capture systematic errors and apply them to our evaluation datasets.

False Accept Rate

As discussed above, the false accept rate should ideally spread out evenly across attackers and therefore minimize systematic errors. In order to reflect systematic false negatives it might be an obvious choice to report the maximal FAR observed, this would then allow to give estimates of the maximal time it takes to find an attacker. However, Table 4.3 shows that this measure is 1 for the vast majority of datasets, suggesting at least some degree of systematic errors for most biometrics. In addition, it would unfairly penalize larger datasets, as the probability of the set including two very similar users increases with the sample size. This could be mitigated by reporting the fraction of undetected attackers (i.e., the fraction of user-attacker pairs with an FAR of 1, given as "1's" in Table 4.3). However, given the relatively small number of samples per user for each dataset, there might not be a statistical difference between an FAR of 1, and one very close to 1, suggesting that this feature would also be overly sensitive. Another candidate metric is the standard deviation of the sample. Table 4.3 shows that the standard deviation varies between 0.05 and 0.37. However, the standard deviation quantifies the variation in a dataset, but does not reveal whether this variation is due to a few extreme outliers (which would be problematic) or a high number of moderate outliers (which would be a less severe problem). This limitation can be mitigated by also taking into account the kurtosis of the sample. Kurtosis is the fourth standardized moment and is a measure of the tailedness of a distribution. As such, a high kurtosis indicates that the distribution tends to produce more extreme outliers. Combining standard deviation and kurtosis (i.e., an ideal distribution being low standard deviation and low kurtosis) seems to fit our required profile. Figure 4.7 shows datasets with similar standard deviation but different kurtosis. The first gait dataset shows systematic errors, indicated by a high kurtosis of 11.53 while the second one exhibits more random errors, leading to a lower value of



(a)



(b)

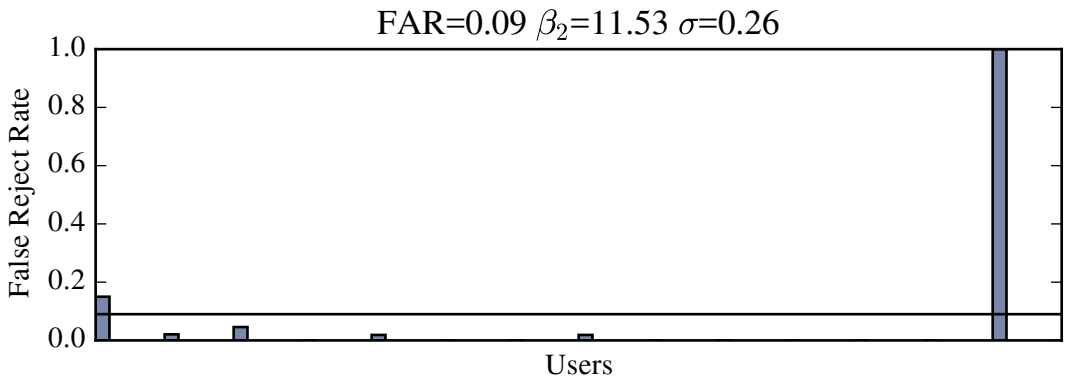
Figure 4.6: False rejects are spread evenly for the touch input biometric and are focused on very few users for the eye movement biometric. This is reflected in the difference in Gini coefficients (0.55 vs 0.93).

2.16. Despite this combination seeming fit for purpose, it would be difficult to use to accurately rank biometrics as any total ordering (i.e., preferring kurtosis over standard deviation or vice-versa) would be somewhat arbitrary.

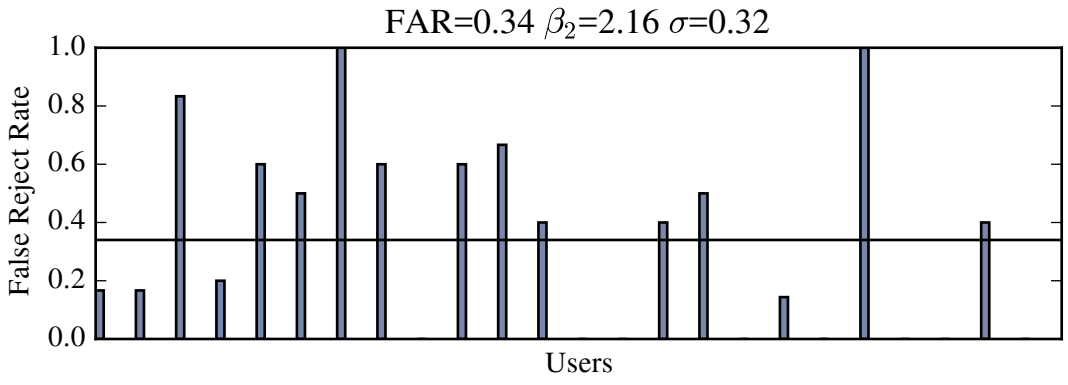
The Gini Coefficient (GC) has been proposed in 1912 as a measure of statistical dispersion to reflect the income distribution of a nation's residents [107]. A GC of 0 indicates a maximal equality of values (i.e., every resident having the same income), while a value close to 1 represents maximal inequality (i.e., one resident earning all the income). As a measure of inequality, the GC is also intuitively applicable to capture types of error distributions, with a high GC reflecting more systematic errors. An intuitive geometric representation of the Gini Coefficient is the area between the Lorenz Curve (which, in our scenario, measures the total error contributed by the bottom x % of users) and the Line of Equality (which is the Lorenz curve of a system where all users contribute identical error rates). The GC is shown as the shaded area in Figure 4.6. The GC has two important properties that makes it a suitable metric: Its scale independence means that it does not depend on the total or average error of a system, only the distribution of values. As such, it can be used to compare systems with different error rates. Conveniently, the GC always lies between 0 and 1, unlike standard deviation and kurtosis, which can take arbitrarily high values. In addition, it is population independent and does not depend on the number of samples in the dataset. This is of crucial importance, as the number of subjects in biometric datasets varies greatly and using only subsets of equal size seems infeasible due to authors rarely publishing their raw data.

Figure 4.8 shows the Gini Coefficient for the two most extreme cases we observe in our datasets. For the touch input biometric many attackers contribute to the overall FAR, while the eye movement biometric's intra-session dataset FAR is caused by very few extremely successful attackers.

Reducing security through strong features: It is interesting to note that the distribution of errors, and thereby the GC, does not simply depend on the biometric modality, but also the type of features used. When removing the pupil diameter, one of the most distinctive features of the eye movement biometric, the average error rates rise, but at the same time the GC decreases. This suggests that the pupil diameter is actually one of the

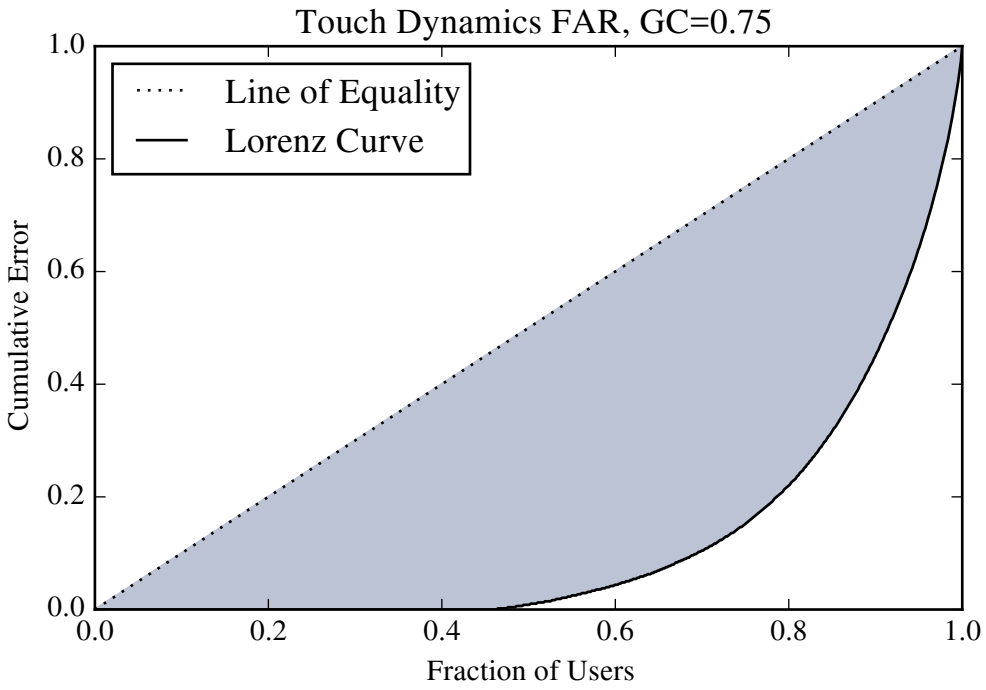


(a) First Gait Dataset

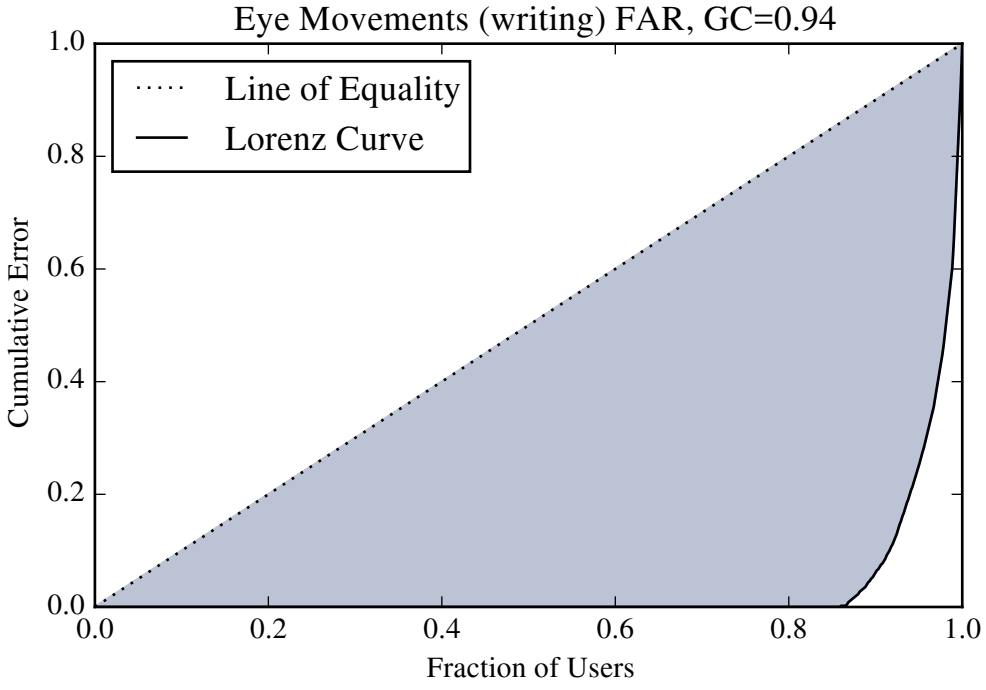


(b) Second Gait Dataset

Figure 4.7: Distribution of the FRR between users for the two gait datasets.



(a)



(b)

Figure 4.8: The different Gini coefficients draw attention to different error distributions. The touch biometric has a comparatively low GC of 0.75, which indicates largely random errors, while the eye movement biometric's higher GC of 0.94 suggests systematic errors which will lead to attackers consistently fooling detection.

			FAR					FRR			
Biometric	**Dataset**	**EER**	σ	β_2	GC	**max**	**1's**	σ	β_2	GC	**0's**
Eye Movements all features	Intra-Session	6.90%	0.22	13.05	0.92	1.00	0.02	0.25	12.59	0.93	0.93
	Inter-Session	7.99%	0.21	11.50	0.90	1.00	0.02	0.25	8.48	0.90	0.89
	2-weeks	8.43%	0.20	9.39	0.87	1.00	0.01	0.15	2.81	0.77	0.74
Eye Movements without pupil diameter	Intra-Session	19.83%	0.34	3.58	0.77	1.00	0.09	0.39	3.41	0.80	0.74
	Inter-Session	17.11%	0.30	4.10	0.74	1.00	0.03	0.27	6.21	0.77	0.50
	2-weeks	17.52%	0.29	4.45	0.74	1.00	0.05	0.27	4.78	0.74	0.58
Eye Movements II	Reading	1.17%	0.03	23.57	0.95	0.21	0.00	0.03	4.26	0.79	0.70
	Writing	4.80%	0.11	51.07	0.94	0.93	0.00	0.11	2.96	0.74	0.40
	Browsing	0.89%	0.04	34.68	0.96	0.29	0.00	0.03	8.11	0.90	0.90
	Video I	3.93%	0.09	15.20	0.88	0.57	0.00	0.09	5.21	0.83	0.80
	Video II	1.86%	0.07	33.59	0.96	0.49	0.00	0.04	3.85	0.74	0.60
Gait	Dataset I	8.44%	0.22	9.57	0.87	0.96	0.00	0.26	11.53	0.87	0.57
	Dataset II	28.4%	0.37	1.94	0.59	1.00	0.12	0.32	2.16	0.87	0.33
Touchscreen Input	Inter-Session	2.99%	0.05	15.01	0.75	0.40	0.00	0.04	6.74	0.55	0.05
Mouse Movements	Own machine	9.22%	0.21	11.98	0.89	1.00	0.02	0.24	5.57	0.85	0.82
	Lab machine	9.98%	0.23	8.96	0.86	1.00	0.02	0.15	2.01	0.69	0.57

Table 4.3: Results of applying the new metrics to our datasets. As evidenced by the Gini coefficient, random errors are particularly prevalent for the touch input biometric, while eye movements are prone to systematic errors. We can also observe that not using the pupil diameter results in fewer systematic errors, as evidenced by a lower GC and lower kurtosis.

key features that contributes to systematic errors especially because it is, on average, a very distinctive one. Due to the pupil diameter's relative stability it is suitable to separate most users, but leads to the consistent confusion of users with a similar baseline pupil diameter. As such, using the feature helps to further distinguish users that were relatively well-separated before, but does little to reduce systematic errors or might even make them more significant. This data supports the idea that, in some scenarios, adding distinctive features could actually *reduce* the security of a system, despite the lower average error, by adding systematic false negatives. As a result, researchers should take great care to not blindly strive for the lowest average EER but to also take into account how changes to features or classifiers influence their system's error distributions.

False Reject Rate

For the FAR, it is easy to agree on the fact that systematic errors are more problematic, as it leads to some attackers perpetually escaping detection. Determining the most favourable error distribution is not quite as obvious for the FRR. If most of the FRR is due to extreme outliers it might suggest that this is due to erratic user behaviour, such as a bad calibration for eye tracking. In that sense, this scenario might be preferable, as this indicates a problem with a small number of users, rather than an overall problem of the system which manifests itself in all users. When the deployed system shows high error rates for some users, it might be possible to further explore the root cause of the errors (which could involve educating the user, but could also aid in improving the system itself). Reporting the fraction of users perfectly recognized by the system (given as "0's" in Table 4.3) would be an obvious approach to reflect this property, but Figure 4.7 shows why it would be quite noisy in practice. Using a combination of kurtosis and standard deviation would also suffer from the same problems as for the FAR, namely the difficulty of establishing a total order between systems.

Following the shortcomings of the other metrics, the Gini Coefficient can again be used to quantify where exactly a biometric recognition system lies between the extremes of purely systematic and purely random errors. Our data shows that the touch input biometric has the most even distribution of false rejects, exhibiting a GC of 0.55. The

eye movement biometric generally shows the highest GC, with little change due to feature sets, time distance or tasks used. This might be explained by the fact that the biometric strongly relies on controlled user behaviour, specifically requiring a good calibration and as few head movements as possible. If some users are better at achieving this optimal behaviour, it would explain this rather extreme concentration of errors. In addition, this type of behaviour would likely occur regardless of the feature set used or increased time distance between sessions.

4.4.3 Lessons Learned

The previous subsections have shown that error distributions vary wildly across different datasets. This observation is valid for both the FRR and the FAR, leading to different consequences. Out of the set of the metrics we analysed to augment the FAR/FRR, the Gini Coefficient is the most promising due to its compactness and ability to provide an absolute ordering of systems. For the FAR, systems with a lower GC are desirable as this indicates false accepts that are spread relatively evenly across attackers, rather than enabling few attackers to perpetually escape detection. Our data shows that adding distinctive features, such as the pupil diameter for eye movement biometric, decreases the EER, but at the same time increases the GC. This suggests that features that change little over the system's operation might be suitable to tell users apart in general, but confuses similar users more consistently, thereby leading to the aforementioned systematic errors. This insight is crucial during feature selection, at which point some distinctive features should even be dropped completely to avoid this scenario. As such, it is important to remember that not every change to a system that lowers the average error is actually beneficial to its security. For the FRR a high GC indicates erratic user behaviour for a small number of users, an insight that can help improve either the system design or aid in avoiding this behaviour during system operation. Overall, we recommend to closely monitor changes to the GC when experimenting with different feature sets to evaluate whether any of them consistently lead to systematic errors. When publishing results, the GC should always be reported together with the mean EER/FAR/FRR in order to allow readers to take error distributions into account during their evaluation.

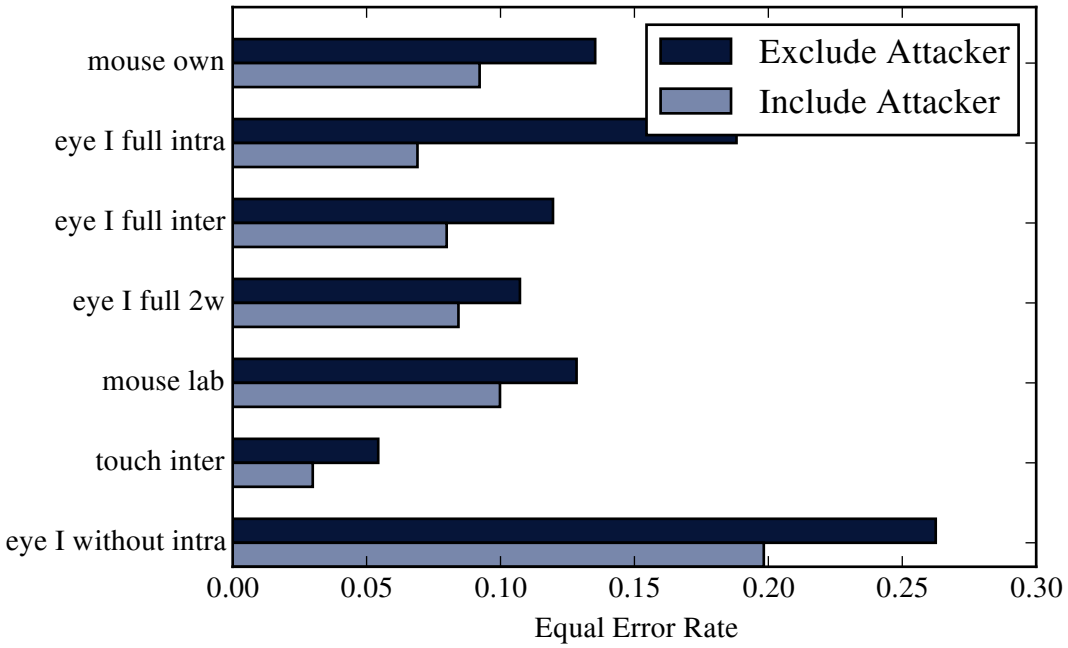
4.5 Influence of Methodology on Error Rates

Section 4.2 has shown that different methodologies are used across related work when building the training set. As evaluation is performed on a static dataset, the entire process is always a simulation of the real-world system. In this section, we evaluate how these different design decisions affect the reported EER. We will focus on the two types of decisions that are likely to have the biggest effect on error rates: the temporal makeup of the training data and the inclusion of attacker data in the training set.

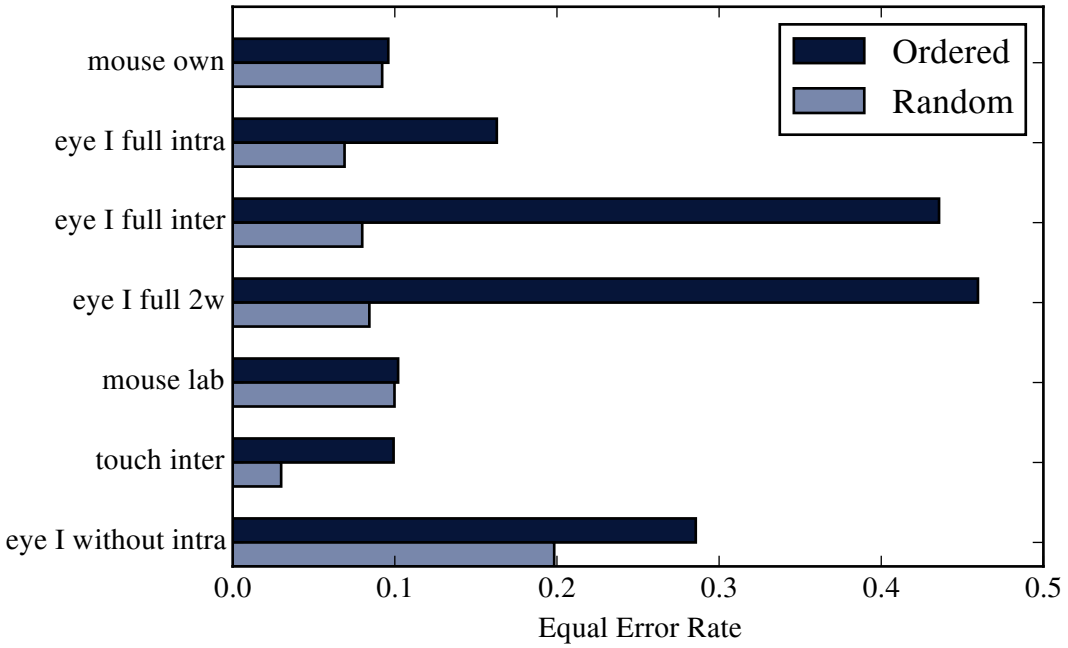
Temporal makeup. We consider two cases: ordered and random training data selection. These are the most common approaches used in related work (Section 4.2). For ordered training data selection, we use the first half of all available data for training and the remainder for testing. For random selection, we randomly select have the samples for training, without regard for their temporal distribution. We use a stratified approach to preserve the relative proportion of classes to avoid bias to any particular user. We repeat this sampling 20 times to increase confidence in the statistical robustness of the results. The reasoning behind not using stratified cross-validation (as in some papers in Section 4.2) is that the fraction of training samples can not be set independently from the number of "folds".

