# A Performance Assessment Framework for Mobile Biometrics

Matthew Boakes

A Thesis Submitted to the University of Kent for the Degree of Doctor of Philosophy in the Subject of Electronic Engineering


School of Engineering

Institute of Cyber Security for Society (iCSS)

March 2022

Pages: 213

# *Abstract*

This project aims to develop and explore a robust framework for assessing biometric systems on mobile platforms, where data is often collected in non-constrained, potentially challenging environments. The framework enables the performance assessment given a particular platform, biometric modality, usage environment, user base and required security level.

The ubiquity of mobile devices such as smartphones and tablets has increased access to Internet-based services across various scenarios and environments. Citizens use mobile platforms for an ever-expanding set of services and interactions, often transferring personal information, and conducting financial transactions. Accurate identity authentication for physical access to the device and service is, therefore, critical to ensure the security of the individual, information, and transaction.

Biometrics provides an established alternative to conventional authentication methods. Mobile devices offer considerable opportunities to utilise biometric data from an enhanced range of sensors alongside temporal information on the use of the device itself. For example, cameras and dedicated fingerprint devices can capture front-line physiological biometric samples (already used for device log-on applications and payment authorisation schemes such as Apple Pay) alongside voice capture using conventional microphones.

Understanding the performance of these biometric modalities is critical to assessing suitability for deployment. Providing a robust performance and security assessment given a set of deployment variables is critical to ensure appropriate security and accuracy. Conventional biometrics testing is typically performed in controlled, constrained environments that fail to encapsulate mobile systems' daily (and developing) use.

This thesis aims to develop an understanding of biometric performance on mobile devices. The impact of different mobile platforms, and the range of environmental conditions in use, on biometrics' accuracy, usability, security, and utility is poorly understood. This project will also examine the application and performance of mobile biometrics when in motion.

# *Acknowledgements*

I wish to take this opportunity to express my sincere gratitude to everyone who has supported and guided me both professionally and emotionally throughout the entire process of my PhD. The COVID-19 pandemic caused rapid changes to many of our lives and research, but we managed to get through it together.

I want to take this opportunity to thank my two supervisors, Richard Guest and Farzin Deravi, for their continued guidance and support throughout the PhD process. Without their knowledge and wisdom, I likely would not have managed to make it anywhere near completing a PhD project.

My family, Mother Sally, Father Nick, Sister Emma, and Brother Stephen, and my girlfriend Ayda, for always being around for the good and the bad, for always believing in me and being proud no matter what the outcomes of life may be and reminding me to remember to ground myself and take the perspective of situations.

Finally, to all my “Big Brain” friends and colleagues who have come and gone over the years or are still around, without whom, I do not think I would have been able to complete this thesis. I am not used to having so much support from people who are willing and wishing me to succeed, so I would like to thank you all and wish you every success and joy in life, and for the many of you working towards a PhD good luck. Extra special shout-out to my housemates in the “PhD House” for keeping me fed and watered during those final weeks of thesis writing.

# Table of Contents

# List of Tables

# List of Figures

# 1 Introduction

## 1.1 Introduction

Biometric systems use automated methods to verify or identify an individual and have seen widespread deployment over the past two decades. Increasingly these technologies are being ubiquitously utilised on mobile platforms such as smartphones and tablets. In the 2020 Biometrics Institute Industry Survey [1], 82% of respondents agreed that standardised biometric testing is crucial to the industry's future. Respondents also said that digital identity (14%) was the most significant development last year. The respondents then followed this development with mobile identity (12%), privacy and ethical issues (11%) and biometrics capture via smartphone (10%).

Following up in the 2021 Biometrics Institute Industry Survey [2], respondents again said that digital identity topped the list of significant developers, increasing from 14% to 33% in the year. "A high 86% agreed that standardised biometric testing is crucial to the industry's future, with minimal disagreement that this was the case. An even higher proportion (94%) believed that testing is essential to understand an algorithm's performance and how risks are managed". "Virtually all [the] industry professionals (90+%) agreed that biometrics will be the key enabler for anchoring digital identity and that there will continue to be significant growth in mobile remote identity verification systems and remote onboarding technology".

The FIDO alliance also comments, stating that "the lack of an industry-defined program to validate performance claims has led to concerns over variances in the accuracy and reliability of these solutions" [3]. The work presented within this thesis aims to explore and address these concerns within a mobile context by exploring the existing area and constructing and analysing a potential performance framework concerning mobile biometric systems.

Section 1.2 briefly introduces biometrics and why they help solve problems associated with traditional authentication methods. Section 1.3 explores the core elements of biometric performance testing, although Chapter 2 aims to explore this in further detail. Section 1.4 provides the research motivations for the work presented within this thesis, and Section 1.5 explores some of the backgrounds towards moving from traditional static biometric systems to mobile biometric systems. Section 1.6 presents the research questions this thesis aims to explore and answer. Section 1.7 provides a table of the key definitions and acronyms and reader is likely to encounter in this thesis. Finally, Section 1.8 and Section 1.9 provide a breakdown of the chapters presented within this thesis and a summary of this chapter.

## 1.2 What Are Biometrics?

From a cyber security perspective, the term authentication validates whom users claim to be. There are three primary ways to achieve this, and they can be categorised as something we know, something we have or are [4]. Something we know includes elements such as passwords and pins. Something we have includes access (smart) cards and keys. Something we are where biometrics play their role in the authentication space.

The ISO and IEC standards have defined biometrics as "The automated recognition of individuals based on their biological and behavioural characteristics" [5]. The biometric characteristic is the "biological and behavioural characteristic of an individual from which distinguishing, repeatable biometric features can be extracted for biometric recognition". In this thesis, the biometric characteristics are usually referred to as modality, with specific consideration for Fingerprint, Face, Voice, and Iris. Section 1.7 provides a complete list of definitions. Unlike passwords and smartcards, biometric attributes cannot be lost or forgotten. Even in the case of identical twins, although a face recognition system will struggle to tell them apart, a fingerprint or iris system can still provide high identification accuracy.

Jain *et al.* [6] identified seven factors that determine the suitability of a physical or a behavioural modality in a biometric application. The factors identified will serve as an insight into what the modality and, in turn, the biometric systems should be capable of achieving and will be investigated further in Chapter 3, where the traditional concept of performance is explored and modernised for the introduced performance framework.

1. **Universality**: Every individual accessing the application should possess the trait.
2. **Uniqueness**: The given trait should be sufficiently different across individuals.
3. **Permanence**: The biometric trait of an individual should be sufficiently invariant over a period for the matching algorithm. A trait that changes significantly over time is not a valid biometric.
4. **Measurability**: It should be possible to acquire and digitise the biometric trait using suitable devices that do not cause undue inconvenience to the individual. Furthermore, the acquired raw data should be amenable to processing to extract representative feature sets.
5. **Performance**: The recognition accuracy and the resources required to achieve that accuracy should meet the constraints imposed by the application.
6. **Acceptability**: Individuals in the target population that will utilise the application should be willing to present their biometric traits to the system.
7. **Circumvention**: This refers to the ease with which the trait of an individual can be imitated using artefacts (e.g., fake fingers) in the case of physical traits, and mimicry, in the case of behavioural traits.

Figure 1.1 shows the general components of biometric systems as defined by ISO. The system is broken up into four subsystems: data capture, storage, comparison, and decision, which comprise the biometric system. The biometric system is split into three paths, forming the enrolment path where a biometric reference is captured and stored. A verification path is a 1:1 comparison between the stored reference and the presented biometric probe. Finally, the identification path is a 1:N comparison between the stored references and presented probe to find the primarily likely biometric candidate.

Biometric identification is the "process of searching against a biometric enrolment database to find and return the biometric reference identifier(s) attributable to a single individual". In contrast, biometric verification is defined as the "process of confirming a biometric claim through biometric comparison". Since mobile is often associated with a single user, the thesis mainly concerns biometric verifications. Although it is noted that some smartphone devices are capable of storing multiple references, however, the intention is that they should all belong to a single user.

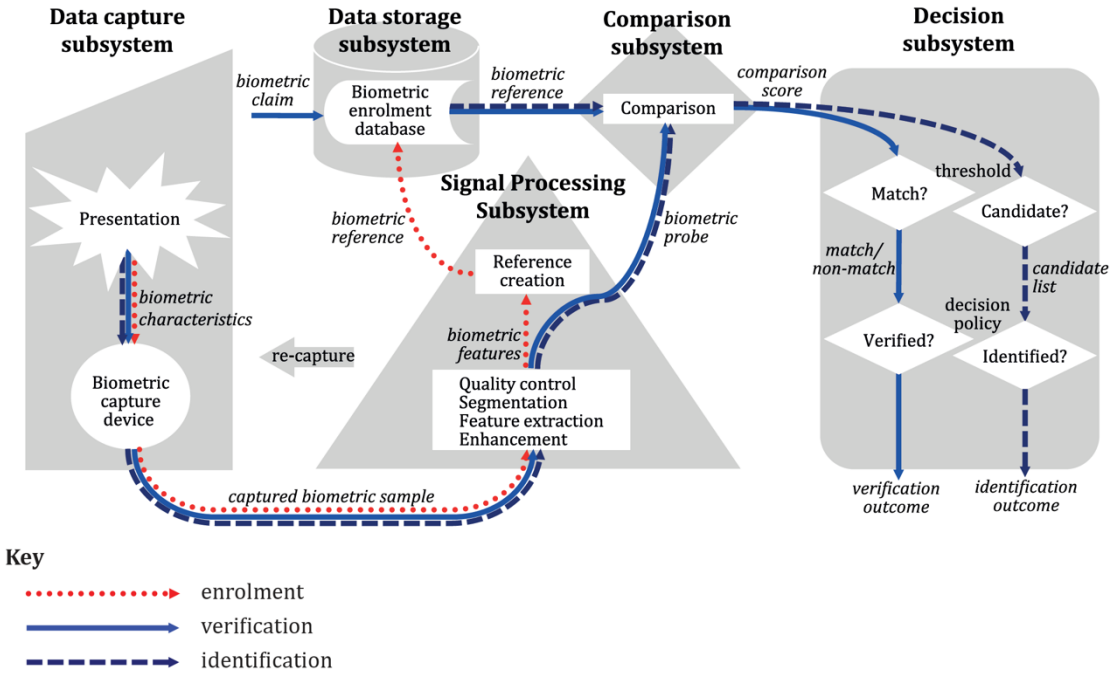



*Figure 1.1: Components of a General Biometric System [7]*

Biometrics aims to solve persistent issues around authentication and the usability and frustrations of users that result in poor security practices regarding authentication, mainly reusing passwords. The FIDO Alliance states some of the current problems with traditional authentication methods [8] from various sources, including:

- 284% growth in authentication credential loss in 2019
- 49% password-driven cart abandonment rate
- 55% of IT leaders reuse a single password

- 51% of passwords are reused across services
- 20-50% of helpdesk calls are for password resets
- 80-90% of e-commerce sites' attempted logins are compromised by stuffing
- $25 billion financial loss caused by online payment fraud by 2024 in Europe
- 1300 years collectively spent by humans each day entering passwords
- 18 million COVID-19-themed malware and phishing emails blocked per day by Google
- 7098 breaches in 2019, exposing 15.1 billion records

## 1.3 Biometric Performance Overview

There are three types of biometric performance testing:

- Technology Evaluation - A technology evaluation compares competing algorithms from a single technology.
- Scenario Evaluation - Scenario testing aims to determine the overall system performance in a prototype or simulated application.
- Operational Evaluation - Operational testing aims to determine the performance of a complete biometric system in a specific application environment with a specific target population.

"Manufacturers tend to test their products under optimal conditions. Unfortunately, most implementations will not reflect these conditions. Consequently, manufacturer claims for biometric system performance are unlikely to match what is seen in day-to-day use" [9]. Mansfield and Kelly [10] stated the difficulties in allowing meaningful performance comparisons between devices and listed five factors for this:

1. The performance of a biometric system can depend heavily on the type of application. For example, if the end-users are familiar with the system, willing to use it, and supervised, one would expect performance to be better than unsupervised, unwilling, untrained end-users.

2. Measures that apply to some biometric devices are meaningless to others. For example, in the case of behavioural biometric systems (such as signature or voice), ease of forgery is essential. However, there is no direct analogy for physiological biometric systems.

3. There are trade-offs between the various performance measures. By relaxing the acceptance criteria, the false rejection rate can be improved at the cost of increasing the false acceptance rate. Allowing multiple attempts can also decrease the false rejection and worsen the throughput rate. Devices can give the best possible performance for one application but will be less than optimal in different circumstances.

4. Manufacturers' quoted performance figures, obtained from in-house laboratory tests, are often not easy to relate to actual live performance.

5. Moreover, there are different interpretations of how to make the measurement, present the results, and what the results mean for many possible measurements.

Additionally, they also state that "the performance figures of most interest are those that tell how the whole system will operate in practice" and list some key questions that should be answered from biometric testing:

- For example, how accurately will the system verify a claimed identity?
- Will end-users find the system manageable, and will they be happy to use it?
- Will the system be fast enough in operation?
- Is the system secure enough to protect against attempted fraudulent use?
- In addition, there are the usual considerations regarding cost, interfaces, and capacity.

Mansfield and Kelly note that "on their own, the false rejection and acceptance rates are not very useful in predicting performance. For a complete view of performance, measurements of usability and security are also needed. The operational false rejection rate is dependent on several usability factors, while the operational false acceptance rate is affected by security issues".

## 1.4 Research Motivations

The world is experiencing a shift towards rapid growth in mobile technology, specifically geared towards the consistent use of smartphones to manage our daily lives. As a result, biometrics are now at the forefront of smartphone authentication, and it is relatively uncommon for a new smartphone to lack any biometric technology.

Currently, manufacturers and developers can claim their system's biometric performance and security without understanding how the evaluation was designed and performed. For example, Apple claimed the probability of someone else unlocking a phone with Face ID is 1 in 1,000,000 as opposed to Touch ID at 1 in 50,000 [11]. However, how useful is this information, and just how accurate is it?

Mobile devices increasingly incorporate biometrics as their primary authentication mechanism to access devices and services, including sensitive information such as financial and commercial data. However, mobile biometric systems are conducted mainly by the device provider, which causes issues comparing results between devices. Work is currently to standardise mobile biometric testing with various institutions, including the international organisation for Standardization (ISO), FIDO Alliance,

Android and academic researchers providing input into conducting and measuring mobile biometric performance.

However, issues remain with these approaches when considering testing commercial devices. First, this paper presents a summary of the existing approaches. Then, it introduces a new approach to testing mobile biometrics by amalgamating the existing ideas, approaches and inspirations and combining additional solutions and recommendations to mitigate the identified issues. Finally, we identify our approach as more universally applicable by comparing the ideas to an existing device and presenting a proposal for a universal testing framework for the performance evaluation of mobile biometrics.

## 1.5 Mobile Biometrics

The proliferation of mobile biometrics is a significant reason for updating the current understanding of 'performance' for biometric systems assessment. Throughout this thesis, the reference to mobile biometrics focuses on smartphone devices. Although the traditional uses of biometrics remain (such as border control systems and national ID cards), there has been widespread adoption into the mass market due to incorporating specific biometric sensors into current smartphone devices. "Consumer acceptance of biometrics is being driven largely by smartphone usage and adoption, which will only increase" [12].

Wojciechowska *et al.* [13] explored the trends and challenges in mobile biometrics and, noting work from Jillela and Ross [14], presented some critical points in a mobile scenario:

- Data privacy: the identification template is usually stored in the mobile phone memory. Thus, it must be encrypted carefully to protect the data from leaking.
- Ease of multi-biometric data acquisition: a smartphone is equipped with various sensors, which can help produce a multimodal system that is more reliable.
- Low operational cost: due to the reduced size and increasing computational power of processing units, the identification cost seems minimal.
- Market penetration: mobile phone popularity is enormous and still increasing. In highly developed countries, even children own a mobile phone.
- Multi-factor authentication: thanks to the specific construction of mobile phones, biometric identification may be combined with traditional kinds of protection like a password or geospatial data (GPS).
- Portability: the mobile phone is carried by its owner to different places and locations.
- Remote identification: according to the lower computational power, the mobile system may be implemented for verification purposes (1:1 matching) rather than identification (1:N matching). However, identification is possible when the biometric data is securely transferred to the server or the cloud.

Wojciechowska *et al.* provided some other trends and challenges in mobile biometrics, including template protection, the effect of ageing, vulnerability, computing performance, user acceptance, and databases. Although it was noted that classic biometric systems suffer from the same limitations, the risk increases with mobile devices because, in cases like smartphones, the user is always in possession of the device, meaning these challenges need to be overcome remotely. For example, asking a user to send a photo to a smartphone manufacturer to update the face reference is not a practical solution to overcome the ageing problem, so the manufacturer employs clever techniques to update the template automatically over time [15] or allow the user to enrol themselves again.

Mobile biometrics provides many novel exploration opportunities and a convenient and more secure authentication method. An example of the adoption of this technology is mobile banking to enable payments through services such as Apple Pay and Google Pay. Additionally, applications on mobile devices can use embedded biometric sensors to provide convenient access to services without the need to enter and remember a password to access personal information securely.

This adoption of biometrics within a mobile market requires a rethink on testing and verifying that the system *fits its purpose*. Moving from the more traditional fixed (static) system to a mobile (dynamic) one can increase the range of environments and scenarios in which they will operate. This increase has a knock-on effect on how users perceive and use the system. Figure 1.2 highlights a range of typical performance influences for a mobile context [16], [17]. For example, screen size and user posture can influence system usability and impact touch-based mobile biometrics [18].

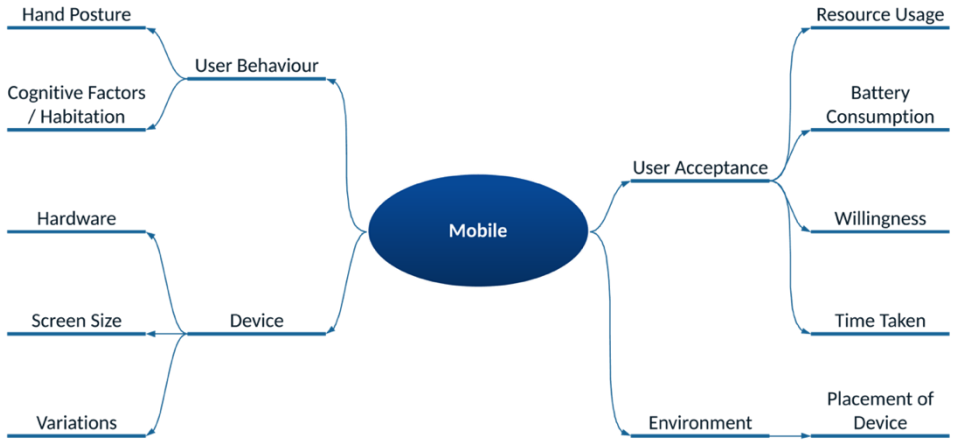



***Figure 1.2: Examples of Mobile Influences Expected to Impact Performance***

Although several studies have begun to explore these factors [17], [19], there are still limited resources on how the performance of mobile biometrics changes in various scenarios and environments, thus requiring further exploration, including the definition of a suitable assessment methodology. This work draws on existing conventional systems' evidence to establish knowledge deficits.

## 1.6 Research Questions

A series of research questions were developed to help explore this field in greater depth and help to uncover the possibilities and limitations moving forward in developing a performance assessment framework. The thesis aims to formulate a framework wherein the performance of a modality (or combination of modalities) can be assessed and quantified on mobile platforms across a range of environments, usage tasks and scenarios. This thesis seeks to answer the fundamental question, "how good is a particular biometric modality?" There are three critical questions that the project aims to answer:

- How can mobile biometric performance be measured?
- How do environment and motion affect biometric performance and security?
- How can any performance deterioration be mitigated?

Some research questions were developed alongside the critical questions to formulate a background into the area of mobile biometrics, with an emphasis on smartphone devices, including:

- What is the current smartphone security (locking) habits among users?
- Do they primarily use the biometric authentication method available on their device?
- Is there an overwhelming preference towards a particular modality?

These first questions all centre around user behaviour and attitudes towards mobile biometrics and will allow for uncovering information about the acceptance of mobile biometrics. The intent is to uncover the current state of mobile biometrics (considering smartphones) and apply this information when testing the performance assessment framework.

- Is it possible to achieve reliable performance metrics with a commercial (off-the-shelf) device with limited knowledge of its internal workings?

One of the known limitations of commercial systems is limited access to the biometric system for sensible security reasons. However, this makes testing the biometric performance tricky for an outsider and means end-users must take the manufacturer at their word when they state biometric performance statistics. This work will look to see what information can be achieved when working with systems with limited access to biometric systems and will allow the performance framework to be built around the knowledge of the amount of access. No system is excluded from the testing framework, even if the amount of data available at the end is limited.

- How does the quality/performance of <modality> change when placed in diverse scenarios and Environments?

One aspect that requires investigation when creating the performance assessment framework is to explore if and how the performance metrics and quality scores of probes change when the system is utilised in a way that is likely to be less common for traditional static systems. For example, it is considered particularly necessary to explore Motion vs Stationary and Indoor vs Outdoor.

- What effect does usability have on the performance of biometrics on a mobile platform?

Combining the users' satisfaction and timings to complete the transaction and whether the authentication was successful will allow some exploration of usability on commercial devices.

- Is it possible to produce a definitive score or ranking regarding the device's performance?

Currently, manufacturers state biometric claims as "FRR 1% @ FAR 1 / 10,000", but what does this mean, and how was it calculated [20]. How useful is this to the consumers? Therefore, one thing that should be considered as part of the performance framework is the reporting and hopefully in such a way that it can be helpful and meaningful to the consumers.

## 1.7 Definitions and Acronyms

Table 1.1 lists the definitions and acronyms a reader will encounter across the thesis and serves as a reference point. The aim was to include the critical definitions for a reader who may be less familiar with biometrics. However, this thesis aims to comply with the biometrics definitions defined by ISO. A complete set of definitions can be obtained from the following ISO/IEC 2382-37:2017 Information technology — Vocabulary — Part 37: Biometrics [5]. The performance metrics mentioned in the table can be turned into metric rates indicating the proportion of a specified set of biometric enrolment transactions or comparison trials that resulted in the metric. These include failure-to-acquire rate (FTAR), failure-to-enrol rate (FTER), false match rate (FMR), and false non-match rate (FNMR).

*Table 1.1: Biometrics Vocabulary*

| Term | Definition [5] |
|---|---|
| ISO | International Organization for Standardization (https://www.iso.org/) |
| FIDO | FIDO Alliance (https://fidoalliance.org/) |
| Biometrics | The automated recognition of individuals based on their biological and behavioural characteristics. |
| Biometric Characteristic | An individual's biological and behavioural characteristics from which distinguishing, repeatable biometric features can be extracted for biometric recognition. |
| Biometric Feature | Numbers or labels are extracted from biometric samples and used for comparison. |
| Biometric Recognition | Automated recognition of individuals based on their biological and behavioural characteristics. |
| Biometric Reference | One or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used as the object of biometric comparison. |
| Biometric Probe | Biometric sample or biometric feature set input to an algorithm for biometric comparison to a biometric reference(s). |
| Comparison | Estimation, calculation or measurement of similarity or dissimilarity between biometric probe(s) and biometric reference(s). |
| Comparison Score | The numerical value (or set of values) resulting from a comparison. |
| Threshold | The numerical value (or set of values) at which a decision boundary exists. |
| Biometric Mated Comparison Trial | Comparison of a biometric probe and a biometric reference from the same biometric capture subject and the same biometric characteristic as part of a performance test. |
| Biometric Non-Mated Comparison Trial | Comparison of a biometric probe and a biometric reference from different biometric data subjects as part of a performance test. |
| Failure To Acquire (FTA) | Failure to accept for subsequent comparison the output of a biometric capture process, a biometric sample of the biometric characteristic of interest. |
| Failure To Capture (FTC) | Failure of the biometric capture process to produce a captured biometric sample of the biometric characteristic of interest. |
| Failure To Enrol (FTE) | Failure to create and store a biometric enrolment data record for an eligible biometric capture subject following a biometric enrolment policy. |
| False Match | Comparison decision of match for a biometric probe and a biometric reference from different biometric capture subjects. |
| False Non-Match | Comparison decision of "non-match" for a biometric probe and a biometric reference from the same biometric capture subject and of the same biometric characteristic. |
| Quality Score | The quantitative value of the fitness of a biometric sample to accomplish or fulfil the comparison decision. |
| Quality | A measure of the fitness of a biometric sample to accomplish or fulfil the biometric comparison decision. |

## 1.8 Thesis Structure

This thesis is divided into a total of eight chapters. The background of mobile biometric testing and reporting is presented in Chapter 2. It is the background (state-of-the-art) of current and best biometric

testing and reporting practices. Chapter 3 presents a scoping style review that results in the organisation of groups defining core factors that result in the critical explorative areas for a performance assessment framework. Chapter 4 then presents the performance assessment framework using the background from the previous chapters and existing research to propose a novel assessment suitable for mobile biometric systems to provide the 'fit for purpose' assurance required.

The first half of the thesis will detail all the theoretical work produced in designing the performance framework. The remaining chapters focus on proving and applying the performance framework using more practical means. Chapter 5 introduces the methodology and approach for experimental data collection, which explores the collection of a range of modalities using commercial smartphone devices and a survey to ascertain user behaviour towards biometrics on mobile devices. Chapter 6 explores the core factors and performance framework using the obtained data to inform the framework and analysis the various effects on the performance to strengthen the credibility of the performance framework.

Chapter 7 uses the information gathered from creating the performance framework to establish a potential novel adaptive threshold approach to mobile biometric authentication making practical use of the embedded sensors present within smartphone devices to help mitigate performance degradation on mobile devices. Finally, chapter 8 concludes the work present within this thesis and summarises the observations and results, addressing the potential impact of the work while providing considerations for future work.

## 1.9 Summary

This chapter has introduced the topic of biometrics and mobile biometric and provided justification and motivations behind the work presented in the remainder of this thesis. With the background highlighted here, the next chapter will explore the existing approaches and methodologies to biometric performance testing and reporting, emphasising mobile biometric performance testing developments. Taking inspiration from these entities will help form a mobile biometric performance framework.

# 2 Mobile Biometric Testing and Reporting

## 2.1 Introduction

This chapter will look at existing approaches and methodologies to biometric testing and reporting by exploring existing approaches to testing traditional and mobile systems and paying particular attention to approaches considering a mobile system. In doing so, the aim is to answer the question, what are the existing approaches to biometric performance testing and where possible? By doing this, the intention is to identify the recommended values and apply them to inform the development of the performance framework.

There are a couple of primary areas that will be considered when examining these approaches:

- Performance Metrics
- Procedure
- Sample Size and Test Crew
- Enrolment
- Verification/Identification Transactions
- Performance Rate Requirements
- Reporting

The chapter will take methodologies and examples from the ISO standards, the FIDO alliance, Microsoft (Windows Hello), Google (Android), Apple (iOS), and current research to form a cohesive overview of the current state-of-the-art for mobile biometric performance testing. It will also influence what is meant by performance for a (mobile) biometric system and how it is measured. To achieve this, existing recommendations in this area were surveyed, including how they approach different elements of biometric testing.

Section 2.2 will explore some research that discusses mobile biometric performance and evaluation. Section 2.3 will show the National Cyber Security Centre's guidance for measuring biometric performance. Sections 2.4 (Microsoft), 2.5 (Google), and 2.6 (Apple) will explore how three major tech companies present biometric performance and any specific testing requirements and guidance they provide. Section 2.7 and Section 2.8 will discuss the guidance and specification provided by ISO and FIDO for biometric testing and reporting. Finally, Section 2.9 and Section 2.10 will discuss and summarise the current trends for (mobile) biometric testing and reporting.

## 2.2 Research

Investigating and exploring mobile biometric testing is relatively new in the academic literature, with most current work focusing on a technology evaluation of data. Fernandez-Saavedra *et al.* [21] provided a detailed account of some of the issues surrounding testing smartphone devices. They applied ISO/IEC 19795-1:2006 [22] to mobile devices to analyse what would work and what would not in this context and identified special features and conditions of mobile devices:

- Changing Ambient Conditions
- Special Interaction of the users with the biometric system
- Restricted biometric functions
- Impossibility of obtaining the captured biometric sample
- The result of the authentication is a Pass/Fail decision

The authors have highlighted the significant issues in mobile biometric performance testing and provide a concluding recommendation to "analyse the biometric functions and methods provided by the mobile device to know the restrictions of the evaluations in advance". Fernandez-Saavedra *et al.* [21] provided input into a methodology for the environmental evaluation of biometric systems. It includes ambient factors known to affect specific biometric modalities, including Temperature, Humidity, Illumination, Noise and Pressure.