Attacker modelling. We consider two of the approaches outlined in 4.2, the all-users and exclude-attacker models. For the former, we build a single model per user. In this model's negative class, we combine the training data of each user under a common label. For the exclude-attacker approach, we build $a - 1$ models per user, where a is the number of attackers. Each of these models is then "attacked" by the attacker which was excluded from the training data. For both approaches, we balance the positive and negative class to avoid classifier bias (otherwise, the negative class would greatly outweigh the positive samples).

The results of our analysis are shown in Figure 4.9. Selecting training data randomly provides the biggest improvement, relative to the original EER. This effect is particularly pronounced for datasets that are collected over larger timespans (such as the inter-session



(a) Attacker Model



(b) Data Division

Figure 4.9: EERs decrease up to 80% when randomly selecting training data. Including the actual attacker in the negative class provides a reduction of up to 63%. The impact of random training data selection is particularly strong for datasets collected over longer time spans.

and 2-weeks eye movement datasets). This strong effect is most likely due to the classifier being unable to observe and account for any changes in user behaviour over time, leading to underfitting when considering the dataset over the entire time period. The mouse movement datasets, which are collected over a short period, are only marginally affected, which further supports this explanation. Another interesting insight is that the EER's variation is very high, depending on the training data selection. This suggests that the training and testing process has to be repeated a number of times to ensure statistical robustness of the result. The distribution of errors is virtually unaffected by the change, which suggests that it mainly leads to shifting the mean.

The effects of the two different attacker models significant, albeit less extreme than those of the training set selection. Across all datasets, including the attacker in the training data results in a relative improvement between 22% (mouse movements) and 63% (intra-session eye movements). It is somewhat counter-intuitive that the effect is bigger for the larger datasets, even though the attacker data only accounts for a smaller fraction of the overall negative class.

These results show that simply looking at the EER of a proposed system is insufficient, as it is skewed greatly by non-functional parameters that would not affect the performance of the system in a production environment. For example, if the exact same dataset (i.e., identical features and classifiers) were evaluated with random and ordered training data selection, one might favour one over the other (even though their practical performance would be identical). This is particularly alarming as our analysis (see Section 4.2) shows that out of 25 papers, 13 use at least one of the methodologies that we have shown to lead to systematic underestimation of error rates. In addition, a further 6 do not report how the error rates were obtained, which not only decreases confidence in the results but also impedes reproducing them and comparing them to related work. In order to inspire the highest confidence in their results researchers should exclude attackers from the negative class in their training data and choose the first part of their entire dataset for training, rather than sampling it randomly. In order to allow an easier comparison with some earlier work it would also be advisable to report error rates for different methodologies (such as random sampling) as well.

4.6 Lessons Learned

The main takeaway of this chapter is that based on current evaluation practices, different papers and systems are extremely hard to compare. While common metrics (such as the EER) are used in most papers, their comparison gives a rough indication of relative quality at best. The reason for this is two-fold: average error rates hide the precise distribution of errors and different simulation methodologies can greatly under-estimate error rates. We have shown that the prevalence of systematic errors varies between different datasets, which underlines that they can not be easily compared using the average EER alone. Different simulation methodologies can lead to underestimating the ”real“ EER by as much as 63%.

The remedy we propose for the issue of systematic errors centres around the Gini Coefficient. This metric, which we propose should be reported along with conventional error rates, gives a measure of how evenly errors are distributed between users and attackers. For the FAR, a high GC (i.e., uneven distribution) is problematic as it indicates that some attackers are consistently successful, which leads to systematic false negatives. Conversely, a high GC may be desirable for the FRR, as this indicates a system that works well for most users while the few ”problematic“ users can be authenticated through other means.

Against the average user, anything works; there's no need for complex security software. Against the skilled attacker, on the other hand, nothing works.

— Bruce Schneier

5

A Cross-Device Attack against ECG Biometrics

Contents

5.1 Introduction

The focus of the previous chapters has been the design and evaluation of continuous authentication systems. In Chapter 4, we have discussed some implications of attacker models, specifically whether including attacker samples in the training data can be judged as realistic. However, the previous chapters all assume a zero-effort threat model, i.e., an attacker that does not actively attempt to circumvent the system and rather relies on accidental biometric closeness to the victim.

In this chapter, we develop an attack against an existing authentication system. Aside from the security analysis of this particular system, this work also yields insights into how a variety of real-world systems can be attacked, and how the severity of these attacks can be mitigated in the future. The reasoning behind choosing a commercial system (rather than an academic paper) as the target is that this allows us to both explore the security of both the hardware components and underlying machine learning classifier. If we were to choose a paper as the target, we would likely have to make a number of assumptions about the system's implementation, thereby questioning the attack's validity. It is important to note that we design the attack presented in this chapter to be as general as possible in order to be able to draw conclusions about the security of other systems as well.

We provide a systematic attack against electrocardiography (ECG) biometrics and demonstrate its effectiveness by applying it to the Nymi Band. The Nymi Band[1] is a multifactor authenticator that authenticates its wearer through an ECG-based biometric template. We first demonstrate our capability of spoofing arbitrary ECG signals. To present the spoofed signals we use three different devices, two Arbitrary Waveform Generators (AWG), one software- and one hardware-based, as well as the audio playback of ECG signals encoded as .wav files using an off-the-shelf audio player. As such, the technological barriers for the attacker are extremely low. We collect ECG data from a total of 41 users and 5 devices. The data shows that the morphology of the ECG signal depends greatly on the device that was used, similar to the task-specific feature distributions evaluated in the previous chapter. This difference constitutes a major challenge, as a

[1] https://nymi.com

signal collected on one device can not easily be used to carry out an attack on another. We tackle this challenge by developing a novel cross-device mapping function derived from population statistics. The purpose of this function is to transform an ECG signal collected on one device such that its morphology matches that of another. After enrolling users in the Nymi Band we use the mapping function to impersonate the user by presenting the (transformed) ECG signals collected on different devices.

5.2 Background

In this section, we will give an overview of ECG in general and ECG-based biometric recognition in particular.

5.2.1 Electrocardiography

The electrocardiogram (ECG) is a measurement of the electrical activity of the heart. It is acquired through electrodes placed on the patient's skin, which are used to capture voltage changes due to depolarization and repolarization of cardiac cells, respectively provoking contraction and relaxation of the cardiac muscle. The ECG is commonly used in clinical practice for its crucial diagnostic capabilities [108]. In addition, the present availability of low-cost ECG sensors has enabled numerous applications in the area of wearable devices and fitness monitoring [109], leading to pervasive acquisition of ECG data.

Figure 5.1 shows an example ECG for one cardiac cycle, together with the duration and amplitude features typically extracted for authentication purposes (detailed in Section 5.2.2). It comprises five main waves, P, Q, R, S and T, which map to specific heart events: the *P wave* indicates activation of the atria (the upper heart chambers); the *QRS complex* corresponds to the activation of the ventricles (the lower chambers); and the T wave indicates ventricular repolarization.

Most ECG recording systems are based on the so-called Einthoven's lead system, where each lead records the difference of potential between two electrodes. Einthoven's leads consist of:

$$\begin{array}{ccc} \text{Lead I:} & \text{Lead II:} & \text{Lead III:} \\ V_I = \Phi_{LA} - \Phi_{RA} & V_{II} = \Phi_{LF} - \Phi_{RA} & V_{III} = \Phi_{LF} - \Phi_{LA} \end{array}$$

where V_i is the voltage of lead i and Φ_j, with $j \in \{LA, RA, LF\}$, is the potential at the left arm (LA), right arm (RA) and left foot (LF), respectively. In particular, the standard 12-lead ECG used in clinical settings is an extension of the Einthoven's 3-lead system based on using seven additional electrodes placed on the chest. Nevertheless, simpler 1-lead ECG recording systems are increasingly being used in the context of personal ECG monitoring and wearables. One example of this is the Apple Watch 4[2], which allows users to perform an ECG measurement by touching the Digital Crown.

5.2.2 ECG Biometrics

Driven by the distinctiveness and universality of ECG, the body of work in this field has been steadily growing over the past few years. Recent surveys of systems based on ECG-based biometrics can be found in [43, 110, 111]. The most striking difference between approaches lies in the biometric features used. The first class of methods is based on *time domain feature extraction*, and work by detecting the so-called fiducial points, i.e., location, amplitude and width of the main ECG waves, as shown in Figure 5.1. Some biometric systems also consider the ST segment, that is, the length of the isoeletric segment between the S and T waves, as well as the slope of waves. In Figure 5.1, the wave slope is accounted for through the extraction of left and right components of its width. In addition to the above intra-beat morphological features, inter-beat features such as Heart Rate Variability and beat patterns (represented by the RR intervals) can well reflect the specific characteristics of the subject. However, intra-beat features are also easily influenced through stress or physical activity (both of which increase the heart rate and shorten the RR interval).

The second class of methods use *frequency domain feature extraction*, meaning that features are obtained after converting the ECG signal in the frequency domain. Examples include application of wavelet decomposition and Fourier Transform.

[2]https://support.apple.com/en-gb/HT208955

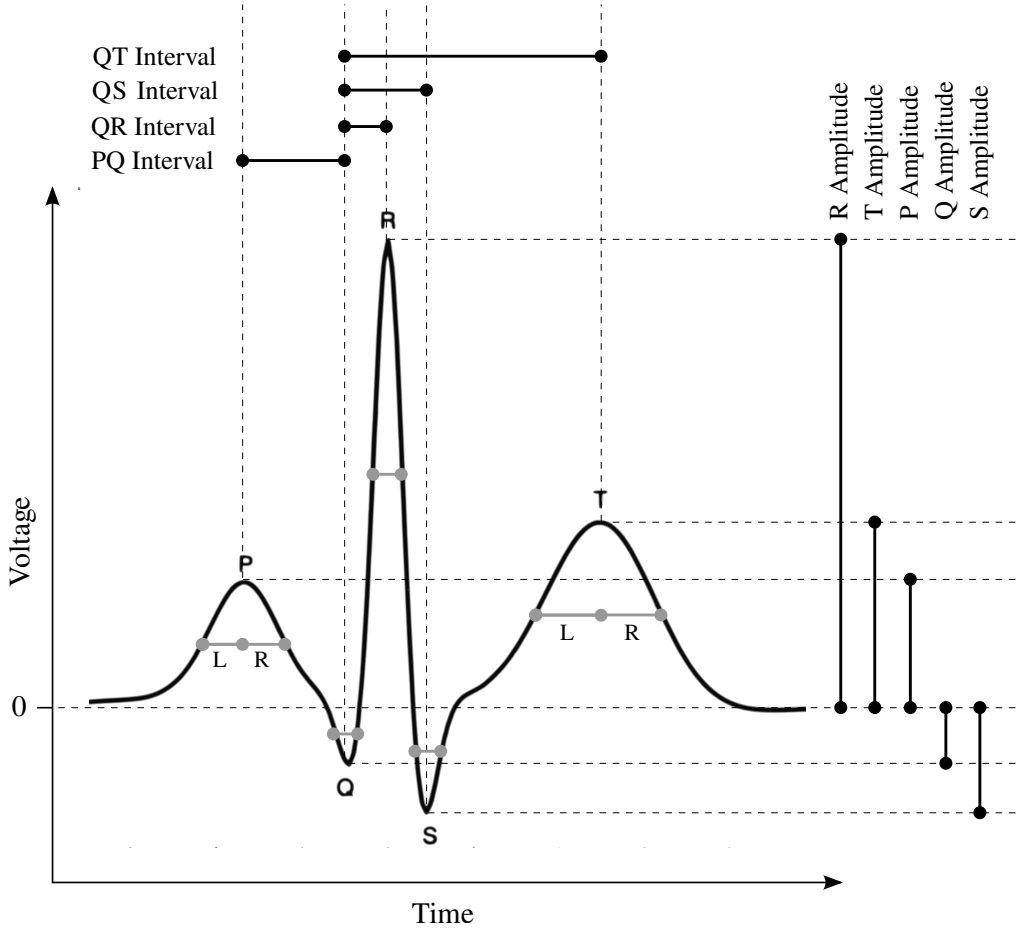


Figure 5.1: Example electrocardiogram and corresponding time-domain features for ECG-based biometrics. Top: duration features given by inter-peak distances. Right: amplitude features. For each wave, we also consider its width at half amplitude (grey solid lines). To account for asymmetric curves, the width of each wave is split into left and right components (L and R segments shown in P and T waves), i.e., before and after the wave peak.

In alternative to time and frequency domain methods, some biometric systems employ statistical approaches for computing the distance between enrolment ECG and recognition ECG directly at the signal level, or analysis of the ECG's trajectory in the phase space.

5.2.3 The Nymi Band

The Nymi Band (see Figure 5.2) is a wristband that incorporates an ECG sensor with two electrodes. The bottom electrode constantly touches the user's wrist while the band is worn. In order to allow ECG measurements (most commonly for enrolment and authentication), the user touches the top electrode with the index finger of their other hand. As such, the signal morphology can be expected to be similar to Lead I of a medical ECG

Figure 5.2: The Nymi Band

(which measures the potential difference between the left and right arm). Besides the actual band, the Nymi ecosystem consists of the Nymi Companion App (NCA) and Nymi Enabled Applications (NEAs). The NCA is provided as an app that runs on the user's smartphone or tablet. The NCA performs two main functions, enrolment and activation.

During *enrolment*, the Nymi Band is paired to an NCA. The correct pairing is confirmed by displaying a pattern on the Nymi Band which the user has to verify against a pattern shown by the NCA (similar to the numerical codes used in Bluetooth device pairing). The Nymi Band and NCA then agree on a shared key that binds the Nymi Band to this NCA. Following pairing, the user is prompted to touch the band's top electrode with her index finger, after which their ECG is measured until a specific amount of ECG data of sufficient quality is captured. The resulting biometric template is then encrypted and stored by the NCA on the phone or tablet. Besides the shared secret, no information is stored on the band at this time.

Activation is performed when the Nymi Band is taken off and put back on again. Specifically, the de-authentication event is detected by the contact between two pins on the inside of the buckle being interrupted (see Figure 5.2). As such, the Nymi Band does not truly perform continuous authentication in the biometric sense, but authenticates the

user once and then detects a possible change in user identity through the band being taken off. However, assuming the band is used as intended, these two approaches lead to the same outcome as it is virtually impossible to take off the band without triggering de-authentication. The activation process is started by the user selecting the appropriate action in the NCA, after which they can choose to either perform ECG authentication or use their backup password. If they choose ECG, they are again prompted to touch the top electrode to begin ECG measurement. Unlike enrolment, which runs until a certain number of seconds of valid ECG data is collected, activation runs until the NCA is sufficiently convinced of the wearer's identity. Once one or several heartbeats are observed that match the owner's template, the Nymi Band is put into the activated state by the NCA. If no matching heartbeats are observed after 60 seconds, the user is automatically rejected by default.

Once the Nymi Band is activated, it can be paired with NEAs. Examples of NEAs include desktop computers (that can then be unlocked without using a password), wearable devices like smart watches and even more complex systems like cars. Ultimately, any device that supports Bluetooth Low Energy (BLE) communication can be setup as an NEA. At the time of writing, the Nymi Band is being trialled for contactless payments. Initially, the band is paired with the NEA through a process similar to regular Bluetooth pairing. The Nymi Band displays a pattern using the five LEDs (leading to only 32 possible combinations), which the user is meant to confirm before proceeding. The Nymi Band and the NEA then use a Diffie-Hellman key exchange to negotiate a shared key, which is stored directly on the band. After pairing, the possession of the shared key (i.e., the presence of the unlocked band) can then be confirmed using a standard challenge-response protocol.

There is one additional capability of NEAs that is relevant to the remainder of the paper: The Nymi SDK grants NEAs direct access to the band's ECG sensor. Once the band and an NEA are paired, the NEA can request the collection of an arbitrary amount of raw ECG data. While this data collection does not have to be explicitly approved by the user, the sensor design requires the user to touch the top electrode with their finger, thus making covert data collection virtually impossible. It is noteworthy that this

functionality has been removed from the official SDK from version 2.0 onwards. The consumer version of the band also lacks this functionality.

The Nymi Band's threat model is described in the Nymi Whitepaper. The band is designed as a three-factor authentication system. In order to communicate with NEAs, an attacker has to be in possession of the Nymi Band and the NCA (typically the user's phone) and be able to bypass the biometric authentication. It is noteworthy that the latter, while not explicitly stated in the Whitepaper, can also be achieved by using the user's backup password (e.g., through guessing a weak password or social engineering). This is particularly dangerous, as the presence of a second authentication factor often leads to users choosing weaker passwords [112]. In terms of bypassing ECG authentication (rather than using a password), the Nymi Whitepaper claims that

> *"There is currently no known means of falsifying an ECG waveform and presenting it to a biometric recognition system."*

In the following sections we will investigate the validity of this claim.

5.3 Spoofing ECG Signals

In this section, we show that fake ECG signals can be injected into ECG enabled recognition systems. We start out with the hypothesis that captured ECG measurements can be reproduced at the biometric sensors without the benign user having to be present.

5.3.1 Motivation

Like any other physiological trait, ECG signals can be captured and (digitally) stored for an indefinite amount of time. Biometric samples from physiological traits do not lose validity and, if the fidelity of the stored signal is sufficiently high, it is possible to physically reproduce the actual biometric signal at a later time. This process does not require the individual from whom the biometric measurements originate to be present.

In ECG recognition, biometric readings are usually acquired with the help of an electrocardiograph, which works by measuring the minute voltage differences of the human heart over time. With today's technologies in signal synthesis and digital to

analog conversion, artificially creating electrical signals that exactly represent stored ECG signals is feasible.

While forging ECG signals is not a concern in the medical domain, it is potentially problematic for ECG-based authentication systems. If a biometric system does not feature an agent or overseer — or other provisions against someone not using the biometric sensors as intended — it is susceptible to so-called presentation attacks. In a presentation attack, the attacker tries to spoof the biometric sensors with an artefact or contrivance. In case of ECG based recognition, the attacker would have to fake the (time-dependent) voltage levels at the electrodes interfacing the user with the help of an electrical device that outputs an ECG signal.

In the remainder of this section, we show that it is indeed possible to replay previously captured ECG signals. To that end, we built three hardware contrivances of varying degrees of sophistication that successfully create and inject ECG signals into the sensing electrodes of a biometric system based on ECG. In order to estimate difficulty and likelihood of a presentation attack on ECG recognition, we additionally evaluate our injection methods along the following non-technical dimensions:

- **Cost**: What is the overall cost for building the contrivance and executing the injection? Although high cost does not deter every attacker, it can discourage less determined ones.
- **Knowledge**: Is expert knowledge required to build and use the contrivance or can it be put together following simple instructions?
- **Size**: Physical size is a very important factor. If the contrivance used for signal injection is sufficiently small, an attacker can covertly spoof the biometric sensors and might even circumvent a guarded biometric system.
- **Signal quality**: We quantize and compare the resulting signal quality of each approach when applied to injecting ECG signals into the Nymi Band. Obviously, signal quality directly correlates with the probability of success for a presentation attack. The conversion from the (stored) biometric data to the physical biometric signal should introduce as little noise as possible.

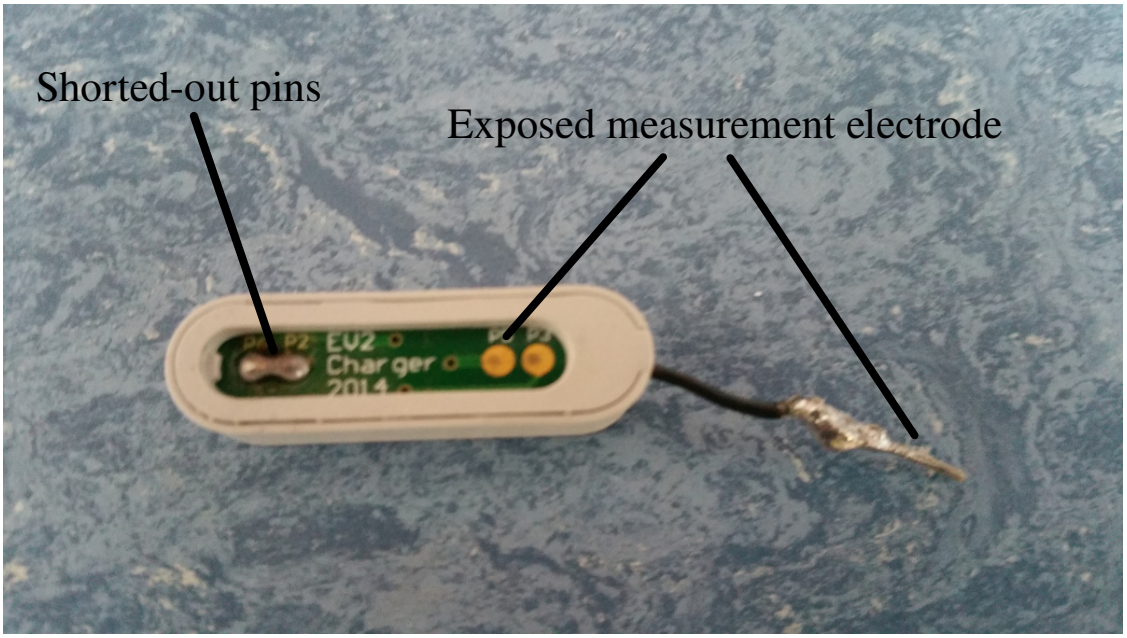


Figure 5.3: Modified charger lead. The two pins on the left are shorted out to put the Nymi Band into "closed" state as soon as the modified charging lead is attached to the band.