Buriro *et al.* [16] present some guidelines to researchers for evaluating smartphone user authentication methods. They acknowledge how "most publicly available frameworks did not discuss or explore any other evaluation criterion, usability and environment-related measures except the accuracy under zero-effort". They identify the potential issues of in-lab testing as not accurately reflecting the reality of the performance meaning "their baseline operations usually give a false sense of progress". Their guidance is firmly aimed towards academia and researchers; the proposed performance framework incorporates many of the guidelines Buriro et al. [16] proposed to help reduce the issues raised in existing methodologies. Figure 2.1 showcases a mind map documenting the guidelines presented by Buriro *et al.* with a couple of additions added in red.

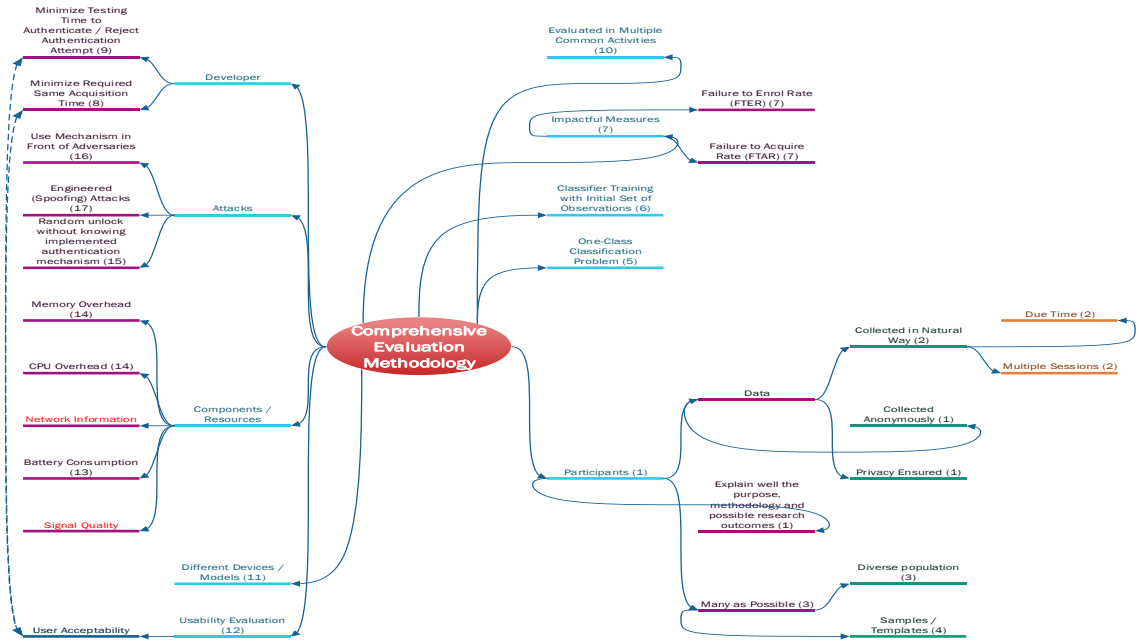



*Figure 2.1: Mind Map Documenting Buriro et al. [16] Guidelines*

Ellavarason *et al.* [23]–[25] have investigated the usability and performance of a behavioural biometric modality (swipe) across usage environments and scenarios on a mobile platform. The authors developed an "evaluation framework for analysing factors influencing user interaction in mobile devices concerning touch interactions. This data collection framework can be used further to perform the performance assessment of various touch-based behavioural biometric modalities using mobile devices". When analysing the swipe data, the authors noted how "the rise in EER values for the dynamic scenario was seen across all three classification algorithms. These results show the extent of the impact of the usage scenarios on verification accuracy. Furthermore, the results raise questions about the stability of swipe gesture authentication when used on a mobile device in real-life situations".

Bhagavatula *et al.* [26] showcased a biometric system's usability and perception concerns. In addition, they identified how specific scenarios impacted usability due to performance issues, notably how Android's Face Unlock would cease to function in dark conditions. Finally, Eglitis *et al.* [27] investigated how the influence of test protocols can affect recognition performance assessment. The authors performed a technology evaluation over a finger vein dataset and stated that "the details regarding such employed [testing] protocols are often not provided with due care, making it hard to compare the new results against those previously achieved in literature, even when performing tests on the same data". The authors demonstrate the "the need to accurately describe comprehensive test protocols when evaluating the recognition performance on a given biometric database".

Dube *et al.* [28] proposed a new framework evaluation for biometric authentication, although not with a specific aim to evaluate mobile systems. The authors identified three core parameters for evaluation security, privacy, and performance. The focus was to present an idea for incorporating biometric template

protection (BTP) metrics into evaluating biometric systems for a more comprehensive analysis. This suggestion is aimed at incorporating it into the performance evaluation framework.

## 2.3 National Cyber Security Centre (NCSC)

The National Cyber Security Centre (NCSC) is a UK Government organisation whose purpose is to support the UK (both public and private entities) in the role of all things cyber with the motto "helping to make the UK the safest place to live and work online". As part of this, they include guidance on the responsible use of biometrics and measuring performance [9].

As part of the guidance provided, it is stated that "in a biometric system, error rates will vary widely depending on the environment in which a system is operating. So, accuracy cannot be expressed independently from other factors". They also state four key performance metrics that should be considered for understanding performance: failure to acquire, failure to enrol, false non-match, and false match.

As part of the guidance, the NCSC mentions the metrics for evaluating biometric performance as failure-to-acquire, failure-to-enrol, false non-match and false match.

The NCSC also recommends that the evaluation occurs in an environment similar to where the system will operate and recommend that test claims and performance from manufacturers be treated as a starting point. "The performance of a biometric system is specific to its operating environment, so it is better to use the results of such evaluations as a starting point, to be followed by further testing of possible solutions in increasingly realistic test environments". They also recommend examining the source of the data and listing the requirements that should be examined for requirements that differ from what is expected:

- The choice of biometric data used in the evaluation
- The operating conditions used
- Test subject population
- Desired security posture

Some recommendations about managing performance to the desired performance level are provided with mentions that inevitable trade-offs will likely be required. The recommendation is fallback, how the system handles failures, throughput, and ergonomics, stating how "ergonomic design and the usability of human-computer interfaces are essential to the successful implementation of any biometric system". Finally, a recommendation regarding consistent performance monitoring of a biometric system is provided to identify any performance degradation within and across an enterprise using the biometric system and hardware and technology changes.

## 2.4 Microsoft (Windows Hello)

Windows Hello is the brand name used and promoted by Microsoft regarding password-less authentication to access the Windows operating system and is recommended for enterprise use. A significant aspect of Windows Hello is the biometric capability, and Microsoft states that “Windows Hello is the biometric authentication feature that helps strengthen authentication and helps to guard against potential spoofing through fingerprint matching and facial recognition” [29], [30].

The biometric reference is stored locally on the device within an encrypted database. Microsoft states that even if the data is stolen and decrypted, it cannot be converted into the raw biometric reference used on the sensor. Microsoft uses the False Accept Rate and the False Reject Rate to measure biometric performance, ensure the hardware meets the requirements set by Microsoft, and have some anti-spoofing measures.

For fingerprint authentication, the FAR requirements are <0.002% and FAR (with anti-spoofing or liveness) <10%. The FAR requirements are <0.001% for face authentication and FRR (with anti-spoofing or liveness) <5%. However, for face, Microsoft has also set a practical, real-world FRR requirement of <10%. It is not entirely clear what this effective real-world FRR is and why there is no similar requirement for fingerprints. The assumption is that this is an in-lab (scenario testing) and real-world (operational testing) requirement. Therefore, Microsoft concludes that Windows Hello face authentication is not recommended for mask wearers.

Microsoft also provides calculations for the number of comparisons and subjects required to reach a particular confidence level. For example, with the desired FAR of 0.001%, at a confidence of 96%, 2,500,000 comparisons would be required to reach the desired confidence with about 2,237 unique biometric samples to verify the confidence in the claimed FAR. In addition, Microsoft guides for transactional time state that it should take less than two seconds to authenticate a user.

Microsoft also provides two core scenarios where they expect users to use Windows Hello: authentication, where users will access the operating system regularly, and re-authentication, where users will enrol themselves into the device. Again, the expected occurrence of this is low. However, no detailed testing plan or requirements were found for Microsoft above what is presented here regarding how the tests should be performed.

## 2.5 Android (Google)

Android is the most extensive operating system for the smartphone market, with two billion users worldwide. As one of the leading mobile device operating systems, they have provided standards for every

device operating system to conform to this standard. This information is present within the Android Compatibility Definition Document (CDD) [31], which specifies the requirements for a device manufacturer to meet the standard required for the use of the Android operating system, including the use and performance of a biometrics implementation [31]. Android employs a tiered approach to authentication, meaning developers can choose how strong the performance can be if it meets the minimum requirements. In addition, more robust security (measured using the false accept rate) allows developers to use the biometric system to handle more sensitive features within the device.

Android favours security over usability, and therefore the minimum (weak) biometric performance must meet the following requirements:

- FAR: 1/50000 (0.002%)
- FRR: 10%
- SAR: 7-20% (Weak)

Android splits its biometric security under two subcategories, **architectural security** and **spoofability**. Android emphasises the security of the biometric functionality and is arguably putting more pressure on this area than the overall usability, as seen in the recommended rates for the various metrics.

'Architectural security' defines the security of the internal pipeline against kernel and platform attacks and is stated as not allowing the reading of biometric sample data or the injection of synthetic data or otherwise into the system. 'Spoofability' is measured against the Spoof Acceptance Rate (SAR) and measures the system's resilience against a dedicated attacker.

### 2.5.1 Metrics

Android emphasises three specific metrics to analyse biometric performance to enhance security and prevent impostors.

- Spoof Accept Rate (SAR): Defines the metric of the chance that a biometric model accepts a previously recorded, known good sample.
- Imposter Accept Rate (IAR): Defines the metric of the chance that a biometric model accepts input that mimics a known good sample.
- False Accept Rate (FAR): Defines the metrics of how often a model mistakenly accepts a randomly chosen incorrect input.
- False Reject Rate (FRR): Defines the metrics of how often the model mistakenly rejects a randomly chosen correct input.

Alternatively, this can be viewed as exposing replay attacks (SAR), exploring targeted attacks (IAR) and defending against passive impostors (FAR). Less emphasis should be placed on FAR as this does not provide necessary information about how well the model holds up to targeted attacks. The FRR initially has no mention. However, it does appear when discussing acceptable performance rates to meet standards.

### 2.5.2 Tiered Authentication

A unique approach taken by Android is to propose the idea of having tiered levels of authentication that allow manufacturers to provide more functionality if they can prove their device performs to the standards set out in Table 2.1. The tiers improve security from '**convenience**' to '**weak**' and '**strong**'. These tiers have been named Class 1, Class 2, and Class 3, respectively.

One purpose of the tiers allows how much API access is allowed to the biometric process. For example, the API must not be exposed with' convenience'. With 'weak', the biometric process can integrate with Android's BiometricPrompt API and with 'strong' access is allowed via the BiometricPrompt and the Keystore on the device.

*Table 2.1: Android Tiered Authentication Metrics*

| Biometric Tier | Metrics |
|---|---|
| **Strong (Class 3)** | SAR: 0-7%<br>FAR 1/50K<br>FRR: 10% |
| **Weak (Class 2)** | SAR: 7-20%<br>FAR: 1/50K<br>FRR: 10% |
| **Convivence (Class 1)** | SAR: >20%<br>FAR: 1/50K<br>FRR: 10% |

### 2.5.3 Evaluation Process

Android provides specific guidance for the face, iris, fingerprint, and voice modalities. Android combines its evaluation process into two stages. One is called the '**calibration phase**' to find the optimal presentation attack for a given solution. A '**test phase**' takes the results of the calibration phase to evaluate the system and determine how many times the attacks were successful. The main point here is to attack the system using the most significant know or expected weaknesses. "The calibration phase is used to find optimal parameters that maximise the chances of spoofing the authentication solution." [31].

Android gives some recommendations for the 'calibration phase' for the face and iris modalities and investigates three areas. First, the presentation 'medium' is the output medium for the spoof. The 'format' is considered a manipulation of the medium or environment to improve spoofing chance. Finally, the consideration across 'subject diversity' primarily refers to picking specific impostors across gender and ethnicity groups for facial-based authentication. Android considers an attempt as the window between a probe presented to the system and receiving feedback. Feedback can be a successful unlock or a message presented to the user.

### 2.5.4 Evaluation Phase

For the enrolment, remove all existing biometric profiles and add the target face in a brightly lit room at 20cm to 80cm. For face recognition, the suggestion is to test the spoof medium (photo or mask) in a vertical and horizontal arc around the device until the medium is no longer visible within the device's field of view, in increments of 10 degrees marking positions that successfully unlock the device. The results of this will determine the calibration positions. The suggestion for fingerprints involves lifting a latent copy of the target fingerprint and creating a mould against the sensor using at least four different materials (gelatine, silicone, wood glue). A 5% margin of error (with 95% confidence) requires 385 test iterations per subject.

### 2.5.5 Common Considerations

The article finishes by considering some common considerations that can apply to all modalities, including the advice that tests should occur using the actual device with the hardware installed to capture accurate metrics. Most modalities have a successful spoof attack, and there exist documented techniques for them, and these existing attacks should be used in the calibration phase when planning the test. The final advice is to anticipate new attacks, and Android acknowledges that not all attacks may involve a suitable setup with publicly known attacks. Existing modalities may need their existing methodology to alternate with discovering a new attack. A process must be in place to adapt to new information in any testing methodology.

## 2.6 iOS (Apple)

Apple is the manufacturer behind the iPhone range of smartphones and introduced Touch ID [32] and Face ID [33] (fingerprint and face, respectively). However, Apple has generally been somewhat secretive about the technology that goes inside the iPhones to the point where they will buy out companies and house them under the Apple name. Such was the case for the Touch ID using AuthenTec technology [34].

Since Apple does not distribute its mobile operating system to any external smartphone manufacturers, it does not produce a biometric testing requirement and guides as Google does for Android. However, they still provide information through whitepapers about how the technology works.

Regarding Touch ID, Apple claim that it “reads fingerprints from any angle and learns more about a user’s fingerprint over time, with the sensor continuing to expand the fingerprint map as additional overlapping nodes are identified with each use”, indicating information relating to reference template updating over time. Furthermore, “With one finger enrolled, the chance of a random match with someone else is 1 in 50,000. However, Touch ID allows only five unsuccessful fingerprint match attempts before the user can enter a passcode to obtain access”.

Regarding security and privacy, “the 88-by-88-pixel, 500-PPI raster scan is temporarily stored in encrypted member within the Secure Enclave while being vectorised for analysis, and then it is discarded. The analysis utilises subdermal ridge flow angle mapping, a lossy process that discards minutia data required to reconstruct the user’s fingerprint. The resulting map of nodes never leaves iPhone 5S, is stored without any identifying information in an encrypted format that can only be read by the Secure Enclave and is never sent to Apple or backed up to iCloud or iTunes” [32].

Apple handles Face ID similarly and claims that “the probability that a random person in the population could look at your iPhone or iPad Pro and unlock it using Face ID is approximately 1 in 1,000,000 with a single enrolled appearance. The statistical probability is different for twins and siblings that look similar and children under 13 because their distinct facial features may not have developed fully. Face ID matches against depth information not found in print or 2D digital photographs. It is designed to protect against spoofing by masks or other techniques through sophisticated anti-spoofing neural networks” [33].

Apple provides no information regarding how they performed the biometric performance evaluation and information relating to the test cohort involved. The only information provided is related to the false match rate for Touch ID, 1 in 50,000, and Face ID, 1 in 1,000,000, meaning consumers only have the manufacturer's word that their claims are valid.

## 2.7 International Organization for Standardisation (ISO)

The International Organization for Standardisation develops and publishes International Standards. ISO has a dedicated Biometrics subcommittee to develop standards within the field known as ISO/IEC JTC 1/SC 37. Within the subcommittee is a dedicated working group for the specialised area of Biometric Testing and Reporting (WG 5). Standards for test protocols are defined within ISO/IEC 19795, “Information technology – Biometric performance testing and reporting” documents, currently consisting of ten parts.

Eglitis *et al.* [27] provide a summary of the recommendations provided by ISO/IEC 19795-1:2006 "Part 1: Principles and framework" [22], including:

- The test phase should be conducted on data unavailable during algorithm development
- Collection of enrolment and probe data should be separated at least by days
- The "rule of 3" and "rule of 30", which relate the number of probes with the achievable error confidence intervals, should be considered when reporting error rates. It is remarked that handling ten probes for ten subjects is not equivalent to having a hundred subjects, each with only a single probe, although, for specific protocols, this produces an equal number of comparisons
- Data from the same subject and the same modality, yet different instances (e.g., distinct eyes, fingerprints, finger veins) can be used to represent distinct users
- Collected samples should be excluded from the database only if a predetermined criterion is violated
- Each test subject should be enrolled only once
- Impostor comparisons involving data captured from the same subject (e.g., vascular data from different fingers of the same person, representing different virtual users) should not be performed because intra-individual data are likely to contain more similarities than data from different individuals
- Zero-effort impostors can be selected by randomly choosing biometric templates or by making a full cross-comparison
- Enrolment templates can be used as impostor data in case different feature extractors are applied to enrolment and probe samples

Mansfield and Wayman [35] proposed a guide for the best practices for the performance evaluation of biometric systems. Their work was incorporated into the framework proposed in ISO/IEC 19795-1:2006 Part 1: Principles and framework [22]. ISO/IEC 19795-2:2007 "Part 2: Testing methodologies for technology and scenario evaluation" [36] describes the recommended scientific practices for technical performance testing. WG5 have also produced an updated technical standard to deal with specific issues presented for testing mobile biometrics, ISO/IEC TS 19795-9:2019 Part 9: Testing on mobile devices [37]. It is stated that the standard aims to guide "performance assessment at a full system level of biometric systems embedded in mobile devices with an offline evaluation of false accept rate (FAR) claims". It is noted that the standard is designed for a complete system level; therefore, elements will not be applicable if biometric system access is unavailable.

ISO/IEC TS 19795-9:2019 provides recommendations and requirements for mobile biometric performance testing, including how "evaluation of biometric performance … should be consistent and follow the same guidelines, methodologies and requirements" and that "ISO/IEC TR 30125 [38] recommends scenario evaluation as the most proper type of evaluation for testing biometric performance on mobile devices".

The guidance also provides a host of considerations for mobile biometric evaluations, including:

- Biometric authentication process
    - Explicit authentication
    - Passive authentication
- Biometric capture sensor
    - Embedded sensors
    - Dedicated sensors
- Uncontrolled environment
- Challenges in storing references and generating comparison scores
- Adaptation of the biometric references
- Biometric application is a black box for security and privacy reasons.
- A third-party evaluation requires that the system provider deliver a customised version that provides access to biometric data or detailed transaction logs.
- Factors that increase the time and cost of biometric performance evaluations
    - Minimum error rates
    - Inability to store a large amount of biometric data
    - Access to the captured biometric samples
    - The number of conditions to evaluate.
- Third-party evaluation of FAR would quickly be impractical and time-consuming if the evaluator only has access to an unmodified mobile device
- Baseline - The conditions suggested for this scenario are indoor conditions, with no noise in which the user hand-held mobile device.

The standard also notes that “devices are not designed to store multiple references and generate comparison scores from the submission of a probe against these references.” “NOTE Some mobile devices allow the enrolment of several users for biometric identification in a small dataset related to low-security features, e.g., device unlocking. This document only considers verification use cases related to secure transactions, which can vary depending on the risk, policy and legislation that applies to the transaction” [37].

The guidance states that an evaluation of a mobile device shall at least report FTE, FTA, FRR and FAR. The standard contains detailed information regarding biometric testing and reporting for mobile devices, and only the minimum information has been presented here. These and related standards will be referenced at appropriate opportunities within the thesis.

## 2.8 FIDO

The FIDO Alliance is an open industry association with a focused mission: authentication standards to help reduce the world's over-reliance on passwords. In addition, they act as a certification authenticator where devices can be labelled as FIDO compliant if they are shown to meet certain conditions, including biometric performance [39]. The FIDO process takes heavy influence from the ISO standards.

The FIDO Alliance is an open industry association with a focused mission: authentication standards to help reduce the world's over-reliance on passwords [40]. Biometrics help in this mission by providing an alternative to the "something we know" passwords with the "something we are" authentication approach. FIDO recognises this and acknowledges that "biometric user verification has become a popular way to replace passwords and PINs, but the lack of an industry-defined program to validate performance claims has led to concerns over variances in the accuracy and reliability of these solutions" [3]

The FIDO Alliance is an open industry association with a focused mission: authentication standards to help reduce the world's over-reliance on passwords. They act as a certification authenticator where devices can be labelled as FIDO compliant if they are shown to meet certain conditions, including biometric performance [39]. The FIDO process takes heavy influence from the ISO standards.

FIDO relies on the following metrics to assess performance:

- False Accept Rate (FAR) - 1/10000 (0.01%)
- False Reject Rate (FRR) - 1/500 (5%)
- Impostor Attack Presentation Accept Rate (IAPAR) - <15% (Level 1), < 7% (Level 2)

- Number of Subjects 245 (Min)

FIDO and Android differ because the FAR performance is based on zero-effort non-genuine transactions. In contrast, Android includes guidance on using impostors to exploit weaknesses (same gender and ethnicity). Therefore, the corresponding decision threshold should be reported with the FRR, and FAR can be shown using a ROC or DET curve.

The Biometric Component Certification Program was created to "utilise accredited independent labs to certify that biometric subcomponents meet globally recognised performance standards for biometric recognition performance and Presentation Attack Detection (PAD) and are fit for commercial use" < Cite>. Part of the goals for this certification is to provide the industry with a testing baseline for biometric component performance< Cite>.

### 2.8.1 Metrics

The FIDO Biometric Component Certification Program relies on three key metrics to evaluate performance:

- False Accept Rate (FAR): The proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed
    - SHALL meet the requirement of less than 1:10,000 for the upper bound of an 80% confidence interval. FAR is measured at the transaction level.
- False Reject Rate (FRR). The proportion of verification transactions with truthful claims of identity that are incorrectly denied
    - SHALL meet the requirement of less than 3:100 for the upper bound of an 80% confidence interval. FRR is measured at the transaction level.
- Impostor Attack Presentation Accept Rate (IAPAR) is the Proportion of presentation attacks in which the target reference is matched
    - SHALL be performed by the FIDO-accredited independent testing laboratory on the TOE provided by the vendor. The evaluation measures the Impostor Attack Presentation Match Rate for each presentation attack type, as defined in ISO 30107 Part 3.

In general, “FIDO-accredited independent testing laboratory performs live subject scenario testing on the TOE provided by the vendor using a combination of online/offline testing and presentation attack testing, based on ISO 19795-1 and ISO 30107-3” < Cite>. A transaction should not exceed 30 seconds.

Following form Android’s example, emphasis is placed on security and defence against impostors.

### 2.8.2 Performance

The following performance rates are required to meet the standards for FIDO certification. All require an upper bound confidence level of 80%:

- FRR: 3/100 (3%)
- FAR: 1/10000 (0.01%)

FIDO and Android differ because the FAR performance is based on zero-effort non-genuine transactions. In contrast, Android includes guidance on using impostors to exploit weaknesses (same gender and ethnicity). Therefore, the corresponding decision threshold should be reported with the FRR, and FAR can be shown using a ROC or DET curve.

Self-attestation of performance is allowed along as test data support it and fully documented in a report submitted by the vendor.

### 2.8.3 Common Test Harness

A proposal for a standard test harness is used for vendors to supply to FIDO to allow open access to a system's biometric component for certification. This standard test harness calls for specific components:

1. Configurable Enrolment system
2. Configurable Verification online system
3. Configurable Verification offline software
4. Logging capabilities

This proposal highlights the difficulty of effectively and accurately the claims of biometric systems without access to the internal workings.

Not as much emphasis is given to the security of the architecture as Android specified. However, it is mentioned that "enrolment templates and verification transactions should be confidentiality and data authentication protected using cryptographic algorithms" [39].

### 2.8.4 Test Procedures

FIDO recommends that testing be carried out using the scenario test approach. The minimum number of subjects is 245, although, for face and iris, 123 unique subjects can be used if four-finger or two eyes are enrolled, respectively.

Regarding age distribution, the requirements indicate that no one under 18 or over 70 should be included. The remaining age groups (18-30, 31-50, 51-70) should be distributed evenly within 25%-40%. Similarly, gender (Male, Female) distributions should be as evenly split as possible within 40%-60%.

A bootstrapping sampling technique with replacement can be used to estimate FAR and FRR distribution curves. If zero errors occur, FIDO defaults to the "Rule of 3" for determining a test size. However, the test can be conducted in one visit for each participant due to the importance of testing the FAR rate.

FIDO acknowledges recent systems employing template adaption techniques that adapt the template after successful verifications. However, for adequate testing, the number of required correct matches to adequately train the system should be used, after which the adaption technique should be turned off to perform the tests.

Enrolment procedures are reported along with any failure-to-enrols that occurred.

### 2.8.5 Report to FIDO

- Summary of the FIDO Biometric Certification and Requirements
- The number of individuals tested
- Distribution of Age
- Distribution of Gender
- Description of the Test Environment
- Description of the Test Platform
- Distribution of the time elapsed between Enrolment and Acquisition
- Number of enrolment transactions
- Number of genuine verification transactions
- Number of impostor verification transactions
- Failure to Enrol Rate
- Failure to Acquire Rate
- False Reject Rate
- False Accept Rate
- Distribution of Genuine Verification Transaction Time
- Bootstrap Distribution

### 2.8.6 Presentation Attack Detection (PAD)

A unique addition from FIDO is the introduction of presentation attack instrument (PAI) levels which level presentation attacks based on their sophistication. The accepted IAPMR varies based on the level of attack chosen and the amount performed from each level. It is stated that using all Level A or Level B PAI attacks much to achieve an IAPMR of less than 50%. No more than five attempts are allowed per impostor transaction. An additional study is required to test the system's presentation attack detection, measured by IAPAR), to determine the resistance against minimal expertise attacks. This requirement can be achieved using 15 subjects.

As well as the items included in the testing report, the PAD report requires some additional items:

- Number and description of presentation attack instruments, PAI species, and PAI series used in the evaluation
- number of test subjects involved in the testing
- number of artefacts created per test subject for each material tested

- number of sources from which artefact characteristics were derived
- number of tested materials
- Imposter Attack Presentation Accept Rate (IAPAR)
- Number of impostor attack presentation transactions

Testing the PAD required online testing using the Common Test Harness.

As mentioned, FIDO lists PAI attacks into various levels focused on the complexity of attack potential as defined in ISO/IEC 30107-3:2017 [41] :

- Level A - Simple to carry out and requires little time, expertise, or equipment. (Paper printout of face image)
- Level B - Require more time, expertise, and equipment. (Paper masks)
- Level C - The most brutal attacks. (Silicon masks)

FIDO provides more information on the PAI attack levels per modality basis. Laboratories should select six Level A and eight Level B attacks, resulting in 14 PAI species and 15 enrolled users, meaning 210 instruments. In addition, the lab must perform ten presentation attack transactions for each PAI, and all acquisition failures must be reported during the process. The FIDO Alliance is continually developing their biometric testing standard, which currently sits at v2.2 the guidance is changed with updates from standards, laboratories, and industry partners.

## 2.9 Discussion (Moving Forward)

Current standards depend on performance rates by looking at the statistical hypothesis testing errors, examining type I (rejection of a true null hypothesis) and type II (non-rejection of a false null hypothesis) errors. When applied to biometrics, these become the false reject, and false accept rates. Existing approaches are available from researchers, standards, and organisations to build a common approach to evaluating biometrics. However, few of these consider the mobile approach independently and with a vaster usage in the hands of everyday usage should a common criterion exist and define how to approach mobile biometric testing for defined testing and comparisons.

A common theme among testing strategies is to take a scenario evaluation approach. These existing standards and approaches offer practical planning and a mobile biometric performance evaluation. However, it can be achieved, and the definition of performance can be expanded beyond traditional measures. Furthermore, the approaches mentioned here usually ignore other metrics as they are usually deemed out-of-scope of the specific purpose. For example, ISO/IEC TS 19795-9:2019 [37] puts out-of-scope privacy aspects and presentation attack detection (PAD).

The aim is to develop a standard structured approach to mobile biometric testing and reporting combining the elements for all these standards and additional metrics that may apply to a broader range of devices. This chapter has presented the core entities defining how biometric testing should be carried out. The aim is not to disregard or undermine the work carried out by these bodies but carry it forward by amalgamating their relevant elements and incorporating them into a new biometric testing framework specific for mobile devices to provide a more comprehensive testing structure for various devices with meaningful results.

However, one aspect is to look beyond these traditional metrics in today's world and look further. As well as this, privacy and security concerns are rising, and biometric data protection is vital as the GDPR has categorised it as 'sensitive' data [42]. Furthermore, cancellable biometrics is one of the major categories for biometric template protection purposes besides biometric cryptosystems [43]. Therefore, assessing the security and privacy-preserving techniques implemented by the mobile biometric system should be part of the evaluation process to help assess architectural security.

Eglitis *et al.* [27] also state that "several of the ISO/IEC 19795 recommendations …, e.g., enrolment and probe data being captured at different days, or computing a minimum number of comparisons to validate error rates, are often not respected in the employed test protocols, thus affecting the reliability of the reported performance".