5.3.2 Hardware Considerations

We build three contrivances that allow us to forge fake ECG signals. Since all approaches inject the resulting signal at the sensing electrodes of the Nymi Band, we design a contrivance that allows us to interface those electrodes in a convenient way. As described in the previous section, the Nymi Band has a locking mechanism, which deauthenticates the wearer of the band immediately if the band is taken off. When the band is closed, one of the two sensing electrodes is located on the inside of the band and faces the wearer's wrist whereas the other electrode is accommodated at the outside of the band, available to be touched with a finger of the other hand. This way, the electrical circuit is closed and the ECG measurement can start. If the wearer of the band is recognized, the band goes to and remains in "authenticated" state as long as the band is worn.

For ease of use and to allow many successive injection attempts without opening and closing the Nymi Band, we modified a genuine charging cable that is included in the delivery of the band. As can be seen in Figure 5.3, the part of the charging lead that interfaces with the band has two shorted-out pins to let the band think it is closed and conveniently exposes one of the sensing electrodes in a separate wire.

This approach does not necessitate any modifications of the Nymi Band itself. This is desirable, as any permanent modifications of the band would be easily detected by its owner following the attack.

Hardware Waveform Generator

Our first and the most obvious approach to artificially create an ECG trace is to use an arbitrary waveform generator. The purpose of waveform generators is to generate electrical waveforms for testing and analysing electronic devices. The generated signal is injected into the device under test while the device's (electrical) behaviour is observed and analysed. The waveform itself is defined as a time series of voltage levels, based on which the waveform generator produces the corresponding signal. The resulting signal exactly matches the predefined voltage levels at the given points on the time axis and interpolates values in between.

We use a Rigol DG4062 Arbitrary Waveform Generator that is capable of generating signals of up to 60MHz at a sampling rate of 500 Mega-Samples per second. These specifications enable us to transform stored ECG signals to their physical counterpart with high accuracy. Electrocardiographs commonly operate at a sampling rate of less than 500 samples per second when acquiring an ECG trace. This means that higher frequency components, i.e., more than 250Hz, can not be registered and hence are not part of the measured signal. Such a frequency limited signal can easily be synthesized by most of today's off-the-shelf hardware waveform generators.

Our signal generator's two output leads are directly connected to the electrodes of the Nymi Band. In order to optimally match electrical impedances between signal generator and the electrodes of the band, we inject the signal through a 75Ω coaxial cable (see Figure 5.4). Electrocardiographs most often feature an instrumentation amplifier with high input impedance as the first step in the signal acquisition pipeline. The Nymi Band does not differ in that regard and requires a relatively low impedance input.

We wrote a software library that loads stored ECG signals directly into the memory of the Rigol DG4062 Arbitrary Waveform Generator, sets the necessary parameters and starts/stops the signal generation. The program code is available upon request.

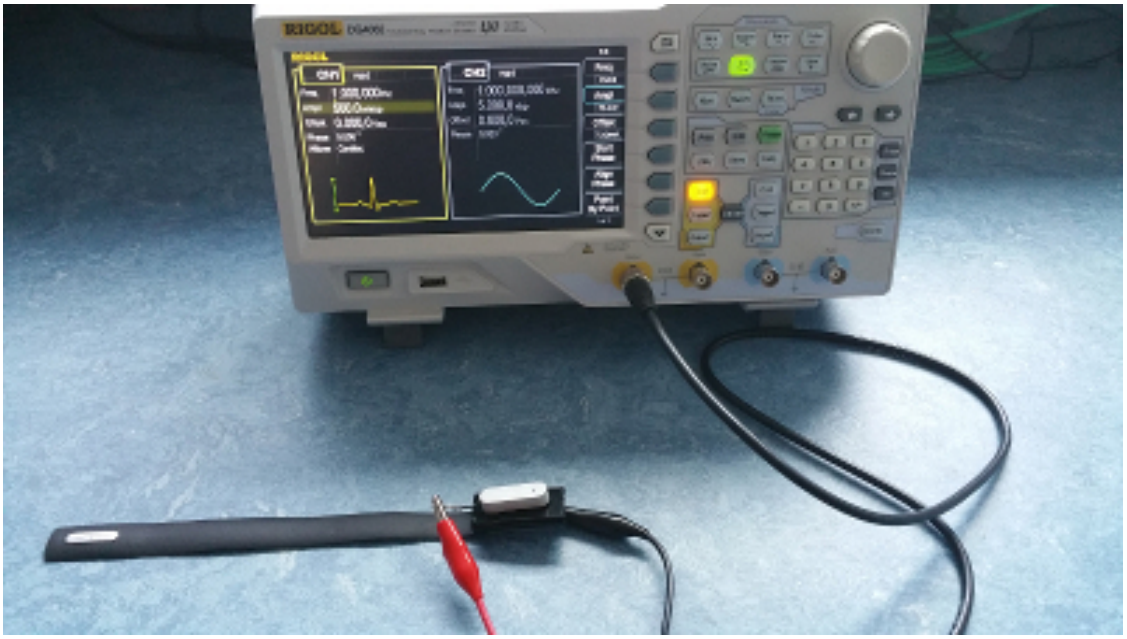


Figure 5.4: Arbitrary waveform generator connected to the Nymi Band via the modified charging lead. The negative output of the waveform generator is clamped to the electrode facing the wrist and the positive output is attached to the second electrode of the band using the modified charging lead.

Software Waveform Generator

Nowadays, almost every personal electronic device, be it mobile or stationary, possesses a dedicated sound card or integrated sound functionality to facilitate analog audio output. Audio signals are an electrical representation of sound, i.e., a mechanical wave that propagates through a medium. Thus, sound cards need to be able to output relatively high-frequency signals. This capability can be harnessed and lets a sound card be utilized as a low-frequency waveform generator. In most cases, no hardware modifications are needed and arbitrary electrical signals can be readily generated, provided that the sound card is driven with the right software components. Naturally, a sound card based waveform generator is not as capable as a dedicated hardware solution and has many limitations such as a narrow range for the generated voltage. However, the nature of ECG signals, which are inherently low-frequency and on the order of a few hundred microvolts, can be generated by a sound card without any problems The majority of dedicated sound cards as well as devices with integrated sound support have output frequencies of up to 20kHz, which is sufficiently high to generate the relatively low-frequency components

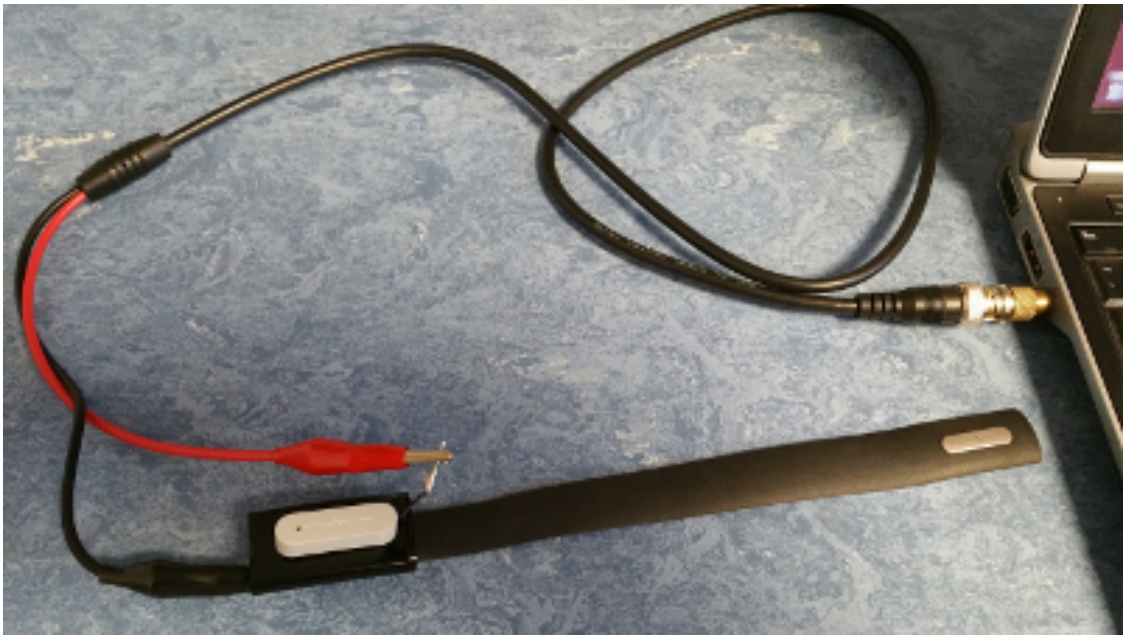


Figure 5.5: A laptop is connected to the Nymi Band via the modified charging lead. Setup is analogous to the configuration involving the hardware waveform generator, apart from the coaxial cable being plugged into the audio output port of a laptop. The laptop either runs a software waveform generator or is used to play back an ECG signal that is encoded in an audio file. The laptop may be replaced by any electronic device with audio playback capability.

of ECG signals. A software waveform generator based on a sound card is therefore a viable option for signal injection. It not only drastically reduces cost, but also simplifies the injection method. Figure 5.5 depicts a possible setup where a software waveform generator is run on a laptop that injects the generated signal though its audio output port.

Audio Playback

Instead of using a software waveform generator and changing the function of the sound card, we explore the possibility of playing back stored ECG signals on the sound card as actual sound. Such an approach does not require specialized software, i.e., a software waveform generator, and might be executed on any device capable of outputting analog audio signals. This could reduce effort and complexity of a presentation attack to a great extent.

The challenge of replicating an ECG signal directly as audio output consists of transforming the digital representation of an ECG signal into an audio file that can be played back on the sound card. We wrote software that filters the ECG signal, applies the

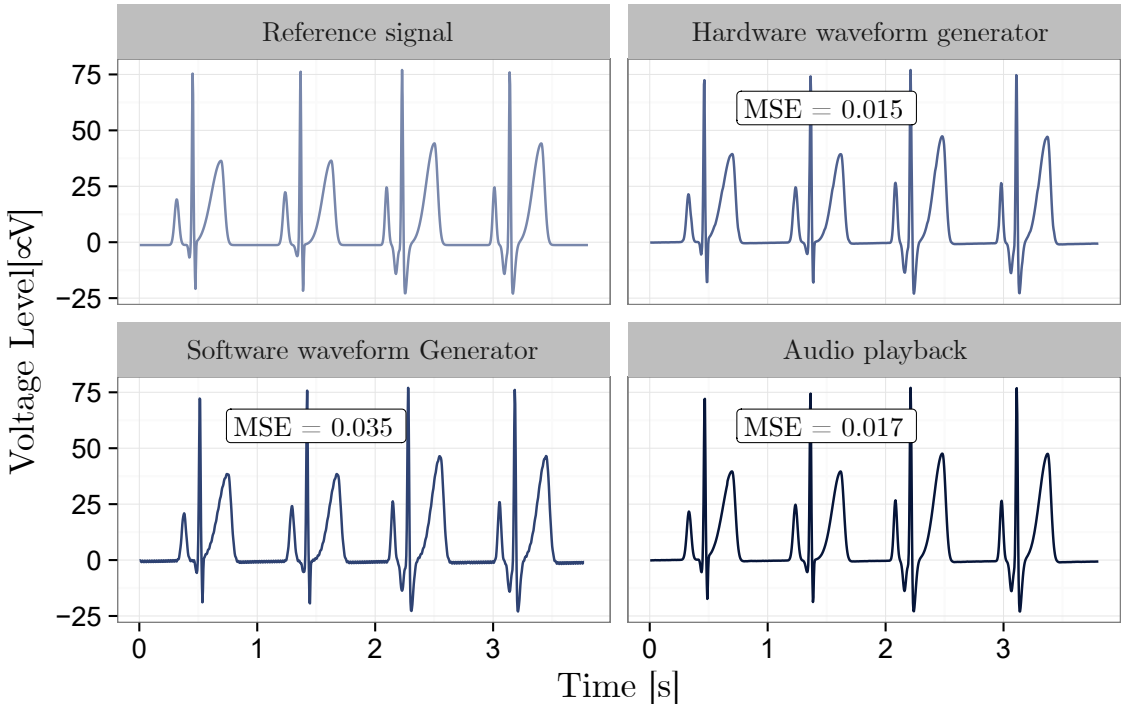


Figure 5.6: Reference ECG signal compared to the ECG traces measured when the reference signal is injected using three different injection methods. The traces are captured by the Nymi Band and read out with the Nymi software development kit.

correct scaling of voltage levels, sets the sampling rate and finally stores the signal as an audio file (WAV format). The resulting file can then be played back on almost any device and potentially injected into the sensors of a biometric system based on ECG recognition. The contrivance we used for the evaluation of the audio playback as injection method is identical to the hardware setup in Figure 5.5. Nevertheless, the attack can be carried out with any device capable of analog audio output.

5.3.3 Injection Quality

In order to validate the presented signal injection methods and assess their quality, we select a stored reference signal, reproduce and inject it using each of the three approaches. We then compare the reference signal to the traces the electrocardiograph measures while injection takes place. In case of the Nymi Band, the captured traces can be accessed and read out with the Nymi software development kit (SDK).

In addition to a visual comparison between the stored signal and the extracted ECG traces, we verify their similarity numerically. The distance metric for comparison is the

(per sample) mean squared error (MSE). We make sure the reference signal's sampling rate matches that of the Nymi Band. Also, the measured traces might be linearly translated and require alignment before the calculation of the distance metric. We determine the constant shift between the stored reference signal and captured traces by aligning the peaks of the R waves.

The results of this comparison are shown in Figure 5.6. All three proposed injection methods manage to reproduce the reference signal remarkably well and we conclude that the contrivances are effective. It is evident that signal quality is proportional to sophistication and cost of the contrivance used for signal injection: The hardware waveform generator and audio playback achieve the smallest error (mean squared error of 0.015 and 0.017, respectively), outperforming the software signal generator with an MSE of 0.035.

5.3.4 Comparison of Injection Methods

In Table 5.1, we present a comparison between our three injection methods along the criteria outlined above (see Section 5.3.1). Not surprisingly, the hardware waveform generator achieves the highest signal quality, but at the same time entails the highest cost. Entry-level arbitrary waveform generators retail at around £250. They are, however, fairly bulky and only designed for stationary use.

Software waveform generators are not only available for personal computers, but even smart phones. A low-end smart phone equipped with an analog audio output costs around £50 as of spring 2016. An ECG signal encoded as an audio file can even be played back on a cheap portable audio player which can cost less than £10 and has a tiny form factor.

5.4 Experimental Design

In this Section we will outline a number of approaches that can be used by an attacker to obtain data for a presentation attack. Based on these attack vectors, we will then discuss our data collection methodology.

	Approximate cost	Required knowledge	Physical size	Signal quality
– Hardware waveform generator				
	£240	high	large	very high
– Software waveform generator				
	£50	moderate	small	high
– Audio playback				
	£10	moderate	very small	very high

Table 5.1: Comparison of injection methods

5.4.1 Obtaining Data for a Presentation Attack

In Section 5.3 we have demonstrated our capability to inject arbitrary signals into the Nymi Band. However, the attacker still requires an input ECG signal that is sufficiently close to that of the victim. There are multiple conceivable approaches to obtain this data:

Medical records often contain printouts of a patient's ECG. A conventional hospital ECG uses 10 adhesive electrodes to simultaneously record 12 leads (see Section 5.2 for details). An attacker could obtain these records either electronically (e.g., through social engineering or a compromised medical database) or on paper (e.g., by taking a photo of the plots). In addition, mobile devices that allow patients to monitor their health at home are becoming more widespread. Besides ECG monitors for medical use, a number of devices are marketed for fitness, for example in the form of heart rate monitors used during cardiovascular exercise. In these cases, the data could be intercepted during transmission (e.g., to the victim's smartphone) or leaked through an insecure or malicious mobile app.

The Nymi SDK allows NEAs to collect arbitrary amounts of raw ECG data (see Section 5.2). There are two conceivable ways in which this might pose a security risk: NEAs have to be actively paired with the NCA by the owner of the band. However, this only means they are trusted by the user, not that they are inherently trustworthy. A rogue NEA could trick the user into providing (a sample of) her ECG, which would then allow the owner or developer of the NEA to later use this data to carry out a presentation attack. A second, probably more severe, way of abusing the SDK data collection is through a social engineering attack. Given the novelty of the Nymi Band an attacker might ask the victim to wear the attacker's Nymi Band to test whether it

is possible for the victim to unlock it. Instead of performing regular activation, the attacker could instead activate the band through the backup password and then collect the victim's ECG data through a previously paired NEA (e.g., on a laptop or the attacker's smartphone). This attack is particularly dangerous, as the data is collected directly on a Nymi Band, rather than a different device that requires the application of the mapping function (see Section 5.5). As an authentication attempt only fails after a fixed number of non-matching heartbeats have been recorded, this provides the attacker with an amount of data similar to a complete enrolment.

5.4.2 Data Collection

The previous subsection has outlined several scenarios that might allow an attacker to obtain a victim's ECG data. We collect data using a variety of devices to reflect them:

The first device we use is a lightweight medical ECG monitor, the Heal Force Prince 180B (see Figure 5.7). The device has two main measurement modes which use either the built-in electrodes on the sides of the device or an external 3-lead ECG cable with disposable electrodes. The first measurement (which we will refer to as the Palm measurement throughout the paper) uses the built-in electrodes. Participants were asked to hold the device as pictured in Figure 5.7 and to remain still during the measurement as the ECG recording is highly sensitive to device movements. When using the built-in electrodes, the device always records data for a fixed duration of 30 seconds. Following the palm measurement, we use the 3-lead cable to record Lead I and Lead II (see Section 5.2 for details), which involves attaching the disposable electrodes to both arms and the left leg. Unlike hospital ECGs, which capture all 12 leads simultaneously, this monitor is limited to recording a single lead. However, by switching the position of the electrodes, all standard leads can be recorded in sequence. We chose to record Lead I, which has measurement points similar to the Nymi Band, and Lead II, which measures the potential difference between the right arm and left leg. Due to practical reasons we choose not to collect a full 12-lead ECG. Both Lead I and Lead II are recorded as part of a medical ECG. As such, an attacker could obtain them by taking a photo of the patient's medical

Figure 5.7: ECG monitor in palm measurement mode

files. In order to reflect this, we don't extract raw data from the device, but instead obtain it by performing image analysis on the plots displayed by the software (see Figure 5.8).

The second device, pictured in Figure 5.9, is a mobile ECG monitor that can be attached to a smartphone and transmits the recorded data via Bluetooth. Following the measurement, the device provides an instant assessment with regard to a number of heart disorders. Participants were again asked to remain still during the 30-second measurement period to avoid the introduction of additional noise. Following the measurement, the device creates a pdf report which can be automatically sent to an arbitrary e-mail address (such as the patient's physician). This report contains, aside from the patient's personal information, a plot of the recorded ECG data. The report is sent via e-mail without any encryption or other security features. Similar to the ECG monitor data, we use image analysis to extract the raw data from the pdf file.

Lastly, we collect data using the Nymi Band itself. As outlined in Section 5.2, the Nymi SDK allows Nymi Enabled Applications (NEAs) to collect raw ECG data once the Nymi Band is unlocked and paired with the NEA. As a result of the band's hardware design, this requires the cooperation of the user as they have to touch the top electrode of the band to enable ECG recording. Following the user enrolment and activation

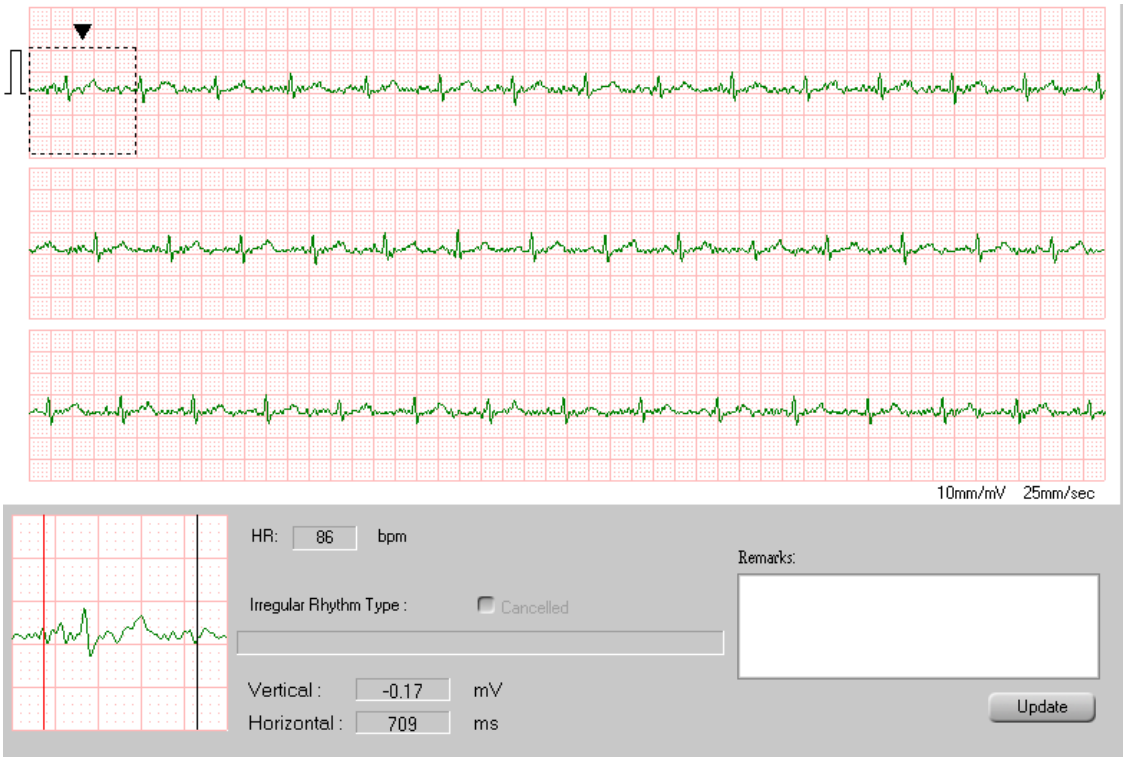


Figure 5.8: The ECG monitor data is extracted through image analysis of this plot shown by the software. This models the threat of an attacker obtaining a photo of the victim's ECG.

of the Nymi Band, we collect 60 seconds of raw ECG data using the SDK. We used three (identical) developer kits of the Nymi Band running the SDK version 1.03. It is important to note that the capability of the band to report raw ECG data to NEAs has been discontinued from SDK version 2.0 onwards. However, it is still possible to collect data by using the legacy SDK on the developer bands. The final consumer version of the band has been released in September 2016 and also lacks the capability to record raw ECG data. We can't make any claims on the success of our attack on this new version, although, based on the published changes in the consumer version, we can not see any structural obstacles to the attack still being successful.

5.4.3 Participant Recruitment and Ethical Considerations

This project has been reviewed by and received clearance from the Central University Research Ethics Committee of the University of Oxford, reference number R42894. The main ethical concern when gathering ECG data is the sensitive nature of the data itself. This sensitivity stems from the fact that a variety of heart disorders can be diagnosed using

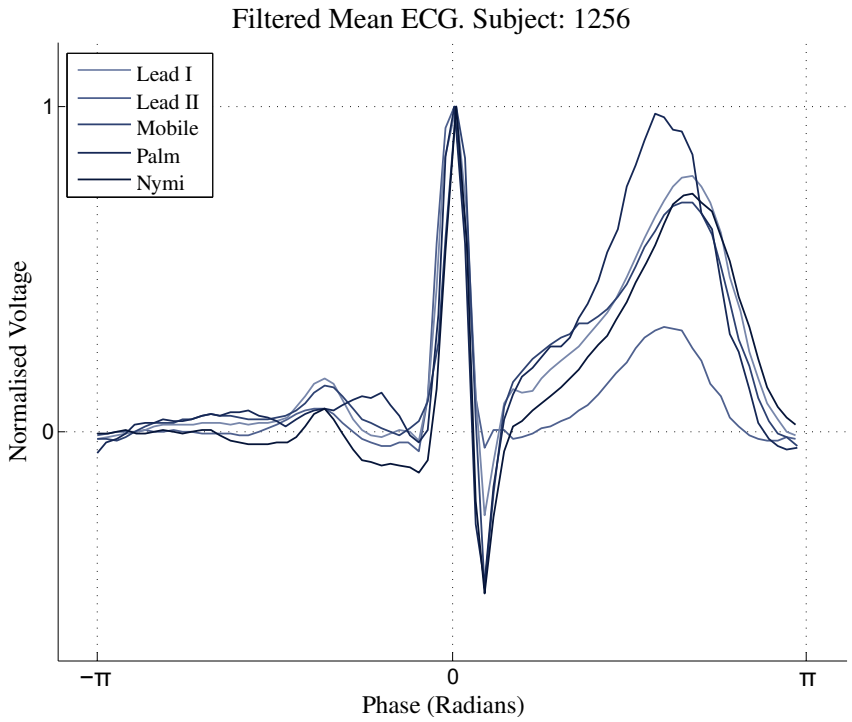
Figure 5.9: Mobile ECG monitor

ECG, including disorders the participant may not have previously been aware of. The possibility of false positives (i.e., the incorrect diagnosis of a condition) in conjunction with the fact that none of the researchers are trained medical professionals led us to the decision of disabling all diagnostic capabilities of the devices used and inform all participants accordingly. Since a future diagnosis based purely on the data is theoretically possible, we store all datasets anonymised to erase any links between a potential condition and a single participant. Given the above concerns, we required participants to be at least 18 years old as the only criteria for inclusion in the study.

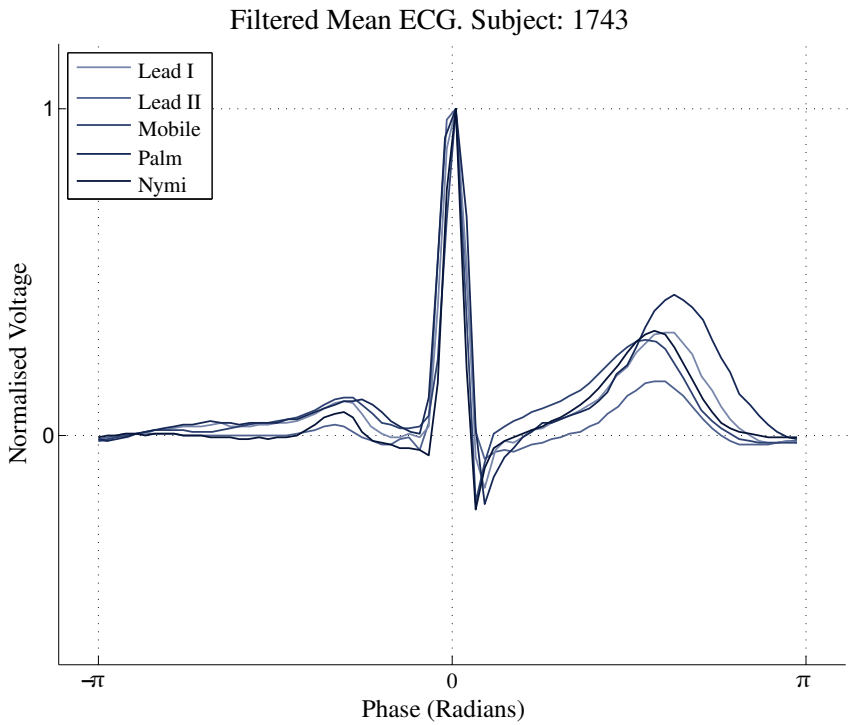
We recruited a total of 41 participants (21 female, 20 male) through mailing list adverts and social media. Participants were made aware that the research involves ECG biometrics, but were only told the specific purpose of the data collection afterwards. Participants were compensated in cash for their time and inconvenience.

5.5 Developing a Cross-Device Mapping Function

In this section, we show how to generate attack signals for the Nymi Band using ECG data from different sources. The method is based on the derivation of *mapping functions*, i.e., functions that transform signals recorded from one device, the *source*, in order to



(a)



(b)

Figure 5.10: Comparison of mean ECGs from same subjects and different leads/devices. Mean ECGs are computed after a linear phase assignment [113], assigning a periodic phase value to each sample in the ECG, starting from one R-peak (phase 0) and ending with the next R-peak (phase 2π). For each heart cycle, amplitudes are scaled by the amplitude of the corresponding R-peak.

resemble the morphology of the signals from a *target* device. In this case, the target device corresponds to the Nymi Band. In other words, we aim to find a function that, given in input an ECG signal from a source device, is able to produce the "same" signal as if it was recorded on the Nymi Band.

In this way, the mapping function can mitigate and even eliminate the statistical differences in biometric features that arise from using different measurement devices and modes. In Figure 5.10, we compare the mean ECG signals recorded for the same individuals but from different devices, showing that the signals exhibit device-specific morphologies, especially as far as amplitude features are concerned. For instance, we observe that for both subjects the palm measurements yield a more prominent T wave, while Lead II signals yields the lowest T wave peak. Similarly, in both cases the P waves obtained from Lead I and the mobile ECG monitor stand above those of the Nymi Band and Lead II, and the R wave of Lead II has the least amplitude. These observations demonstrate that many discrepancies in the ECGs are device-specific and thus can be addressed by the application of mapping functions.

Let S and T be the source and target devices, respectively. Let J be the set of ECG features described in Figure 5.1, and I be the set of subjects we use for computing the mapping. The training dataset consists of the sets $\{\mathsf{ECG}_i^S\}_{i\in I}$ and $\{\mathsf{ECG}_i^T\}_{i\in I}$ of ECG signals recorded, for each subject $i \in I$, with the source and target device, respectively. The method is based on the following steps:

1) **Feature extraction.** From the input ECG data, we extract the relevant biometric features. The outputs of this step are, for each subject $i \in I$, sets of discrete probability distributions $\mathcal{D}_i^S = \{D_{i,j}^S\}_{j\in J}$ and $\mathcal{D}_i^T = \{D_{i,j}^T\}_{j\in J}$, where $D_{i,j}^S$ ($D_{i,j}^T$) is the distribution of ECG feature j for subject i in the source (target) signal.

Specifically, we consider the time domain features summarized in Figure 5.1 and apply the algorithm of [114] for their detection.

2) **Mapping estimation.** This boils down to an optimisation problem (described in Section 5.5.1) where we seek to find an optimal mapping, i.e., a set of transformation functions $\mathbf{f} = \{f_j\}_{j\in J}$ with $f_j : \mathbb{R} \to \mathbb{R}$, such that, for each feature j and subject i, they minimise the statistical distance between the transformed source distribution $f_j(D_{i,j}^S)$ and

the corresponding target distribution $D^T_{i,j}$ [3]. In other words, f_j transforms values of feature j from device S in order to be as close as possible, statistically speaking, to the values of the same feature from device T. We restrict the search to linear functions, of the form:

$$f_j(x) = a_j x + b_j \tag{5.1}$$

Linear mappings are adequate in this context because the amplitudes of the ECG wave peaks along different leads are linearly related [108]. Moreover, unlike more complicated transformation functions (e.g. polynomial or logarithmic), linear mappings are less likely to suffer from over-fitting, especially for smaller datasets.

Once estimated, the mapping $\mathbf{f}$ between S and T is used to generate attack signals for device T starting from new signals recorded with S. Let $i' \notin I$ be our victim, for which we possess an S-signal $\mathsf{ECG}^S_{i'}$. The procedure consists of the following steps:
1) Extract the feature distributions $\mathcal{D}^S_{i'}$ from the source signal $\mathsf{ECG}^S_{i'}$.
2) Apply the estimated mapping function $\mathbf{f}$ to derive the transformed features distributions: $\mathbf{f}(\mathcal{D}^S_{i'}) = \{f_j(D^S_{i',j})\}_{j\in J}$.
3) Produce an attack signal by generating a synthetic ECG signal out of the transformed features $\mathbf{f}(\mathcal{D}^S_{i'})$ and present it to the Nymi Band, as explained in Section 5.3.

5.5.1 Optimization Problem

We formulate the problem of finding the best mapping function as a non-linear constrained single-objective optimization problem that we solve using a genetic algorithm [115]. The problem is defined as follows:

$$\underset{(a_j,b_j)_{j\in J}}{\text{minimize}} \quad \sum_{i\in I^*} d\left(\mathbf{f}(\mathcal{D}^S_i), \mathcal{D}^T_i\right) \tag{5.2}$$

$$\underset{i\in I, j\in J}{\text{subject to}} \quad a_j, b_j \in [k_j^{\perp}, k_j^{\top}] \tag{5.3}$$

$$a_j \cdot D^{S,\min}_{i,j} + b_j \in \left[D^{T,\min *}_{i,j}, D^{T,\max *}_{j}\right] \tag{5.4}$$

$$a_j \cdot D^{S,\max}_{i,j} + b_j \in \left[D^{T,\min *}_{i,j}, D^{T,\max *}_{j}\right] \tag{5.5}$$

[3] Technically, for discrete distribution D, $f_j(D)$ is the distribution whose support is the image of $\mathsf{supp}(D)$ under f_j ($\mathsf{supp}\,(f_j(D)) = f_j[\mathsf{supp}(D)]$) and with probability mass function defined, for $x \in \mathsf{supp}\,(f_j(D))$, by $f_j(D)(x) = \sum_{x'\in f_j^{-1}[x]} D(x')$ where $f_j^{-1}[x]$ is the preimage of x under f_j, that is, all the elements $x' \in \mathsf{supp}(D)$ such that $x = f_j(x')$.

The decision variables are, for each feature $j \in J$, the linear coefficients a_j and b_j characterising the transformation f_j we seek to estimate (see Equation 5.1). The objective function is the sum over subjects $i \in I^* \subseteq I$ of the distance between the transformed source distributions of i, $\mathbf{f}(\mathcal{D}_i^S)$, and the corresponding target distributions $\mathcal{D}_i^T$. Here d is a generic distance measure (discussed later). In particular, I^* is a subset of the training set I and is obtained as follows: 1) we compute the distances $d\left(\mathbf{f}(\mathcal{D}_i^S), \mathcal{D}_i^T\right)$ for all $i \in I$; 2) we apply the Grubbs' test [116] to detect the set of outliers $I' \subseteq I$ on these distances; 3) we remove the identified outliers: $I^* = I \setminus I'$. These three steps are repeated until a maximum number of outliers is removed or no further outliers are identified. The rationale for considering I^* instead of I in the objective function is that we do not want to penalise mappings that perform well for most subjects and poorly for few of them (the outliers). This approach has the added advantage to bypass subjects with inaccurate input ECG data, e.g., through noise introduced through excessive movement. These cases are indeed very likely to be identified as outliers.

Regarding the feasible region of the optimization problem, Equation 5.3 ensures that the linear coefficients are bounded in some real-valued interval $[k_j^\perp, k_j^\top]$. The purpose of Equations 5.4 and 5.5 is to constrain the ranges of the transformed source features, in a way that they are similar to the ranges of the target features. Preliminary results showed that these constraints are crucial to ensure that the corresponding attack signal resembles a biologically realistic ECG. For subject i and feature j, let $D_{i,j}^{T,\min}$ and $D_{i,j}^{T,\max}$ be the minimum and the maximum values of distribution $D_{i,j}^T$, respectively[4]. We define the lower and the upper bounds for the transformed features as:

$$D_j^{T,\min *} = (1-q) \cdot \min_{i \in I} D_{i,j}^{T,\min} \text{ and } D_j^{T,\max *} = (1+q) \cdot \max_{i \in I} D_{i,j}^{T,\max}$$

where $q \in (0,1)$ is a factor for relaxing the range width. The resulting range constraints are given, for all points x in the support of the source distribution $D_{i,j}^S$ by:

$$a_j \cdot x + b_j \in \left[D_j^{T,\min *}, D_j^{T,\max *}\right]. \tag{5.6}$$

[4]With abuse of notation, the minimum and maximum of a discrete distribution D are meant as the minimum and maximum of its support.

Note that the number of such constraints quickly explodes with the number of subjects, features and distinct data points per feature. However, by the monotonicity of the linear mappings, it suffices to check Equation 5.6 only for the minimum and maximum values of $D^S_{i,j}$, denoted respectively by $D^{S,\min}_{i,j}$ and $D^{S,\max}_{i,j}$, thus yielding Equations 5.4 and 5.5. Importantly, this implies that our estimation method supports not just linear functions, but general monotonic functions.

Statistical distance. The distance function of Equation 5.2 is defined as the mean of the statistical distances between the transformed and the target distributions over all the features:

$$d\left(\mathbf{f}(\mathcal{D}^S_i), \mathcal{D}^T_i\right) = \frac{1}{|J|}\sum_{j\in J} d_s\left(f_j(D^S_{i,j}), D^T_{i,j}\right).$$

where d_s is a generic statistical distance. Among the possible candidates for d_s, we chose the L^2 distance between distributions. Let $\mathcal{F}^{\overline{S}}_{i,j}$ and $\mathcal{F}^T_{i,j}$ be the piece-wise linear estimations of the cumulative distribution functions of $\mathbf{f}(D^S_{i,j})$ and $D^T_{i,j}$, respectively. Then, we define d_s as the L^2 distance between functions $\mathcal{F}^{\overline{S}}_{i,j}$ and $\mathcal{F}^T_{i,j}$:

$$d_s\left(f_j(D^S_{i,j}), D^T_{i,j}\right) = w_j\left(\int_{D^{T,\min *}_j}^{D^{T,\max *}_j}\left(\mathcal{F}^{\overline{S}}_{i,j}(x) - \mathcal{F}^T_{i,j}(x)\right)^2 dx\right)^{\frac{1}{2}}$$

where $w_j = D^{T,\max *}_j - D^{T,\min *}_j$ is introduced as a normalisation factor. In the implementation, the above integral is approximated using a composite mid-point quadrature formula.

5.5.2 Synthetic Signal Generation

Synthetic ECG signals are generated as the sum of Gaussian functions, used to reproduce the typical bell-shaped curves of the ECG waves and parametrised by sampling values from a given set of feature distributions. As previously explained, for attack signals we consider ECG features after the application of some mapping function.

The method extends [114, 117] in order to support asymmetric ECG waves, which are physiologically more accurate, thus leading to attack signals that better emulate the Nymi Band's ECG, as discussed in Section 5.5.3.

Let $(\mathsf{PP}_1, \ldots, \mathsf{PP}_{n-1})$ be the sequence of PP intervals detected from the source signal. The sequence is used to determine the beginning of each heart cycle such that, for $h = 1, \ldots, n$, the h-th heart cycle starts at time $T_h = T_0 + \sum_{k<h} \mathsf{PP}_k$, where T_0 is the offset of the first P wave.

For each ECG wave kind $w \in \{P, Q, R, S, T\}$ and heart cycle $h = 1, \ldots, n$, the considered features are: wave amplitude, $A_{w,h}$; wave peak location relative to the start of the h-th cycle, $L_{w,h}$; and left and right components of the wave width at half amplitude, $W^l_{w,h}$ and $W^r_{w,h}$. Note that peak locations are easily derived from the interval features shown in Figure 5.1. The synthetic ECG at time t is defined as follows:

$$s(t) = \sum_{h=1}^{n} \sum_{w \in \{P,Q,R,S,T\}} G\left(t, T_h + L_{w,h_1}, A_{w,h_2}, W^l_{w,h_3}, W^r_{w,h_4}\right) \tag{5.7}$$

where $G\left(t, T_h + L_{w,h_1}, A_{w,h_2}, W^l_{w,h_3}, W^r_{w,h_4}\right)$ is the value at point t of an asymmetric Gaussian curve centred at $T_h + L_{w,h_1}$, with amplitude A_{w,h_2}, and full width at half maximum made of left component W^l_{w,h_3} and right component W^r_{w,h_4}. G is given by:

$$G\left(t, L, A, W^l, W^r\right) = A \cdot \exp\left(-4 \cdot \log 2 \cdot \frac{(t-L)^2}{W(t)^2}\right)$$

where $W(t) = W^l$ if $t \leq L$ and $W(t) = W^r$ otherwise.

Note that in Equation 5.7, the features used to generate the Gaussian curve are not necessarily drawn from the same heart cycle h. Specifically, for each cycle h, we randomly sample the heart cycles $h_1, \ldots, h_4$ from which location, amplitude and width features are extracted. Based on preliminary results, among the possible sampling strategies, we choose peak location and widths from the same heart cycle, i.e., $h_1 = h_3 = h_4$.

Importantly, such generated synthetic ECGs account for the specific inter-beat patterns of the subject (another common ECG biometric feature), since we use the same PP sequences detected from the source signal.

5.5.3 Evaluation

In this subsection, we perform an in-depth evaluation of the methods for estimating mapping functions and generating synthetic signals. The aim of the following experiments is to obtain insight into HeartID, the Nymi Band's authentication and biometric

recognition library, in order to devise the best design choices for our methods. These choices include the ECG features to be used in the mapping or the filtering algorithm in the ECG detection procedure. There is only very little information available about the algorithms used in HeartID apart from [118] which proposes a continuous biometric recognition system based on ECG called HeartID.

Unfortunately, it remains unknown to what degree the techniques and algorithms described in [118] have been included in the authentication library currently being used by the Nymi Band's companion app. As such, we do not have any prior knowledge of the classifiers or features used in the authentication process. Nevertheless, we can use the NCA as an oracle by querying it with an ECG signal and recording the response (i.e., an accept or a reject of the signal).

The main obstacle to such an extensive analysis is the time needed to inject attack signals using a waveform generator or a sound player. To overcome this limitation, we devised another kind of attack, called *offline attack*, which is instantiated by directly interfacing with HeartID through the API calls of the NCA (the Nymi Band's companion app). We implemented a simple Android app that allows setting up previously stored biometric templates and performing authentication for arbitrary attack signals, at a rate of hundreds of signals per minute, without requiring a waveform generator or Nymi Band. Our devised Android app does not require physical ECG input, but accepts biometric template and biometric samples in digital form via a command line interface and forwards them to the HeartID library. For every authentication attempt, the library's authentication decision is stored in a database by our Android app for later analysis.

As this attack necessitates modifications of the NCA before enrolment, it would not be feasible to execute in an actual attack. As such, we only use it for the preliminary development of the attack and obtain the final results of Section 5.6 online, i.e., using the actual Nymi Band.

The experiments below were performed on a selection of 8 subjects from the training set. For each experiment, subject and device, we tested 20 randomly generated attack signals. We recall that the synthetic ECG is generated by sampling features from distributions, hence the reason of their randomness. The reported success rate (SR)

is the ratio between successful authentications and total attempts. As expected, the enrolment signals used to build the biometric templates yielded an SR of 100%.

Filtering. We evaluate the adequacy of our ECG filtering algorithm, which builds on a Savitzky-Golay (SG) smoothing filter [119]. We applied the SG filter to the enrolment signals, obtaining for them an SR of 100%. Stronger SG smoothing parameters resulted in SRs below 63%. We also wanted to assess how noise affects authentication chances. For this purpose, we added white Gaussian noise with standard deviation computed from the enrolment signal to the filtered signal. These signals still yielded an SR of 100%, demonstrating that the filtering method of HeartID can equally support noisy and filtered ECG data. Note that the above experiments did not require the generation of synthetic ECGs.

Synthetic signal generation. Here, we assess the fit of our synthetic ECG signals. To this purpose, we generated signals drawing on the features detected in the enrolment ECGs, and reproduced in the same order. These signals resulted in an overall SR of 75%, with an SR of 100% for 5 out of 8 subjects. To understand if HeartID is sensitive to all the ECG waves, we further produce signals where one of the waves is systematically suppressed by setting its amplitude to 0. For each wave kind, the obtained SR was 0%, suggesting that all ECG waves are used in the biometric template. In a variation of the first experiment, we test signals with symmetric waves, resulting in an SR drop from 75% to 15%. This motivates our claim that asymmetric waves are needed in order to produce realistic and successful attack signals. Finally, we report that subject-specific ECG features are indeed central for authentication: signals generated using default parameters from literature [120] produced an SR of 0%.