## 2.10 Summary

This chapter has explored what currently exists regarding biometric performance testing and considering mobile biometric systems where possible. One thing that became apparent from the requirements present here is the test sizes. The ISO standards emphasise "Rule of 3" and "Rule of 30", but how practical is this? Admittable, it is hard to withdraw from the statistical significance using these numbers provided, but even 245 participants seem a high requirement for all settings, such as academia. Could an alternative approach be developed to reduce the need for such an extensive test size?

The next chapter will explore the core factors identified to be crucial to the performance of a mobile biometric system and showcase how significant a user's influence over the performance of a mobile biometric system can be, deeming it essential that some form of usability testing is incorporated into the testing framework.

# 3 Core Factors and Relationships

## 3.1 Introduction

In this chapter, the intent is to take an approach to separate the core factors affecting mobile biometric systems' authentication performance. Jain *et al.* [44] described seven factors to help assess the suitability of a human trait for biometric authentication, one of which defines *performance* in that it "relates to the accuracy, speed, and robustness of technology used". There is a vast field of research about performance claims for biometric systems; however, these claims usually predict them by observing a changed factor(s) and noting the resulting performance deterioration or improvement demonstrated. This chapter aims to identify the core factors that need consideration to specification, evaluate, and report biometric systems on mobile devices. The work described in this chapter is an adapted version of previously published work [45] addressing this theme.

Section 3.2 discusses how performance is viewed, looking at critical areas. Sections 3.3 - 3.9 will discuss each influencing factor in more detail. Section 3.10 will introduce the relationships between factors, and the final Sections 3.11 and 3.12 will provide a discussion and summary.

## 3.2 Discussing Performance

Mansfield and Wayman [35] produced a comprehensive list of factors that can affect the performance of a biometric system. Outlined in Table 3.1, these '*influencing factors*' were later included within the ISO/IEC 19795-1:2006 [22] international standard on biometric performance testing and reporting, including strategies for mitigation against performance degradation, such as a section on 'controlling factors that influence performance. Although the procedures discussed in the standard are primarily relevant to modern biometric devices, including mobile systems, any alternative approach must observe them to the current state of the art regarding developments in biometric technologies. For example, one strategy in the standard suggests that "enrolment conditions should model the target application enrolment", but this is not so straightforward when moving from a statically implemented system to a mobile one where the enrolment conditions could be a plethora of different environments and scenarios. Furthermore, the ISO standard only seeks to assess performance through very generalised metrics, mainly failure-to-enrol, failure-to-acquire, false match rate and corresponding false non-match rate.

*Table 3.1: List of Influencing Factors as Defined in ISO/IEC 19795-1:2006 [22]*

| **Factor** | **Description** [22], [35] |
|---|---|
| Population Demographics | Characteristics of a population, such as age, gender, ethnic origin, and occupation |
| Application | The overall system itself, such as enrolment and verification elapsed time, user familiarity and user motivation |
| User Physiology | Physical properties of a person, such as a beard, skin tone, height, and disability |
| User Behaviour | Behavioural properties of a person, such as a dialect, movement, stress, and facial expressions |
| User Appearance | How a person looks, such as clothing, hairstyle, bandages, and tattoos |
| Environmental Influences | Factors of the environment, such as background, lighting level, weather, and reflections |
| Sensor and Hardware | The factors affecting the device's correct operation, such as dirt, focus, sensor quality and transmission channel |
| User Interface | Means by which the user and a computer system interact, such as feedback, instruction, and supervision |

Further research [46]–[48] discusses assessing performance for traditional biometric systems, and these have usually been in the form of exploring *influential factors*. This current work aims to expand on these previously defined factors and identify new areas that need extra consideration when applied to a mobile biometric system. Part of this will be informed by assessing the role of the 'Users' factor within a biometric system and increasingly considering the importance of usability when discussing performance. Furthermore, this study aims to illustrate how the 'Environments' factor has a greater context when considering a mobile setting that links closely with the newly introduced factor of '*Scenarios*'.

This work aims to develop a model for assessing the performance of mobile biometric systems and implementations by investigating factors that are likely to affect the performance. The seven identified factors are '*Users*', '*Modality*', '*Environments*', '*Diversity of Scenarios*', '*System Constraints*', '*Hardware*' and '*Algorithms*', and they form 'The Core Factors Affecting Mobile Biometric Performance'. Utilising these factors allows for illustrating the practicalities of mobile biometric implementations, enabling the binning of performance alterations within one of these factors.

Along with '*Scenarios*', the model establishes '*Algorithms*' and '*System Constraints*' as new factors that have not previously been considered explicitly within a performance assessment context. The Oxford Dictionary defines a factor as "a circumstance, fact, or influence that contributes to a result" [49]. The core factors defined by this study are the fundamental elements of mobile biometric performance, acting as a foundation layer for future performance assessment development. Even as biometric systems evolve, additional properties and areas are discovered, and new relationship connections are forged, they will always link back to one (or more) of these core factors. Furthermore, the core factors form unique connections, meaning that an impact on one core factor can cause a performance alteration in another.

This study will allow developers to concentrate efforts more effectively when devising ways of testing and analysing the performance of mobile biometric systems in the future. Building on previous studies

from the community, the developed model demonstrates each factor's existence and the impact on the overall performance when applied to a mobile context.

While examining the definition of biometric system performance, it is necessary to include further input from Jain *et al.* [44] regarding *acceptability*, how well individuals "accept the technology such that they are willing to have their biometric trait captured and assessed". Additionally, the concept of *circumvention* which "relates to the ease with which a trait might be imitated using an artefact or substitute". Both issues are critical when considering the overall performance of a biometric system. Acceptability is essential because public perception and acceptance within the technology sector will be one reason to prevent the uptake of a biometric system.

Biometric technology has trust issues among consumers [12], with a Paysafe Group survey revealing that 81% prefer passwords for online payments due to security concerns. A recent example of driving public opinion through security concerns happened in 2019 in San Francisco, where the public administration banned facial recognition for local services [50]. The cited reason was over-perceived bias in facial recognition technology regarding ethnicity. More recently, the European Union MEPs backed a motion to ban facial recognition technology for mass surveillance [51], paving the way for future laws and restrictions regarding AI technology, citing privacy rights.

Although the examples are rather extreme, it highlights that it is necessary to consider the user's acceptance of a system before rushing ahead with the implementation; otherwise, it will alienate the users. Circumvention primarily refers to spoofing the biometric system using artificial means. Any system that offers a biometric solution will be required to have some level of resistance to these attacks. For example, a fingerprint system must be resistant to artificially produced finger attack specimens or the insertion of images into the biometric pipeline. Failure to attack specimens highlights a significant flaw with the system, leading to decreased overall performance.

By considering *acceptability* and *circumvention,* the model challenges the conventional approach and definition of *performance*. The focus within the model begins to shift towards the end-user, how they perceive the use of the system and how both *good* and *bad* actors may attempt to circumvent the system using various methods. These are vital areas when considering the performance of the overall system. The metrics discussed in ISO/IEC 19795-1:2006 [22] also ignore the usability aspect, which the model illustrates as vital to a system's performance.

## 3.3 Factor #1: Modalities

With existing fixed biometric modalities, evaluators can target the strategy to consider known '*influencing factors*' likely to affect the performance and include these within the testing strategy. Mansfield and Wayman [35] provided a list detailing many of these '*influencing factors* in Table 3.1. Each

modality directly introduces a set of '*influencing factors*' caused by choosing a particular modality, for example, an assessment of how wearing glasses will likely affect the assessment of an iris recognition system.

Table 3.2 shows illustrative examples of *influencing factors* for a range of common biometric modalities and is by no means a completely comprehensive list.

***Table 3.2: Examples of Influencing Factors per Modality***

| **Modality** | **Sample of *Influencing Factors*** [51] | |
|---|---|---|
| Face | • Movement<br>• Age | • Facial Expression<br>• Skin Tone |
| Fingerprint | • Fingerprint Condition<br>• Arthritis | • Weather<br>• Offsets and Rotations |
| Voice | • Ethnic Origin<br>• Colds or Laryngitis | • Noises<br>• Misspoken Phrases |
| Iris | • Lightning Level<br>• Blindness | • Eyelashes<br>• Reflections |
| Signature | • Age<br>• Sensor Pressure | • Injuries<br>• Motivation |

It can be observed from these factors in both Table 3.1 and Table 3.2 that the users are a significant influence that affects the performance of a particular modality.

The remaining core factors, defined in the model, will uniquely link to this *modality* core factor, for example, the *hardware* and *algorithms core factors*, as each will have a performance impact that could affect the other. Therefore, it is also essential to know the '*influencing factors*' specific to each modality when used in various scenarios. This knowledge will help identify the areas requiring increased attention when testing a biometric system. For example, a user's appearance will have little impact on voice but will likely significantly affect a facial recognition system. In addition, the understanding that a single modality may perform with error is apparent as researchers have utilised multimodal biometric systems to combine separate modality systems with improving overall recognition accuracy performance. Part of these multimodal systems tries to overcome the '*influencing factors*' issues by combining the results of another trait that will hopefully not be affected and reduce the error that would otherwise have happened [52].

Jain *et al.* [52] reported that a unimodal system (using a single trait/modality) might experience several problems, including "noisy sensor data, non-universality and lack of distinctiveness of the biometric trait, unacceptable error rates, and spoof attacks". All of which can affect the overall performance of a system. In comparison, He *et al.* [53] explored the performance of a multimodal system using three traits: fingerprint, face, and finger vein. The experiments concluded that a "multimodal biometric system can achieve significantly better performance compared to a single biometric system" and that adding a finger vein "results in a verification system with very high accuracy". Thus, this research demonstrates how performance changes when using various biometric traits, confirming that modalities affect performance.

Gafurov *et al.* [54] assessed the use of gait as a biometric trait. They concluded that it is better suited as a "complementary biometric" and not a "replacement for traditional authentication mechanisms". The work also noted "several factors that may negatively influence the accuracy". They classed the factors for gait as '*External*' (viewing angles, lighting conditions) and '*Internal*' (sickness, physiological changes). They identified how gait was "robust against minimal effort impersonation attacks". The authors concluded by noting that an "investigation of these factors is critical towards developing robust systems", which identifies how necessary it is to appropriately select a modality for a particular scenario in a way that will try to mitigate issues (caused by *influencing factors*).

Ito *et al.* [55] commented on how researchers seek "new biometric traits to enhance the accuracy and convenience of biometric recognition", suggesting that modalities differ in their performance. Each modality is unique, bringing an assortment of *influencing factors* to contend with, although some are common between specific traits. It is undoubtedly true that a modality holds extraordinary power over the system. It can define how the rest of the system develops around it, meaning selecting the suitable modality for the job is a priority. An example of a lesser-researched biometric includes the 3D ear shape explored by Yan *et al.* [56].

Furthermore, external factors relating to the target population may cause modality-specific performance alterations. For example, the sample quality of the elderly [57] and issues of accessibility [58] will influence the performance of the modality. Elliott *et al.* [59] investigated how fingerprint sample quality can affect a biometric system across age groups. They concluded that "more emphasis should be placed on an individual's age, rather than the moisture of the finger when developing a fingerprint recognition system" as the image quality became more variable for an older population (aged 62 and over).

Due to the '*influencing factors'* present, the modality themselves impact the performance of a system. It is for this reason, therefore, that it is one of the factors.

## 3.4 Factor #2: Environments

The environment can significantly impact the performance of a biometric system. These systems are becoming less fixed and mobile, and the environments they will operate in are nearly impossible to predict.

Research produced by Lunerti *et al.* [60] examined the environmental impact factors on smartphone face recognition. They assessed facial image quality (FIQ) in indoor and outdoor conditions. They concluded that "[biometric] scores obtained with the images taken from the smartphone are higher with the images taken indoors", showing that the environment impacts a system's performance.

One of the main aspects that allows a mobile biometric system to differentiate itself from a traditional one is the sheer range of environments and conditions the device will be required to operate within. Consequently, the performance could be affected at both the enrolment and authentication phases due to the various environments and situations that could occur during the process. For example, in practice, a robust enrolment template captured under 'optimal' conditions may not produce accurate matches with samples collected under certain circumstances at later verification attempts. Furthermore, an enrolment template captured under poor conditions may not be functional.

Different environments that have an impact on the performance include:

- Indoors vs Outdoors
- Lighting
- Weather Conditions
- Terrain -- physical features of the environment, from the ground, being walked over to the type of location (e.g., city, countryside, ocean)

Previously Elliott *et al.* [59] had also shown how illumination could significantly affect the performance of facial recognition systems. They concluded that "enrolment illumination level is a better indicator of performance than the illumination level of the verification attempts". Furthermore, they found that the "enrolment light level should be as high as possible" when the "lighting conditions are not constant for verification".

Regarding behavioural biometrics, voice recognition performance severely degrades when ambient noise is present, as shown by Gong [61]. Therefore, researchers have researched to mitigate and detect this noise and hence the performance of a voice recognition system. For example, Yamada *et al.* [62] "described a method for estimating the performance of a speech recognition system using a distortion measure".

Applying the environment concept to conventional modalities has been known to affect performance. These conditions include:

- **Face** – Background (Multiple Faces)
- **Fingerprint** – Weather
- **Voice** – Noise
- **Iris** – Illumination

These studies show how the environment impacts biometric performance, and it is deemed worthy as another factor.

## 3.5 Factor #3: Diversity of Scenarios

The 'Diversity of Scenarios' relates to how an individual uses a device within an environment. Classing these scenarios under two category headings: 'motion' and 'stationary' is possible. Table 3.3 shows a brief table of example scenarios.

*Table 3.3: Examples of Scenarios Under the Categories of 'Motion' and 'Stationary'*

| Motion | | | Stationary |
|---|---|---|---|
| **User** | **Transportation** | **Dual** | |
| Walking | Bus | Walking on a train | Sitting |
| Running | Train | Walking on a boat | Standing |
| Cycling | Earthquake | Swimming | Lying Down |

In this classification, 'motion' refers to the scenario of being in movement relative to the environment. Likewise, 'stationary' is where the device is at rest (no action), corresponding to the environment. Furthermore, 'transportation' is defined as any scenario where the device is in motion caused by an external influence from the environment. For example, when being driven around, such as on a bus. This splitting of 'motion' scenarios into 'user' and 'environment' concepts introduces an overlap scenario where the cause is both a user and environmental factor—defined as a 'dual' motion scenario. Figure 3.1 illustrates a flowchart to aid in assigning scenarios to a potential category. As a result, minor movements that will likely occur in stationary situations, such as shaky user hands while holding the device, are overlooked.

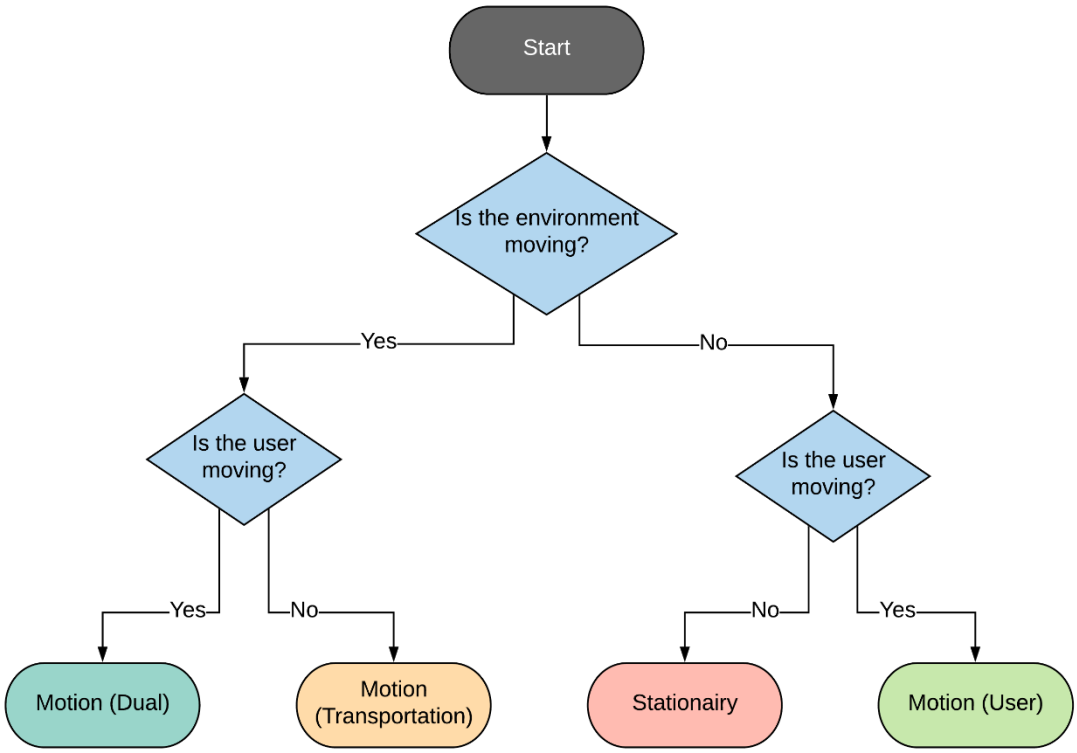



*Figure 3.1: Flowchart to assign a scenario to a category of 'Motion' or 'Stationary'*

There is currently limited research assessing various scenarios under which biometric authentication can occur, including the conventional modalities: Face, Fingerprint, Voice, and Iris. However, different scenarios can alter the performance of a biometric system. For example, Blanco-Gonzalo *et al.* [17] explored performance changes across conditions when signing using dynamic signature verification (DSV) systems. Their results showed that although there is "not an ideal scenario for signing", the observed performance improved when using a stylus device with "the user sat on a chair, and the device is resting on a table" and for finger-stylus devices when "the user has to handle the device without support".

The location of a scenario plays a role in affecting the user's behaviour and state of mind. For example, "stress negatively influences performance and usability" [17]. For example, consider signing within an outlet such as a post office. Here, the scenario encourages users to sign quickly and carelessly to avoid causing long delays, introducing stress and anxiety, and negatively affecting performance. Whereas, in a ceremony-based scenario, such as signing a legal document, the "user typically signs with greater care, striving for enhanced quality and clarity", as demonstrated in research by Guest *et al.* [63], which causes an increase in performance.

The scenario links closely with the user's interaction with a system in a particular environment and how adjustments may be needed to account for these changes. These adjustments can come from the system itself or how the user interacts with it. The development of the Human-Biometric Sensor Interaction (HBSI) model [64] investigated how scenarios can modify performance. Brockly *et al.* [65]

concluded: "the development [of HBSI] reveals the complexity of the potential interactions and the changes of those interactions when digitisers change, as well as when the ceremony changes".

The number of scenarios in that biometric authentication can occur increases dramatically when introduced to a mobile environment. Bhagavatula *et al.* [26] assessed the usability of a range of mobile biometrics systems. Firstly (and probably the least surprising) was the Android "face unlock completely unusable in a dark room". They also explored Apple's Touch ID and found that the Touch ID and face unlock "mechanisms fail in specific scenarios, wet fingers and dark rooms, respectively". They also conducted a series of walking experiments, one merely walking and another walking while carrying a bag in one hand. The authors performed the experiments in laboratory conditions (indoors). Participants "did not find unlocking difficult for any authentication scheme in either of the walking scenarios". The experiment was mainly conducted from a usability perspective, concluding that participants preferred the Android face unlock in the walking scenarios as they could handle the phones in their desired positions. In contrast, Apple's Touch ID had to hold the phone more precariously from the bottom as this is the fingerprint sensor's location.

Sitová *et al.* [66] introduced "hand movement, orientation, and grasp (HMOG)" to authenticate smartphone users continuously, and their work investigated two conditions, sitting and walking. The results showed that "HMOG improves the performance of taps and keystroke dynamic features, especially during walking". They theorised that this improvement was "attributed to (a) the distinctiveness in hand movements caused by tap activity and (b) the distinctiveness in movements caused by walking".

With increased development in mobile technology, research is looking into ways to help capture and authenticate a biometric trait while in motion, including work to perform long-range iris recognition, also known as iris-on-the-move, as surveyed by Nguyen *et al.* [67].

Given the proliferation of biometrics on mobile systems, this indicates that researchers should conduct further work in this area, which should be considered a factor.

## 3.6 Factor #IV: Users

Users of a biometric system need to have confidence in the authentication process. Therefore, biometric systems must be universal (applicable to all) for the system's end-users to gain this confidence. Therefore, implementers chose modalities due to their uniqueness in identifying between individuals. However, developing a system to accurately extract all the small features within a modality to achieve uniqueness is difficult; this is how false positives can occur.

One of the most prominent ideas in biometric testing is the concept of the '*Biometric Zoo*' proposed by Doddington *et al.* [68]. In this work, they used voice recognition to show that users of a system could

directly impact the overall performance. In addition, this work proved the existence of different categories of users:

- **Sheep** – Sheep dominate the population, and systems perform nominally well for them.
- **Goats** – Goats are those users who are particularly difficult to recognise.
- **Lambs** – Lambs are those users who are particularly easy to imitate.
- **Wolves** – Wolves are those users who are particularly successful at imitating others.

They state that "goats have the greatest performance effect", adding a considerable amount of false-negative data. In contrast, the wolves and lambs attribute more to the false-positive data, affecting overall performance. Finally, Yager and Dunstone [69] extended the biometric zoo or menagerie to include different groups of users that cover the extreme ends of the spectrum and explore the existence of the menagerie within other modalities:

- **Worms** – Worms are the worst conceivable users and match poorly against themselves.
- **Chameleons** – Chameleons always appear like others and receive high match scores.
- **Phantoms** – Phantoms always receive low match scores regardless of the comparison template.
- Doves are the best possible users, matching well against themselves and poorly against others.

The framework presented here highlights weaknesses in the system and whether any of these user groups exist, whether within the algorithm itself, the enrolment quality, or data integrity. They conclude that the "biometric menagerie is a diagnostic tool that takes a more user-centric approach".

The biometric menagerie is not without critics. Popescu-Bodorin *et al.* [70] claim the concept is 'fuzzy' regarding whether the categories refer to the users themselves or the templates. Part of their claim highlights that the category of users can change based on the system's calibration. Although this may be true, the biometric menagerie concept is still beneficial for highlighting how users can affect performance or how users can be used to identify potential flaws and weaknesses within a system. In either of these cases, the users directly affect the performance.

The user interaction with the interface and user acceptance of the modality and scenario is a factor of performance that is often not considered. However, if users encounter a bad experience in using the system, it may result in an unwillingness to use the technology on an ongoing basis. As well as examining the environmental impact, Lunerti *et al.* [60] also examined the effect of user interaction with face recognition on smartphones. Through a questionnaire given to the participants, the authors found that users' ease and confidence with the system increased with each session when operated indoors. However, when used outdoors, the confidence remained relatively constant throughout.

As noted previously, research has also shown that a user's physiology can affect the performance of a biometric system, including age [59] and accessibility [71].

Another important factor when discussing how users affect system performance is to examine users' acceptance and satisfaction with biometric systems. These factors will drive the overall performance and indicate the willingness to use such a system. El-Abed *et al.* [72] proposed that "taking into account users' view … is beneficial to the end-users, but it will also help to improve performance and effectiveness". Manufacturers can use the thoughts and opinions of users to influence the design and interface of the biometric system. For example, respondents of a survey conducted by El-Abed *et al.* found that "biometric-based technology is more appropriate than secret-based solutions against fraud" and that the "trust factor has been identified as a major [aspect] that affects their general appreciation". It is also worth noting that the user's culture can influence the acceptance and use of biometrics [13], [73].

It has become clear that users' acceptance and willingness are crucial factors to consider when using a particular biometric system. It is possible to imagine a system with excellent performance results; however, this does not guarantee that the users will be inclined to engage with the system. Thus, usability's effect on performance is becoming an increasingly relevant research topic. Unfortunately, traditional performance metrics have relied only on error rates, which generally do not consider usability concerns.

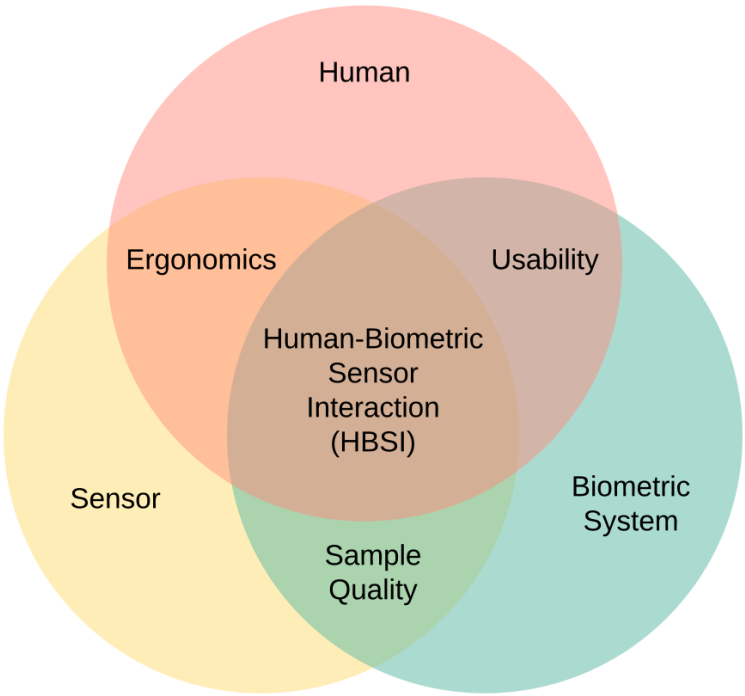



*Figure 3.2: Human Biometric Sensor Interaction (HBSI) Model [74].*

Shown in Figure 3.2 is a thematic outline of the Human-Biometric Sensor Interaction (HBSI) model to address this issue. Guest *et al.* [63] identified that an interaction error might cause performance deterioration of a tuned biometric software system with a biometric capture device. Furthermore, they discovered that when a biometric system is "deployed within a public setting … [the] performance of a system drops, not because of a change in the algorithmic implementation". This discovery points to the

need to include *Algorithms* as a separate factor that can affect performance. Finally, they concluded that in the case of dynamic signature verification, "these problems can be solved through the design of appropriate on-screen user interfaces and hardware", which strengthens the argument for having *System Constraints* and *Hardware* as different factors affecting performance.

Miguel-Hurtado *et al.* [75] assessed voice using the HBSI model for a smartphone's mobile authentication system. Their results concluded that "the learnability of the application needs to be improved by better guidance … thus, better user interfaces and participant guidance within the application have been recommended". They noted that this would improve the overall performance by "avoid[ing] user's assistance requests and reduce the user's errors. Hence, it will help to reduce the number of incorrect presentations and raise the rate of successful enrolments".

Jain *et al.* [44] compared biometric traits using data from the perception of three biometric experts. This comparison showed the acceptability given to the common modalities on a high, medium, and low scale as follows:

- **Face** – High
- **Fingerprint** – Medium
- **Voice** – High
- **Iris** – Low

As mentioned, acceptability is essential, but the concept may have grown into public perception along with privacy and security concerns. Therefore, acceptability is one reason that will prevent the uptake of a biometric system's use and highlights the necessity to consider the user's acceptance before rushing ahead with the implementation.

When assessing the performance of mobile biometrics, it will be necessary to identify the user's familiarity with the device in question. The reason for this is habituation. Users more familiar and comfortable with a device are likely to perform better than someone handling it for the first time. Therefore, the suggestion is that all users have time to adjust and familiarise themselves with the device before any formal testing begins to mitigate any performance impact from different habituation levels from the users involved.

Users are a significant factor affecting biometric systems' performance, so it is little surprise to include them within the core set.

## 3.7 Factor #V: System Constraints

It is necessary to consider that systems often need to meet their own needs and demands for the scenario they require to operate. These needs will include requirements such as:

- Verification or Identification?
- Throughput Rate
- Required Error Rates
- On-Device vs Off-Device
- Time to Enrol or Authenticate
- Privacy Protection / Control
- Latency

On-device refers to the processing of the biometric algorithm occurring on the hardware of the device of the biometric system. In contrast, off-device is where some or all the processing gets delegated to external equipment for processing, most likely a server. Each introduces security concerns that will need exhaustive testing and ties into the privacy protection and control provided.

Privacy protection and control refers to the security for storing all captured samples, even when the authentication algorithm compares samples. Securing these templates is crucial to gain user confidence and ensure the system is not vulnerable to attacks. However, there will be a trade-off between security and performance, as more significant restrictions will generally mean slower functionality. Latency is the delay introduced by transferring data around. This aspect could be an issue for off-device systems where the captured sample needs to be transmitted for processing or providing visual cues to users.

These specific requirements will introduce their constraints and impact the system's performance. For example, whether authentication takes place off-device or on a server, evaluators must note how the performance is affected under different network setups (including Wi-Fi, 4G, and 3G).

Many factors within *System Constraints* link closely to that overall user experience. For example, a system will likely function as intended, but due to the constraints placed on the system by its requirements, it now takes a long time to perform verifications, and then its performance will suffer. Thus, the model defines a *System Constraint* as introducing a constraint on how the system can function. This constraint could be due to many factors, including external (corporate) restrictions, device requirements and scenarios.

Users enjoy convenience and ease, and hopefully, when implemented correctly, this is something that biometrics can offer as a service. A classic use case for biometrics is airport security to process many people as quickly and efficiently as possible. Sasse *et al.* [76] investigated various biometric processes at

various airports. This scenario contains many constraints and requirements that any biometric system must adhere to, thus creating *System Constraints*. The study mentions how a significant factor is how the users react to the system and that any implementation should "emphasise usability's importance in successfully operating biometric systems".

Universal access is also one of the primary requirements associated with an airport border control system. Early tests showed how disabled users struggled to enrol and later verify with the system. The other issue was the experience of the "*bendy shuffle*", as Sasse *et al.* defined the scenario when trying to position the body correctly for the verification sensor. This scenario was due to the interaction being entirely different from that of the enrolment phase caused by the fixed position of the sensors. Thus, there is a hardware issue present here. However, a different verification process from the enrolment has equally caused problems, which was a flawed system design.

While exploring recent advancements in biometric recognition, Ito and Aoki [55] state, "biometric techniques [that are] to be used in the practical system depend heavily on application requirements".

Research has also explored visual feedback and how this affects performance. Visual feedback is where visual cues help guide the user through the biometric process and provide suggestive feedback. As expected, "the better [the] visual feedback, the better performance and usability" was demonstrated by Blanco-Gonzalo *et al.* [17]. However, they also showed that "users do not feel comfortable when [no] visual feedback is provided". Furthermore, while exploring dynamic signature verification (DSV), Blanco-Gonzalo *et al.* found that "latency … involves annoyance in users, and it also affects the performance". Here the latency refers to the digital ink appearing on display, providing a visual aid to the user.

However, latency raises questions about future concerns regarding the performance of systems that require off-device processing and how network latency can cope with data movement, thereby affecting performance.

Besides the dedicated biometric sensor, exploring all the available resources in a system may improve the recognition system. For example, using the available sensors and hardware can provide a continuous authentication mechanism on a smartphone. An example is using the available touchscreen gestures explored by Feng *et al.* [77]. When observing keystroke dynamics on a mobile device, Buschek *et al.* [78] were able to "improve implicit authentication accuracy through new features" available on a smartphone. They were also able to "improve usability with a framework to handle changing hand postures".

It is again necessary to mention *circumvention* from Jain *et al.* [44]. Here measures will explicitly need to be incorporated into a system to prevent security concerns and vulnerabilities. In addition, there will likely be the introduction of trade-offs between having a secure system and error rates along with the time to enrol/authenticate. All of which will mean balancing the performance. It is for these reasons that *System Constraints* belong as a factor.

## 3.8 Factor #VI: Hardware

A biometric system can only be as good as the hardware it has to function on, regarding both the speed and functionality for processing and the resources available to be exploited. The sensors, both dedicated biometric and otherwise, can affect the system's overall performance. For example, Jain *et al.* [79] explored the performance of smartphone touchscreens with traditional hardware keyboards using the same modality of keystroke dynamics. Given that touchscreen sensors "provide considerably richer data", they could exploit this data to generate results demonstrating that "touchscreen data has considerably greater biometric value than that available on hardware keyboards".

Obtaining a detailed hardware description helps better understand performance, especially when considering authentication involving multimodal biometrics. For example, one sensor may capture multiple modalities, or various sensors capture a single trait each.

Elliott *et al.* [59] noted that "various devices used in studies demonstrate different physical and measurement characteristics". This message is still relevant in a mobile context that provides a more excellent range of hardware devices to implement biometric authentication. Developing a testing framework for this purpose will involve considering device variability to ensure consistency.

"Hardware properties can affect the variables collected in the data acquisition process, and therefore the quality and performance of the device" [59]. Likewise, how available sensors collect the acquisition features will also affect the performance of a system. Elliott *et al.* again concluded that "there are significant differences in the variables across devices, yet these variables are not significantly different within device families".

The introduction of biometrics into the mobile market is a relatively recent event, with the first smartphone featuring biometric technology appearing in 2004 [80]. However, significant adoption was due to the introduction of Touch ID into the Apple iPhone series [81].

Hwang and Verbauwhede [82] explored the implications of *portable biometric authentications*. Although this research is from 2004, it is interesting to see how the problems and experiences are still relevant today. They stated that a potential scenario is "financial and commercial transactions as a replacement for (biometric) smart cards" – it is now observed today that a significant end-use for biometrics on smartphones is the introduction of mobile payments, including Apple and Google Pay. Factors also discussed, including where to store the biometric template within the system (when portable), are crucial in the design and overall performance of the final product, and this is one of the critical arguments for the trade-offs surrounding on-device vs off-device:

"Performing the biometric processing on the server provides performance benefits with significant security problems. On the other hand, performing all the biometric processing locally [on the device] provides the best security but requires a relatively larger amount of energy and latency" [82].

This statement concerns examining a system's hardware and the limitations and benefits it provides. It is hard to estimate whether people expected the current advancements in computing and how powerful smartphones have developed. Nevertheless, Moore's law, the observation that "Manufacturers … [have] been doubling the density of components per integrated circuit at regular intervals" (every two years) as surveyed by Schaller [83], has been used as a reliable method for calculating and predicting future trends. Simplistically, this law's application has allowed for higher-performance computers. The same is true for smartphones and the mobile market; with a continuing drive from the industry to exploit hardware resources, current predictions show that Moore's Law is still likely to be accurate until around 2050.

As stated previously, it can be challenging to predict the future outcome of the computing industry. However, the hardware used to support the system will impact the system's performance within the realm of biometric systems. For example, Cantó-Navarro *et al.* [84] developed a floating-point accelerator "specially designed for accelerating biometric recognition algorithms" for embedded systems. They achieved this by exploring accelerating the stages that usually proved the most time-consuming for biometric systems, such as Support Vector Machines, Gaussian Mixture Models and Dynamic Time Warping. As a result, they obtained "acceleration factors ranging from x7 to x22" on two complete biometric algorithms.

Emerging developments in cloud computing and mobile systems have shown that processing can be moved from a mobile device to save on the demand for power consumption and storage capacity by effectively using the cloud. Smartphones today can currently store and run a biometric system without the need to offload resources. However, the same is not valid for mobile IoT devices with limited resources.

Hu *et al.* [85] explored cloud computing and the Internet of Things (IoT) to create a functioning biometric system, as IoT devices typically do not have the same processing and storage capacity as modern smartphones. They created a face identification system that could "meet the growing demands of computation power and storage capacity in the current big data era" by utilising the advantages of cloud computing with the parallel resolution mechanism. "In this scheme, resolution services and identity information management services are deployed in the cloud, which can fully use the high reliability, high scalability, powerful computing and storage capacity of cloud computing to provide efficient and accurate face resolution services". However, they admitted that the system was not without drawbacks, including storing templates in a third-party data centre and the privacy and security associated with the overall system.

The captured sample quality from the biometric sensor must be high enough to satisfy the biometric system. Poor-quality images may result in more false rejects and cause the performance to suffer. Metrics exist to measure sample quality, including Face Image Quality (FIQ) and NIST Fingerprint Image Quality (NFIQ) algorithm.

For all these reasons, hardware is a factor affecting the performance of biometric systems.

## 3.9 Factor #VII: Algorithms

Algorithms are the computational backbone of a biometric system. Recent advances in machine / deep learning have allowed for significant progress in computer vision, speech recognition, natural language processing, and many more. It has, therefore, also found its way into biometric authentication. Conventional machine learning methods, including Support Vector Machines, Principal Component Analysis and Linear Discriminant Analysis, have previously provided the backbone algorithm for biometric systems. However, now with more 'deep learning' approaches discovered, it is likely that a rise in performance will occur as researchers produce more accurate machine learning models.

Taigman *et al.* [86] have experimented with deep learning on a 3D face model and have developed a system called 'DeepFace' which claims to "reduc[e] the error of the current state of the art by more than 27%".

Examining the conventional algorithms, He *et al.* [53] explored the performance comparison of sum rule-based score level fusion and support vector machines (SVM)-based score level fusion for multimodal systems. They discovered that SVM "could attain better performance … provided that the kernel and its parameters [were] carefully selected".

There is a close link between the algorithm and hardware whereby the algorithm must perform on the available device to ensure the performance is usable. This relationship involves carefully considering the amount of memory available to run the algorithm without causing a significant delay for the algorithm users. Cantó-Navarro *et al.* [84] proved this as they achieved higher acceleration performance regarding execution time by altering the hardware components to be more efficient for a biometric system.

Algorithms are designed to produce higher accuracy results and overcome certain environmental factors. Therefore, face image processing is an important research topic. It covers many fields, including computer vision, pattern recognition, image processing and biometrics, as surveyed by Ito and Aoki [55]. Here the authors state, "a variety of face image processing methods has been proposed since the performance of face image processing is significantly influenced by environmental changes such as head pose, expression and illumination changes".

Mobile biometrics provide another challenge for the development of algorithms for those devices. They will need to contend with an array of unconstrained conditions to maintain a high level of operation. Biometrics is an exercise in pattern recognition, and machine learning algorithms have proved extremely useful in this area. Similarly, it has been shown that "machine learning offers several advantages over other approaches for biometric pattern recognition", as discussed by Ortiz *et al.* [87]. At the same time, it also "satisfies an increasing need for security and smarter applications". Similarly, Blanco-Gonzalo *et al.* [17] stated that "the objective of the algorithm is to decide whether the user is the one who claims to be or not". Thus, the whole system's functionality depends on the algorithm, which is why it is a factor.

## 3.10 Modelling Factor Relationships

Figure 3.3 shows the interaction model between the factors. These relationships (connections/links) show an association between factors. The model forms these relations in several ways, such as being constrained or influencing the behaviour of each other.

Many connections demonstrate that an alteration in performance may be caused by several factors discussed here. It is plausible that adjusting to one of these factors could incur a knock-on effect on another. For example, should the hardware be modified, this could cause the algorithm's functionality to change, producing a poor implementation for feature extraction and causing more false positives. Similarly, a relationship may connect more than two nodes.

Figure 3.3 interprets where connections are forming; however, that is not to say that this is a definitive model, and more relationships may exist. The model is a first attempt at forming relationships between the factors, not a comprehensive list. Therefore, there will be missing relationships (links), and others are encouraged to find links and continue to use and adapt the model to form a complete model. The model here begins to present the metrics for reporting the performance of a mobile biometric system, with the connections being some of the key features that a report should include to provide the assurance users need. Table 3.4 provides the relationship definitions (mainly from the Oxford Dictionary [88]).

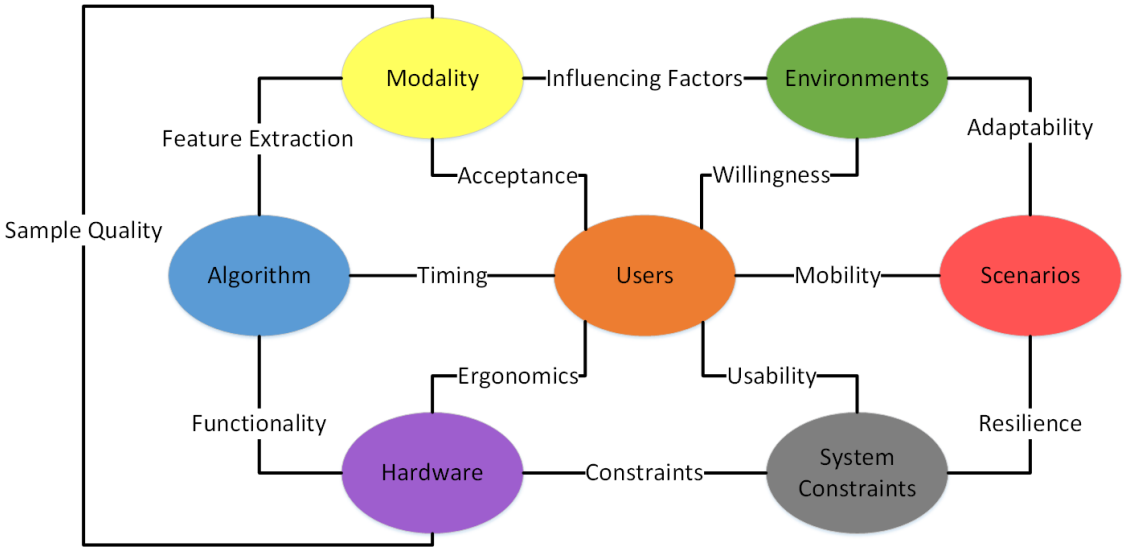



*Figure 3.3: Model Showing the Potential Relationships (Connections) between Factors.*

*Table 3.4: Defining the Relationships Identified within the Model*

| **Relationship** | **Definition** [5], [88] | **Measure** |
|---|---|---|
| Influencing Factors | Any factor that affects the observed performance | Data Binning |
| Adaptability | The quality of being able to adjust to new conditions | Quantitative + Qualitative |
| Resilience | The capacity to recover quickly from difficulties | Qualitative |
| Constraints | A limitation or restriction | Data Binning |
| Functionality | The quality of being suited to serve a purpose well | Quantitative + Qualitative |
| Feature Extraction | The process of extracting information from data intended to be informative | Quantitative |
| Acceptance | Allowing a transaction using a specific modality | Qualitative |
| Willingness | The state of being prepared to operate within a particular environment | Qualitative |
| Mobility | The ability to move freely and easily | Quantitative + Qualitative |
| Usability | The measure of effectiveness, efficiency, and satisfaction | Quantitative + Qualitative |
| Ergonomics | The efficiency of the solution when being operated and handled by the user | Quantitative + Qualitative |
| Timing | The time that is taken by a process, activity, or a person doing it to complete | Quantitative |
| Sample Quality | The fitness of a biometric sample to accomplish the comparison decision | Quantitative |

The connection model highlighted *Users* as one of the most influential factors—an important factor in determining the performance of a system, the model attributes to user satisfaction. However, "the users' satisfaction is often put aside", as highlighted by Blanco-Gonzalo *et al.* [17]. However, its importance is evident as "a non-usable system has not only repercussions in performance but users' acceptance of the technology also".

The aim is that using this approach will generate values that can universally express performance by finding a way of qualitatively or quantitatively defining these relationships. Some identified relationships

already have robust research methods for obtaining a quantitative value, such as retrieving sample qualities of biometric traits discussed in current ISO standards [89]. However, the same is not valid for all the relationships identified here and gathering all this information would be impractical for testing one system. Therefore, the proposal is that a subset of the data will be sufficient. The model theorises that it is possible to build an overall picture of performance by treating each relationship separately and becoming more confident in its value with each newly added piece of information. The aim would be to compare different devices more consistently, but this practicality will require further research.

Table 3.5 begins with suggestions about a possible collection approach to collect the relationship data defined here, along with some suggested basic examples. This table is not a complete list, but it is currently forming some initial ideas that will require further analysis of best methods and practices.

*Table 3.5: Examples of Suggested Methods for Collecting Model Relationship Data*

| Relationship | Collection Suggestion | Explanation | Examples |
|---|---|---|---|
| Influencing Factors | Literature | Explore current influencing factors | [Illumination, Noise, Wearing Glasses] |
| Adaptability | Algorithmic | Measure standard performance rates[1] in various environments and scenarios | While in "Environment 1", FAR increased to 9% |
| Resilience | Algorithmic | Measure standard performance rates[1] across a range of challenging conditions | In challenging conditions, FRR was 34% |
| Constraints | Literature | Explore current hardware that can be supported and usable | [Identification, Off-Device Processing, 2 Seconds to Authenticate] |
| Functionality | Statistical | Explored with analysis of using different algorithms and hardware | "Algorithm 1" achieved 88% successful matches, and "Algorithm 2" achieved 92% |
| Feature Extraction | Algorithmic | The measure of how well the algorithm performs at extracting features | Extraction was able to find a total of 8 total features |
| Acceptance | Questionnaire | Survey of users | The survey revealed that 80% of users would allow a transaction to happen with chosen modality |
| Willingness | Questionnaire | Survey of users | Only 20% of surveyed users would be happy to use this verification method in the chosen environment |
| Mobility | Statistical | Explored with analysis of performance in motion scenarios | While in motion, the FRR was 13% |
| Usability | Questionnaire + Interaction | Survey of users and interaction measures | 74% of users were satisfied, and it took 10 seconds to read each of the task prompts |
| Ergonomics | Questionnaire + Interaction | Survey of users and interaction measures | 67% of users were comfortable and managed to complete the task within 35 seconds |
| Timing | Experimental | The device in operation should be capable of capturing timings | Authentication took an average of 7 seconds to complete |
| Sample Quality | Algorithmic | The measurement of Sample quality to ISO and similar standards for some modalities | The sample quality score achieved was 81 |

[1] Standard Performance Rates = FNMR, FMR, FTA, FTE

## 3.11 Discussion

As noted, these seven factors have significant overlap with one another. For example, algorithms will require a particular hardware setup to function as expected. Also, the willingness of users to engage with a system will be profoundly affected by the modality and environment used. El-Abed *et al.* [72] stated that "evaluating biometric systems constitutes one of the main challenges in this research field". They also conclude that "the main drawback of the widespread use of biometric technology is the lack of a generic evaluation methodology that evaluates biometric systems taking into account: performance, users' acceptance and satisfaction, data quality and security aspects".

Each defined factor could easily extend and expand to incorporate more detail. However, the aim here is to highlight the central concepts that create a foundation acting as a *parent node* to use tree terminology. For example, the model identifies *Users* as a core factor, but many subsections will occur, including interaction and acceptance. Both are significant areas that could arguably be a factor of their own, but the model will still identify *Users* as the central area incorporating these.

Interestingly, some of the factors identified fit directly into the HBSI model proposed by Elliott *et al.* [59]. Here, Human, Sensor and Biometric Systems become *Users*, *Hardware*, and *System Constraints*. In identifying factors that affect performance differently, it is reassuring to note that the HBSI model is still present and preserved within this updated model, which goes beyond usability aspects to define performance.

Comparing the factors presented here with the ones provided initially by Mansfield and Wayman [35] and included within the ISO/IEC 19795-1:2006 [22], they are similar.

- Population Demographics ↔ *Users*
- Application ↔ *System Constraints*
- User Physiology ↔ *Users*
- User Behaviour ↔ *Users*
- User Appearance ↔ *Users*
- Environmental Influences ↔ *Environments*
- Sensor and Hardware ↔ *Hardware*
- User Interface ↔ *System Constraints*

Mansfield and Wayman's factors only cover four of the seven factors presented here. However, it confirms that users are one of the most influential factors. Furthermore, it demonstrates how the *Users* consideration can extend into subsections incorporating what Mansfield and Wayman have previously identified. The factors added are *Modality*, *Diversity of Scenarios* and *Algorithms,* and these are the core factors for mobile biometric performance.

Looking back at Jain *et al.* [44] seven factors for assessing biometric traits, the model takes influence from them in terms of *performance*, *acceptability* and *circumvention*. The model incorporates *acceptability* within the *Users* factor and *circumvention* within the *System Constraints* factor. The definition of *performance* talks about "accuracy, speed, and robustness," which are fundamental concepts. However, there is a need to update and incorporate the other factors presented here to provide more assurance for a mobile context.

## 3.12 Summary

This chapter has identified seven core factors that form binning categories of performance alterations in mobile biometric systems. These factors are Users, Modality, Environments, Diversity of Scenarios, System Constraints, Hardware, and Algorithms. In defining the factors, an informative overview was provided to developers, implementers, and testers of biometrics systems, enabling the binning of performance alterations within one of these factors. The model expects categorical overlaps, so a performance alteration will likely have many factors contributing to the observed effect.

Research is shifting to accommodate this change from a fixed to a more mobile system and exploring new opportunities and situations for mobile biometrics. Hopefully, the identified factors will help pave the way for future research to focus on some of these critical areas and allow future biometric systems to have a high level of performance and assurance that it fits their purpose. This work is the first step in designing a suitable framework to assess performance. The next chapter will combine all the current theories into a potential mobile biometric performance framework.

# 4 Towards A Flexible Performance Assessment Framework

## 4.1 Introduction

This chapter aims to develop a performance assessment framework for mobile biometrics and finds ways to suitably measure and mitigate the effects of the core factors identified in Chapter 3. However, testing every usage outcome within a mobile context is impossible. Therefore, the requirement is to develop an approach providing confidence that the system *fits its purpose*.

Adopting this change will likely mean modifying the current testing strategy that the community is familiar with within specific areas. The changes will require thorough testing that includes more available situations in a mobile context. Other metrics for performance will need to be added to the standard testing procedures, including measuring usability from the HBSI model and data quality. These changes will bring more confidence into results from biometrics studies and allow users to feel more comfortable interacting with a biometric system. Researchers will need to conduct more research to identify the quality of biometric samples under various conditions with more significant influence given to the collection environment.

This chapter will present a proposal for an evaluation framework to establish the suitability of mobile biometric systems for applications. Chapter 2 showcased the existing methodologies and approaches, and Chapter 3 identified the core factors known to affect mobile biometric performance. Finally, the concepts are amalgamated to provide mobile devices with an extensive and extendable performance testing framework. The framework is designed to be tuneable to best suit the needs of the end-users and evaluators by providing a more comprehensive testing structure for various devices with more meaningful results.

The organisation of this chapter is as follows: Section 4.2 provides some information relating to the approach taken when producing the guidance for the performance framework, followed by Section 4.3 to introduces the proposed seven-stage mobile biometric performance framework. Section 4.4 discusses the various evaluation approaches based on the identified access level to the device. Section 4.5 showcases the framework as a flow diagram. Section 4.6 discusses usability requirements, and Section 4.7 introduces and explains the approach to non-mated testing using a ‘Tailored Impostor’ approach. Finally, Section 4.8 and Section 4.9 compares the proposed framework and summarises the chapter.

## 4.2 Performance Framework Approach

There is generally considered a trade-off between usability and security [90], and finding an appropriate balance depends on what would be considered acceptable for any given situation. There are two crucial issues to consider here: firstly, performing exhaustive testing is unfeasible, and secondly, what is considered an acceptable amount of testing will vary for each use case. The aim here is to introduce an approach to a testing framework that will hopefully provide some confidence and trustworthiness in a mobile system under evaluation to ensure that it fits its purpose. The proposed framework acknowledges that testing every possible scenario and eventuality may be impossible. However, the intention is to create a process to certify that the device fits its purpose.

The question then becomes how to determine if a device is fit for purpose, and the answer here is by taking existing theories and approaches from similar disciplines. The basis of the proposed approach is utilising 'worst-case analysis' from system design and 'boundary value analysis' from software testing with concepts of 'equivalence class partitioning (ECP)'.

Boundary value analysis is concerned with using input values (test cases) around the boundary of and including the expected values and the extreme cases (minimum and maximum) values [91]. Finally, the worst-case analysis uses statistical analysis to identify the worst possible input parameters' values and then analyse whether the system meets its specifications under such circumstances. "Worst-case analysis (or testing) is most useful when you have a few parameters that are known to be troublesome" [92].

Biometric testing is not the same as system and software testing primarily due to the incorporation of people, as the test space for a mobile biometric system is not comparable to a software system. However, the above ideas can still be applied and could help achieve a testing approach that scales down the amount of testing required by not relying on a random testing approach. The framework applies boundary value analysis and worst-case testing ideas to formulate an approach to scenario testing that allows evaluators to examine the 'troublesome' scenarios. These scenarios are known as influencing factors (independent variables and experimental conditions), defined as "user and environmental factors that have been found to affect performance" [35]. This approach would demonstrate how the system will perform at its worst and provide a guideline for a more informative testing regime.

Similarly, the framework proposes an alternative approach regarding non-mated testing, also known as a biometric non-mated comparison trial [5]. Implementing a technique based on equivalence class partitioning (ECP), and more specifically 'edge testing' described as a "hybrid of boundary value analysis and equivalence class testing" [91], to identify the "faults near the boundaries of the classes, and edge testing will exercise these potential faults". In addition, the approach is to consider the use of 'tailored' impostors to force the system into showing its worst False Match Rate when exposed to 'doppelgangers' (similar looking people).

The framework does not provide recommendations for sample sizes partly because this will not affect the process of using the proposed performance evaluation framework. The “rule of 3” and “rule of 30” are required for statically significant results. The framework does present an alternative to non-mated comparison testing, which is hoped could provide a method for reducing the required sample size and still provide meaningful results.

## 4.3 Performance Evaluation Framework

This chapter outlines the design of the seven-stage performance assessment framework. The framework presents a complete performance evaluation, and the design of each stage is to operate as a continuous flow from one stage to the next. However, not every stage must be completed (but it should be for a complete system evaluation), which means the framework can adapt to individual requirements. However, the approaches adopted in each stage will help execute a meaningful evaluation.

The framework stages relate to a particular outcome, allowing for a selective approach using only the required stages. For example, evaluators whose concerns are only related to scenario performance will only need to follow the guidelines specified in Stage One, Stage Three and Stage Four (plus reporting in Stage Seven). This outcome generally relates to one or more of the core reporting requirements: (Scenario) Recognition Performance, (Operational) Recognition Performance, Usability, Spoofability and Presentation Attack Detection (PAD) and Privacy.

### 4.3.1 Stage One – Determine Evaluation Parameters

The initial stage is to set up the evaluation parameters to determine how the evaluators should run the evaluation. Determining the best approach to provide the most meaningful outcome requires understanding the requirements. An interview may be needed to gather this information and determine the requirements for the evaluation between the key stakeholders, including users, manufacturers, and evaluators. The framework includes three parameters to set up the testing process and aid: modality or modalities under examination, the level of access, and the desired security level. The joint outcome from the stakeholders should decide upon these three parameters and the system's desirable scenario and operational conditions. This information-gathering phase is similar to Mansfield and Wayman’s “Determine Information About You System” [35]. The questions asked should include the following:

- What is the modality or modalities under evaluation?
- How much access does the evaluator have to the system?
- What level of performance (security level) are you hoping to achieve?
- What would be the appropriate and suitable scenario conditions?
- What are the operational conditions for the system?

A similar approach will involve looking at requirements from a use case classification point of view to determine the risk analysis when defining the evaluation criteria and parameters; such examples may include:

- A low classification where a commodity approach consistent with current implementations is acceptable.
- A low classification includes specific requirements and sensitivities, such as cases involving VIPs where individuals are more likely to be specifically targeted. The evaluators would want to protect against a lower-end attacker.
- A high classification where often the environment is simple, but spoofing protection is a higher requirement due to more skilled attackers.
- A high classification under challenging environments where the spoofing requirement is still vital must operate in unpredictable locations.

The first parameter to consider is the modality or modalities in a multimodal system. The modality will influence what aspects to test by considering and exploiting known influencing factors and spoofs. The testing and spoof attack process will depend on whether the evaluators deal with a physiological or behavioural biometric. The ambient environmental conditions and the necessary presentation attack instrument (PAI) will need adjusting.

Fernandez-Saavedra *et al.* [21] recommend evaluators "analyse the biometric functions and methods provided by the mobile device to know the restrictions of the evaluations in advance". The framework encompasses this recommendation within the evaluation. Therefore, to provide a universally acceptable testing framework, the evaluators must establish what functions and methods (access) they have to the device.

The framework introduces a tiered system to indicate the access level, as shown in Section 4.7.

*Table 4.1: Evaluators' Levels of Access to a Device*

| Level of Access | Aspects |
|---|---|
| **Closed (Blackbox)** | • No internal access |
| **User** | • System-controlled biometric functions<br>• Impossibility of obtaining captured biometric samples<br>• The result of the authentication is a Pass/Fail decision |
| **Developer** | • Restricted biometric functions<br>• Ability to create a custom application with biometric API<br>• Access to logging capabilities<br>• Impossibility of obtaining captured biometric samples<br>• The result of the authentication is a Pass/Fail decision |
| **Tester** | • Greater access to biometrics functions<br>• Offline access to biometric functions<br>• Ability to obtain a captured biometric sample<br>• The result of the authentication is a match score (with a known system decision threshold) |
| **Open (Whitebox)** | • Full internal access<br>• Algorithm and source code access |

The framework will adapt to the level of access determined by the evaluator. The closed (Blackbox) level showcases the extremes in access level. The framework deems this Blackbox level as only theoretical, implying that the evaluators obtain no indication of the biometric decision outcome. Therefore, this access case is untestable.

Any system that indicates the biometric decision, such as unlocking a smartphone or opening e-Gates at borders upon presenting a passport and the user's face, is considered testable. The same logic applies here, and any system with an access level above a Blackbox is considered testable. Again, however, access to the system will impact the extent of the testing.

An evaluator can assess a device with functionality that crosses the aspects of the levels defined in the framework. For example, a commercial device where the evaluators can obtain match scores. In those instances, the evaluator will need to adapt the framework to determine the best approaches based on the framework's guidance. Ideally, an evaluator should aim for a high level (open) of access for the evaluation until external restrictions force a downgrade in access.