Statistical distance. Now that we have proved the adequacy of the filtering and synthetic signal generation algorithms, the next step is evaluating the mapping function. In this experiment we want to assess how different statistical distances in the optimisation problem affect the success rate. Here, we restrict to *one-to-one mappings*, i.e., mappings estimated over data from a single subject. In contrast to the default mappings, one-to-one mappings are tailored to the subject, meaning that the resulting attack signal better mimics the target ECG. However, these cannot be used to instantiate an attack, since they require

the victim's enrolment signal. We compare the previously introduced L^2 distance with the χ^2 index for empirical distributions and the Kolmogorov-Smirnov statistic (KS). Using L^2, we get an overall SR of 49.9%, while for χ^2 and KS, the SR is 29.5% and 39.6%, respectively. This supports our choice of L^2.

Mapping strategies. We finally evaluate a number of alternatives to the default mapping functions. Signals generated using the default settings, where the mapping is estimated from the whole training dataset and considers the full set of ECG features, resulted in an SR of 26.2%. Slightly worse success rates are obtained when some of the features are not transformed, i.e., left as detected in the source signal: the overall SR is 22.8%, 21.1% and 20% if excluding, respectively, widths, peak intervals and both. As expected, the exclusion of amplitude features has shown no significant success. We get comparable results when considering sub-optimal mappings in place of the best mapping function found by the optimisation algorithm: for the second and third best mappings, we have SRs of 21.5% and 21.9%, respectively. We also learnt that one-to-one mappings estimated from one single subject are mostly ineffective for attacking a different subject (SR = 4.4%), thus confirming the importance of gathering a substantial training dataset. Another alternative that performed poorly (SR = 11.5%) is building a mapping from a single "super-individual", obtained by merging the features distributions across all subjects.

5.6 Results

We conduct the signal injection attack by combining the building blocks described in the previous sections. After enrolling a user into the Nymi Band, we first inject the raw data that was collected using each of the devices (see Section 5.4). Following that, we apply the mapping function, which has been trained on the initial set of users (i.e., not including the user that is being attacked) to obtain the transformed signals. As outlined in Section 5.5, the mapping function employs random sampling of features to generate the attack signal. We use between 2 and 4 repetitions of this sampling process and report whether at least one of them was accepted by the Nymi Band. As the Nymi Band does not limit the number of authentication attempts a user can make, the only cost of repetitions

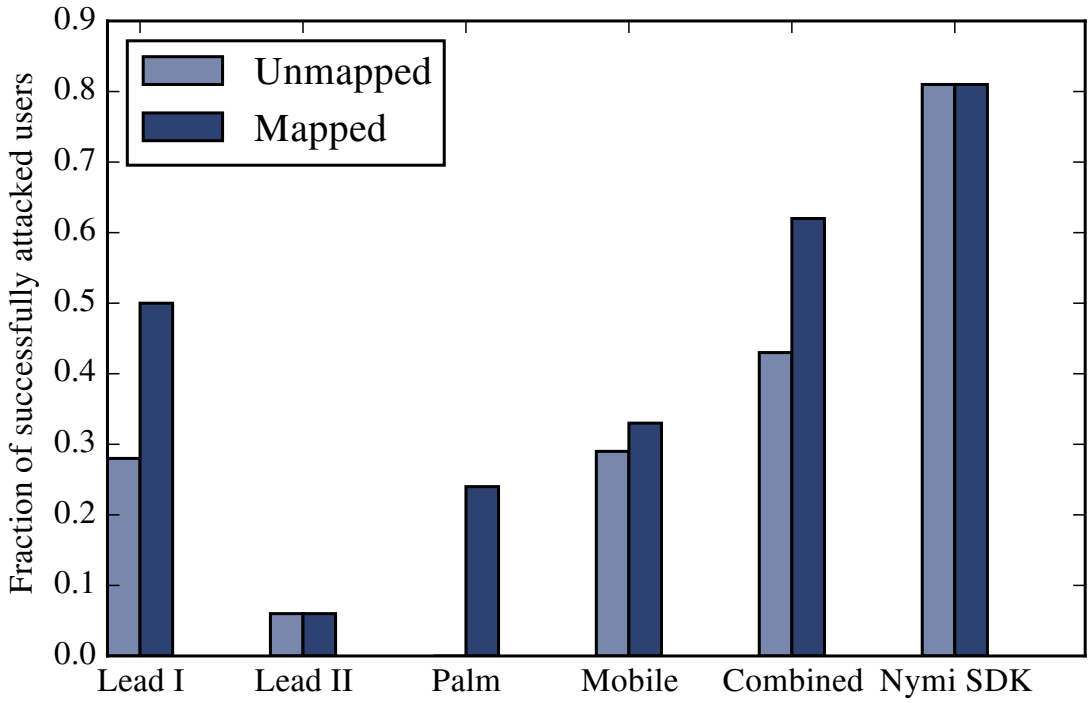


Figure 5.11: Attack success rate for different devices. Results are shown for raw data and the data obtained by applying the mapping function.

lies in an increased time required to carry out the attack. Due to the extensive time required to manually perform signal injections for all users, we only report the results of injecting the ECG signals by playing back .wav files using an off-the-shelf music player. As we have discussed in Section 5.3, the performance that can be expected is similar for all three devices, and the .wav playback poses the lowest technological barrier for the attacker, both in terms of cost and compactness. The raw results of our analysis are shown in Table 5.2, a summary is shown in Figure 5.11. For 3 out of 21 users we were unable to collect high-quality signals using the external electrodes of the ECG monitor. This is most likely a result of hair impeding the contact between the electrode and the participant's skin. Due to practical reasons, we have not asked participants to shave the respective areas (as is done before medical ECGs). As a result, we are left with 18 attack attempts for the Lead I and Lead II data sources.

Not surprisingly, the success rate of injecting data collected with the Nymi Band is the highest. With the exception of four users, using this data resulted in unlocking the band, leading to a success rate of 81%. However, two out of the four unsuccessful

User	Lead I	Lead II	Mobile	Palm	Combined
1	✗	✗	✗	✓	✓
2	✗	✓	✗	✗	✓
3	✗	✗	✗	✗	✗
4	✗	✗	✗	✗	✗
5	✓	✗	✓	✓	✓
6	✗	✗	✗	✗	✗
7	✗	✗	✗	✗	✗
8	N/A	N/A	✓	✗	✓
9	N/A	N/A	✗	✗	✗
10	✗	✗	✓	✗	✓
11	✗	✗	✗	✗	✗
12	✓	✗	✓	✗	✓
13	✓	✗	✗	✗	✓
14	✓	✗	✗	✗	✓
15	✓	✗	✓	✓	✓
16	✗	✗	✗	✗	✗
17	N/A	N/A	✗	✗	✗
18	✓	✗	✗	✗	✓
19	✓	✗	✗	✓	✓
20	✓	✗	✓	✗	✓
21	✓	✗	✓	✓	✓

Table 5.2: Results of injecting the data generated by applying the mapping function. The shaded cells show the results that changed by applying the mapping function. Due to source and target device being identical, the mapping function is not applied to the Nymi SDK data. The tick marks signify users that were successfully attacked.

users failed to authenticate themselves following enrolment. This hints that the failure of the attack is most likely a consequence of erratic or noisy data being collected during the enrolment phase. Injecting the Lead II measurement succeeded only for a single user. Intuitively, this makes sense since Lead II measures voltage between the left leg and right arm (rather than the left and right arm, which is approximately what the Nymi Band observes). Conversely, Lead I performs relatively well in the attack, succeeding in five out of 21 users (24%), most likely being a result of the similar measurement points. Based on this intuition, one would expect the palm measurements to perform similarly well, as the measurement points (palm of the left hand and index finger of the right hand) are even closer to those used by the Nymi Band (wrist of left hand and index finger of right hand). However, using these measurements caused the attack to fail for all users (and is the only data source to do so).

Applying the mapping function considerably improves these results, particularly for those data sources that performed poorly initially. The success rate for using Lead I data improves from 28% to 50%. The Palm measurements, which were initially unsuccessful for all users, could be used for successful attacks on 5 users, thus increasing the success rate from 0% to 24%. The mobile ECG monitor data source is the only one where applying the mapping function causes the attack to fail for one user. This is most readily explained by the unmodified data just being on the edge of being accepted. The only data source for which the mapping function had no effect (positive or negative) is the Lead II data obtained from the ECG monitor. This could be either due to limitations of the mapping function (such as the precise feature set of the Nymi Band being unknown), or due to important biometric information simply not being present in this lead. The former case could possibly be remedied by obtaining a better understanding of the biometric features involved (which is difficult in a blackbox scenario as presented by the Nymi Band). In the latter case, it is not possible to find a mapping between the feature distributions, thus requiring the attacker to obtain a different source of ECG information. In the medical domain it is necessary to measure multiple ECG leads as they contain different kinds of information, so it is not implausible that this is similar for identifying (biometric) information.

Overall, the attacker's chance of success is 81% assuming they have obtained a measurement through the Nymi Band, and 62% if they have only obtained data from the remaining sources. The latter is computed as the fraction of users for which at least one of the four data sources led to a successful activation of the band (shown as the "Combined" column in Table 5.2.

5.7 Discussion and Countermeasures

In the previous sections we have outlined an effective presentation attack against ECG biometrics in general and have shown its effectiveness when applied to the Nymi Band. Most generally, there are two main approaches to mitigate the attack:

The first approach is liveness detection. Liveness detection attempts to detect an injected signal and distinguish it from a signal originating from a human. This technique

has been applied with varying degrees of effectiveness for other biometrics. The case of fingerprint readers in particular showcases that this is most likely an arms race between system designers considering more indications of liveness (e.g., the presence of skin oils or moisture for fingerprint readers) and attackers spoofing these indicators.

The second approach would be to keep the biometric data secret, thus attempting to prevent the attacker from obtaining any of the victim's ECG data. One could make the case that the Nymi Band's SDK should not allow for ECG data to be recorded, either through NEAs or otherwise. Disabling this functionality would make it at least significantly harder for the attacker to directly obtain data with the correct ECG morphology. As creating the mapping function requires training data from the target device, making this data hard to obtain would also raise the bar to carry out the attack. However, it might still be possible to determine the device-specific effects on the signal by analysing the hardware, rather than through empirical analysis of recorded signals. As we have demonstrated, the mapping function allows the attacker to obtain data from an arbitrary device and use it to attack the authentication system. While the probability of success somewhat depends on the device, we have demonstrated successful attacks using all of the devices we analysed. Due to the large variety of devices recording ECG (e.g., medical ECG monitors, fitness devices, e-health) for purposes other than authentication, we do not consider keeping the ECG data secret a viable strategy.

In this chapter, we have applied the mapping function in a way that allows us to attack an ECG-based authentication system with data collected on a different device. However, the same approach can be used to improve the interoperability of a biometric across different devices. For example, once a user is successfully enrolled for one device (e.g., a smartphone for the touchscreen input biometric), she could be authenticated on any other device without the need for re-enrolment, provided a mapping function between these devices has been trained beforehand. In the next chapter, we will present a generalised version of this attack. This generalisation will then allow us to not just carry out attacks, but also to quantify the security of biometrics and individual features.

La vie n'est facile pour aucun de nous. Mais quoi, il faut avoir de la persévérance, et surtout de la confiance en soi. Il faut croire que l'on est doué pour quelque chose, et que, cette chose, il faut l'atteindre coûte que coûte.

Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something, and that this thing, at whatever cost, must be attained.

— Marie Curie

6

Cross-Context Attacks Against Behavioural Biometrics

Contents

6.1 Introduction

In Chapter 5, we first faced the problem of device-specific distributions of ECG features. Our results show that ECG exhibits device-specific feature distributions, i.e., the distribution of each individual feature depends not just on the individual, but also the measurement device. This property constitutes a challenge for an attacker, as data obtained through one device can not be simply presented to the authentication device. We solved this by adapting the mapping function methodology to the attack and combining it with the generation of synthetic ECG signals (i.e., generating an ECG signal that matches the feature distributions generated by the mapping function). This approach has greatly improved the attack's success rate.

The goal of this chapter is to further generalise this methodology outside the context of a specific attack. To this end, we extend our previous work to cover a wider variety of biometrics. We also go beyond the cross-device scenario presented in the previous chapter and extend it to a *cross-context* setting. A context is identified through a number of factors, such as measurement device, sensor placement, selected task and environment. While the previous chapter has used a binary metric of success (i.e., whether an individual attack succeeded or failed), we now derive an *unpredictability score*. This score, which intuitively reflects the error of the mapping function, is a more fine-grained measure and more accurately reflects the security of a biometric without being affected by arbitrarily chosen accept/reject thresholds. Deriving this score serves three main purposes: (a) overall comparison of biometric systems, (b) identifying vulnerable target contexts and (c) selection and engineering of secure features.

Drawing from the insights gained in the two previous chapters, we formulate a number of research questions:

- How can the cross-context predictability of biometric features be quantified?
- How do different biometrics perform according to this metric?
- How do individual features contribute to (un-)predictability?
- How do different population sizes affect feature predictability?

6.2 Threat Model

In this chapter, we focus on adversaries that attempt to bypass a biometric authentication system using incomplete biometric information about the victim from another *context* and combining it with population data.

Overview Figure 6.1 shows an example of such a scenario. The victim is enrolled into a gait authentication system through their phone. We refer to the system and the context used by the system as *target*. The system maintains a confidence in the user's identity based on their gait patterns and allows certain sensitive operations (e.g., authorising payments) only when the confidence is above a threshold. The victim's biometric template is stored on the phone in a trusted module that cannot be accessed by the adversary.

The adversary knows that the victim uses a smartwatch that monitors their gait, for example for health or sport reasons. We refer to the context of the smartwatch as *source*. Either the smartwatch, its connected smartphone application or the wireless link are insecure and the adversary exploits the smartwatch to obtain the victim's gait data. However, similar to the attack presented in Chapter 5, the smartwatch data cannot be used directly to impersonate the user at the *target* system, because of feature differences caused by the different context. Therefore, the adversary collects biometric data from a population (which excludes the victim), reproducing the *source* and *target* contexts. Using only population data, the adversary attempts to learn how to transform gait data from *source* to *target* and uses this information to transform the stolen victim's gait. The transformed data allow the adversary to impersonate the victim at *target*.

Assumptions The victim is enrolled into a biometric authentication system (*target*). The biometric data used by the *target* system is measured in a pre-defined context (*target* context). The attacker wants to impersonate the victim at *target* system. We assume the following:

- Obtaining the victim's biometric data usable in *target* context is hard, because the devices that process *target* system data are highly protected;

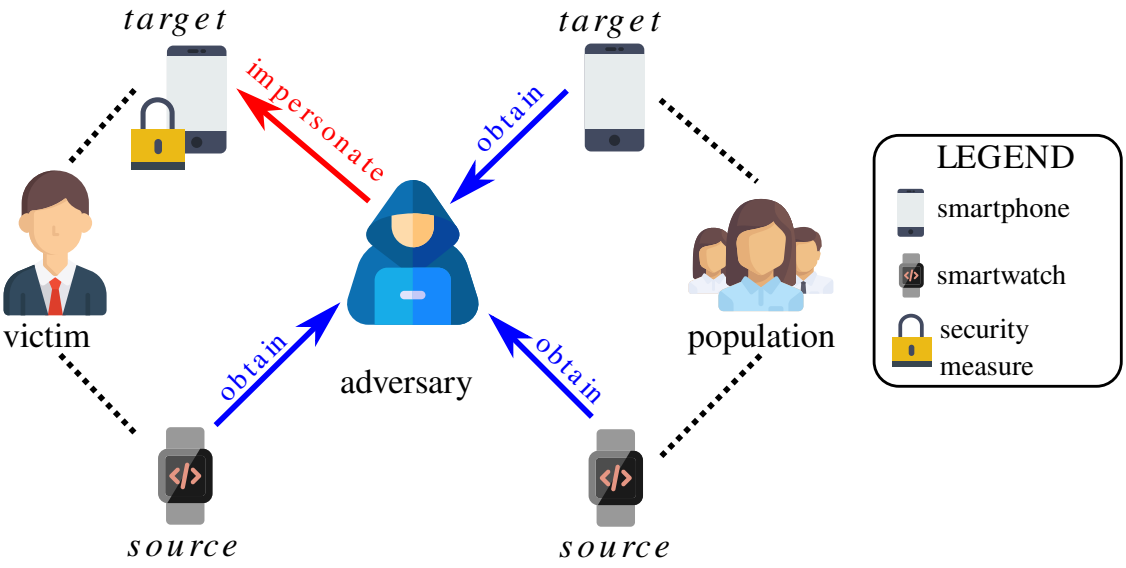


Figure 6.1: Example of threat scenario.

- The victim uses another system that makes use of the same type of biometric data as *target*, we refer to this as *source* system. Data from *source* are more easily obtainable, but are measured in a different context (*source* context);
- The adversary can obtain biometric data from a population for *source* and *target* contexts (i.e., same sensor, task, environment, etc.);
- The adversary knows the biometric features used for recognition by *target* system, but does not know any other detail used by the recognition algorithm;
- The adversary can reconstruct biometric signals from biometric templates and can inject forged biometric data into *target* system.

It should be noted that the biometric data for *source* and *target* can either be raw biometric signals or vectors of biometric features. In fact, since the adversary knows the feature extraction algorithm used by the system, they can easily compute features from raw signals.

The adversaries may obtain population data in different ways. As an example, they could ask their friends to provide their biometric samples, or invite members of the general public for a lab study. For some biometrics, it might also be possible to use publicly available data (e.g., medical databases for ECG). Although the adversary may need to

invest time and effort in collecting the population data, it is a worthwhile investment. In fact, once the universal transformation is learned, the adversary can use it to impersonate potentially anyone. In the case of local authentication, the adversary will have to obtain physical access to the device and, depending on the method of injection, bypass liveness detection. On the other hand, if authentication is performed remotely (i.e., on a server), the adversary can perform the attack in a more scalable way.

In the following subsections, we motivate the threat model by presenting different scenarios for each of the biometrics. In each scenario, we outline how different factors contribute to the feature differences between contexts.

6.2.1 Gait

Little attention is being paid to the confidentiality of accelerometer-based gait data. At the time of writing, accessing the accelerometer does not require a permission in the Android Manifest file (Android v8.0), and can be accessed directly by websites, through the DeviceMotion API. This means that adversaries might obtain control of an application (or make an application or website themselves) and silently collect data from oblivious users. Furthermore, most fitness trackers have been proven vulnerable to exploits, both in the wireless channel [121] or in the firmware [122].

With the adversary being more likely to obtain data from a fitness tracker (or a fitness application running on the smartphone), two main factors should be considered. The first one is the on-body location of the accelerometer sensor and the second one is the type of movement: either walking or running. The rationale behind the location is that different parts of the body are subject to different accelerations (e.g., arms, chest, or wrists). On the other hand, the use of fitness trackers is more popular while running than walking (e.g., to monitor work-out statistics). Attackers need to consider that running data looks extremely different from walking due to the stronger forces generated by the run and the shorter timing between steps.

Biometric	Factors	Considered Scenarios	Devices
Gait	Activity Sensor Location Input Device	walking, running arm, chest, hand, pocket, wrist smartphone, smartwatch, fitness tracker	BLU VIVO 6, Movisens ekgMove, Garmin Vivoactive HR
Touch dynamics	Input Device	low-, mid-, high-end phone	TTSim M5 Smart, Motorola MotoG3, BLU VIVO 6
ECG	Sensor Type Activity	mobile monitor, medical monitor, fitness tracker, authenticator resting, walking, running	AliveCor KardiaMobile, Heal Force Prince 180B, Movisens ekgMove, Nymi Band
Eye movements	Task Calibration	reading, watching video, writing, browsing calibrated, uncalibrated	SMI RED500
Mouse movements	Input Device	trackpad, mouse	MSI GT72 6QE Dominator Pro G trackpad, Dell Laser USB mouse

Table 6.1: Factors of feature distribution differences considered for each biometrics and devices used for the measurements.

6.2.2 Touch Dynamics

In the case of touchscreen data, there are two main ways in which the adversary could obtain users' biometric data: a malicious mobile application, or a malicious website. Adversaries could create applications that silently monitor the touchscreen inputs and trick victims (e.g., through social engineering) into installing and using these applications on their smartphones. Similarly, touchscreen data collection could be carried out on a website through simple Javascript[1].

For touch devices, we focus on the scenario where victims use at least two phones, one of which is highly protected. An example for this scenario is where victims have a company-issued smartphone that contains company-sensitive information and is secured with different means (e.g., trusted modules, touchscreen lock, no installation of arbitrary apps allowed). In particular, the device continuously authenticates the user using touchscreen biometric while they are using sensitive applications.

The adversary will try to obtain the user biometric from a less protected device (e.g., their personal smartphone). In this case, the first factor to account for in the transformation is the dimensions of the touchscreen, as these changes the span/shape of the swipe gesture. Additionally, the sampling rate of the touchscreen has a significant effect, as less fine-grained information changes the meaning of features based on a subset of the swiping gesture (e.g., initial acceleration of swipe). Other sensor data, such as pressure or area covered, might also be different in terms of scale, resolution, precision and granularity.

6.2.3 ECG

Similarly to gait (Section 6.2.1), insecurities in the communication channel or the device firmware can both be a point of attack for the adversary that is attempting to obtain ECG data. In addition, computerized medical records are often handled poorly in terms of their confidentiality. Reports show that large amounts of sensitive healthcare data are vulnerable to leakage or theft, or have already been compromised because of security lapses at hospitals, insurance companies or government agencies [123]. Adversaries

[1] https://developer.mozilla.org/en-US/docs/Web/API

may also easily obtain raw ECG signal from photos of ECG printouts, as we have demonstrated in Chapter 5.

Comparably to gait, with fitness trackers being more likely to be exploited, the adversary should also account for the different ECG behaviour due to the activity performed by the user during the measurement. The ECG signal significantly changes when the user is exercising, both due to the physical exertion and noise introduced by imperfect electrode connection.