The second evaluation parameter is the desired security level, which could relate to the desired performance level the evaluators are looking to achieve or a claimed performance level from the manufacturer that needs proving. Finally, a security level is assigned, taking inspiration from the 'Tiered Authentication' classes provided by the Google Android standard [93].

The framework defines three classes of security levels: Convenience, Balanced and Secure. The 'convenience' level is associated with a greater emphasis on usability (low false non-match rate). Whereas a 'secure' level considers the overall security more critical, potentially at the cost of usability (having a low false match rate), a 'balanced' level aims to achieve an equal level between usability and security. Table 4.2 shows the concept of this tiered authentication system.

*Table 4.2: Security Levels*

| Security Levels | | |
|---|---|---|
| **Convenience** | **Balanced** | **Secure** |
| • High Usability<br>• Low Security<br>• Low FRR | • Medium Usability<br>• Medium Security | • Low Usability<br>• High Security<br>• Low FAR |
| • Low Sophisticated Spoof Attacks<br>• Zero Effort Attacks | • Medium Sophisticated Spoof Attacks | • High Sophisticated Spoof Attacks |

The framework identified a link between access, the desired security level, and what would be realistically achievable. For example, it would be deemed unfeasible with a 'user' access device to evaluate to a 'secure' level primarily due to a lack of offline testing capabilities. Ultimately the decision will come down to how much time and resources the evaluator will spend on the device. However, as a guiding rule of thumb, the framework foresees the maximum security obtainable by access level as:

- Closed (Blackbox) → None
- User → Convenience
- Developer → Balanced
- Tester → Secure
- Open (Whitebox) → Secure

The framework includes three evaluator parameters: the level of access defined in Table 4.1, the desired security level highlighted in Table 4.2, and the modality or modalities involved. These three pieces of information establish the framework parameters, and the idea is to view their contribution as follows:

- **Modality** = What the evaluator(s) assess
- **Level of Access** = How the evaluator(s) assess
- **Security Level** = Why the evaluator(s) assess

### 4.3.2 Stage Two: Algorithmic Evaluation

An algorithmic evaluation tests the performance of the biometric algorithm, generally by using existing or pre-collected data. The approach here is the same as the Technology Evaluation defined in ISO/IEC 19795-1 [6], an "offline evaluation of one or more algorithms for the same biometric modality using a pre-existing or especially collected corpus of samples". However, an algorithmic evaluation would only be possible if the access level was that of a 'Tester' or higher and would allow for data injection into the system.

Academic research prioritises algorithmic evaluations to create new and enhance existing algorithms [94]. Using customised or existing datasets, researchers can mimic expected performance. For example, using these samples, they can evaluate False Non-Match Rates and False Match Rates, resulting in an Equal Error Rate calculation. Although there is no general rule or standard comparing performance between algorithms, the current approach in the literature seems to be using Equal Error Rates and Receiver Operating Characteristic (ROC) curves to showcase a low Equal Error Rate.

"ISO/IEC 19795-1 [22] describes three biometric performance evaluations: technology, scenario and operational evaluations. ISO/IEC TR 30125 recommends scenario evaluation as the most proper evaluation for testing biometric performance on mobile devices" [38]. ISO recommends that a technology evaluation is inappropriate for mobile biometrics as it does not provide sufficient information about its operational performance. The framework acknowledges this recommendation; however, performing an algorithmic evaluation can be an excellent early indicator of the performance outcome. An algorithmic evaluation is appropriate for catching early warnings before continuing with an expensive evaluation when data is available, and the system offers offline testing capabilities. For this reason, the framework does not include the results of this algorithmic evaluation as part of the main reporting stage, owing to the lack of operational performance information.

Stage two exists to assess the performance outcome of an algorithm before undertaking extensive evaluations. The aim is to avoid a costly evaluation process for a system with identifiable, early issues from an algorithmic evaluation, such as an unacceptable false match rate. A possible future extension to this section would be utilising the knowledge of statistical models to predict performance without the need to carry out extensive testing incorporating the work of the currently under development ISO standard "Biometric performance estimation methodologies using statistical model" [95]. However, even with statistical models to improve algorithmic evaluations, it is likely that the requirements to perform a practical evaluation will remain.

This stage's outcome will provide the evaluators with performance scores consisting of a false non-match rate and a false match rate. The evaluators should then decide if the obtained scores meet the system's required specifications sufficiently to allow the evaluation to continue or if further work is required to improve the algorithm's performance.

### 4.3.3 Stage Three: Perform Baseline Evaluations

Having established that the algorithm works using existing data, the next step is to test the system under various scenarios in a scenario evaluation. The baseline scenario is where the evaluators create their comparable benchmark for comparisons between scenarios highlighting performance alterations by creating the most controlled and 'optimal' conditions for the mobile biometric system to operate. The framework uses this scenario for two aspects, enrolling the modality and the initial verifications.

The baseline scenario conditions take inspiration from the conditions specified in ISO/IEC TS 19795-9:2019 [37] as indoor conditions with no noise with the device handheld by the user. The framework extends this to specify that the user should be seated (if appropriate). The enrolment conditions can harm overall system performance. For example, if the captured enrolment template's quality is considered low, it will cause an impact on subsequent comparisons. Modi *et al.* showcased this by looking at the effect of image quality and age on biometric performance. They concluded by noting how the "removal of lower quality images from … the datasets showed that the number of false non-matches decreased, which shows that performance of the system can be significantly improved by removing images of lower quality" [96]. However, as the enrolment process is generally considered a one-time action, it seems reasonable to perform the rest of the evaluation against optimal high-quality enrolment templates.

Each stage of this framework aims to expose problems with the system under evaluation. For example, suppose the evaluators discover performance issues affecting baseline evaluations. In that case, the evaluators should stop the evaluation process. This pause would allow the developers to make the necessary changes to the system and algorithm to mitigate the performance defects before continuing. This stage's outcome should ensure that the proposed system works and will be used to gain the necessary baseline performance relating to recognition accuracy, the false rejection rate, and the false acceptance rate for comparison against the scenario and operational testing. For 'Developer' access and lower, this is the main stage that involves non-mated comparisons because of the high cost involved in performing large-scale, non-mated testing.

Each enrolled user should perform five verifications against their stored template for five attempts per verification as specified in ISO/IEC TS 19795-9:2019 [37]. Next should follow the non-mated comparisons. The framework employs a novel approach to non-mated testing called 'Tailored Impostors' or 'Tailoring' as a counter approach to the traditional random selection process for the methodology approach here. The 'Tailored Impostor' theory is discussed further in Section 0 below. In addition, the effect of a user's 'habitation', defined here as the user's familiarity with the operations and control over a device, and previous experience towards mobile technology could impact the obtained performance. Therefore, the suggestion is that the first verification is used as a 'habitation' to allow the user to get comfortable, meaning six verifications are necessary.

Where the system allows (i.e., the level of access is that of a 'Tester' or more), the evaluators should collect the biometric samples and store them securely for offline non-mated comparison testing. For systems with lower access levels than 'Tester', the suggestion is to conduct online non-mated comparison testing. The recommendation is to utilise the presented 'Tailored' impostor methodology for both methods

The framework encourages the collection of usability metrics, especially when extensive scale evaluations are impossible, and the desired security level aims for 'convenience'. Suggested usability requirements are discussed further in Section 4.6. The evaluators should report the results obtained from the false non-match rate (FNMR) and false match rate (FMR).

### 4.3.4 Stage Four: Targeted Scenario Evaluations

The evaluators have established that the algorithm works, enrolment can capture features, and the system performs optimally. The framework enters the entire scenario evaluation stage, defined as an "evaluation in which the end-to-end system performance is determined in a prototype or simulated application" [22]. The recommendation is to target specific environmental conditions and weaknesses of the system and limit generic scenarios, defined as those unlikely to impact performance. However, the evaluators should consider selecting scenarios representing standard system use cases, taking inspiration from boundary value analysis by including typical data for the system. The number of evaluation scenarios will vary based on the system under evaluation. It will partially be down to the evaluators to ensure they have explored a diverse range.

As this framework aims to capture mobile biometric performance, the recommendation is that at least one of the scenarios involves the motion of the user while holding the device, and another explores a known influencing factor for the modality. The suggested scenario is walking in a straight line (track) at a steady pace that is comfortable for the user indoors with no noise, distractions, or obstacles. At the same time, they operate the biometric subsystem of the device in their hands. Ideally, the tasks should be achievable before the user runs out of the track; otherwise, they should turn around and walk back down along the same track they came.

The evaluators should base the remaining scenarios on known weaknesses in the literature and desirable use cases for the system. Table 4.3 shows which ambient factor affects each biometric modality specified in ISO/IEC TR 19795- 3 [97]. The ambient factors displayed here represent examples of environmental influencing factors.

*Table 4.3: Ambient factors that affect each biometric modality [97]*

| Ambient Factor | Biometric Modalities |
|---|---|
| Temperature | Face, Fingerprint, Vascular, Voice and Hand Geometry |
| Humidity | Face, Fingerprint, Vascular, Voice and Hand Geometry |
| Illumination | Face, Vascular, Iris and Fingerprint (optical sensors) |
| Noise | Voice and audio guides for all modalities |
| Pressure | Signature |

Precisely defining what scenarios are worth testing will come down to an agreed specification between the stakeholders, including users, manufacturers and evaluators, based on the criteria defined in stage one. The framework utilises the identified 'Core Factors' [98] to evaluate conditions that affect the system's performance. For example, if the evaluators were testing a facial recognition system embedded into a smartphone, one possible suggestion would likely be to look at evaluating the following scenarios:

- Walking (Motion)
- Standing
- Transportation (Motion)
- High Illumination
- Low Illumination
- Distance to Camera
- Field of View

Here, the chosen scenarios include some common, practical use cases (standing), along with some motion-based scenarios (walking and transportation), some influencing factors (high illumination, low illumination) and some hardware-specific considerations (distance to camera, field of view).

Likewise, to the evaluation approach conducted for the baseline evaluation, each user should perform a minimum of five verification transactions with a maximum of five attempts allowed per transaction for every chosen scenario. If allowed, the evaluators should conduct offline non-mated testing (i.e., if the access level is 'Tester' or more). Subject to resource constraints, it may be recommended not to proceed with online non-mated testing for these scenarios.

### 4.3.5 Stage Five: Presentation Attack Detection and Architectural Security

At this point, the evaluators have hopefully addressed performance regarding scenario recognition accuracy and usability. At this stage, the aim is to address cybersecurity and privacy concerns with the system. Spoofability refers to vulnerability to presentation attack detection (PAD). It involves defending

the system against attacks, usually at the sensor level, to trick the system into accepting the spoof as a mated probe.

The framework incorporates the FIDO Alliance's work in PAD criteria within the FIDO Biometrics Requirements documentation [99]. Attacks are generally ranked based on their sophistication and execution time. The desired security level will determine the recommendations for what level of sophisticated attacks to consider. For example, suppose the evaluators are aiming for usability over security. In that case, the evaluators will test their system against low sophisticated attacks, for example, a 2D paper-based mask for a facial recognition system and replaying a voice recording for a voice recognition system.

The framework utilises the same criteria as defined in ISO/IEC 30107-3:2017 [41] and the FIDO Alliance [99] to define the sophistication of an attack. The sophistication is also known as the attack potential, which is the "measure of the capability to attack a [Target of Evaluation] TOE given the attacker's knowledge, proficiency, resources, and motivation". The framework mirrors FIDO's simplified three-level approach to classifying presentation attack instruments (PAI) into levels based on time, expertise, and equipment. These levels will inform the framework's knowledge regarding PAD sophistication. For example, a common sophisticated approach would come under FIDO's Level A. A highly sophisticated approach would come under Level C, meaning the medium-level approaches would comprise Level B. Table 4.4 provides examples of these levels of sophistication.

*Table 4.4: Spoof presentation attack examples separated by levels [99]*

| | **Fingerprint** | **Face** | **Iris** | **Voice** |
|---|---|---|---|---|
| **Level A (Laymen)** | A paper printout | A paper printout of a face image | A paper printout | A replay of the audio recording |
| **Level B (Proficient)** | Fingerprints made from artificial materials | A paper mask | A video display of an iris | A replay of an audio recording of a specific passphrase |
| **Level C (Expert)** | A 3D-printed spoof | A silicon mask | A contact lens/prosthetic eye with a specific pattern | A voice synthesiser |

FIDO requires 12 PAI, six from level A and eight from level B, using 15 participants. The framework mirrors these requirements but alters them to match the desired security levels to best match the evaluator's needs. The recommendations are:

- Convenience – 10 Level A, 4 Level B
- Balanced – 6 Level A, 8 Level B (FIDO)

- Secure – 3 Level A, 7 Level B, 4 Level C

The 14 PAIs combined with the 15 participants resulted in 210 transactions for PAD testing. The evaluators should report the Impostor Attack Presentation Accept Rate (IAPAR).

Architectural security is concerned with privacy and resilience against compromises. Android defines architectural security as "how resilient a biometric pipeline is against kernel or platform compromise. A pipeline is considered secure if kernel and platform compromises do not confer the ability to read raw biometric data or inject synthetic data into the pipeline to influence the authentication decision" [93]. It will be challenging to thoroughly assess the architectural security independently unless the level of access is close to that of 'open' (Whitebox). Otherwise, the evaluation will rely on acknowledging the published claims of the manufacturer.

The evaluators can look at the employed underlying algorithm and encryption techniques when access allows. This task will uncover any personal or sensitive biometric data leakage at any process stage. It is expected for commercial and functioning biometric systems to inhibit unauthorised tampering and exposure of sensitive biometric data. In other words, when the system is fully functional, it should be as locked down as possible, and the level of access restricted to that of a developer or below. For this proposed framework, architectural security also relates to template protection. Template protection helps to ensure the irreversibility [100] and unlikability (cancellable biometrics) [43], [101] of the biometric data.

PAD also incorporates liveness detection to check whether the authentication probe belongs to an 'alive' being. Part of PAD testing aims to confirm whether the system can distinguish between' alive' actors. Ideally, a system should not accept a probe that is not live. For example, face recognition could be achieved by capturing a live video feed instead of a static image and checking for movement within the captured sample [102]. The report should include details of any liveness detection components and the corresponding Impostor Attack Presentation Accept Rate (IAPAR).

Based on the criteria defined in stage one, the general requirements will be to meet a set of standards as agreed upon and defined between the stakeholders, including users, manufacturers, and evaluators, based on the criteria defined in stage one to establish what behaviour is appropriate for the given system. A commercial application will need to fulfil specific international standards, most notably the EU General Data Protection Regulation (GDPR) [103], [104]. The GDPR classes biometrics as sensitive personal data requiring additional user permissions to obtain, store and process. As an extension, the developers should guarantee the ethical use of the data. Whiskerd *et al.* [42] identified privacy-sensitive attributes from biometric data. The attributes include race, gender, language, nationality, age, and sexual orientation.

Under the heading of 'privacy' are biometric template protection schemes such as cancellable and privacy-preserving biometrics. Under operational conditions, access to stored references and probes

during authentication should be restricted and unobtainable. Cancellable biometrics refers to a process where “the biometric template of a person is distorted in such a manner that the original data is not available to the intruder, but still identity recognition can be performed” [101] (also referred to as feature transformation).

Biometric encryption (cryptosystems) “securely bind a digital key to a biometric or generate a digital key from the biometric so that no biometric image or template is stored. Therefore, it must be computationally difficult to retrieve either the key or the biometric from the stored biometric encryption template” [105] (also referred to as data-based helper schemes). Both cancellable and cryptosystems meet some of ISO/IEC 24745:2011 biometric information protection [106].

It is required that privacy and template protection be employed in any system that operates commercially or in the wild. Therefore, the final report should detail all the techniques from the evaluators and manufacturers.

### 4.3.6 Stage Six: Operational Evaluations

This stage allows evaluators to test the system under operational conditions and observe any performance alterations from the previously achieved scenario performance. Operational testing aims to allow the system to run (operate) as intended and measure errors. ISO/IEC 19795-1:2006 defines operational evaluation as an “evaluation in which a complete biometric system’s performance is determined in a specific application environment with a specific target population” [22].

This stage will vary by the system’s requirements under evaluation, and the operational conditions will need to be set based on the stakeholder criteria defined in Stage One. However, the suggestion is that the evaluators incorporate some outdoor scenarios into this stage (if appropriate). The main objective is to create an experiment with conditions outside the evaluator’s control (e.g., weather, terrain). For example, one possible suggestion for a smartphone system would be to devise a walking route or trail for the participant to follow. At designated locations, the user should perform an authentication matching the exact requirements before, a minimum of five transactions for each location with a maximum of five attempts allowed per transaction.

### 4.3.7 Stage Seven: (Final) Reporting

While reporting should be occurring on an ongoing basis, this stage comprises the final report. The final report will cover the findings throughout the process and, more specifically, will divide the performance into distinct categories:

- Baseline Recognition Accuracy

- (Scenario) Recognition Accuracy
- (Operational) Recognition Accuracy
- Usability
- Presentation Attack Detection and Spoofability
- Privacy and Architectural Security

The reporting requirements will closely follow the requirements defined by ISO/IEC 19795-1:2006 [22] and FIDO [99], including the participant demographics. The general and minimum recognition accuracy includes the false reject and accept rates (baseline). This framework only requires the false match rate for the baseline recognition accuracy. The access level does not allow offline testing (where the access level is less than 'Tester').

The scenario and operational evaluations should include a detailed description of the scenario involved and the recognition accuracy achieved. The usability should contain a combination of satisfaction, efficiency, effectiveness and performance-related usability metrics as covered within the Human-Biometric Sensor Interaction (HBSI) model. Further details are covered in Section 4.6. The Spoofability and PAD will contain a PAI list given the security level and the obtained Impostor Attack Presentation Accept Rate (IAPAR). The privacy requirements will form a report featuring the techniques implemented to protect the user's privacy and avoid data leaks explaining any cancellable and cryptographic techniques employed or incorporating a published manufacturer report detailing the techniques used.

The reporting approach aims to simplify how biometric performance results are presented and potentially allow for a more functional approach rather than quoting one figure to cover the entire depth and breadth of what the performance will likely be. In addition, by showcasing the baseline, everyone can see what the system's optimal performance is likely to be, allowing for comparisons between the targeted scenarios and operational conditions.

Ideally, one aim was to incorporate a colour-coded ranking system for each category defined to allow easier comparisons between systems in a user-friendly approach. However, this approach was discarded due to no formal agreements as to what constitutes "good" or "bad" performance rates, as shown in Chapter 2, with a range of performance rates and requirements. Equally, having a ranking system would potentially introduce the problem of assuming one system is superior. The intention is to allow users to decide the system they want to use. However, this concept should be revisited and perhaps incorporate a unique ranking rate for each desired security level.

## 4.4 Evaluation Approach

The above stages have walked through the approach that a complete system evaluation should include for a thorough analysis. The specific evaluation parameters (the modality, level of access, and security level) will impact or impede the approach taken when assessing the performance of a biometric system.

The approach to conducting the evaluations will depend on the level of access defined in stage one. As the access level increases, the evaluation moves from a manual process to a more automated one. For example, the process will be manual for a 'User' level, with the evaluator making notes of successful and unsuccessful authentications and user interaction errors. It would generally be unfeasible to do large-scale impostor testing with this level of access [21].

The process becomes more automated when moved into the 'Developer' access realm. Here the evaluators can use specific tools that the manufacturers allow to aid in testing. The biggest will likely be the ability to call the biometric authentication components and log analysis. This ability would allow for developing a custom application for testing purposes. For 'Tester' and higher, the assumption is that a pre-made testing application can access the biometric functions for both online and offline use.

*Table 4.5: Testing approach associated with the level of access*

| Level of Access | Approach |
|---|---|
| **Closed (Blackbox)** | • Untestable |
| **User** | • System defined |
| **Developer** | • Custom application to trigger the system's biometric API functions<br>• Logging capabilities |
| **Tester** | • Pre-made application with access to unlimited attempts<br>• Offline access for impostor testing |
| **Open (Whitebox)** | • Same as a tester with added benefits of looking into greater privacy and architectural security concerns. |

Table 4.5 summarises these approaches and shows that with 'User' and 'Developer' access, the approach taken is greatly impeded by what access the manufacturer provides. Although testing the system entirely with this lower level of access, other approaches may be adaptable to indicate likely performance. For example, the evaluators can use a device's camera to collect facial images through third-party face recognition software to see likely performance and quality information using their camera for such purposes. Similarly, voice samples could be recorded from a device's microphone to explore voice recognition.

## 4.5 Framework Flowchart

Figure 4.1 shows a high-level flow diagram demonstrating the path through the performance evaluation framework corresponding to the available access level. It starts with defining the three parameters that drive the performance assessment framework. Then, based on the level of access provided, it determines whether an algorithmic evaluation is plausible and directs the flow accordingly. Afterwards, the baseline evaluation is performed. Then, the process splits, again, based on the level of access, ranging from a more manual 'user' process to a more automated 'tester' process.

Once the process in which to run the evaluation is determined, the process again splits based on the modality the system uses. Using known influencing factors affecting the modality means that the targeted scenario evaluation phase can begin targeting a range of scenarios involving at least one motion-based scenario. Next, we can begin to examine the presentation attack detection phase and, using the desired security level, can aim the sophistication of attacks to that level. All that is left is the operational evaluation now, knowing that, hopefully, all the previous stages are performing at a level expected by the evaluator.

So, the processes defined within the framework are guidelines, and the evaluator may have alternative ways to evaluate the device's performance with the level of access and security level defined. Although this is the recommended approach, it is not static, and evaluators are encouraged to use judgment against the performance framework developed here to identify their best approach. The aim is to help in producing a suitable evaluation framework.

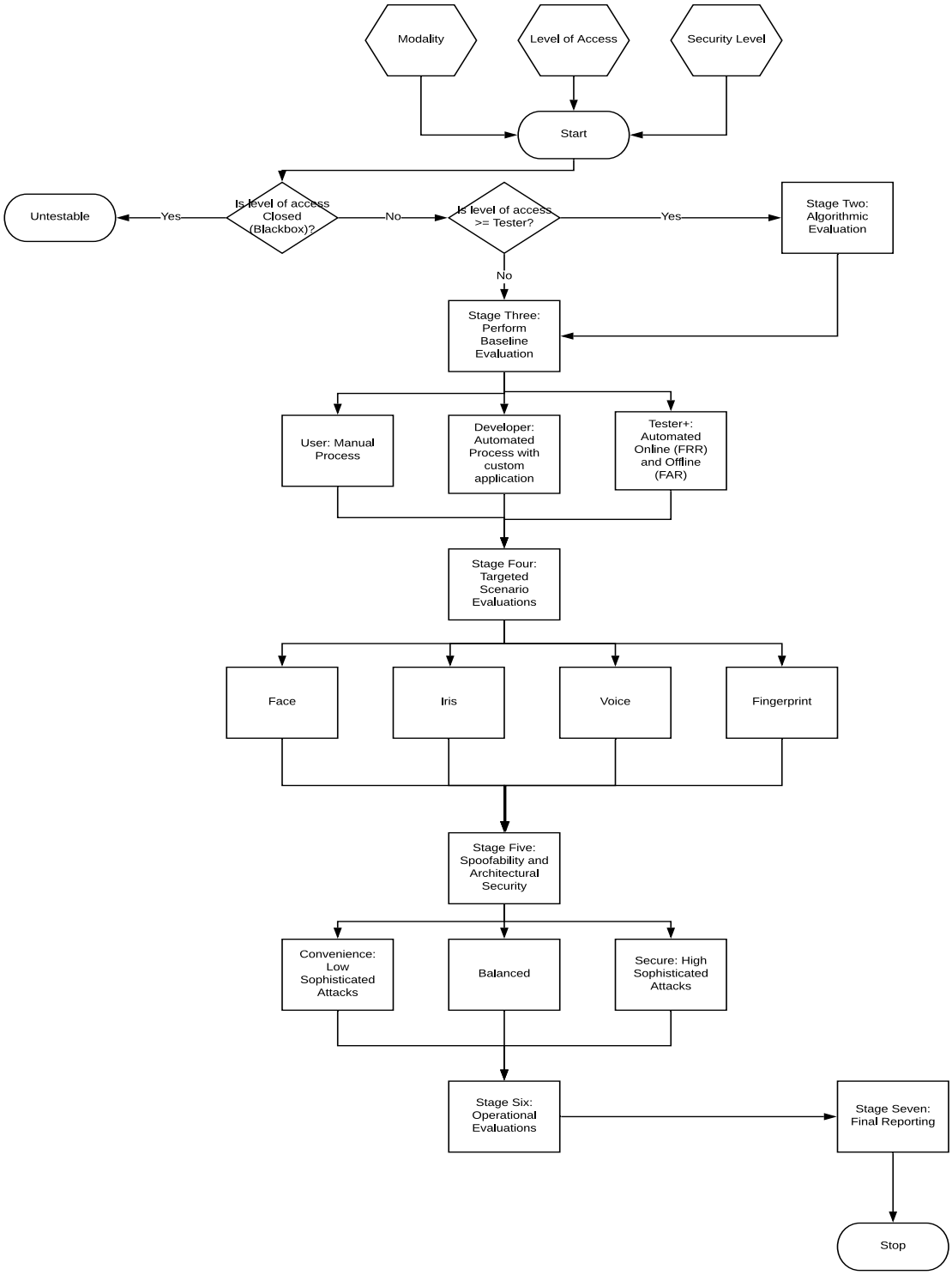



*Figure 4.1: Performance Framework Flowchart*

## 4.6 Usability

Usability is defined as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO 9241-11:2018 [107]).

- **Effectiveness**: accuracy and completeness with which users achieve specified goals
- **Efficiency**: resources used to concern the results achieved
- **Satisfaction**: the extent to which the user's physical, cognitive, and emotional responses that result from the use of a system, product or service meet the user's needs and expectations

The core factors [98] identified 'Users' as the cornerstone of mobile biometric performance and considered usability more in line with how Human-Computer Interaction (HCI) influences performance where the main goal is to improve performance results. One model that addresses usability from a performance perspective is the Human-Biometric Sensor Interaction (HBSI) model. The HBSI model "focuses on the interaction between the user and the biometric system to understand the individual details during this time. Including detecting and classifying both user and system errors" [108].

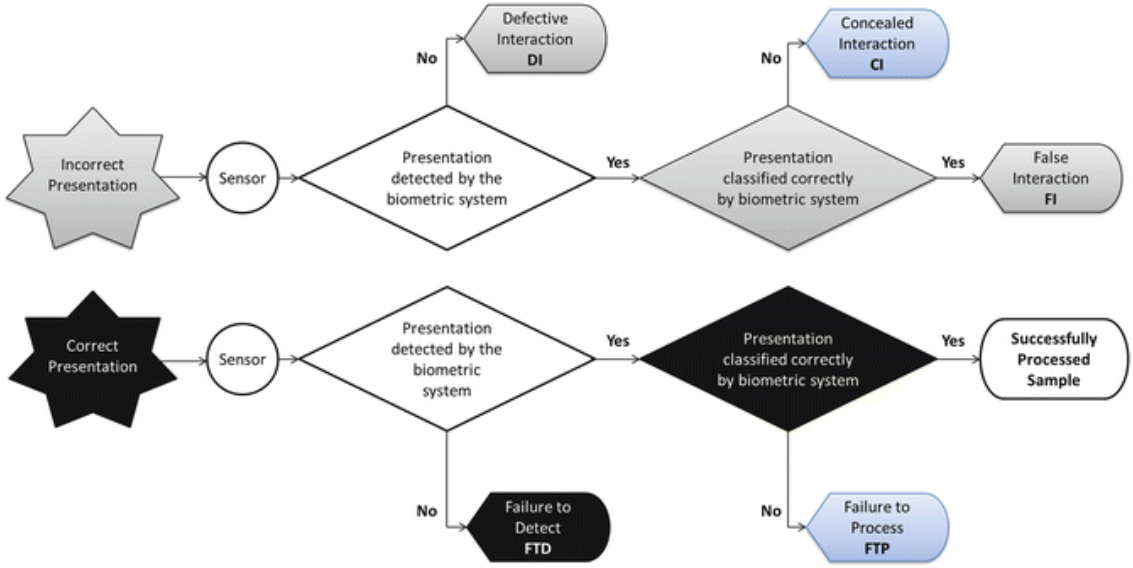



*Figure 4.2: HBSI Error Framework [108]*

- A **defective interaction** (DI) occurs when a user makes an incorrect presentation that is not detected by the biometric system [109].
- A **concealed interaction** (CI) occurs when the biometric system detects an incorrect presentation but is not classified correctly as an error [109].
- A **false interaction** (FI) is an incorrect presentation detected by the biometric system but, unlike a CI, is correctly handled as an error [109].
- A **failure to detect** (FTD) is a correct presentation made by the user that is not detected by the biometric system [109].
- A **failure to process** (FTP) is a correct presentation made to the biometric system that encounters an error when processed [109].
- A **successfully processed sample** (SPS) is a correct presentation detected by the biometric system and successfully processed as a biometric sample [109].

The HBSI model should be incorporated into the usability testing of the biometric system where appropriate and resources allow, and the metrics should be reported within the final report. However, it is noted that these metrics may not always be possible to use, mainly when the level of access is less than "developer", and it will be difficult to distinguish between individual events and causes for failure.

The user's satisfaction is encouraged to be measured using a questionnaire with a similar question to "how satisfied are you with the ease of use of <modality> in the proposed system?" which can be measured on a Likert scale to gauge an indication of satisfaction and further interviews with the users can be used to explore low scores further.

Transactional timings are another metric that should be recorded and examined where possible. This requirement will likely need to be set within the requirements as to recommendation range from around two seconds (Android) to 30 seconds (FIDO). However, low transactional time will increase user satisfaction and adoption of the system and should be measured.

## 4.7 Tailored Impostors

This section introduces the methodology for selecting 'tailored' impostors as an alternative approach to selecting impostors at random to allow for more informative results utilising the approaches discussed earlier relating to 'edge case' testing from software engineering [91]. The approach involves exploiting previously identified know weaknesses with biometric systems around users' physiological and behavioural characteristics [98], [110], [111].

The definitions for the work presented here include the following:

- **Tailored Impostor**: An individual chosen to serve as a passive impostor (probe) to a genuine reference based on similar characteristics determined from physiological and behavioural attributes.
- **Tailoring**: The algorithmic process selects the tailors suitable for a given reference based on determined and appropriate characteristics from physiological and behavioural attributes for the modality or modalities under evaluation. (Past Tense: Tailored)

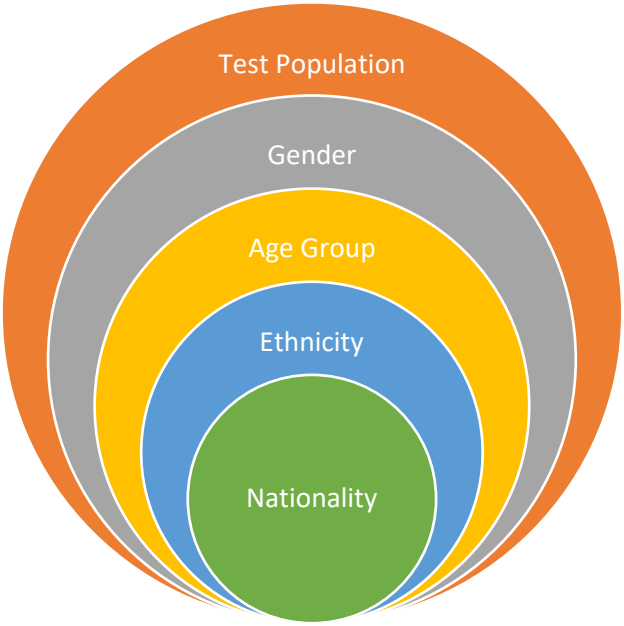



*Figure 4.3: Example Tailored Impostor Diagram (Tailoring)*

In simple terms, the approach involves selecting the most similar or 'lookalike' participants for each genuine participant to draw out and explore potential increases in the false match rate (and highlighting potential bias in the process). The factors chosen for 'tailoring' will depend on the modality under investigation. Table 4.3 shows an illustrative potential tailoring model for a facial recognition system. The algorithm will select impostors from the innermost grouping from the test population before moving into the outer groups. The suggestion is to remove identical twins from the test population and keep the age groups to a maximum of 10 years from either side of the subject and ideally five years in young adults (18-34).

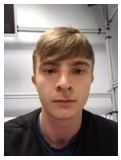
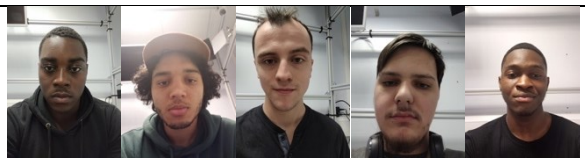
This concept of using 'homogenous' people is becoming more prevalent, as was explored by Rathgeb *et al.* [112], who used a database of doppelgangers, and concluded that "many face recognition evaluation protocols randomly pair face images to obtain non-mated comparisons. Obtained non-mated comparison score distribution may be used to set up decision thresholds at fixed FMRs. Consequently, it may be concluded that FMRs (and decision thresholds) obtained in such a way overestimate the security of the underlying face recognition system". Additionally, Popescu *et al.* [113] explored face verification with challenging imposters and diversified demographics and found that their "evaluation shows that model comparison is more meaningful when using challenging imposter pairs" when exploring gender, origin (ethnicity) and age.

Recently ISO also published a technical report, ISO/IEC TR 22116:2021 [114], which looks at how demographic factors can affect the performance of a biometric system. The report found that "if the demographic distributions for a given biometric system change significantly, the system's overall performance can also change. However, the magnitude of this variation in biometric performance depends on the specific demographic factors and the modality of the system". It is worth noting that the

report was not focused on non-mated comparisons but found that for FMR for face recognition, FMRs for men and women differ. In addition, FMR is more significant for the very young and very old. Regarding ethnicity, FMR varies globally, with high FMR in East Asia, South Asia and sub-Saharan Africa.

Table 4.6 shows a possible tailoring process for the face modality. This approach will be tested further in Chapter 7, although it was impossible to extensively test it due to a relatively small dataset (consisting primarily of students).

*Table 4.6: Example of an impostor selection process showcasing a possible 'Tailoring' algorithm*

| Genuine (Reference) | Impostor (Probes) (5 Probes) |
|---|---|
|  | |
| **Random** |  |
| **Gender** |  |
| **Gender + Age Group** |  |

## 4.8 Comparing the Framework

The work presented here was a comprehensive testing framework for mobile biometrics to provide users and businesses with the assurance they need to satisfy their requirements. The framework utilises existing standards and approaches to bring together current industry standards and recommendations, including the likes of ISO, FIDO Alliance, and Google. Furthermore, the framework aims to go further and improve upon the existing standards in certain aspects by incorporating more performance areas into the framework, most notably usability.

The FIDO alliance biometric testing protocol (V2) can be broadly split into three components. These components are still maintained within the proposed framework within stages two, three and five.

1. Online Testing (FNMR)
2. Offline Testing (FNMR and FMR)

3. Presentation Attack Detection (IAPMR)

Table 4.7 highlights the critical comparisons and differences between the existing methodologies and our proposed framework. One of those key differences is the consideration of varying access levels to a device, as all existing approaches assume a reasonable level of access to the system. It should be noted that only ISO/IEC 19795-1 is considered for the comparison present. It is acknowledged that some components are present within an additional standard, like presentation attack detection.

Several vital differences are present in the proposed framework, including:

- Separate FTA from FRR
- Consider FNMR and FMR (attempt-based) (for usability purposes)
- Considers options for when the 'common test harness' is not available
- Use targeted impostors (the current suggestion is to use random for attempt-based)
- Defined enrolment scenario
- Defined multiple verification scenarios (incl. motion) + operational
- Various levels of security testing

*Table 4.7: Comparing the Key Features of Available Performance Methodologies*

| Identity | FIDO | Android (Google) | ISO/IEC 19795-1:2006 | Proposed Framework (Boakes) |
|---|---|---|---|---|
| **Usability** | ✗ | ✗ | ✗ | ✓ |
| **Tiered Authentication (Security Levels)** | ✗ | ✓ | ✗ | ✓ |
| **Enrolment Scenario** | ✗ | ✗ | ✗ | ✓ |
| **Motion Scenario** | ✗ | ✗ | ✗ | ✓ |
| **Presentation Attack Detection** | ✓ | ✓ | ✗ | ✓ |
| **Technology Evaluation** | ✗ | ✗ | ✓ | ✓ |
| **Scenario Evaluation** | ✓ | ✓ | ✓ | ✓ |
| **Operational Evaluation** | ✗ | ✗ | ✓ | ✓ |
| **Designed for Varying Access (Commercial)** | ✗ | ✗ | ✗ | ✓ |

As noted, the ISO standards covering (mobile) biometrics are split across multiple working groups. Therefore, they are not the most accessible to gather a complete picture as each standard will often reference another. However, Table 4.8 compares some of the critical ISO standards when considering a

complete mobile biometric system performance evaluation. The proposed performance evaluation framework takes tremendous influence from these existing standards, as highlighted in Chapter 2, and seeks not to reinvent the wheel from the work of existing experts in the biometrics field.

Additionally, it is worth noting the contribution and influence of ISO/IEC TR 29156:2015 [116], "Information technology — Guidance for specifying performance requirements to meet security and usability needs in applications using biometrics" and ISO/IEC 29197:2015 [117] "Information technology — Evaluation methodology for environmental influence in biometric system performance". ISO/IEC TR 29156:2015 provides guidance and recommendation around usability and security requirements. It presents a potential classification by security level (High, Medium, Basic), providing target rates for FMR of 1%, 0.01%, and 0.0001%, respectively. It considers the following factors for usability requirements: accessibility, throughput, the authentication failure rate for authorized users, ease of use at point of authentication, ease of user for enrolment. Most of these concepts have been replicated into the proposed performance framework. ISO/IEC 29197:2015 discusses evaluating environmental influences on biometric performance and suggests to express the environment-specific performance relative to baseline performance however it is worth noting that both of these standards are not mobile specific and therefore the requirements for setting scenarios, such as for enrolment, is generally down to the evaluator or the likely operating conditions for the system. Influences from these standards have been incorporated as approach such as including environmental evaluation as part of the operational testing.

*Table 4.8: Comparing the Key Features of Available ISO Standards*

| Identity | ISO/IEC | | | | |
|---|---|---|---|---|---|
| | 19795 | | | | 30107 |
| | 1:2021 [7] | 2:2007 [36] | 3:2007 [97] | 9:2019 [37] | 3:2017 [41] |
| **Usability** | ✗ | ✗ | ✗ | ✗ | ✗ |
| **Tiered Authentication (Security Levels)** | ✗ | ✗ | ✗ | ✗ | ✗ |
| **Enrolment Scenario** | ✗ | ✗ | ✗ | ✓ | ✗ |
| **Motion Scenario** | ✗ | ✗ | ✗ | ✗ | ✗ |
| **Presentation Attack Detection** | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Technology Evaluation** | ✓ | ✓ | ✓ | ✗ | ✗ |
| **Scenario Evaluation** | ✓ | ✓ | ✓ | ✓ | ✗ |
| **Operational Evaluation** | ✓ | ✗ | ✓ | ✗ | ✗ |
| **Designed for Varying Access (Commercial)** | ✗ | ✗ | ✗ | ✗ | ✗ |

## 4.9 Summary

This chapter has introduced the potential new framework for evaluating the performance and applicability of a mobile biometric system. The framework incorporates the existing standards and state-of-the-art research presented in Chapter 2 and the core performance factors identified in Chapter 3. The aim was to look at some of the unique aspects of mobile systems when considering performance. The chapter presents a seven-stage performance framework for a complete system evaluation which aims to provide the 'fit for purpose' assurance required by users and evaluators. Several procedural approaches are identified and considered for a system's varying access levels that an evaluator could have. This chapter also introduces the concept of tailored impostors to draw out a system's worst-case scenario false match rate.

The next chapter will introduce a comprehensive data collection experimental procedure further to explore mobile biometric systems first-hand from commercial smartphone devices and begin evaluating the possibilities of the proposed performance evaluation framework. As part of the data collection exercise, a survey was provided to gauge the demographics of the resulting dataset, understand the current state of how users are utilising mobile biometrics as a security option on smartphones, and analyse usability.

# 5 The Experimental Data Collection

## 5.1 Introduction

To fully understand the intricacies of mobile biometric performance testing, an experimental data collection was designed and performed to help understand the framework's possibilities and limitations by trialling parts of it. For this purpose, a two-session experimental data collection was designed. The experimental data collection was performed between April and June of 2019. Therefore, the questionnaire data reflects the time, particularly regarding the habits and trends of the participants. This timing is worth highlighting due to the rapid pace of technology changes. Therefore, if the same questionnaire were presented today, the responses would likely differ from when the experiment was conducted.

This chapter introduces the experimental data collection, including the scenarios and devices chosen. It explores the breakdown of our collected database, including our participant demographics, trends in smartphone security locking habits among our participants, and usability metrics. Upon completing the experimental data collection design, ethical approval was sought and granted for human participation from the University of Kent Faculty of Sciences Research Ethics Committee.

Section 5.3 discusses the demographic breakdown of the participants in the data collection. Next, section 5.4 introduces the decision behind the chosen commercial devices and the specifications. Next, sections 5.5 and 5.6 describe what the participants must do within sessions one and two, respectively, and Section 5.7 discusses the mobile application developed to assist with the data capture process. Next, section 5.8 will explore the results of our pre-experiment questionnaire, which will provide a background into our participants and highlight the trends of smartphone habits among them. Next, section 5.9 will seek to showcase the results of the post-experiment questionnaire providing an insight into the satisfaction and thoughts of how the participants respond to biometric authentication on mobile devices. Section 5.10 will then show the initial results from the data collection by looking at the false non-match rates achieved for each device's in-built modalities. Next, section 5.11 provides some analysis and reflection, and a final Section 5.12 will summarise the chapter.

## 5.2 Evaluation Design and Dataset to be Collected

Although there exist several facial recognition datasets [117]–[119], there are limitations, with most including not enough subjects or images per subject. Furthermore, most datasets consist of images scraped from various sources on the internet. Therefore, the source cannot be guaranteed when seeking a dataset containing only images from mobile sources. These datasets are also limited in the scenarios

involved, where most are confined to stationary scenarios. The same limitations are also present when seeking a dataset of other physical modalities (fingerprint, iris, and voice). For these reasons and to test the framework, a bespoke data collection process and dataset were produced.

When evaluating biometric systems, it is best to test on a population that best matches the target population for the system for the evaluation to be an accurate reflection of the performance. Unfortunately, that can prove not easy with most commercial systems as essentially the entire planet's population could be viewed as a potential target. The aim of this data collection is similar in that the aim is to reflect a generic but balanced population sample. Ideally, the experiment would contain gender, age, and ethnicity demographic balance.

Achieving a balanced demographic is ideal for commercial biometric testing, and we achieved a good gender balance. However, due to the constraints placed upon the process (location, time, finances), the population consists mainly of a university cohort. Hence, although a mixture of nationalities is present, the balance is not met, and the age is skewed to a young adult population. However, this is not a problem, provided it is acknowledged that the target demographic is not a balanced general population but a balanced university population.

Although there exists a bias within the dataset, the collected data still allows for the examination and evaluation of the proposed performance framework and begins to evaluate its suitability. The emphasis is that this is a test sample and not a comprehensive database that will allow for a thorough proof-of-concept. The intention is that the results of this thesis will provide sufficient merit for further investigation, where researchers can perform evaluations with comprehensive databases. However, at the time of writing, no existing datasets provided sufficient depth and scope to explore our approach fully, hence the need to create our own to meet this purpose, despite certain limitations.

## 5.3 Participant Demographics

Our trial was split across two sessions, whereby 60 participants completed the first session, and 56 completed the second session. They were awarded a £15 Amazon voucher for their time. Figure 5.1 shows these participants' gender split, showing that 35 participants were male and 25 were female.

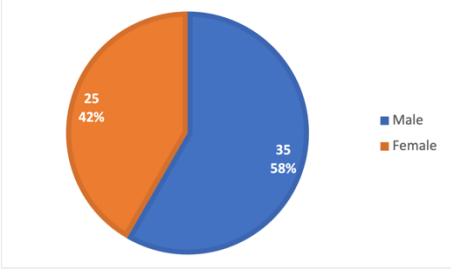



*Figure 5.1: Gender Split of Participants*

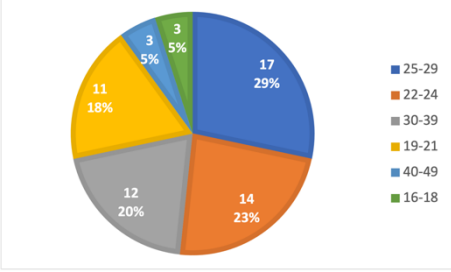



*Figure 5.2: Age Split of Participants*

Regarding age, the general population cohort was a university population; therefore, 75% of the participants were under 30. The nationality of our participants was collected, and most of our participants (40%) identified themselves as 'British', with the following two most significant groupings consisting of 'Indian' (14%) followed by an equal amount of 'Italian' and 'Romanian' (5%). The questionnaire also asked participants about their eyewear, with 35% saying that they wore glasses and 5% wearing contacts for the first session of the study.

## 5.4 Devices

Four commercial smartphone devices were chosen to serve as the mobile devices used for this experimental data collection. They were partly chosen based on their high profile within the UK mobile device market, as shown in Figure 5.3. For this reason, two iOS and two Android devices were selected.

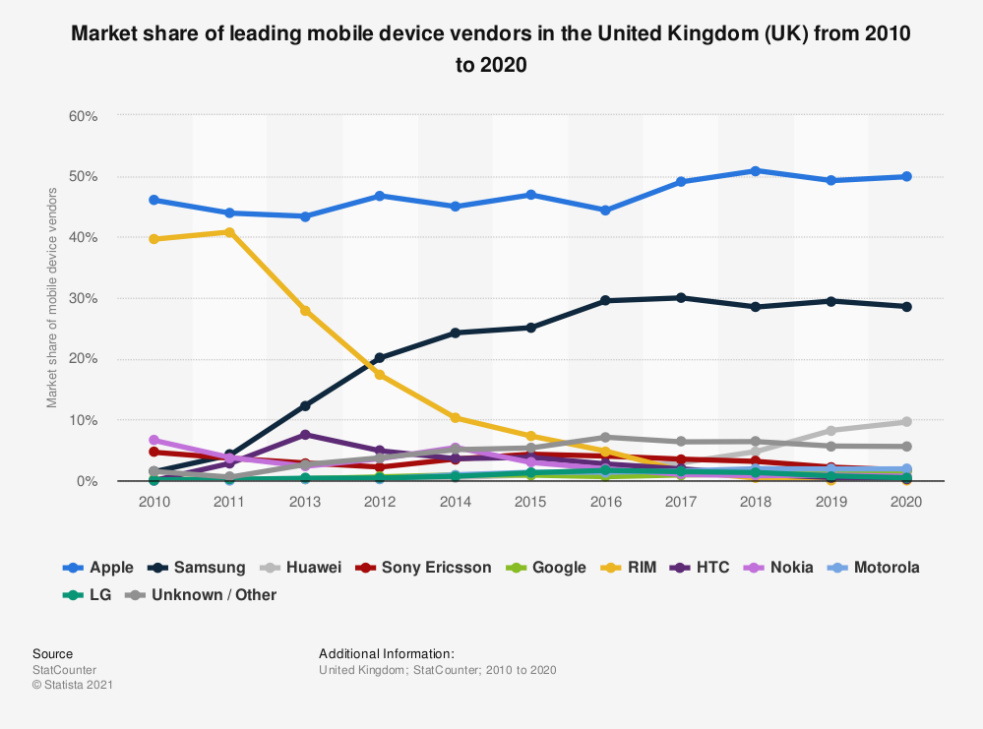



*Figure 5.3: Market Share of UK Mobile Device Vendors [120]*

*Table 5.1: Experimental Smartphone Device Specification*

| | | | | |
|---|---|---|---|---|
| **Device** | **Samsung Galaxy S9** | **Google Pixel 2** | **Apple iPhone 8** | **Apple iPhone X** |
| **Modality** | • Fingerprint<br>• Face<br>• Iris | • Fingerprint | • Fingerprint | • Face |
| **Fingerprint Sensor** | Egis Technology (ET510) | Fingerprints (FPC1075) | AuthenTec (TouchID) | |
| **Camera (Front)** | | | | |
| **Resolution** | 8 MP | 8 MP | 7 MP | 7 MP True Depth |
| **Aperture** | ƒ/1.7 aperture | ƒ/2.4 aperture | ƒ/2.2 aperture | ƒ/2.2 aperture |
| **Focus** | Fixed Focus | Auto Focus | Auto Image Stabilisation | Auto Image Stabilisation |
| **Flash** | | | Retina Flash | Retina Flash |
| **Pixel Size** | 1.22 µm | 1.4 µm | | |

The Samsung Galaxy S9 provides users with three biometric modalities: fingerprint (located at the rear), face, and iris. The Google Pixel 2 provides users with fingerprint recognition capabilities that can be used for authentication. However, it is worth noting that the device can perform both face and voice authentication using the 'trusted face' and 'trusted voice' features, respectively but not with a "developer" level of access. The Apple iPhone 8 features a frontal fingerprint sensor, and the Apple iPhone X features a face recognition system.

These devices are all commercial, so access to the biometric system is limited. First, the data will be analysed with a "developer" level of access, meaning the results will be obtained from recording the Boolean verification results, and Section 5.10 provides the discussion. Furthermore, biometric samples, including images (face and iris) and recordings (voice), are obtained to analyse further what might be possible with greater access. The results of this will be discussed in Chapter 6.

Along with the four commercial smartphone devices, a separate mobile device was used to collect iris images. The IriTech IriShield (MK 2120U/UL) [121] can capture infrared images of the iris and be connected to an Android smartphone and incorporated into an Android application through the supplied SDK.


*Figure 5.4: IriTech IriShield (MK 2120U/UL)*

## 5.5 Session One (Indoor)

The first session took place indoors while controlling the conditions and external influences as much as possible. First, participants would arrive, be briefed about the experiment, and be allowed to ask any questions. Once the participant was satisfied, they would be asked to sign a consent form, and the experiment would begin.

One of the scenario tests involved the use of a treadmill. The participants were allowed to walk at a comfortable pace. Before testing, they familiarised themselves with the treadmill's controls and selected a comfortable speed. This speed would then be fixed for the remainder of the trials.

The participant would use two devices for this session: one iOS-based device and one Android. Which device a participant would use and the order in which they would use them was randomised beforehand to minimise any bias from device ordering and habitation that could occur in the results.

With the participant now with a device, the first step was to enrol the participant into the device's in-built biometric system. This enrolment process was device-specific, depending on the biometric capabilities of the device (see Section 5.4 ). The enrolment was performed while the participant was sitting with the device hand-held by the user.

The participant was then tasked with performing verification attempts in four defined scenarios for a minimum of five transaction attempts. The four scenarios were while the user is 'Sitting', 'Standing', walking on a 'Treadmill' (at a personalised speed), and walking down a 'Corridor'. In addition, the session analysed a 'Factor' scenario, introducing extreme conditions to test while the user was sitting. For example, the device was tested in a dark room with low lighting (around 4-5 lx approx.) for face and iris recognition. Before attempting the authentication, the user was asked to dip their finger into a glass of water for fingerprint recognition.

- **Sitting** – The participant sits in a chair while operating the device with their hands.
- **Standing** – The participant stands while operating the device with their hands.
- **Treadmill** – The participant walks on a treadmill at a predefined speed to allow a comfortable simulated walking pace while operating the device in their hands.
- **Corridor** – The participant walks comfortably down a straight corridor while operating the device in their hands.
- **Factor** – Introducing a known 'influencing factor' to expose potential weaknesses in the system (low lighting for face and iris, wet finger for fingerprint).

These scenarios comprised two stationary (Sitting and Standing) and two motion (Treadmill and Corridor) scenarios. Thus, this process complies with the recommendation of the framework by introducing at least one motion-based scenario and incorporating an influencing factor.

## 5.6 Session Two (Outdoors)

The second session took place with a minimum gap of one week from the first session but was, on average, two weeks from the first session. For the second session, the experiment moved outdoors, where the environment became unconstrained and outside the control of the evaluators and participants. Again, the participants were required to follow a predefined trail around the University of Kent campus and stop at selected destinations that formed a circular route. Each participant performed this route with one device, randomly selected from one of the two devices they operated during session one.

At each destination, the participant performed verifications (or data capture) depending on the available modalities on the device they were operating. Before the participant set off, a screen capture showing the weather conditions and the temperature was captured and stored for analysis.

Once the participant had successfully returned, they returned the device for a £15 Amazon voucher for their time.

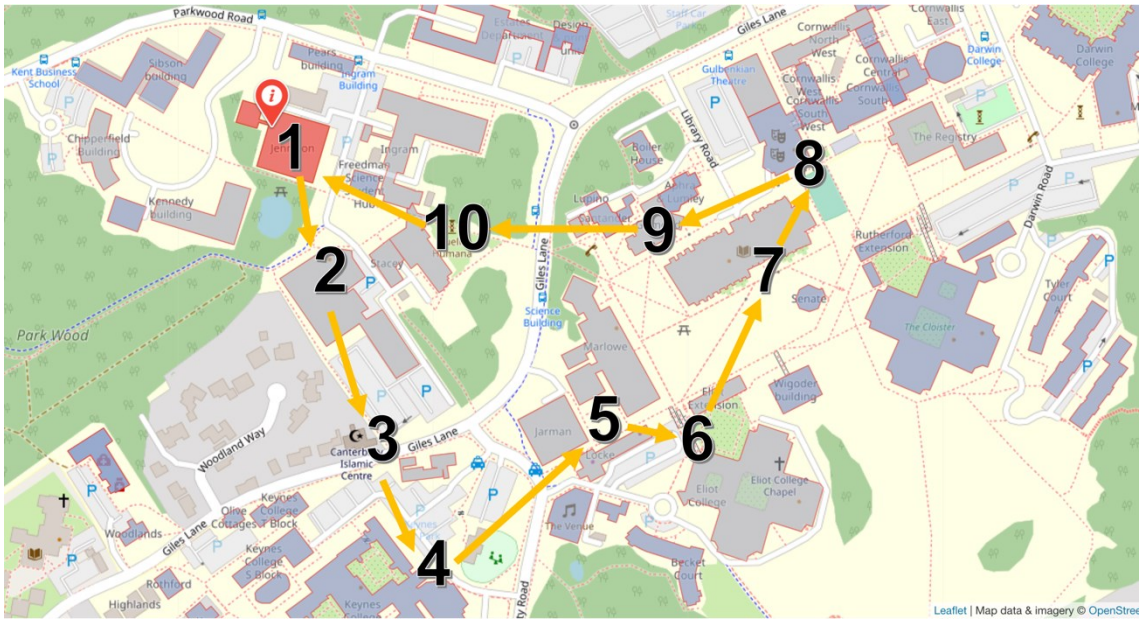
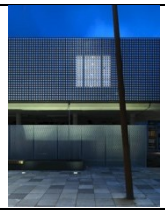



*Figure 5.5: Session Two Route Map of the University of Kent Canterbury Campus*

Ten locations were involved in the circular route, as shown in Figure 5.5 and Table 5.2. The aim was to try and capture as many environments as possible from the university campus, generally areas with a high concentration of student footfall near significant university buildings and locations near roads.

*Table 5.2: Session Two 'Outdoor' Stop Locations*

| | | | | |
|---|---|---|---|---|
|  |  |  |  |  |
| Jennison | Sport Centre | Canterbury Mosque | Keynes College | Central Plaza |
|  |  |  |  |  |
| Elliot College | Library | Gulbenkian | Grimond | Huella Humana |

## 5.7 Application Development

A custom application was developed for Android and iOS platforms to capture the required data, entitled "Biometric DC" (Data Collection). The intent was to mirror the biometric capture process as closely as possible from a developer's perspective. Unfortunately, capturing helpful information directly from the smartphone logs upon unlocking the devices resulted in limited data with a cumbersome process.

A smartphone application was developed using the available API for each operating system, BiometricPrompt [122] for Android and the equivalent LAPolicy.deviceOwnerAuthenticationWithBiometrics [123] for iOS. The result was a device authentic authentication experience with the same user interface prompt as a device user would expect to see.

Alongside the verification attempts, the application collected background sensor data available on the device while the verifications took place.

*Table 5.3: Background Sensors Collected from the Android Devices*

| | |
|---|---|
| **Motion Sensors** [124] | |
| **Accelerometer** | Acceleration force along the x, y, and z-axis (including gravity). |
| **Gravity** | Force of gravity along the x, y, and z-axis. |
| **Gyroscope** | Rate of rotation around the x, y, and z-axis. |
| **Linear Acceleration** | Acceleration force along the x, y, and z-axis (excluding gravity). |
| **Position Sensors** [125] | |
| **Magnetic Field** | Geomagnetic field strength along the x, y, and z-axis. |
| **Orientation** | Azimuth (angle around the z-axis), Pitch (angle around the x-axis), Roll (angle around the y-axis). |
| **Proximity** | Distance from an object. |
| **Environment Sensors** [126] | |
| **Ambient Temperature** | Ambient air temperature |
| **Light** | Illuminance |
| **Pressure** | Ambient air pressure |
| **Relative Humidity** | Ambient relative humidity |
| **Temperature** | Device temperature |

The information from the sensor will be used to help inform any trends about the performance and is used as the basis to build a novel authentication system to alter the threshold depending on the scenario involved dynamically. This branch of work is discussed in-depth in Chapter 7. The application would store and save CSV files containing all the required sensor information. Additionally, another CSV file would store all the verification results, which could then be extracted from the device and stored for analysis.