6.2.4 Eye Movements

The popularity of eye-tracking is increasing and a number of consumer electronics are equipped with eye trackers. With more services implementing eye-tracking, adversaries can use these services to obtain eye movement data (e.g., hijacking browsers and using a Web API, or exploiting application weaknesses). Additionally, we consider the threat of the user being tricked into using an attacker-controlled machine which is equipped with a covert eye tracker.

We have first shown in Chapter 3 that gaze data strongly depend on the type of task performed by the user (e.g., reading, writing, browsing). Since the adversary can not easily force the user into performing a specific task, they might need to adapt the victim's data to the task that is used for authentication. Additionally, eye trackers need to be calibrated before use to provide accurate data. Since it would be considerably more difficult to trick the user into calibration (as this procedure would raise suspicion if there is no legitimate reason for eye tracking being used), we assume that the attacker only possesses data from a device that is not calibrated for the victim.

6.2.5 Mouse Movements

As mentioned in Section 6.2.2, collection of mouse movements data can easily take place on the Web, where it has been shown that mouse tracking is common-place [124]. In order to obtain the victim's data, adversaries may create websites, or hijack existing ones. It could also be possible for the adversary to highjack the victims' browsers (e.g., by installing malicious extensions [125]).

As users interact (and browse) with an increasing number of devices, the adversary needs to account for the different interactions that happen depending on the device hardware. Previous work shows that changing the pointing device hardware causes fluctuations in the measured users behaviour, enough to significantly degrade the recognition performance [18]. Using these observations, we decide to consider the extreme case where the pointing device is either a mouse or a trackpad. This fits well the scenario where mouse data collection happens remotely, that is the most likely to occur online (as mentioned above).

6.3 Experimental Design

In order to evaluate the threat model motivated in the previous section, we conduct a study where we collect participants' biometrics for each of the five biometric modalities. The study is designed to reflect the scenarios presented in the threat model (Section 6.2). The reasoning behind collecting new data is that no publicly available datasets include data in different contexts and different biometrics at the same time. For all biometrics measurements, we stick to state-of-the-art common practices. In the following, we describe the details of the study and briefly comment the processing methodologies that we adopt.

6.3.1 Study Outline

The study consists of two separate but identical sessions which are at least 5 and not more than 30 days apart. In each session, participants undergo a series of tasks designed to collect their biometric traits for a specific context. A single session lasts approximately one hour and 45 minutes. In Table 6.1 we report all the feature difference factors that we accounted for in the analysis and the devices used for the measurements. In the remainder of this section, we present the details of the study procedure for each biometric.

Mouse Movements The first task is carried out on a laptop to collect mouse data [106]. Participants are shown a grid of rectangles and click on the rectangle that contains a picture. After the user clicks, the picture moves to another rectangle and users click on this new rectangle. The task ends after 250 total clicks and is repeated with the trackpad. While mice and trackpads are different in nature, they are used for the same operations

and result in the same type of data. In addition, users may seamlessly switch between the two, which makes this a worthwhile analysis.

Eye Movements The participant is then requested to complete five different tasks on a laptop equipped with an eyetracker. The study is carried out in a lab in controlled lighting conditions (blinds closed and light switched on). The tasks are identical to those in Chapter 3: reading, writing, watching a movie trailer, browsing and watching an educational video. Each task continues for 3 minutes before the next one starts automatically. Differently from [5], we include two different videos to account for the number of scene changes that directly influence the participant's gaze: the movie trailer contains lots of fast-paced scene changes, while the educational video does not. At the end of the session, the five tasks are repeated on an uncalibrated eyetracker. To account for the users getting used to the tasks when they repeat them, we randomly swap the order of the calibrated and uncalibrated tasks.

Touch Dynamics Afterwards, the participant uses a smartphone to complete a "spot the difference" task (similarly to [22]). The smartphone shows two images which contain subtle differences between each other and the user attempts to find them. Only one image is shown on the smartphone at a time and the user swipes (either to the left or to the right) to see the other image. The task lasts 3 minutes in total and is repeated three times, each time with a different phone and a different pair of images. To avoid bias generated by the selection of images and users acclimating to the task, we randomize the order of the phones for each user.

ECG Then, the participants' ECG is monitored for a set of devices: an authenticator (the Nymi Band), a mobile ECG monitor attached to a smartphone and a medical ECG monitor. For the ECG monitor measurement, we collect the palm measurement using the built-in electrodes and use an external 3-lead ECG cable with disposable electrodes, to obtain Lead I, Lead II and Lead III [126]. Additionally, at the beginning of the session, participants wear a chest-strap fitness tracker that monitors their ECG and gait data throughout the session (i.e., including the non-ECG tasks).

Gait Finally, the participant goes for a short walk and a subsequent run in a nearby park (around 700 meters each). During this time, five different sensors monitor the

participant's gait pattern: three smartphones (placed on left arm, right front pocket and held in the left hand), a smartwatch worn on the left wrist and the fitness tracker mentioned above. The fitness tracker also monitors ECG during the walk and the run.

Participant Recruitment We recruited a total of 30 (11 female, 19 male) participants through local announcements and social media. Participants were compensated for their time and inconvenience. This study was reviewed by and obtained clearance from the Inter-Divisional Research Ethics Committee of the University of Oxford, reference number R50977/RE001.

6.3.2 Feature Extraction

We adopt state-of-the-art common practices for biometric data processing and feature extraction. Table 6.2 reports the papers we used. For ECG and gait we use preprocessing steps to allow us to isolate the individual signals (single heartbeat and single gait cycle, respectively), rather than frequency domain analysis. The rationale behind this choice is that feature representation based on frequency domain does not have a direct and understandable meaning, while providing similar (if not weaker) performance results. For gait, we ignore the use of dynamic time warping, as this is only necessary during template matching and does not have an effect on the raw signal behaviour. The details of the feature extraction for each biometric can be found in the cited papers.

6.4 Computing Unpredictability Scores

Based on the data collection process described in Section 6.3, we compute the source-target mappings using the methodology introduced in Chapter 5.

In order to evaluate the effectiveness of the mapping, we measure the prediction error on a per-feature base. Let v be a victim user, $\{u_i\}_{i=1,\dots,n}$ a population of users and g_j the *j*-th feature used by the biometric algorithm. For feature g_j, we compute the optimal cross-context mapping $f_{\theta^*}^{(g_j)}$ (using the population) and the prediction error for the victim's

Biometric	Paper(s)	Description
Gait	M. Derawi et al. [96]	magnitude of acceleration features (based on cycle detection)
Touch dynamics	M. Frank et al. [22]	pressure, spatial, speed and acceleration features
ECG	A. Fratini et al. [110]	temporal, amplitude, morphology features (based on fiducial points)
Eye movements	S. Eberz et al. [5]	pupil, temporal and spatial features
Mouse movements	N. Zheng et al. [16] A. Weiss et al. [106]	stroke curvature, speed and acceleration features

Table 6.2: Description and original paper of the pre-processing and feature extraction methodologies used for each biometric.

source observations to the victim's target observations as $\epsilon^{(g_j)}_{f_{\theta^*}}(v)$. This gives an *unpredictability* score U for feature g_j and victim v in the source-target context transformation:

$$U_v^{(g_j)} = \epsilon_{f_{\theta^*}^{(g_j)}}(v)\,. \tag{6.1}$$

A small value of $U_v^{(g_j)}$ implies that for feature g_j, the cumulative functions of the victim's transformed source random variable and of the target random variable are almost overlapping. This means that (for the *j*-th feature) the cross-context mapping approach is able to accurately map observations from the source context to samples from the target context (the differences are systematic). On the other hand, a value of $U_v^{(g_j)}$ close to 1 implies that for feature g_j the transformed feature values from source random variable and from target random variable have highly non-overlapping distributions. This means that the differences between the *j*-th feature values in the source and target contexts cannot be systematically predicted in this way.

6.4.1 Weighted Score

Following on from the previous section, we know that we obtain an unpredictability score for each feature in the feature-set. We want to aggregate this score to the level of the whole biometric modality (across the features), so that it provides an idea of the resilience of a particular biometric to this transformation. A simple average of the

unpredictability score for each feature is not reasonable, as features contribute differently to the recognition. For example, if a non-distinctive feature is very predictable, it might have a significant negative influence on the overall score. This is not the desired effect, as an attacker would gain very little by correctly predicting that feature.

RMI Weights We weigh features based on Relative Mutual Information (RMI). The reasoning is that more distinctive features should contribute more heavily towards the overall score, as they grant a bigger advantage to an attacker predicting them. To avoid problems with the choice of the number of bins (that may introduce bias in the mutual information), we adopt the non-parametric RMI computation of Ross [127]. In this approach, mutual information is computed based on the relationship between a data point's neighbours and its class neighbours. We weigh each feature mapping result with the feature's RMI and obtain an aggregated score that accounts for feature distinctiveness this way.

Formally, given the set of features for a biometric $\{g_j\}_{j=1,\dots,m}$, the victim user v and each feature RMI value $\{r_j\}_{j=1,\dots,m}$ we compute a RMI-weighted unpredictability score W_v:

$$W_v = \frac{\sum_{j=1}^{m} (\epsilon_{f_{\theta^*}^{(g_j)}}(v) \cdot r_j)}{\sum_{j=1}^{m} r_j}. \tag{6.2}$$

6.4.2 Score Interpretation

The weighted unpredictability score W_v of a biometric modality (Equation 6.2) depends on the scores of the individual features, with distinctive features contributing more to it. It should be noted that the score itself does not directly correspond to a certain success rate of an actual attack, because the cross-context mapping effectiveness also depends on the specific template matching algorithm and false accept and false reject rates thresholds. The main advantage of the unpredictability score lies in its comparative capability, rather than in being an absolute scale. The score can be used to ***compare different biometrics***, with biometrics with higher unpredictability scores across all sources being judged more

secure. Similarly, a system developer can use the scores to ***identify vulnerable target contexts***. For example, a biometric might exhibit low unpredictability scores on specific devices (e.g., due to lower quality sensors). In that case, a developer could change the classifier's decision threshold to account for the increased danger of cross-context attacks.

Lastly, individual feature unpredictability scores $U_v^{(g_j)}$ can be a driving factor in the ***selection and engineering of features***. Higher security can be achieved both by changing the definition of features and by modifying sensor hardware (e.g., by making it less similar to common source contexts).

6.4.3 Evaluation Methodology

Cross-Validation For the evaluation of the cross-context mapping, we operate in a leave-one-out cross-validation fashion. At each step i, we consider one user u_i as the victim and we use the remaining 29 users as the population. With the population, for each feature, we compute the optimal cross-context mapping $f_\theta*$ and the prediction error for the victim's source observations to the victim's target observations. We obtain $U_v^{(g_j)}$ (Equation 6.1) and W_v (Equation 6.2) this way. This step is repeated for each user. If not otherwise specified, the results shown are averages of unpredictability scores over the users in our dataset. The RMI is computed on the feature distribution of the population obtained in the first session, for the target context.

Considered Scenarios In the evaluation, we select a set of sources for each biometric and consider the scenario where the adversary has the information from an individual source, or for the full set of sources (*all*). In the second case, the adversary uses the source with the best performing cross-context mapping (lowest unpredictability) for *each feature*. This scenario constitutes the strongest attacker since some sources may be useful to predict some features but not others. Additionally, we consider two different time scenarios: same session and cross session. The former represents the case in which the victim's source and target data are collected in the same session, which leads to greater similarity. In the latter, the victim's source data were collected in a different session than the victim's target data. Intuitively, this reflects the case of the attacker's source data being older or newer than the victim's template.

Biometric contexts	Same Session avg (min, all)	Cross Session avg (min, all)
ECG		
target: *Authenticator-rest*	**.09 (.07, .06)**	**.12 (.09, .08)**
- *Lead I-rest*	.075 ± .010	.093 ± .014
- *Lead II-rest*	.106 ± .011	.128 ± .015
- *Lead III-rest*	.114 ± .008	.144 ± .014
- *Palm-rest*	.080 ± .007	.110 ± .010
- *Mobile-rest*	.075 ± .007	.092 ± .005
- *Fitness tracker-rest*	.104 ± .010	.134 ± .012
- *Fitness tracker-walk*	.100 ± .012	.123 ± .017
- *Fitness tracker-jog*	.103 ± .011	.122 ± .017
Eye movements		
target: *Calibrated*	**.08 (.07, .07)**	**.10 (.09, .09)**
- *Intra task-uncalibrated*	.068 ± .014	.089 ± .023
- *Cross task-uncalibrated*	.084 ± .017	.103 ± .023
Mouse movements		
target: *Mouse*	**.07**	**.07**
- *Trackpad*	.068 ± .011	.071 ± .010
Touch dynamics		
target: *Mid-end phone*	**.08 (.07, .07)**	**.08 (.08, .07)**
- *Low-end phone*	.084 ± .009	.082 ± .008
- *High-end phone*	.071 ± .008	.075 ± .009
Gait		
target: *Pocket phone-walking*	**.15 (.15, .13)**	**.14 (.14, .13)**
- *Smartwatch-walk*	.155 ± .016	.144 ± .020
- *Hand phone-walk*	.154 ± .021	.145 ± .019
- *Smartwatch-jog*	.148 ± .019	.141 ± .018
- *Cheststrap-jog*	.154 ± .019	.144 ± .020
- *Arm phone-jog*	.156 ± .020	.146 ± .021

Table 6.3: Unpredictability score, for data from the same and cross session. Rows in bold report the aggregated score, introduced in Section 6.4.3. For each source we also show the 95% confidence intervals computed over the unpredictability scores of individual users.

6.5 Results

In this section we present the results of our analysis. We first explain the choice of the source and target contexts and present high-level results. Afterwards, we show a feature-level analysis and discuss the effect of the population size.

6.5.1 Context Choice

In order to present data in a readable way, we select a subset of target and source contexts, following the most relevant attack vectors presented in the threat model. Of the 30 possible target contexts coming from our experimental design (see Table 6.1), we select five possible targets (one for each biometric) and a number of representative sources for each of them. The chosen contexts are the following:

- **Gait –** *Pocket phone-walk*: we select the pocket phone with walking activity as the target. We consider five different contexts: *Smartwatch-walk*, *Hand phone-walk*, *Smartwatch-run*, *Chest strap-run* and *Arm phone-run*.

- **Touch dynamics –** *Mid-end phone*: the middle-end phone represents the reasonable choice, as it allows us to measure the effect of using higher and lower quality devices as sources.

- **ECG –** *Authenticator-rest*: the *Authenticator* (Nymi band) uses ECG for authentication purposes and therefore represent an ideal target. All the remaining ECG sensors are considered as the sources, including the different measurements obtained with the medical monitor: *Lead I*, *Lead II*, *Lead III*, *Palm*.

- **Eye movements –** *Calibrated*: all the calibrated tasks are considered as target. We consider only uncalibrated data as the source and separate between uncalibrated data coming from the same task (e.g., *Uncalibrated-reading* to *Calibrated-reading*) and uncalibrated data coming from different tasks (e.g., *Uncalibrated-writing* to *Calibrated-reading*).

- **Mouse –** *Mouse*: we select *Mouse* as the target and will use *Trackpad* as the source.

Hereafter, results will refer to these target contexts.

6.5.2 Biometrics Overview

In Table 6.3 we report the resulting RMI weighted scores for each target and source context considered in Section 6.5.1. The first rows report the aggregated results over the sources: *average*, *minimum* and *all* weighted score (see Section 6.4.3). In Table 6.3, we can see that biometrics rank differently in terms of unpredictability. The table shows that ECG and gait are on average more resilient to the cross-context transformation, in both the same session and cross session scenarios. Gait in particular is very resilient to cross-context attacks, with an unpredictability score two times higher compared to touch dynamics, eye and mouse movements. This means that the different placement of the sensors provide poor information about the gait signal as measured in other contexts. Comparatively low results are obtained for eye movements, touch dynamics and mouse. Most of these biometric features are easily and consistently mapped across source contexts (see the discussion in Subsection 6.5.3). For the eye movements biometric there are also differences depending on the respective source and target task. Naturally, intra-task mappings produce the lowest unpredictability score (as the only difference is the lack of calibration for the source task), while cross-task mappings perform particularly poorly for some combinations. The results show that an attacker could gain a significant advantage if they are able to choose the source task freely.

Comparing the average, minimum and all score we can see that: (i) by selecting the appropriate source context the adversary can expect an improvement of 10% on average, that is, from *average* to *minimum* score (consistent across same and cross session scenarios); (ii) by combining information from several sources the adversary might obtain a further improvement up to 15% (again consistent across same and cross session scenarios), that is, from *minimum* to *all* score. This means that it might be worthwhile for an adversary to obtain biometric information over a higher number of sources and selectively choose to map individual features from whichever source provides the lowest unpredictability score for that feature.

The results show that same-session scores are lower compared to cross-session scores for ECG and eye movements in particular. As a result, an attack would appear to be more likely to succeed if very recent data (as in the same-session experiment) is used.

However, the authentication system itself has to cope with the (lack of) time stability which causes this difference. Most likely, this will be achieved through either periodic retraining or continuous template updating. While template updating will make false rejects as a result of increasing time distance less likely, it will also enable the attacker to use older data for the attack.

6.5.3 Feature Analysis

In order to understand to what extent individual features contribute to the overall score, we analyse them separately. We report boxplots for the raw (non-weighted) cross-context feature unpredictability scores. Each box shows the unpredictability score for a single feature from the source context to the target context. Features are ordered by decreasing RMI on the x-axis and the RMI value is reported for each feature. For conciseness, for each biometric, we only show a couple of meaningful sources and present just the top-ten RMI-ranked features, as these are the ones that contribute the most to the weighted score.

ECG We notice that the type of sensor used as the source has a significant impact on the weighted unpredictability score, similar to the results in Chapter 5. In Figure 6.2 we can see how *Mobile* consistently outperforms *Lead III* for each feature. This can be explained by closer similarity of the ECG signal when measured at the extremity of the subject's arms (true for *Lead I*, *Mobile* and the target *Authenticator*) compared to for example the *Lead III* measurements, which measures voltage potential between the left arm and left leg. The differences in predictability for different sources shown in Table 6.3 and Figure 6.2 highlight that ECG-based authentication might still be secure if the adversary steals ECG data from dissimilar contexts, but becomes less secure the easier it is to obtain data from similar contexts. Hand-based measurements are convenient and common (as shown by the popularity of e-health devices), this highlights the danger of using the same type of measurement for authentication.

Eye movements Figure 6.3 shows how most eye movements features are highly predictable, both pupil-based and speed- or acceleration-based ones. The boxplot additionally shows how *Intra task* consistently provides relatively lower unpredictability than *Cross task*, which show that each task produces feature changes that depend on the

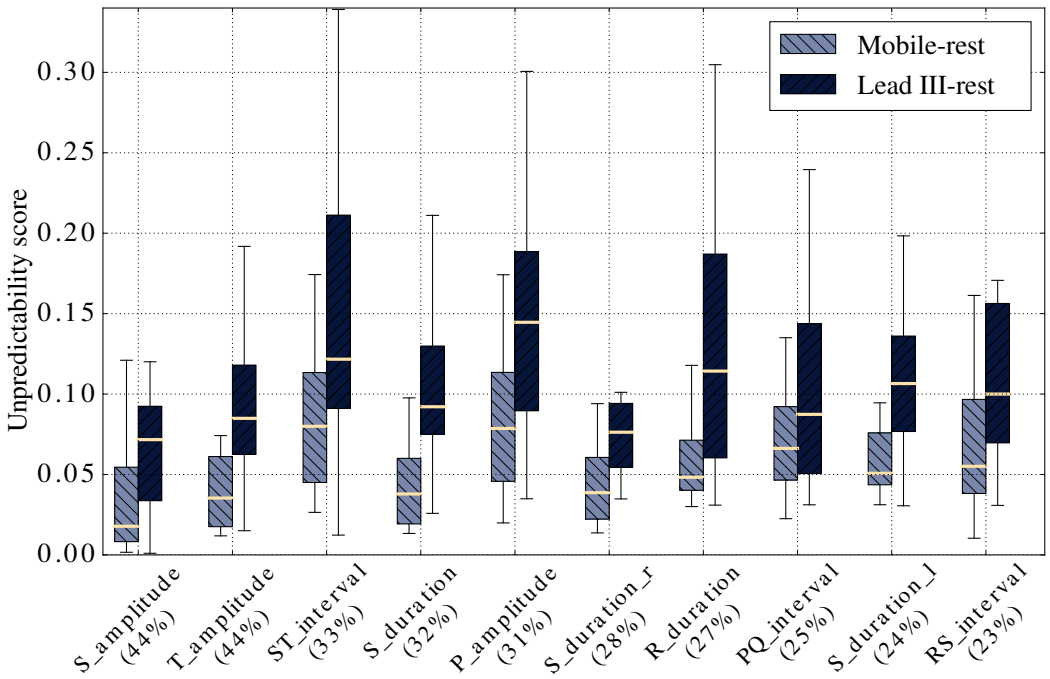


Figure 6.2: ECG features

user. Our threat model considers the case of the victim using a compromised machine with a covert eye tracker (see Section 6.2). The results show that if the attacker can choose the task on this machine freely (i.e., close to that on the authentication machine), he will obtain more useful data.

Touch dynamics In Figure 6.4 we can see that *High-end* phones provide slightly lower unpredictability scores compared to *Low-end* phones. The low result of *stroke_duration* shows that the feature is easily predictable across devices. This is intuitively explained with users adjusting the length of their swipes to the size of the touchscreen. In a feature selection scenario, a system designer might reasonably decide to drop *stroke_duration* from the feature-set. In fact, even if the feature has a decent distinctiveness, it is extremely predictable compared to other similarly distinctive features. Overall, it is evident that the lower-end phone is a less useful source of biometric information. This is mainly due to less precise sensors (i.e., lower sampling rate and resolution), which particularly affects acceleration features (low touchscreen sampling rate) and area covered (low resolution). Conversely, this shows that high-quality sensors can not just improve the baseline error rate, but also reduce the susceptibility of the device to active attacks.