Figure 5.6 shows screenshots taken from the iOS version of the data collection application. The participants' flow can be seen by observing the images from left to right. The first screen asked the participant to select the current session. The second screenshot shows the first session's home screen. Next, the participant would select which modality to collect, moving down the list from 'Fingerprint' to 'Factor' (depending on what modalities are available on the device). The third screen shows the modality collection screen. The modality is bold at the top of the screen, followed by the current scenario. Pressing the "Authenticate" button triggered the collection of the modality with a number on the screen to count the number of transactional attempts. Once the minimum number of transactions was achieved (5), a next button would appear to allow the participant to progress onto the following scenario. The interface was designed to guide the participant with ease and precise information from start to finish.

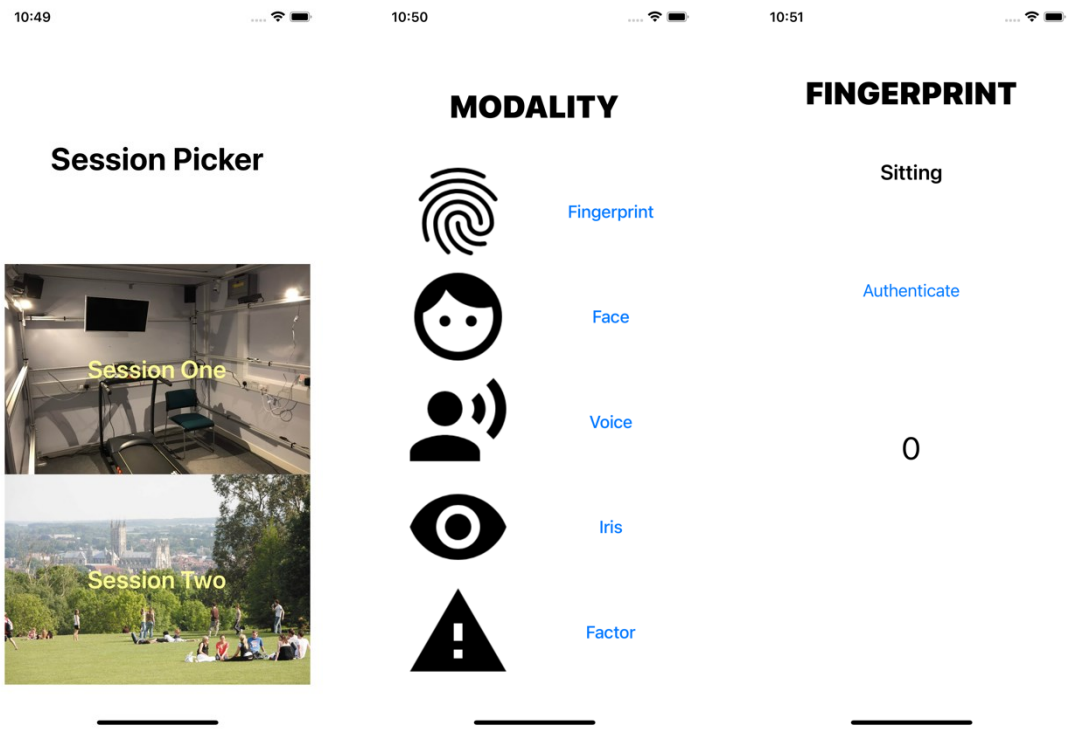



*Figure 5.6: Example Screenshots from the "Biometric DC" Data Collection Application*

## 5.8 Pre-Experiment Questionnaire Results

The participants were provided with a questionnaire (Appendix A: Questionnaire). The first part collected demographic information about the participants before the trials began and information about their current smartphone security habits. The second part, completed after the experiment, sought the participant's opinions regarding satisfaction and usability.

### 5.8.1 Smartphone Habits

The participants were asked, "**Do you currently own a mobile phone?**" All of them said that they did own a personal smartphone device. They were then asked, "**What mobile phone do you own?**". This information was extracted into the device's operating system, allowing the categorisation of Android and iOS users.

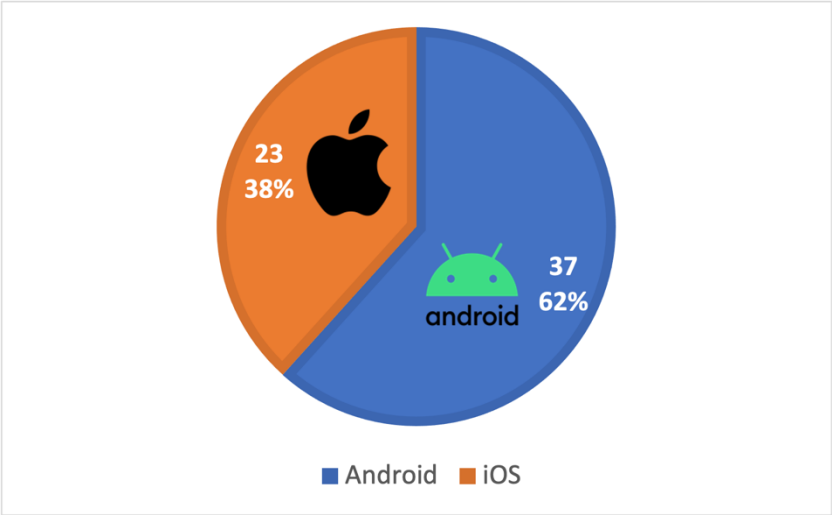



*Figure 5.7: The Operating System for the Participants' Smartphone Device*

The questionnaire then sought the current phone locking habits and whether the participant utilised a biometric locking method. Firstly, in the question "**Do you currently operate a screen lock on your mobile phone?**" most participants (58) said that they did, with only 2 participants stating that they did not currently operate a locking mechanism on their device. Next, those operating a screen lock were asked, "**What type of screen lock do you use primarily?**". Again, biometrics was the most common choice to act as the participant's screen lock security mechanism.

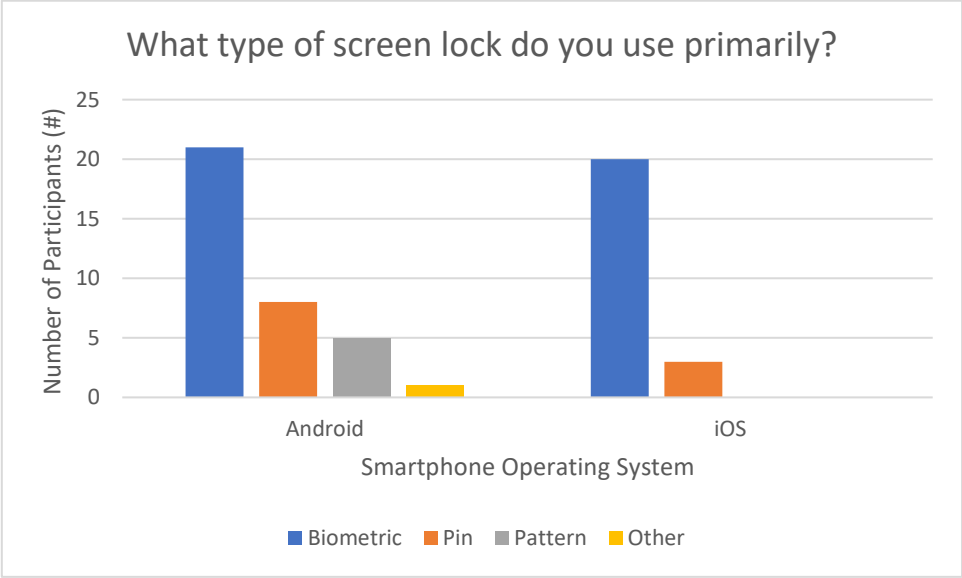



*Figure 5.8: Participants' Primary Smartphone Unlocking Method Categorised by Operating System Users*

Overall, 41 (71%) participants operated a biometric modality as their primary method for unlocking their smartphone. This smartphone unlocking mechanism was followed by 11 (19%) pins, 5 (9%) patterns and 1 (2%) who used another, which transpired to be a password.

Smartphones employ a range of biometric modalities, with manufacturers deciding what modalities to employ in their devices. The participants (41) said they were already using a biometric modality as their primary unlocking method and were further asked, "**If you use a biometric screen lock, please specify which you use?**". From the study, a fingerprint was the biometric modality most used as the primary modality for unlocking a smartphone device, with 80% of the biometric users using this modality, with the remaining 20% using their face.

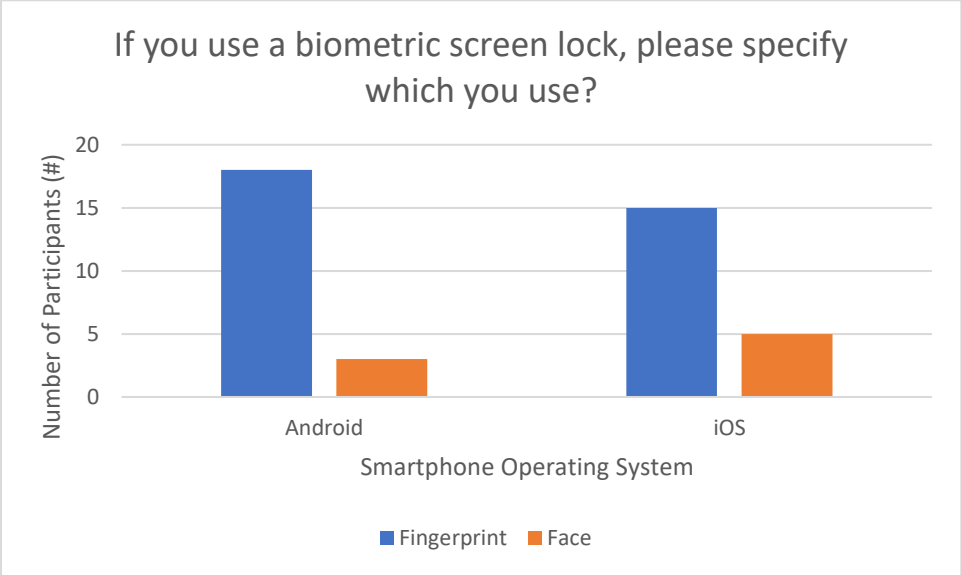



*Figure 5.9: Participants' Primary Biometric Modality for Smartphone Unlocking*

Smartphone biometrics do not currently operate without a backup mechanism for authentication. For example, iOS users must set a pin as a backup mechanism, whereas Android users are usually offered choices between a swipe pattern, pin or even a password. The backup mechanism is used for cases of verification failure to prevent users from becoming locked out of their devices. Next, the questionnaire asked, "**If you operate a secondary (backup) screen lock, what type do you use?**". The most common backup lock was a pin. However, this is the only backup option available to iOS users, with 80% of participants opting for this method, 15% opting for a pattern, and the remaining 5% using a password.

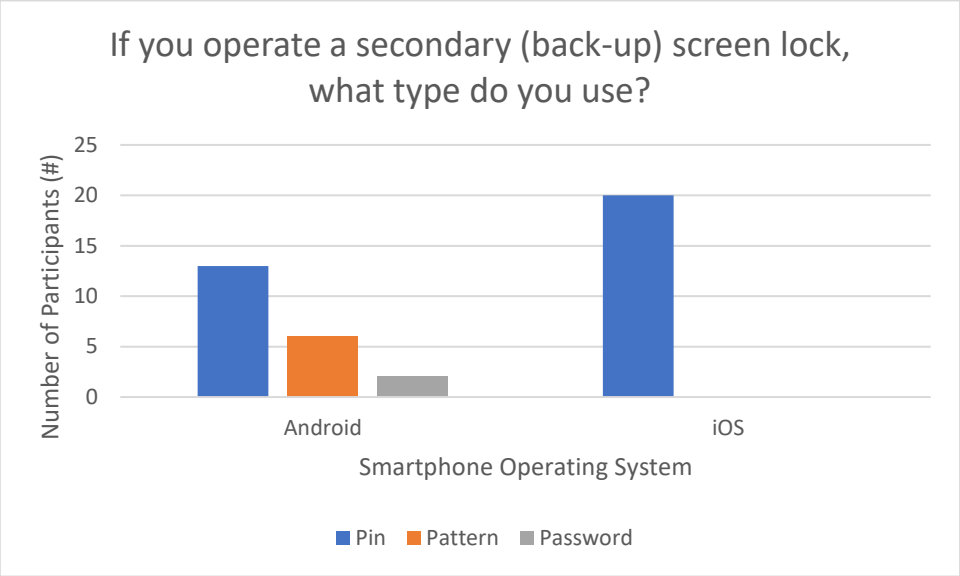



*Figure 5.10: Participants' Backup Mechanism for Biometric Users*

A final pair of pre-experiment questions asked how the participants felt regarding the security unlocking mechanism of their smartphone device. First, they were asked, "**How satisfied are you with the ease of use of your current phone lock?**" by ranking their satisfaction on a seven-point Likert scale, with one being *very dissatisfied* and seven being *very satisfied*.

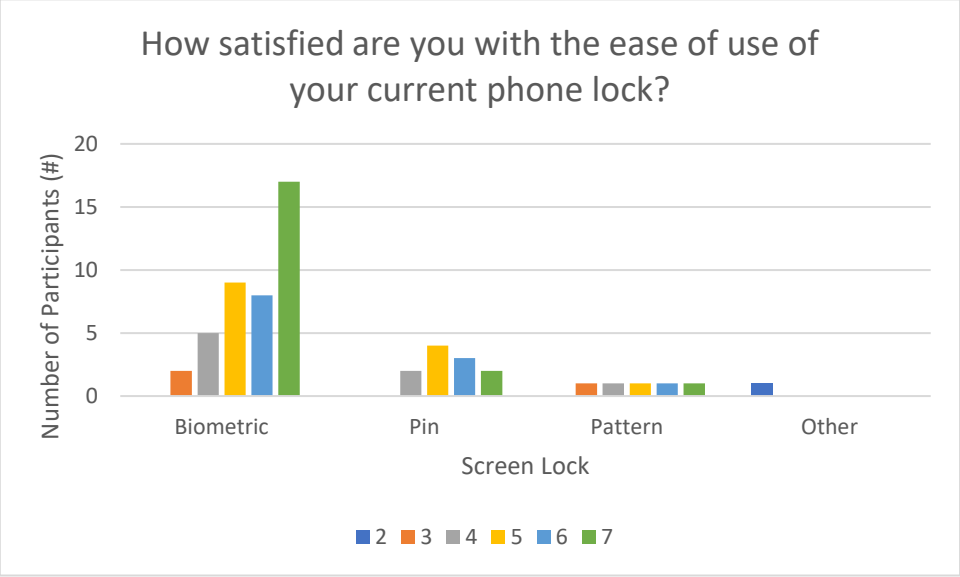



*Figure 5.11: Likert scale showing the participants' perceived 'satisfaction' with their current phone lock based on the lock type*

Figure 5.12 shows this satisfaction organised by the type of locking mechanism the participant primarily uses. For example, the participants that use biometrics as their primary locking mechanism have a satisfaction average of 5.8 (± 1.2). A pin follows this with an average satisfaction of 5.5 (± 1.0) and a pattern of 5.0 (± 1.6).

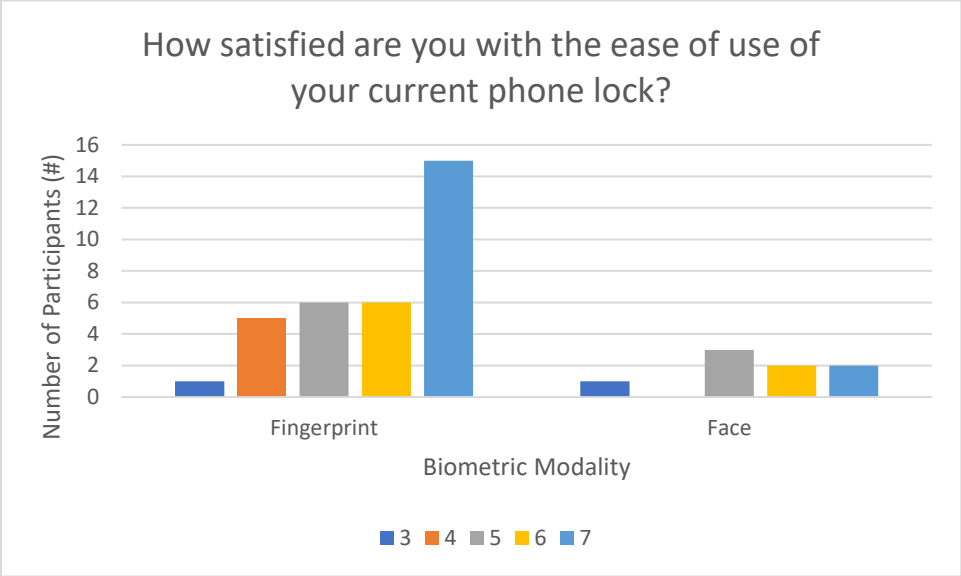



*Figure 5.12: Likert scale showing the participants' perceived 'satisfaction' of their current phone lock based on biometric modality*

Figure 5.12 shows a further breakdown of satisfaction per modality for the participants operating a biometric screen lock. The fingerprint modality shows an average satisfaction of 5.9 (± 1.2), while the face has an average of 5.5 (± 1.3), indicating a preference for the fingerprint modality.

Similarly, the participants were then provided with a definition of reliable in a biometric context as "the ability to identify you promptly consistently accurately", which was followed by a question which asked, "**How reliable do you find biometrics as a form of authentication on smartphones?**" using the same Likert scale as before.

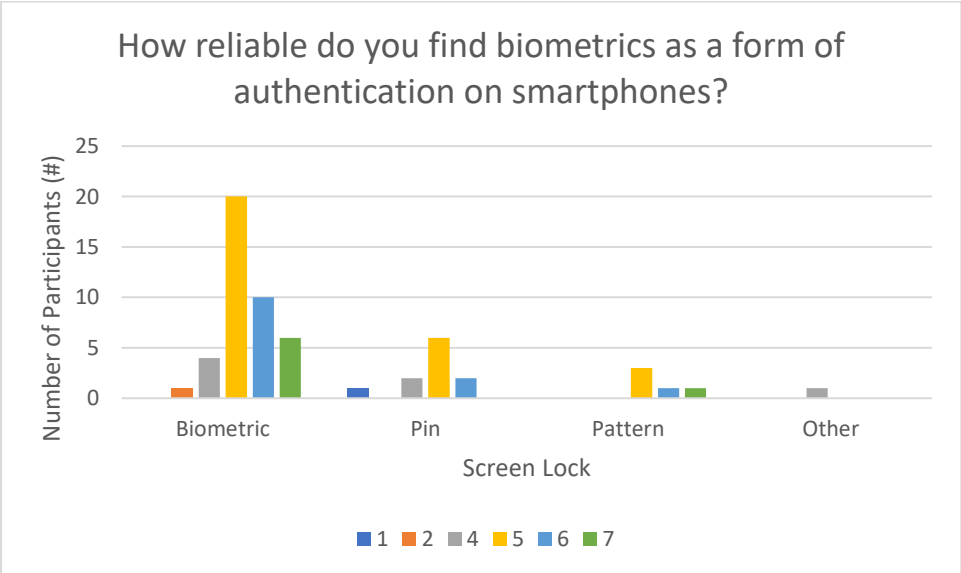



*Figure 5.13: Likert scale showing the participants' perceived 'reliability' of their current phone lock based on biometric modality*

Figure 5.13 shows the participants' reliability scores based on their current primary phone lock. Overall, the participants ranked the reliability of biometric authentication with an average score of 5.2 (± 1.1). Participants who used biometrics as their primary phone lock gave an average reliability score of 5.4 (± 1.0), followed by a pattern of 5.6 (± 0.9) and a pin with an average reliability score of 4.6 (± 1.4). Figure 5.14 shows the reliability breakdown organised by the biometric modality for the participants using a biometric screen lock. The participants gave fingerprints an average reliability score of 5.4 (± 1.0) and face 5.1 (± 1.0).

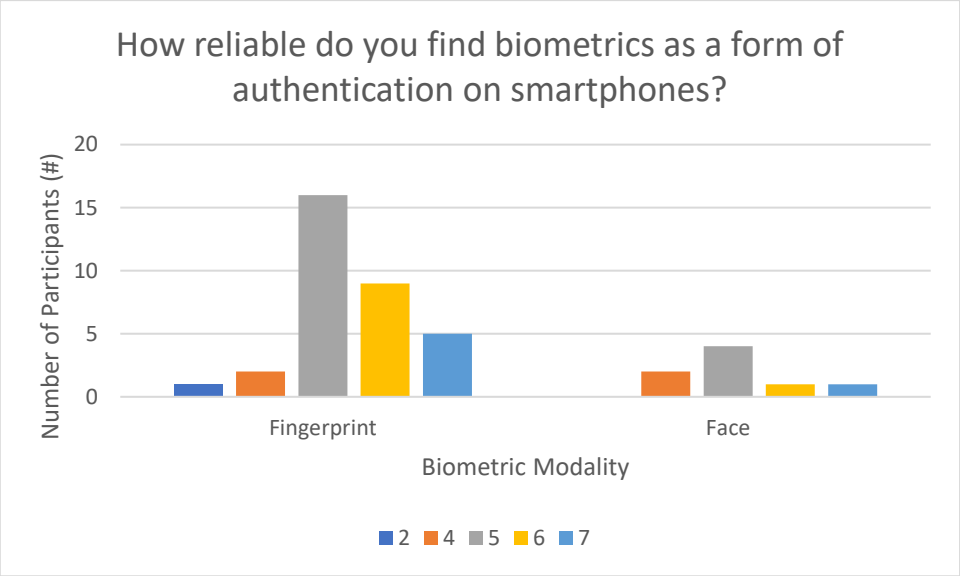



***Figure 5.14: Likert scale showing the participants' perceived 'reliability' of their current phone lock based on biometric modality***

## 5.9 Post-Experiment Questionnaire Results

### 5.9.1 Satisfaction

The participants were asked questions about their experience using a seven-point Likert scale where one was very dissatisfied and seven was very satisfied. The questions took the form:

**"How satisfied are you with the ease of use of <modality> authentication on smartphones?"**

The results shown below represent the entire participant population. They, therefore, do not show separations based on the devices the participant used during the study or personally owned, which could cause some deviation in the observed results.

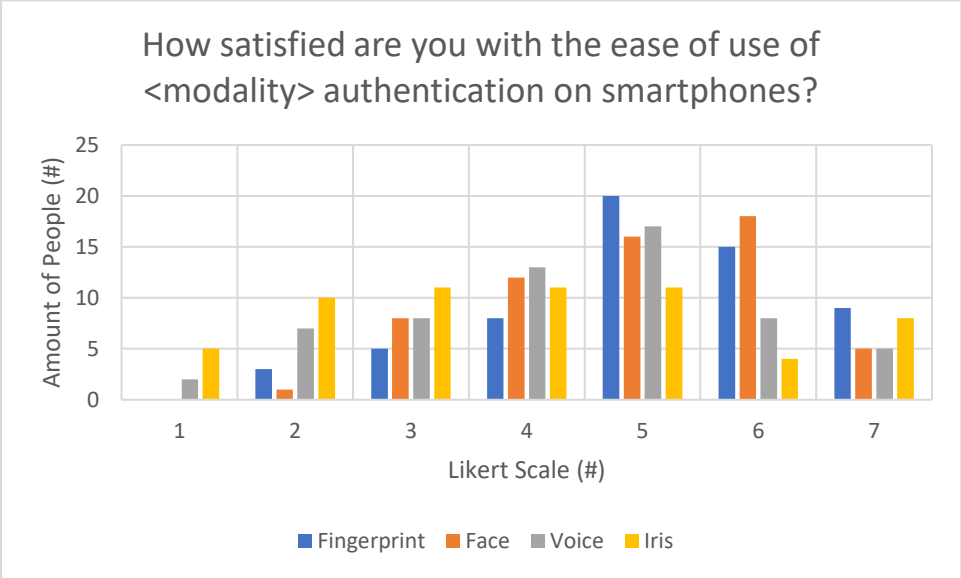



*Figure 5.15: Participants' Satisfaction for Each Modality*

The participants ranked their satisfaction with fingerprint recognition with an average of 5.1 (± 1.3), face recognition at 5.0 (± 1.2), voice recognition at 4.3 (± 1.5), and iris recognition at 4.0 (± 1.8). However, none of the devices used in the trial could perform voice authentication with a "developer" level of access. These results are more of the users' experience towards uttering a passphrase using a mobile device.

The results allow ranking the modalities based on the participant's perceived ease of use using the average satisfaction, resulting in 1: fingerprint, 2: Face, 3: Voice, and 4: Iris. These results highlight a potential preference for fingerprints in mobile devices.

### 5.9.2 Preference

The participants were also asked to select a preferred modality, having now experienced the four trialled sample capture processes. The specific question was

**"Which would be your preferred modality for authentication on a smartphone?"**

Figure 5.16 shows the breakdown of results where it can be shown that fingerprint was the most preferred modality for authentication on a smartphone, with 65% (39) participants selecting it, and voice was the least preferred modality, with only 2% (1) participants choosing voice. Interestingly, despite more people being satisfied with voice over the iris, the same is true for a preference, where more people prefer the iris over voice.

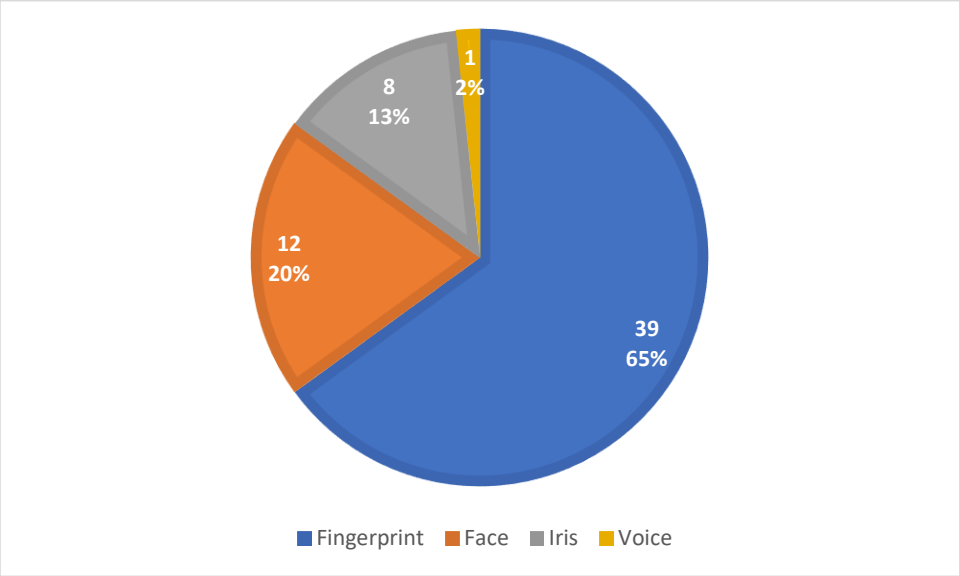



*Figure 5.16: Participants' Preferred Modality*

A further question was asked of the participants who were not already using a biometric modality to unlock their smartphone device (19). "**If you were not previously using a biometric screen lock, are you now considering using it?**". 68% (13) responded "Yes" to indicate they would. The remaining 32% (6) said "No". Chapter 3 identified how much the 'User' could affect the performance of a system, including usability (and satisfaction). The results here indicate that general scepticism towards biometric recognition systems, which misunderstandings could cause, can be overcome by interacting with such systems to improve user confidence and willingness.

### 5.9.3 Reliability

Similarly to Section 5.9.1 regarding the participant's satisfaction, the question was asked regarding the user's perceived perception of the reliability of the biometric authentication process on mobile devices. The participants were provided with a definition for reliability for a biometric system as "the ability to identify you promptly consistently accurately". Using a seven-point Likert scale, where one was very unreliable, and seven was very reliable, the participants ranked the overall concept of reliability for mobile biometrics based on the following question "**How reliable do you find biometrics as a form of authentication on smartphones?**"

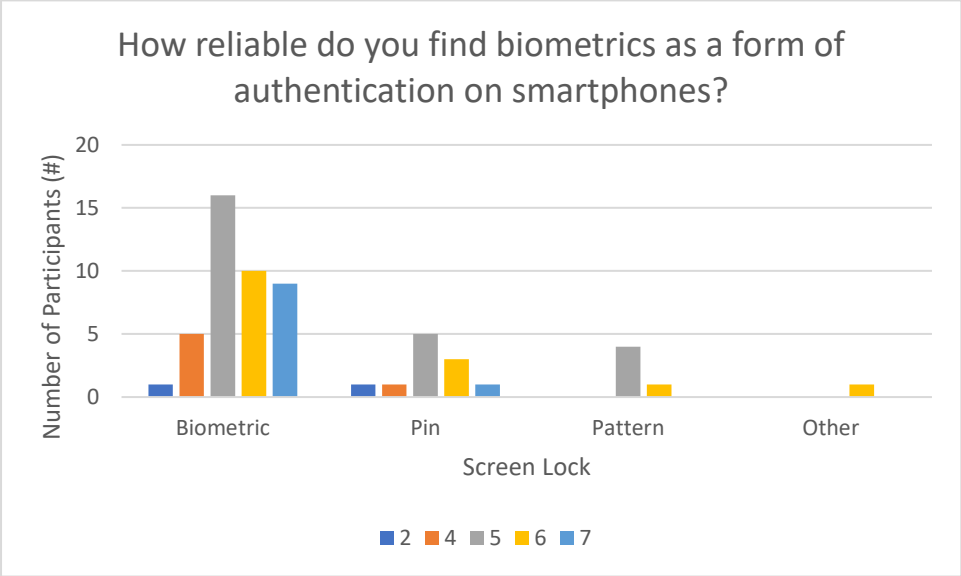



*Figure 5.17: Post-Experiment Reliability Questionnaire Scores Organised by Participants' Primary Screen Lock Type*

Upon completing session one, the participant's perceived reliability for mobile biometric authentication shows an average of 5.4 (± 1.1). For the participants who were already using a biometric screen lock, they scored reliability with an average of 5.5 (± 1.1) for pattern users, 5.2 (± 0.4), and for pin users, an average of 5.1 (± 1.3).