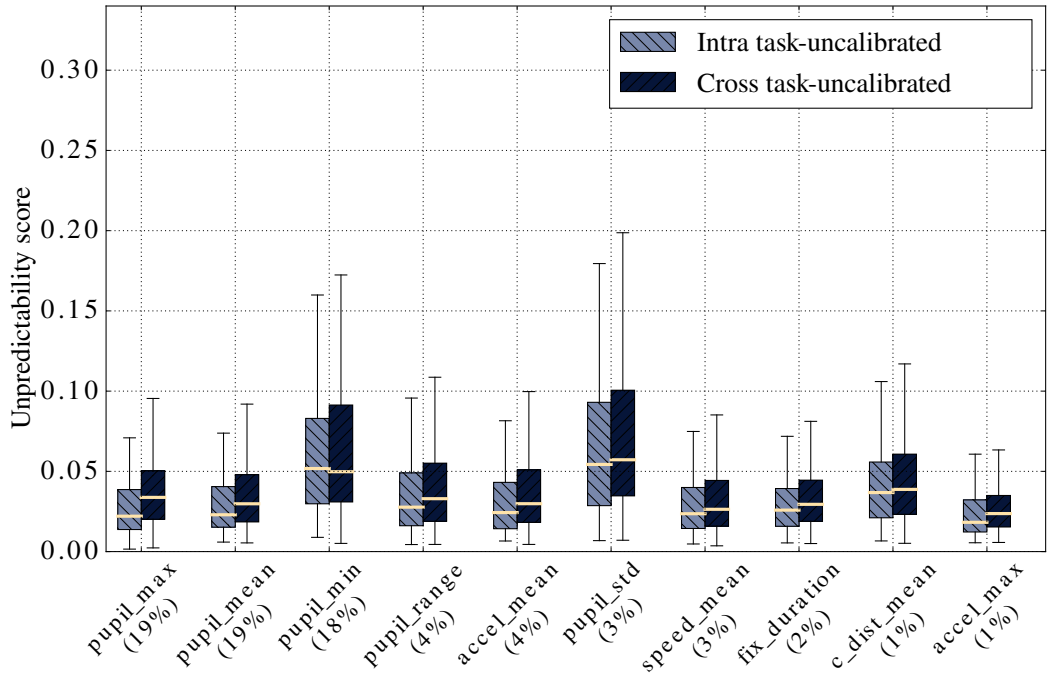


Figure 6.3: Eye movement features

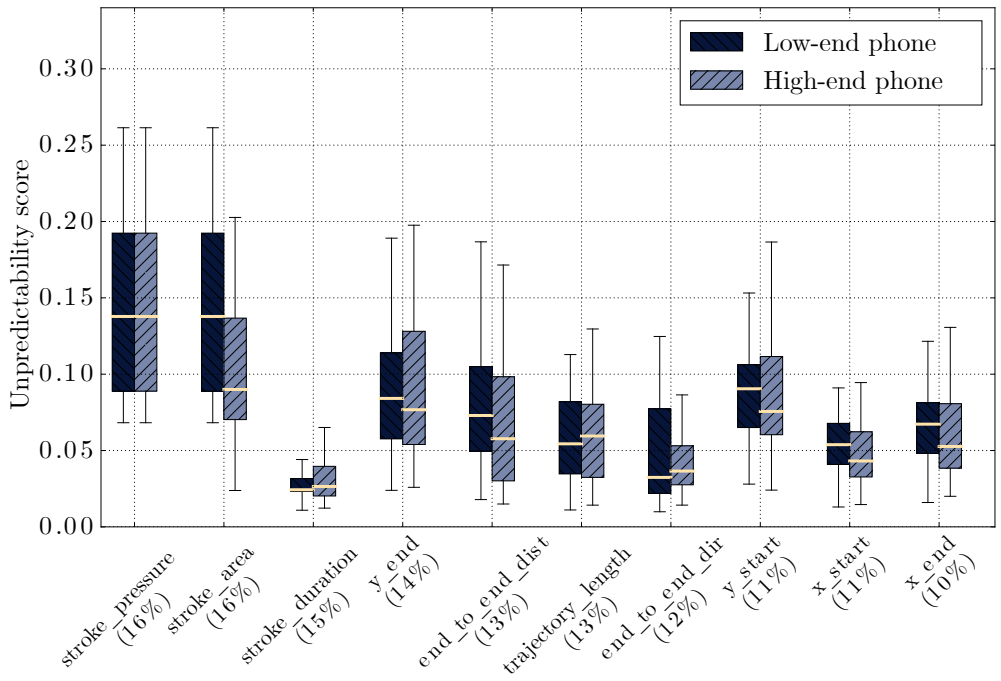


Figure 6.4: Touch features

Mouse movements Figure 6.5 reports on average low unpredictability results for most mouse movements features. Curvature-based features in particular seem highly

predictable, while not carrying significant distinctiveness (they might be dropped in a security-critical scenario). However, the plot shows a high mean and standard deviation for *click_duration*. This is due to the the trackpad API returning a coarse-grained click information, less sensitive than that returned by the mouse. Conversely, if source and target were switched, this feature would be very easy to predict as the set of valid target values would be small. This example highlights that more accurate sensors with higher resolution can thwart attacks coming from lower-quality data sources.

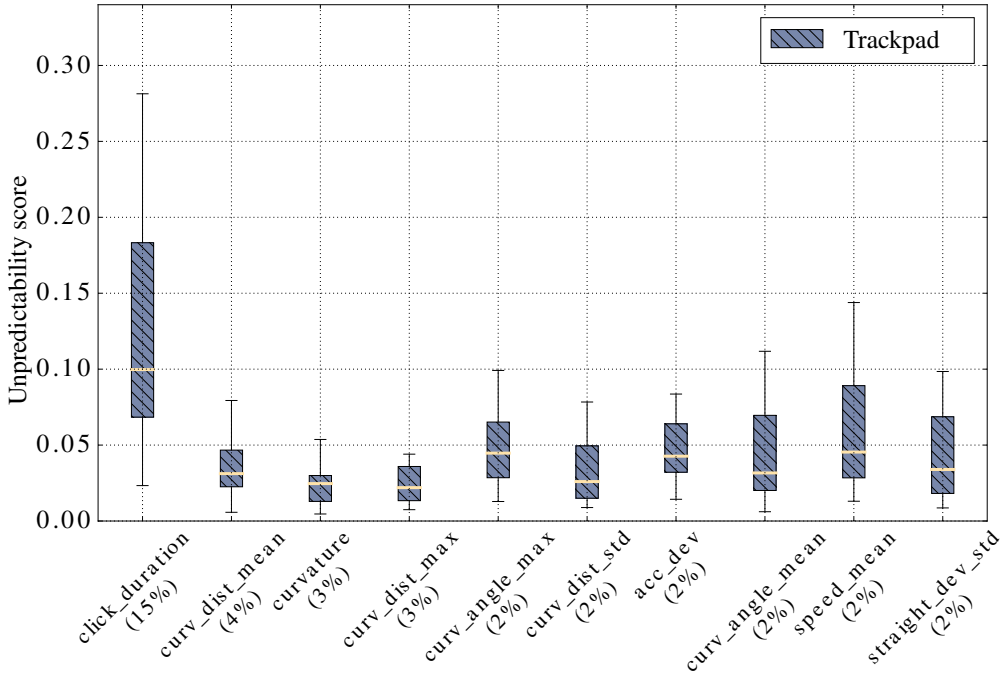
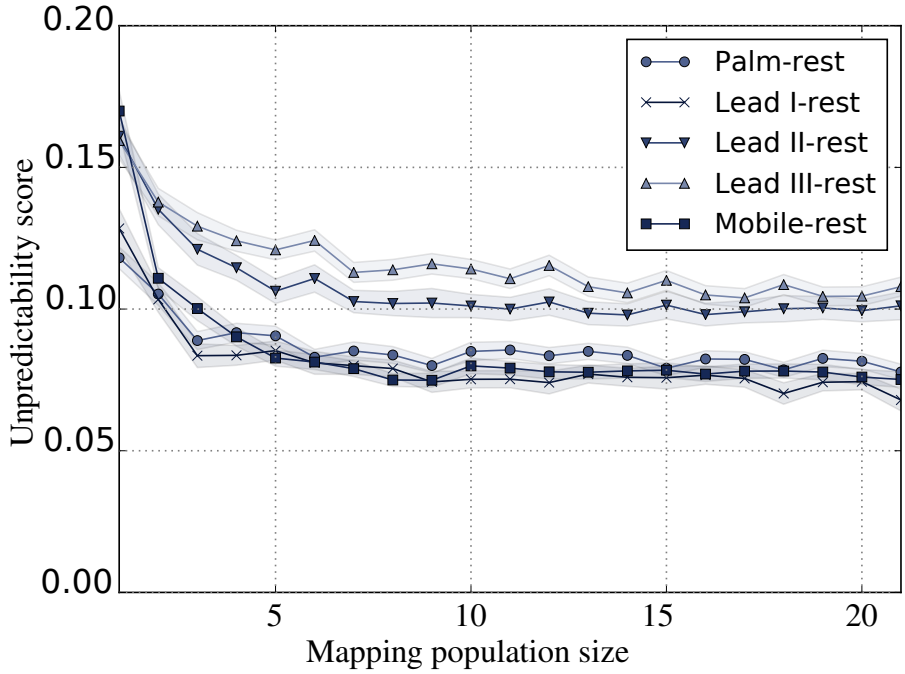


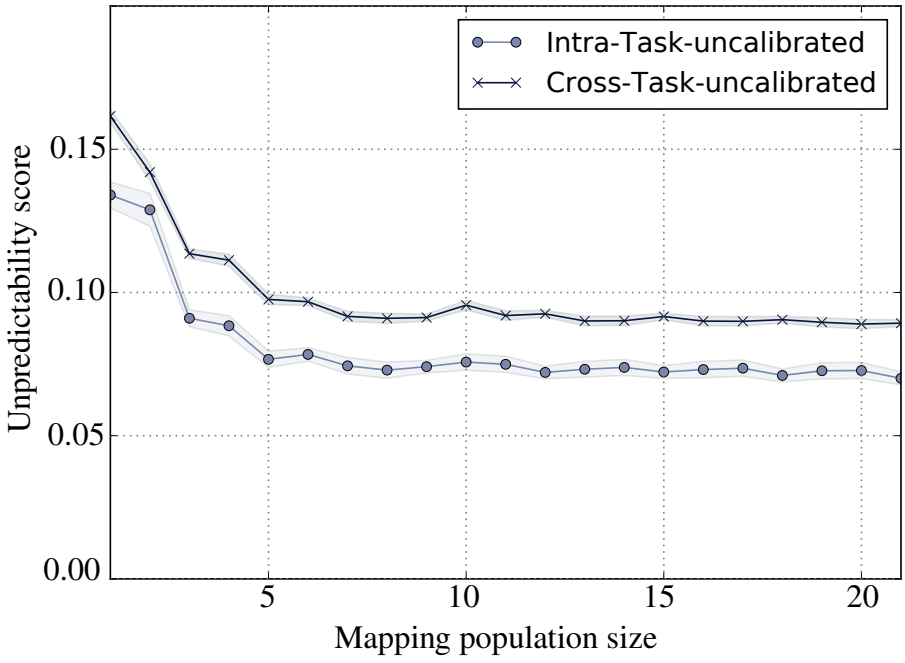
Figure 6.5: Mouse features

6.5.4 Population Size Analysis

Collecting a large number of (pairs of) biometric samples to train the cross-context mapping is a considerable effort. While it is possible to use publicly available datasets (see Section 6.2), this data may not always be available for the victim's target context (e.g., when the victim uses an unusual device). As such, it is important to know how large (in terms of number of users) the population has to be to produce acceptable results. Figure 6.6 shows the relationship between the number of users in the population and the average score of the resulting cross-context mapping. All biometrics show an initial



(a) ECG



(b) Eyes

Figure 6.6: Effect of population size on unpredictability scores. Using five subjects in the population greatly reduces the scores, while including more than 10 only yields a negligible improvement.

sharp drop in the score and exhibit diminishing returns beyond a population of size of 10. These results show that most of the cross-context mapping's predictive power can be achieved with a relatively small population. In addition, Figure 6.6 suggests that the sample size of our study (30 participants) is large enough to demonstrate differences between individual features, contexts and biometrics.

6.6 Conclusion

In this chapter, we have presented an analytical framework that allows us to measure the unpredictability of biometric features across different contexts. We define the notion of an *unpredictability* score, which can be calculated both for individual features and complete biometrics. The score provides fine-grained information about the resilience of biometric systems against cross-context attacks and can be used to: (i) compare biometric systems, (ii) identify vulnerable target contexts and for (iii) the selection and engineering of features. The framework is based on computing a mapping between a source and target context, where the mapping is derived from population data.

Our results demonstrate that the five biometrics evaluated in this paper show different degrees of resilience to cross-context attacks. In particular, we showed that ECG and gait are up to twice as unpredictable across contexts compared to touch dynamics, mouse and eye movements. Our analysis highlights particularly predictable features and suggests that some of can be reasonably dropped from the feature-set to achieve greater security against this attack. Furthermore, our data suggests that improving the quality of the biometric sensor improves the resilience of the authentication system. The fact that some contexts are more useful than others for the prediction shows that the sources of biometric information potentially available to an attacker need to be an integral part of any biometric threat model.

Es gibt in der Welt einen einzigen Weg, auf welchem niemand gehen kann, außer dir: wohin er führt? Frage nicht, gehe ihn.

There is one single path in the world which no one but you can tread. Do not ask, take it.

— Friedrich Nietzsche

7 Summary and Future Work

Contents

7.1 Summary of Results

This work aims to take a systematic look at the security that can be provided by using behavioural biometrics for continuous authentication. We first developed a prototype authentication system that makes use of distinctive eye movement behaviour. This work lays the foundation for our further work on evaluation methodologies and security analysis.

We highlight that commonly used evaluation methodologies do not always give an accurate measure of a system's real-world performance. Researchers almost universally report average error rates over the entire sample size, without accounting for the distribution of errors. This is problematic, as systematic false negatives (i.e., consistently undetected attackers) are far more severe in a continuous authentication system. Another common problem we highlight is that researchers often perform unrealistic training data selection. During real-world operation, the training phase has to be completed in

its entirety before any unknown samples can be classified. This is not reflected in the commonly used practice of cross-validation for training data selection.

Behavioural biometrics are often cited as being less easily observed than their physiological counterparts (e.g., fingerprints), which makes them less susceptible to active attacks. In this thesis we make the case that the type of behavioural modality used for authentication (e.g., gait, ECG and others) also occurs on a daily basis outside the authentication context. We have demonstrated this property by developing an attack against ECG biometrics. Our results show that data gathered from medical devices or fitness trackers can be adapted to impersonate a user. While this attack still requires substantial additional effort compared to photographing fingerprints, it shows that behavioural biometrics are not immune to active attacks.

We further generalise this cross-device attack to a cross-context attack. Based on a novel mapping function methodology we derive a per-feature unpredictability score. This allows us to judge how easily data from a number of accessible sources can be adapted for imitation attacks. We believe this analysis to be a crucial step in more accurately judge the security of behavioural biometrics against sophisticated adversaries. In addition, having an objective measure of resilience against cross-device attacks is of paramount importance when improving features, classifiers and hardware.

7.2 Future Work

Based on the insights gained from this thesis we have identified a number of future research directions to strengthen the security of authentication systems bases on behavioural biometrics.

Secure feature design. Our data shows that some features are inherently hard to predict across contexts. However, it is still unclear what exactly causes this unpredictability. Going beyond the obvious solution of selecting already resilient features, we believe that features can be specifically engineered to be less affected by cross-context attacks.

Creating unpredictability through hardware filters. Chapter 5 has demonstrated the effect that different hardware has on biometric features. The general intuition is that if the source and target device are sufficiently similar, then features will be easier to predict.

The problematic result is that using commodity hardware for biometric recognition (one of the great strengths of behavioural biometrics) makes ”useful“ biometric data available much easier. One approach that avoids using only purpose-built authentication hardware would be to include randomised filters into existing hardware. This concept is similar to Physically Unclonable Functions (PUFs) and would introduce a device-specific signature into the biometric template. As such, the mapping function approach presented in this thesis would have to be specific to each device, as the feature distribution is unique for this target.

7.3 Final Conclusions

Over the past years, we have seen an increasing trend to biometric authentication, ranging from consumer electronics to border controls. Most scenarios employ physiological biometrics, mostly due to their superior uniqueness. Nevertheless, behavioural biometrics have a number of key advantages, including unobtrusive measurements and the potential for continuous authentication.

In this work, we have demonstrated that these biometrics (including the eye movement modality presented in Chapter 3) can be a powerful tool. As with any security system, securing them against powerful adversaries remains one of the biggest challenges. According to Bruce Schneier ”[a]gainst the average user, anything works; there’s no need for complex security software. Against the skilled attacker, on the other hand, nothing works”. We believe that the wide availability of biometric data is the biggest challenge in designing actually secure biometric recognition. As wearable devices loaded with a plethora of sensors are getting more and more common, attempting to keep behavioural data secure may appear to be a losing battle. Nevertheless, the relative difficulty in capturing behavioural data and their strong dependency on measurement devices and environments may be the greatest strength of behavioural biometrics. By combining strong biometric features with increasingly complex liveness detection, biometric authentication will be able to deter all but the most powerful adversaries.

The first kind of intellectual and artistic personality belongs to the hedgehogs, the second to the foxes ...

— Sir Isaiah Berlin

References

[1] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. "Measuring password guessability for an entire university". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*. 2013, pp. 173–186.

[2] Elizabeth Stobert and Robert Biddle. "The password life cycle: User behaviour in managing passwords". In: *SOUPS '14: Proceedings of the Tenth Symposium On Usable Privacy and Security*. 2014, pp. 243–255.

[3] Alex Hern. *Hacker fakes German minister's fingerprints using photos of her hands*. 2014.

[4] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. "Preventing Lunchtime Attacks: Fighting Insider Threats With Eye Movement Biometrics". In: *Proceedings 2015 Network and Distributed System Security Symposium*. 2015.

[5] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. "Looks Like Eve: Exposing Insider Threats Using Eye Movement Biometrics". In: *ACM Transactions on Privacy and Security* 19.1 (2016).

[6] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. "Evaluating Behavioral Biometrics for Continuous Authentication". In: *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security - ASIA CCS '17*. New York, New York, USA: ACM Press, 2017, pp. 386–399.

[7] Simon Eberz, Nicola Paoletti, Marc Roeschlin, Andrea Patani, Marta Kwiatkowska, and Ivan Martinovic. "Broken Hearted: How To Attack ECG Biometrics". In: *Proceedings 2017 Network and Distributed System Security Symposium*. 2017.

[8] Simon Eberz, Giulio Lovisotto, Andrea Patane, Marta Kwiatkowska, Vincent Lenders, and Ivan Martinovic. "When your fitness tracker betrays you: Quantifying the predictability of biometric features across contexts". In: *2018 IEEE Symposium on Security and Privacy (SP)*. 2018, pp. 740–756.

[9] R. Stockton Gaines, William Lisowski, S. James Press, and Norman Shapiro. "Authentication by keystroke timing some preliminary results". In: *No. RAND-R-2526-NSF. RAND CORP SANTA MONICA CA* (1980), pp. 1–51. arXiv: `arXiv:1011.1669v3`.

[10] Luciano Bello and Maximiliano Bertacchini. "Collection and publication of a fixed text keystroke dynamics dataset". In: *XVI Congreso Argentino de Ciencias de la Computacion*. 2010, pp. 822–831.

[11] Hai-Rong Lv and Wen-Yuan Wang. "Biologic verification based on pressure sensor keyboards and classifier fusion techniques". In: *Consumer Electronics, IEEE Transactions on* 52.3 (2006), pp. 1057–1063.

[12] Daniele Gunetti and Claudia Picardi. "Keystroke analysis of free text". In: *ACM Transactions on Information and System Security* 8.3 (2005), pp. 312–347.

[13] Daniele Gunetti, Claudia Picardi, and Giancarlo Ruffo. “Keystroke analysis of different languages: A case study”. In: *Advances in Intelligent Data Analysis VI* 2 (2005), pp. 133–144.

[14] Salil Partha Banerjee and Damon Woodard. “Biometric Authentication and Identification Using Keystroke Dynamics: A Survey”. In: *Journal of Pattern Recognition Research* 7.1 (2012), pp. 116–139.

[15] Hugo Gamboa and Ana Fred. “A behavioral biometric system based on human-computer interaction”. In: *Biometric Technology for Human Identification*. International Society for Optics and Photonics, 2004, pp. 381–392. arXiv: 1411.5179.

[16] Nan Zheng, Aaron Paloski, and Haining Wang. “An efficient user verification system via mouse movements”. In: *Proceedings of the 18th ACM conference on Computer and communications security - CCS ’11*. 2011, p. 139.

[17] Youssef Nakkabi, Issa Traore, and Ahmed Awad E. Ahmed. “Improving mouse dynamics biometric performance using variance reduction via extractors with separate features”. In: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans* 40.6 (2010), pp. 1345–1353.

[18] Zach Jorgensen and Ting Yu. “On mouse dynamics as a behavioral biometric for authentication”. In: *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security - ASIACCS ’11*. 2011, p. 476.

[19] Adam J Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M Smith. “Smudge Attacks on Smartphone Touch Screens”. In: *USENIX conference on Offensive technologies* (2010), pp. 1–7.

[20] Napa Sae-Bae, Kowsar Ahmed, Katherine Isbister, and Nasir Memon. “Biometric-rich Gestures: A Novel Approach to Authentication on Multi-touch Devices”. In: *SIGCHI Conference on Human Factors in Computing Systems* (2012), p. 977.

[21] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. “Touch me once and I know it’s you! Implicit authentication based on touch screen patterns”. In: *Chi 2012* (2012), pp. 987–996.

[22] Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. “Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication”. In: *IEEE Transactions on Information Forensics and Security* 8.1 (Jan. 2013), pp. 136–148.

[23] Georg Essl, Michael Rohs, and Sven Kratz. “Use the Force (or something) - Pressure and Pressure-Like Input for Mobile Music Performance”. In: *Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME 2010), Sydney, Australia* Nime (2010), pp. 182–185.

[24] T. Feng, J. Yang, Z. Yan, E. M. Tapia, and W. Shi. “TIPS: Context-aware Implicit User Identification Using Touch Screen in Uncontrolled Environments”. In: *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications* (2014), 9:1–9:6.

[25] Cheng Bo, Lan Zhang, and Xiang-Yang Li. “SilentSense: Silent User Identification via Dynamics of Touch and Movement Behavioral Biometrics”. In: *Proceedings of the 19th annual international conference on Mobile computing \& networking*. 2013, pp. 187–190. arXiv: 1309.0073.

[26] Maricor Soriano, Alessandra Araullo, and Caesar Saloma. “Curve spreads - A biometric from front-view gait video”. In: *Pattern Recognition Letters* 25.14 (2004), pp. 1595–1602.

[27] Michela Goffredo, Imed Bouchrika, John N. Carter, and Mark S. Nixon. “Self-calibrating view-invariant gait biometrics”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 40.4 (2010), pp. 997–1008.

[28] Jani Mäntyjärvi, Mikko Lindholm, Elena Vildjiounaite, Satu-Marja Mäkelä, and Heikki Ailisto. “IDENTIFYING USERS OF PORTABLE DEVICES FROM GAIT PATTERN WITH VTT”. In: *Proc. IEEE ICASSP* (2005), pp. 973–976.

[29] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. “Biometric gait authentication using accelerometer sensor”. In: *Journal of Computers (Finland)* 1.7 (2006), pp. 51–59.

[30] Heikki J. Ailisto, Mikko Lindholm, Jani Mantyjarvi, Elena Vildjiounaite, and Satu-Marja Makela. “Identifying people from gait pattern with accelerometers”. In: *Biometric Technology for Human Identification II*. 2005, pp. 7–15.

[31] Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. “Reducing shoulder-surfing by using gaze-based password entry”. In: *Proceedings of the 3rd symposium on Usable privacy and security - SOUPS ’07*. 2007, p. 13.

[32] Alexander De Luca, Martin Denzel, and Heinrich Hussmann. “Look into my eyes!: can you guess my password?” In: *SOUPS ’09: Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), pp. 1–12.

[33] Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and P Franti. “Eye-movements as a biometric”. In: *Image Analysis, Proceedings* 3540 (2005), pp. 780–789. arXiv: arXiv:0906.3353v1.

[34] Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. “Towards task-independent person authentication using eye movement signals”. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA ’10*. 2010, p. 187.

[35] Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. “GANT: Gaze analysis technique for human identification”. In: *Pattern Recognition* 48.4 (2015), pp. 1023–1034.

[36] Zhen Liang, Fei Tan, and Zheru Chi. “Video-based biometric identification using eye tracking technique”. In: *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*. 2012, pp. 728–733.

[37] Ivo Sluganovic, Marc Roeschlin, Kasper B. Rasmussen, and Ivan Martinovic. “Using Reflexive Eye Movements for Fast Challenge-Response Authentication”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS’16*. New York, New York, USA: ACM Press, 2016, pp. 1056–1067.

[38] Jr. David C. Sabiston. “Heart Disease: A Textbook of Cardiovascular Medicine”. In: *Annals of Surgery* 194.1 (1981), p. 116.

[39] Zhaoyang Zhang, Honggang Wang, Athanasios V Vasilakos, and Hua Fang. “ECG-Cryptography and Authentication in Body Area Networks”. In: *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE* 16.6 (2012), pp. 1070–1078.

[40] L Biel, O Pettersson, L Philipson, and P. Wide. “ECG analysis: A new approach in human identification”. In: *IEEE Transactions on Instrumentation and Measurement* 50.3 (2001), pp. 808–812.

[41] Steven A. Israel, John M. Irvine, Andrew Cheng, Mark D. Wiederhold, and Brenda K. Wiederhold. “ECG to identify individuals”. In: *Pattern Recognition* 38.1 (2005), pp. 133–142.

[42] A. D. C. Chan, M. H. Hamdy, A. Badre, and V. Badee. “Wavelet distance measure for preson identification using electrocardiograms”. In: *IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT* 57.2 (2008), pp. 248–253.

[43] Fahim Sufi, Ibrahim Khalil, and Jiankun Hu. “ECG-Based Authentication”. In: *Handbook of Information and Communication Security* (2010), pp. 309–331. arXiv: arXiv:1011.1669v3.

[44] Abdul Serwadda and Vir V. Phoha. “Examining a Large Keystroke Biometrics Dataset for Statistical-Attack Openings”. In: *ACM Transactions on Information and System Security* 16.2 (2013), pp. 1–30.

[45] Chee Meng Tey, Payas Gupta, and Debin Gao. “I can be You: Questioning the use of Keystroke Dynamics as Biometrics.” In: *20th Annual Network and Distributed System Security Symposium - NDSS ’13* (2013), pp. 1 –16.

[46] Jonathan Ness. “Presentation Attack and Detection in Keystroke Dynamics”. PhD thesis. NTNU, 2017.

[47] Nan Zheng, Kun Bai, Hai Huang, and Haining Wang. “You are how you touch: User verification on smartphones via tapping behaviors”. In: *Proceedings - International Conference on Network Protocols, ICNP*. 2014, pp. 221–232.

[48] Hassan Khan, Urs Hengartner, and Daniel Vogel. “Targeted Mimicry Attacks on Touch Input Based Implicit Authentication Schemes”. In: *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys ’16*. 2016, pp. 387–398.

[49] Abdul Serwadda and Vir V. Phoha. “When kids’ toys breach mobile phone security”. In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS ’13*. 2013, pp. 599–610.

[50] Abdul Serwadda, Vir V. Phoha, Zibo Wang, Rajesh Kumar, and Diksha Shukla. “Toward Robotic Robbery on the Touch Screen”. In: *ACM Transactions on Information and System Security* 18.4 (2016), pp. 1–25.

[51] Sujit Poudel, Abdul Serwadda, and Vir V. Phoha. “On humanoid robots imitating human touch gestures on the smart phone”. In: *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems, BTAS 2015*. 2015.

[52] Davrondzhon Gafurov, Einar Snekkenes, and Patrick Bours. “Spoof attacks on gait authentication system”. In: *IEEE Transactions on Information Forensics and Security* 2.3 (2007), pp. 491–502.

[53] Rajesh Kumar, Vir V. Phoha, and Anshumali Jain. “Treadmill Assisted Imitation Attack on Gait-based Authentication Systems”. In: *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–7.

[54] Isaac Griswold-Steiner, Zakery Fyke, Mushfique Ahmed, and Abdul Serwadda. “Morph-a-Dope: Using Pupil Manipulation to Spoof Eye Movement Biometrics”. In: Nov. 2018.

[55] Andrew T. Duchowski. *Eye tracking methodology: Theory and practice: Third edition*. Cham: Springer International Publishing, 2017, pp. 1–366. arXiv: arXiv:1011.1669v3.

[56] Barbara Cassin and Sheila Solomon. *Dictionary of eye terminology*. Triad Publishing Company, Gainesville, Florida, 1984.

[57] Susana Martinez-Conde, Stephen L. Macknik, Xoana G. Troncoso, and Thomas A. Dyar. "Microsaccades counteract visual fading during fixation". In: *Neuron* 49.2 (2006), pp. 297–305.

[58] R. V. Abadi and E. Gowen. "Characteristics of saccadic intrusions". In: *Vision Research* 44.23 (2004), pp. 2675–2690.

[59] A. Jones, R. P. Friedland, B. Koss, L. Stark, and B. A. Thompkins-Ober. "Saccadic intrusions in Alzheimer-type dementia". In: *Journal of Neurology* 229.3 (1983), pp. 189–194.

[60] Brett A Clementz, John A Sweeney, Michae Hirt, and Gretchen Haas. "Pursuit gain and saccadic intrusions in first-degree relatives of probands with schizophrenia." In: *Journal of abnormal psychology* 99.4 (1990), p. 327.

[61] Keith Rayner, Caren M. Rotello, Andrew J. Stewart, Jessica Keir, and Susan A. Duffy. "Integrating text and pictorial information: Eye movements when looking at print advertisements". In: *Journal of Experimental Psychology: Applied* 7.3 (2001), pp. 219–226.

[62] Michel Wedel and Rik Pieters. "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings". In: *Marketing Science* 19.4 (2000), pp. 297–312.

[63] Robert J K Jacob. "Eye Tracking in Advanced Interface Design". In: *Virtual Environments and Advanced Interface Design* (1995), pp. 258–290.

[64] W. Leigh Ottati, Joseph C. Hickox, and Jeff Richter. "Eye Scan Patterns of Experienced and Novice Pilots during Visual Flight Rules (VFR) Navigation". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43.1 (Sept. 1999), pp. 66–70.

[65] David Tock and Ian Craw. "Tracking and measuring drivers eyes". In: *Real-time computer vision* (1995), pp. 71–89.

[66] Takehiro Ito, Shinji Mita, Kazuhiro Kozuka, Tomoaki Nakano, and Shin Yamamoto. "Driver blink measurement by the motion picture processing and its application to drowsiness detection". In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. Vol. 2002-Janua. 2002, pp. 168–173.

[67] Mandalapu Sarada Devi and Preeti R. Bajaj. "Driver Fatigue Detection Based on Eye Tracking". In: *2008 First International Conference on Emerging Trends in Engineering and Technology*. 2008, pp. 649–652.

[68] Colleen MacLachlan and Howard C. Howland. "Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years". In: *Ophthalmic and Physiological Optics* 22.3 (May 2002), pp. 175–182.

[69] Daniel Kahneman and Jackson Beatty. "Pupil diameter and load on memory". In: *Science* 154.3756 (1966), pp. 1583–1585.

[70] Sasitorn Taptagaporn and Susumu Saito. "How display polarity and lighting conditions affect the pupil size of VDT operators". In: *Ergonomics* 33.2 (Feb. 1990), pp. 201–208.

[71] Donald R. Jasinski, Jeffrey S. Pevnick, and John D. Griffith. “Human Pharmacology and Abuse Potential of the Analgesic Buprenorphine: A Potential Agent for Treating Narcotic Addiction”. In: *Archives of General Psychiatry* 35.4 (1978), pp. 501–516.

[72] B Winn, D Whitaker, D B Elliott, and N J Phillips. “Factors affecting light-adapted pupil size in normal human subjects”. In: *Investigative ophthalmology & visual science* 35.3 (1994), pp. 1132–1137.

[73] CMU. *CyberSecurity Watch Survey*. 2011.

[74] Michelle Keeney, Eileen Kowalski, Dawn Cappelli, Andrew Moore, Timothy Shimeall, and Stephanie Rogers. “Insider Threat Study: Computer System Sabotage in Critical Infrastructure S ectors”. In: *U.S. Secret Service and CERT Coordination Center/SEI* May (2005), pp. 1–44.

[75] Miltiadis Kandias, Alexios Mylonas, Nikos Virvilis, Marianthi Theoharidou, and Dimitris Gritzalis. “An Insider Threat Prediction Model”. In: 2010, pp. 26–37.

[76] Kenneth Holmqvist, M Nyström, and F Mulvey. “Eye tracker data quality: what it is and how to measure it”. In: *Proceedings of the symposium on eye tracking research and applications* 1.212 (2012), pp. 45–52.

[77] James Dougherty, Ron Kohavi, and Mehran Sahami. “Supervised and Unsupervised Discretization of Continuous Features”. In: *Machine Learning Proceedings 1995*. 1995, pp. 194–202. arXiv: 9809069v1 [arXiv:gr-qc].

[78] Euisun Choi and Chulhee Lee. “Feature extraction based on the Bhattacharyya distance for multimodal data”. In: *Geoscience and Remote Sensing Symposium, 2001. IGARSS ’01. IEEE 2001 International* 1.C (2001), 524–526 vol.1.

[79] Ziad M. Hafed and James J. Clark. “Microsaccades as an overt measure of covert attention shifts”. In: *Vision Research* 42.22 (2002), pp. 2533–2545.

[80] Kristina Herbst, Birgit Sander, Dan Milea, Henrik Lund-Andersen, and Aki Kawasaki. “Test-retest repeatability of the pupil light response to blue and red light stimuli in normal human eyes using a novel pupillometer”. In: *Frontiers in Neurology* FEB (2011).

[81] Ralf Engbert and Reinhold Kliegl. “Microsaccades keep the eyes’ balance during fixation”. In: *Psychological Science* 15.6 (June 2004), pp. 431–436.

[82] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds. “Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation”. In: *National Institut of Standards and Technology Gaithersburg* (1998), pp. 1–4.

[83] Patrick Bours and Soumik Mondal. “Performance evaluation of continuous authentication systems”. In: *IET Biometrics* 4.4 (Dec. 2015), pp. 220–226.

[84] Stefan Axelsson. “The base-rate fallacy and the difficulty of intrusion detection”. In: *ACM Transactions on Information and System Security* 3.3 (2000), pp. 186–205.

[85] Hui Xu, Yangfan Zhou, and Michael R Lyu. “Towards Continuous and Passive Authentication via Touch Biometrics: An Experimental Study on Smartphones”. In: *Symposium On Usable Privacy and Security, SOUPS*. 2014, pp. 187–198.

[86] Stefania Budulan, Elena Burceanu, Traian Rebedea, and Costin Chiru. “Continuous User Authentication Using Machine Learning on Touch Dynamics”. In: *NEURAL INFORMATION PROCESSING, PT I*. Vol. 9489. 2015, pp. 591–598.

[87] Hugo Gascon and Sebastian Uellenbeck. “Continuous Authentication on Mobile Devices by Analysis of Typing Motion Behavior.” In: *Sicherheit* October (2014), pp. 1–12.

[88] Xi Zhao, Tao Feng, and Weidong Shi. “Continuous mobile authentication using a novel Graphic Touch Gesture Feature”. In: *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013*. 2013.

[89] Ben Draffin, Jiang Zhu, and Joy Zhang. “KeySens : Passive User Authentication through Micro-behavior Modeling of Soft Keyboard Interaction”. In: *International Conference on Mobile Computing, Applications, and Services* 130 (2014), pp. 184–201.

[90] Daniel Buschek, Alexander De Luca, and Florian Alt. “Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI ’15*. 2015, pp. 1393–1402.

[91] Zhongmin Cai, Chao Shen, Miao Wang, Yunpeng Song, and Jialin Wang. “Mobile authentication through touch-behavior features”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8232 LNCS. 2013, pp. 386–393.

[92] P Saravanan, S Clarke, DHP Chau, H Zha Proceedings of the Second, and Undefined 2014. “Latentgesture: active user authentication through background touch analysis”. In: *Proceedings of the Second International Symposium of Chinese CHI*. 2014, pp. 110–113.

[93] Aditi Roy, Tzipora Halevi, and Nasir Memon. “An HMM-based multi-sensor approach for continuous mobile authentication”. In: *Proceedings - IEEE Military Communications Conference MILCOM*. Vol. 2015-Decem. 2015, pp. 1311–1316.

[94] Chao Shen, Yong Zhang, Zhongmin Cai, Tianwen Yu, and Xiaohong Guan. “Touch-interaction behavior for continuous user authentication on smartphones”. In: *Proceedings of 2015 International Conference on Biometrics, ICB 2015*. 2015, pp. 157–162.

[95] Kasper Bonne Rasmussen, Marc Roeschlin, Ivan Martinovic, and Gene Tsudik. “Authentication using pulse-response biometrics”. In: *The Network and Distributed System Security Symposium (NDSS)*. 2014.

[96] Mohammad O. Derawi, Claudia Nickely, Patrick Bours, and Christoph Busch. “Unobtrusive user-authentication on mobile phones using biometric gait recognition”. In: *Proceedings - 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2010*. 2010, pp. 306–311.

[97] Elena Vildjiounaite, Satu Marja Mäkelä, Mikko Lindholm, Reima Riihimäki, Vesa Kyllönen, Jani Mäntyjärvi, and Heikki Ailisto. “Unobtrusive multimodal biometrics for ensuring privacy and information security with personal devices”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 3968 LNCS. 2006, pp. 187–201.

[98] Rong Liu, Zhiguo Duan, Jiarizhong Zhou, and Ming Liu. “Identification of individual walking patterns using gait acceleration”. In: *2007 1st International Conference on Bioinformatics and Biomedical Engineering, ICBBE*. 2007, pp. 543–546.

[99] Maja Pusara and Carla E. Brodley. “User re-authentication via mouse movements”. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security - VizSEC/DMSEC ’04*. 2004, p. 1.

[100] Ahmed Awad E. Ahmed and Issa Traore. “A New Biometric Technology Based on Mouse Dynamics”. In: *IEEE Transactions on Dependable and Secure Computing* 4.3 (2007), pp. 165–179.

[101] Douglas A. Schulz. “Mouse curve biometrics”. In: *Biometrics Symposium, BCC 2006*. 2006.

[102] Soumik Mondal and Patrick Bours. “Continuous authentication using mouse dynamics”. In: *2013 International Conference of the Biometrics Special Interest Group (BIOSIG)* 2003 (2013), pp. 1–12.

[103] Kevin Allix, TFDA Bissyande, Jacques Klein, and Y Le Traon. “Machine Learning-Based Malware Detection for Android Applications : History Matters”. In: *University of Luxembourg, SnT* May (2014), p. 17.

[104] Kevin Allix, Tegawendé F. Bissyandé, Jacques Klein, and Yves Le Traon. “Are Your Training Datasets Yet Relevant?” In: *Engineering Secure Software and Systems* (2015), pp. 51–67.

[105] Agata Brajdic and Robert Harle. “Walk detection and step counting on unconstrained smartphones”. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp ’13*. 2013, p. 225.

[106] Adam Weiss, Anil Ramapanicker, Pranav Shah, Shinese Noble, and Larry Immohr. “Mouse movements biometric identification: A feasibility study mouse movement biometric system”. In: *Proceedings of StudentFaculty Research Day CSIS Pace University* (2007), pp. 1–8.

[107] Corrado Gini. “Variabilità e mutabilità”. In: *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi* (1912).

[108] Jaakko Malmivuo and Robert Plonsey. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Ed. by Oxford University Press. 1995. arXiv: 0402594v3 [arXiv:cond-mat].

[109] Alexandros Pantelopoulos and Nikolaos G Bourbakis. “A Survey onWearable Sensor-Based Systems for Health Monitoring and Prognosis”. In: *IEEE transactions on Systems Man, and Cybernetics, Applications and Reviews* 40.1 (2010), pp. 1–12.

[110] Antonio Fratini, Mario Sansone, Paolo Bifulco, and Mario Cesarelli. *Individual identification via electrocardiogram analysis*. Dec. 2015.

[111] F Agrafioti, J Gao, D Hatzinakos Biometrics, and Undefined 2011. *Heart biometrics: Theory, methods and applications*. InTech, 2011.

[112] Hugh Wimberly and Lorie M Liebrock. “Using Fingerprint Authentication to Reduce System Security: An Empirical Study”. In: *2011 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2011, pp. 32–46.

[113] R Sameni, M B Shamsollahi, C Jutten, and G D Clifford. “A nonlinear Bayesian filtering framework for ECG denoising”. In: *IEEE Trans Biomed Eng* 54.12 (2007), pp. 2172–2185.

[114] Benoit Barbot, Marta Kwiatkowska, Alexandru Mereacre, and Nicola Paoletti. “Estimation and verification of hybrid heart models for personalised medical and wearable devices”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9308. 2015, pp. 3–7.

[115] David E Goldberg. *Genetic algorithms*. Pearson Education India, 2006.

[116] Frank E. Grubbs. “Sample Criteria for Testing Outlying Observations”. In: *The Annals of Mathematical Statistics* 21.1 (1950), pp. 27–58.

[117] Patrick E Mcsharry, Gari D Clifford, Lionel Tarassenko, and Leonard A Smith. “A Dynamical Model for Generating Synthetic ElectrocardiogramElectrocardiogram Signals”. In: *Annual International Conference of the IEEE* 2 (2011), pp. 5686–5689.

[118] S. Zahra Fatemian, Foteini Agrafioti, and Dimitrios Hatzinakos. “HeartID: Cardiac biometric recognition”. In: *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2010*. 2010.

[119] Sophocles J Orfanidis. *Introduction to Signal Processing Theory*. Prentice Hall, 1995.

[120] Gari D Clifford. “ECG statistics, noise, artifacts, and missing data”. In: *Advanced Methods and Tools for ECG Data Analysis* (2006), pp. 55–99.

[121] Mario Barcena, Candid Wueest, and Hon Lau. “How safe is your quantified self?” In: *Symantec* (2014), pp. 1–38. arXiv: 0807.2023.

[122] Jakob Rieck. “Attacks on Fitness Trackers Revisited: A Case-Study of Unfit Firmware Security”. In: *arxiv.org* (2016). arXiv: 1604.03313.

[123] Josh Benaloh, Melissa Chase, Eric Horvitz, and Kristin Lauter. “Patient controlled encryption: ensuring privacy of electronic medical records”. In: *CCSW ’09 Proceedings of the 2009 ACM workshop on Cloud computing security* (2009), pp. 103–114.

[124] Dongseok Jang, Ranjit Jhala, Sorin Lerner, and Hovav Shacham. “An empirical study of privacy-violating information flows in JavaScript web applications”. In: *Proceedings of the 17th ACM conference on Computer and communications security - CCS ’10*. 2010, p. 270.

[125] Alexandros Kapravelos, Chris Grier, and Neha Chachra. “Hulk: eliciting malicious behavior in browser extensions”. In: *Proceedings of the 23rd USENIX Security Symposium* (2014), pp. 641–654.

[126] Jakob Hohl and Stanley Rush. “The complete heart-lead relationship in the einthoven triangle”. In: *The Bulletin of Mathematical Biophysics* 30.4 (1968), pp. 615–623.

[127] Brian C. Ross. “Mutual information between discrete and continuous data sets”. In: *PLoS ONE* 9.2 (Feb. 2014). Ed. by Daniele Marinazzo, e87357.