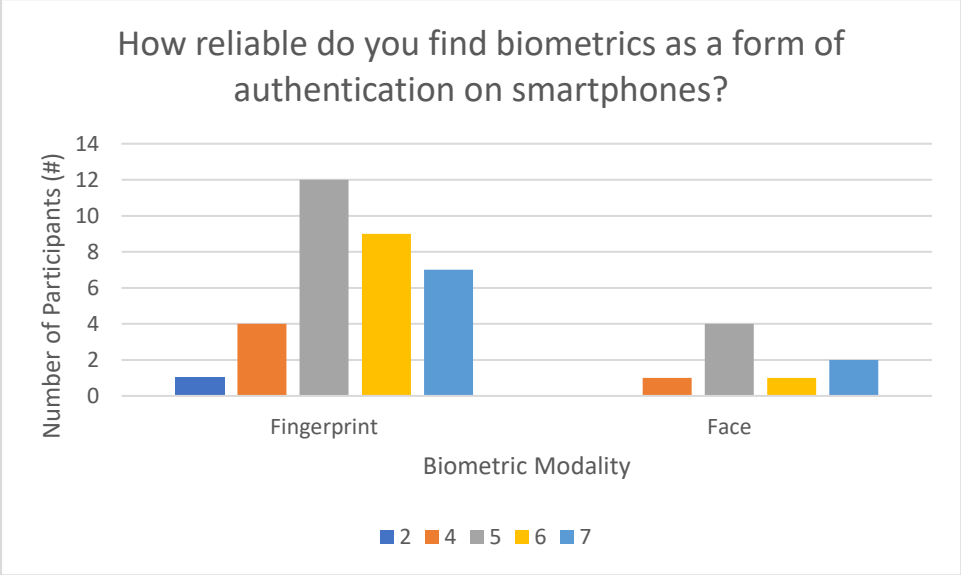



*Figure 5.18: Post-Experiment Reliability Questionnaire Scores Organised by Participants' Biometric Modality*

Figure 5.18 shows the reliability breakdown organised by the biometric modality for the participants using a biometric screen lock. The participants gave fingerprints an average reliability score of 5.5 (± 1.1) and face 5.5 (± 1.1).

Comparing these results to the ones achieved from the pre-experiment questionnaire, we can see that the overall average score increases from 5.2 (± 1.1) to 5.4 (± 1.1), again supporting how interacting with a biometric system can improve users' trust towards it. However, a paired t-test shows that these results are not statistically significant and are included only for comparison purposes.

*Table 5.4: Likert reliability scores pre- and post-experiment*

| Screen Lock | Pre-Experiment Reliability | Post-Experiment Reliability |
|---|---|---|
| All | 5.2 (± 1.1) | 5.4 (± 1.1) |
| Biometrics | 5.4 (± 1.0) | 5.5 (± 1.1) |
| Pattern | 5.6 (± 0.9) | 5.2 (± 0.4) |
| Pin | 4.6 (± 1.4) | 5.1 (± 1.3) |
| Fingerprint | 5.4 (± 1.0) | 5.5 (± 1.1) |
| Face | 5.1 (± 1.0) | 5.5 (± 1.1) |

### 5.9.4 Continuous Authentication

A final set of two questions asked participants about their knowledge of behavioural biometrics. The questions were regarding the concept of continuous authentication to gauge the awareness of modern biometric modalities that could become much more prevalent within a mobile context utilising elements such as swipe analysis on touch screen devices. The first question asked:

**"Are you familiar with continuous authentication? (Definition - It utilises a user's behaviour to continuously verify identity throughout a session, not just at the entry login point. Such as swipe behaviour or movement.)."**

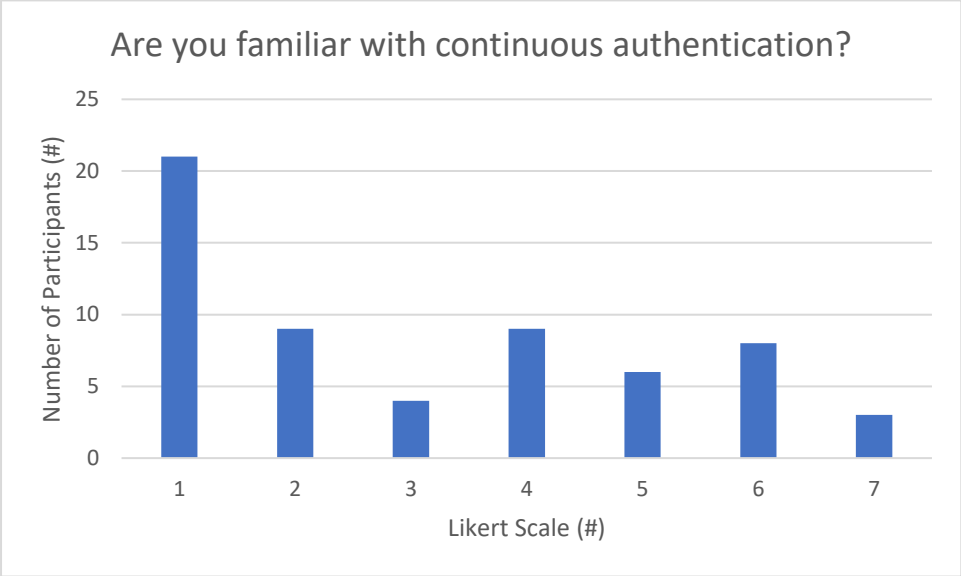



*Figure 5.19: Participants' response to their familiarity concerning continuous authentication*

Most participants were unfamiliar with continuous authentication; on average, they ranked themselves with a 3.1 (± 2.0). Finally, the participants were asked about their perception regarding their privacy and continuous authentication with the following question:

**"Do you feel that continuous authentication would be an invasion of your privacy?"**

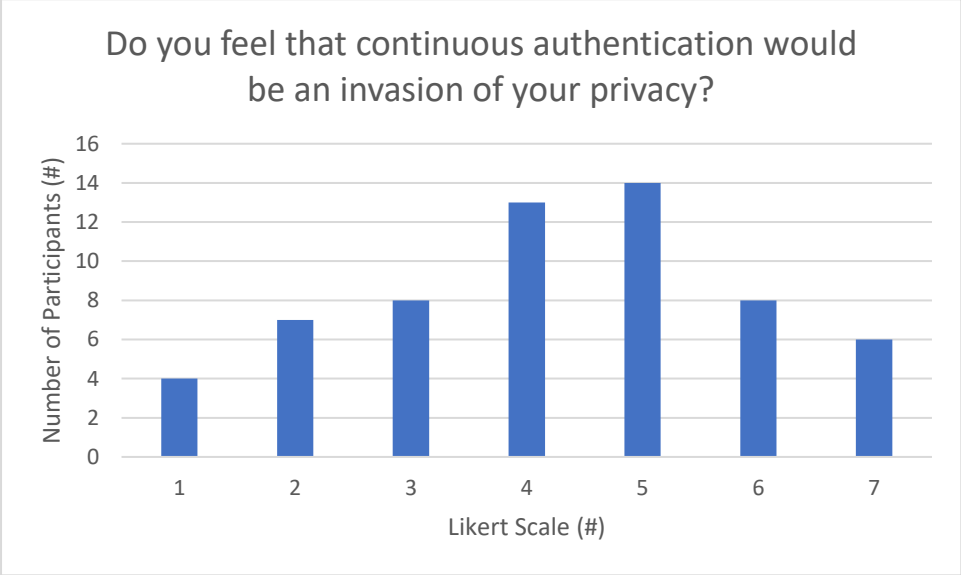



*Figure 5.20: Participants' response to how privacy-invasive they perceive the concept of continuous authentication*

Figure 5.20 shows the results of participants' perceptions towards privacy regarding continuous authentication. The results are captured using a Likert scale where one indicates 'Very Invasive', and seven indicates 'Very Non-Invasive'. On average, the participants ranked the privacy of continuous

authentication with 4.2 (± 1.7), indicating a slight preference for continuous authentication not being an invasion of privacy.

## 5.10 Trialling the Framework Model

To begin to test and apply the modelling and methodology, this was applied to commercial devices, starting with the Samsung Galaxy S9. This device allows users to enrol in three modalities, Fingerprint, Face, and Iris. From session one, 25 users enrolled on the three modalities while seated and holding the device comfortably in their hands.

Table 5.5 presents the total False Non-Match Rates (FNMR), the proportion of times a biometric system fails to grant access to an authorised person, found from each scenario. Here, the False Non-Match Rate is an outcome where the result was unsuccessful authentication. This outcome includes unsuccessful recognition, user interaction errors, user cancelled or invalid sample capture.

*Table 5.5: False Non-Match Rate of modalities on the Samsung Galaxy S9 in a variety of scenarios*

| | **Fingerprint** | **Face** | **Iris** |
|---|---|---|---|
| **Sitting** | 26% | 8% | 28% |
| **Standing** | 6% | 10% | 17% |
| **Treadmill** | 4% | 9% | 27% |
| **Corridor** | 7% | 7% | 29% |
| **Factor** | Wet - 77% | Dark - 1% | Dark - 25% |

The remaining devices had one modality that could be explored with a “developer” level of access. The FNMR results with these are shown in Table 5.6.

*Table 5.6: False Non-Match Rate of modalities on the Google Pixel 2, Apple iPhone 8, and Apple iPhone X in a variety of scenarios*

| | **Google Pixel 2** | **Apple iPhone 8** | **Apple iPhone X** |
|---|---|---|---|
| | **Fingerprint** | **Fingerprint** | **Face** |
| **Sitting** | 14% | 4% | 1% |
| **Standing** | 10% | 4% | 1% |
| **Treadmill** | 13% | 7% | 2% |
| **Corridor** | 5% | 9% | 3% |
| **Factor** | Wet - 80% | Wet - 100% | Dark - 3% |

Under the specifications of biometric testing standards ISO 19795-1 [11], only the first row would be required to meet the requirements set out in the standard (albeit with a more significant number of

participants involved), hiding from view the remaining information in the table. By exploring a more significant number of scenarios and conditions, we can begin to extract more information about the performance of a mobile biometric system.

This preliminary test begins to show the foundations of the framework model by testing several of the factors presented within Chapter 3, including 'Modality', 'Scenarios', 'Environments', 'Users' and 'Hardware'. Part of the increase in FRR seen in the sitting scenario was likely because this was the first scenario the user was presented and asked to authenticate themselves in and highlighted the relationship between the users and hardware before the users adjusted themselves to the current setup.

Introducing a challenging condition (Factor) and altering the environment can cause alternations in the performance, demonstrating a link between the modality and environments. For example, observing how darkening the lighting conditions saw a significant decrease in the FNMR scores (Table 5.5). This difference is likely due to the phone's use of an infrared camera which can focus more without disturbances and influences from external light sources, highlighting the relationships between the modality, environment, and scenario.

The acceptability was investigated by asking users their preferred modality post-experiment. Only 13% of the participants confirmed a preference for the iris, which likely reflects the FNMR found while using the iris modality despite the scenario. For the other modalities under test, 65% of participants preferred fingerprint, 20% for face and 2% for voice. The 'mobility' relationship was examined by introducing 'motion' scenarios, and slightly against expectations, the FNMR tended to drop slightly in these 'motion' scenarios. However, this could be a consequence of the users' habituation of the device.

Utilising "Developer" level features available within Android and iOS platforms can reveal further information about the system's performance as it reveals certain information as to why a transaction or attempt failed by providing details relating to the outcome. For Android, these outcomes are defined with the BiometricPrompt [122] API. The outcomes encountered here include:

- Succeeded – When a biometric is recognised, indicating that the user has successfully authenticated.
- Failed – When a biometric is presented but not recognised as belonging to the user.
- Negative – The user pressed the negative button.
- Exceeded – The operation was cancelled because the API was locked out due to too many attempts.
- User Cancelled – The user cancelled the operation.
- Cancelled – The operation was cancelled because the biometric sensor was unavailable.
- Time Out – The current operation has been running too long and has timed out.

Receiving some “Cancelled” outcomes is likely a consequence of some early development bugs with the application, which were quickly rectified. A couple is device-specific, such as when using the Samsung Galaxy S9:

- No Face Detected – No face detected.
- Irises Not Detected – Irises were not detected.
- Phone Too Close – Phone too close to face.

*Table 5.7: Outcomes from Fingerprint using the Samsung Galaxy S9*

| Scenario | Outcome | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | Succeeded | Failed | Negative | Exceeded | User Cancelled | Cancelled | |
| Sitting | 112 | 24 | 12 | 2 | 0 | 1 | 151 |
| Standing | 125 | 6 | 1 | 0 | 0 | 0 | 132 |
| Treadmill | 125 | 4 | 0 | 0 | 0 | 1 | 130 |
| Corridor | 126 | 7 | 2 | 0 | 0 | 0 | 135 |
| Factor - Wet | 44 | 84 | 43 | 12 | 2 | 0 | 185 |

*Table 5.8: Outcomes from Face using the Samsung Galaxy S9*

| Scenario | Outcome | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | Succeeded | Failed | Negative | Time Out | User Cancelled | No Face Detected | |
| Sitting | 114 | 1 | 2 | 3 | 2 | 1 | 123 |
| Standing | 110 | 0 | 11 | 0 | 0 | 0 | 123 |
| Treadmill | 111 | 0 | 10 | 0 | 0 | 0 | 123 |
| Corridor | 113 | 0 | 8 | 0 | 0 | 0 | 121 |
| Factor - Dark | 127 | 0 | 1 | 0 | 0 | 0 | 128 |

*Table 5.9: Outcomes from Iris using the Samsung Galaxy S9*

| Scenario | Outcome | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | Succeeded | Failed | Negative | Irises Not Detected | Phone Too Close | No Face Detected | |
| Sitting | 92 | 0 | 16 | 12 | 7 | 0 | 127 |
| Standing | 100 | 0 | 14 | 2 | 4 | 0 | 120 |
| Treadmill | 89 | 0 | 15 | 15 | 2 | 0 | 121 |
| Corridor | 87 | 0 | 15 | 16 | 3 | 0 | 121 |
| Factor - Dark | 96 | 1 | 16 | 10 | 2 | 1 | 126 |

*Table 5.10: Outcome from Fingerprint using the Google Pixel 2*

| Scenario | Outcome | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | Succeeded | Failed | Negative | Exceeded | User Cancelled | Cancelled | |
| Sitting | 171 | 22 | 1 | 0 | 4 | 0 | 198 |
| Standing | 173 | 15 | 2 | 0 | 1 | 0 | 191 |
| Treadmill | 171 | 14 | 1 | 2 | 7 | 0 | 195 |
| Corridor | 179 | 5 | 1 | 2 | 1 | 0 | 188 |
| Factor - Wet | 47 | 97 | 49 | 18 | 17 | 3 | 231 |

For iOS, there is a similar approach to examining the authentication outcome. However, one key difference from Android is that iOS will only provide the outcome from the complete transaction and not from the individual attempts, meaning the information presented below is transactional-based. A list of potential errors from the iOS process can be seen from LAError.Code [127]. The ones present here are:

- Success – When a biometric is recognised, indicating that the user has successfully authenticated.
- Cancelled – Cancelled by the user.
- Locked Out – Biometry is locked out.
- Fallback – Fallback authentication mechanism selected.
- Limit Exceeded – Application retry limit exceeded.
- Misbehaving – Misbehaving caller PID:XXX has made too many authentication requests.

It is worth noting that “Cancelled” here does not have the same meaning as the Android version and is instead equivalent to Android’s “Negative” outcome.

*Table 5.11: Outcomes from Fingerprint using the Apple iPhone 8*

| Scenario | Outcome | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | Success | Cancelled | Locked Out | Fallback | Limit Exceeded | Misbehaving | |
| Sitting | 160 | 5 | 0 | 0 | 0 | 0 | 170 |
| Standing | 160 | 3 | 1 | 0 | 2 | 0 | 166 |
| Treadmill | 152 | 6 | 2 | 1 | 1 | 0 | 162 |
| Corridor | 147 | 4 | 0 | 0 | 0 | 9 | 160 |
| Factor - Wet | 1 | 119 | 5 | 4 | 1 | 0 | 130 |

*Table 5.12: Outcomes from Face using the Apple iPhone X*

| Scenario | Outcome | | | Total |
|---|---|---|---|---|
| | Success | Cancelled | Fallback | |
| Sitting | 134 | 1 | 0 | 135 |
| Standing | 135 | 1 | 0 | 136 |
| Treadmill | 134 | 2 | 0 | 136 |
| Corridor | 133 | 3 | 0 | 136 |
| Factor - Dark | 131 | 2 | 2 | 135 |

Once the initial FNMR and FRR results are broken down into the outcomes that make up those results, more in-depth performance analysis can be observed. Intriguingly, most "failed" cases can be attributed to user error or usability issues hidden under the current standard testing protocols that rely only on standard performance rates. For example, although looking at the data above, one of the reasons for failure is the user cancelling the authentication. It is, unfortunately, unclear from this data exactly why the user cancelled. Although some assumptions can be made regarding the factor scenarios when the user struggled to authenticate, they would likely quick the transaction.

## 5.11 Analysis and Reflection

The aim of presenting the questionnaire results is to provide insight into the demographics and behaviours of the test population. This information helps when testing a biometric system to have realistic expectations about the results achieved. As previously acknowledged, the test cohort described here is a student population, so it would be unreasonable to assume the results would extrapolate and mirror the general population. However, by surveying the participants, we can picture what opinions and trends begin to form. For example, comparing the two major smartphone operating system users' opinions, Android users (5.8 ± 1.4) are more satisfied with biometric authentication than their iOS counterparts (5.3 ± 1.3). These factors could cause an inherent bias towards biometric testing.

The first question relating to a participant's mobile device asked if they owned a mobile phone, and everyone involved in the study did. The remaining questions regarding participants' devices aimed to seek out the potential habitational impact that the participants in the trial could experience. In other words, the exposure a participant previously had towards a particular device and operating system would likely impact the practical usability and performance. This effect is already seen in the initial results as the first scenario tested (Sitting) has the most significant false non-match rates observed. This result also highlights a lesson learnt from conducting this data collection as no habitational or practise transactions were performed before starting the trial. A noticeable decrease in false non-match rates would likely be seen if the participants practised how to perform a transaction before starting. This recommendation was added to the framework from this experience.

Looking back towards the core factors of mobile biometric performance, 'users' was identified as one of these factors and sitting at the centre of the groupings of relationships. Meaning users can be viewed as the cornerstone of biometric performance, and their influence could determine the obtained results. A detailed view of the participants' breakdown can help put the achieved results into perspective. Looking at the results, the population is generally young adults who will likely be more familiar with mobile technology. However, there is no knowledge of how the system will perform around different populations, such as the elderly, where more training sessions are usually required [57].

## 5.12 Summary

This chapter introduced experimental data collection, discussing methodology and data collection. Next, the pre-and post-experiment questionnaire results were shown and analysed, including exploring the trends smartphone users used to secure their devices, including locking habits and satisfaction. Finally, the initial results from the experiment were introduced, showcasing the Boolean results from the devices and the breakdown of causes for the FNMR possible from developer access, finally followed by a discussion regarding the potential benefits of the framework method exploring a more dynamic range of scenarios for analysis.

The next chapter will take the samples from this experimental data collection and use third-party open-source algorithms to analyse the performance metrics obtained within the scenarios and environments. The aim is to prove the statistical significance of these choices and demonstrate the potential merit of some of the key concepts and ideas introduced within the performance framework.

# 6 Measuring and Analysing Mobile Biometric Performance Factors

## 6.1 Introduction

This chapter aims to present the research findings of the collected data when applied to open-source algorithms and take elements from the core factors to demonstrate the effectiveness of using these elements in forming the performance framework. The data explored here back up the core factors' findings and strengthen their inclusion within the performance framework. The result is an illustrative run-through applying the data to the observed results.

Using quality score enables further analysis of systems where access is limited, but access to a sample is available. Furthermore, it is possible to use the quality score as a substitute for a match score to obtain further performance information relating to a system. This analysis will focus on genuine verification results and comparisons between core factors.

The results presented throughout this chapter will, at times, where appropriate, be split in terms of the device used to capture the data. This separation is because the hardware was identified as a core performance factor that impacts the observed performance. For example, the device camera used to collect facial images or the microphone used to capture voice recordings should be presented separately. The main exception is the Iris images because an external mobile device (IriTech IriShield) was used to capture the samples. However, both Android devices were used as the controller for this device.

This chapter is split into roughly two parts. The first explores the impact of the scenarios and influencing factors using the data collected from the first data collection session presented in Chapter 5. In comparison, the second half explores the environmental impact using the data collected from the second (Outdoor) session. The statistical tests performed throughout this chapter, specifically ANOVA and pairwise T-test step-down method using Bonferroni adjustments, were performed with the help of the 'scikit-posthocs' Python library [128].

The chapter breakdown is as follows: Section 6.2 explores the habituation effect seen from the data. Section 6.3 introduces the open-source biometric algorithms used for further analysis of the data and the scenario results achieved using them. Section 6.4 explores some quality metrics obtained from the data, and Section 6.5 explores the effect of the motion observed. Section 6.6 looks at the impact of the environment and weather. Section 6.7 briefly explores usability, and Section 6.8 provides an initial investigation into the tailored impostors' technique. Finally, Section 6.9 provides a summary.

## 6.2 Habituation

Within biometrics, habituation refers to the user's familiarity with the system and concerns usability, which influence the system's performance [129]. When performing the experimental data collection, performing habituation transactions to allow the user to get familiar and comfortable with the system was omitted and later acknowledged as a limitation of the data as it skewed the results of the first scenario (sitting), which was intended to act as our baseline and optimal scenario conditions. To briefly explore the impact of this habituation effect, the FNMR for each attempt is shown separately in Table 6.1.

During the data collection, instructions were provided to the participants in the form of verbal instructions with visual aids (screenshots from the application) presented on laminated cards. This method was the primary form of communication presented to the participants. The participant's interaction with the device's biometric system was (where possible) the same as the one provided by the operating system; this was important so possible evaluation of the UI could be explored. However, no complete transaction was performed before the participant's first attempt. Usability increases with use and exposure, so one would expect to see this reflected in the results.

Although the habituation effect was known before the commencement of the data collection, the impact it could have had on the results was underrepresented. It is now acknowledged as a limitation of the test dataset that should have been considered more carefully. However, it is helpful to see the impact of habituation and usability on the UI of these systems. As proposed by the evaluation framework, the intention is to use the 'sitting' scenarios as the baseline. However, not taking careful consideration of the habituation effect beforehand could have an impact on how the results are evaluated.

When comparing scenarios and utilising this baseline scenario, it should be noted that this habituation effect is present. However, lessons can still be learnt from the outcome presented here and show that usability improves with use, partly noted by the transactional times presented later in this chapter. The same can be said for the biometric performance, as shown in Table 6.1. This chapter will explore the scenario's impact on biometric performance. However, it is interesting to see the habituation effect within such raw data, particularly when these tech companies pride themselves on intuitive UI and researchers know how critical human interaction is on biometric system performance [130].

In the context of this thesis, the habituation effect is likely more present within the initial intended baseline. Although the effect provides additional exciting insights, it may not provide the fairest comparison between scenarios as expected. Not taking this into account beforehand was an oversight of the data collection task. However, the data still provides valuable insights and an understanding of the proposed performance framework, as explored in the remainder of this chapter.

*Table 6.1: FNMR Attempt Breakdown*

| Device | Modality | Scenario | FNMR (%) | | | | |
|---|---|---|---|---|---|---|---|
| | | | Attempt # | | | | |
| | | | 1 | 2 | 3 | 4 | 5 |
| Samsung Galaxy S9 | Fingerprint | Sitting | 31 | 25 | 26 | 19 | 25 |
| | | Standing | 15 | 4 | 4 | 4 | 0 |
| | | Treadmill | 8 | 8 | 4 | 0 | 0 |
| | | Corridor | 18 | 0 | 4 | 4 | 4 |
| | | Factor (Wet) | 81 | 84 | 68 | 67 | 71 |
| | Face | Sitting | 5 | 12 | 9 | 8 | 4 |
| | | Standing | 9 | 9 | 9 | 9 | 9 |
| | | Treadmill | 9 | 9 | 9 | 9 | 9 |
| | | Corridor | 9 | 9 | 5 | 9 | 5 |
| | | Factor (Dark) | 0 | 0 | 0 | 0 | 0 |
| | Iris | Sitting | 36 | 28 | 28 | 24 | 24 |
| | | Standing | 14 | 17 | 21 | 13 | 21 |
| | | Treadmill | 21 | 25 | 30 | 30 | 30 |
| | | Corridor | 38 | 34 | 25 | 25 | 21 |
| | | Factor (Dark) | 31 | 20 | 20 | 20 | 28 |
| Google Pixel 2 | Fingerprint | Sitting | 32 | 11 | 6 | 13 | 3 |
| | | Standing | 17 | 3 | 11 | 6 | 11 |
| | | Treadmill | 11 | 16 | 16 | 11 | 6 |
| | | Corridor | 11 | 3 | 3 | 6 | 0 |
| | | Factor (Wet) | 88 | 83 | 71 | 70 | 82 |
| iPhone 8 | Fingerprint | Sitting | 4 | 10 | 0 | 4 | 0 |
| | | Standing | 4 | 4 | 4 | 0 | 7 |
| | | Treadmill | 10 | 10 | 4 | 0 | 0 |
| | | Corridor | 7 | 7 | 4 | 10 | 13 |
| | | Factor (Wet) | 95 | 100 | 100 | 100 | 100 |
| iPhone X | Face | Sitting | 0 | 4 | 0 | 0 | 0 |
| | | Standing | 0 | 0 | 4 | 0 | 0 |
| | | Treadmill | 0 | 4 | 4 | 0 | 0 |
| | | Corridor | 0 | 4 | 0 | 4 | 4 |
| | | Factor (Dark) | 0 | 8 | 4 | 4 | 0 |

Looking at the data on an attempt basis seems to support the theory of habituation impacting performance. The table shows that the first scenario sitting has some of the highest FNMR scores compared to the other scenarios, with the introduced factor(s) having the second-highest FNMR scores.

It is with the support of this data that a recommendation to introduce a trial (habituation) phase before measuring and capturing the results of the baseline scenario should be included, and attempts by the evaluators made to help the users compensate for any user errors that occur from improper use of the system.

## 6.3 Open-Source Biometric Algorithms

The main disadvantage to using commercial devices to perform a biometric evaluation is the closed nature of the algorithm (and sometimes hardware) controlling the biometric system. The proposed framework aims to help alleviate this problem by focusing more on usability when access is limited and expanding upon more significant security testing where access allows.

The experimental data collection focuses on the collection and evaluation of commercial devices. However, that leaves little room to explore further and evaluate the framework. Therefore, to go beyond the 'commercial' and 'developer' level of access, open-source algorithms were obtained to compare and match the obtained biometric samples from the mobile devices, allowing for the simulated evaluation of the framework at the 'tester' level.

The reasoning behind using open-source biometric algorithms, as opposed to commercial algorithms, was partly for the reproducibility of the results and the evaluation of algorithms that anyone could obtain and use in their biometric solutions. Where possible, the option was to use popular open-source algorithms backed by scientific research and recommended for research use. These algorithms will be referred to as Face Recognition (Face), USIT (Iris) and Deep Speaker (Voice) moving forward.

The first captured sample within the sitting scenario was used as the enrolment reference when using the algorithms. Then, all the following probes from the remaining scenarios were compared against this reference. This decision partly followed the framework that we want to have an 'optimal' scenario as the baseline for comparison. Furthermore, it is known that the enrolment can affect subsequent authentications [131]. Therefore, the analysis and the framework are present, assuming we have a high-quality enrolment. This approach allows the evaluators to ensure they have given the system the 'best' start for the evaluation.

The authors of the algorithms themselves admit that the algorithms are not perfect, and accuracy will drop when presented with challenging scenarios or conditions discussed below, demonstrating how the algorithm fits into the system's performance as a core factor. Usually, this is due to the training of machine learning models and the data used is too specific or 'clean' from noise.

### 6.3.1 Face (Face Recognition)

The Python' face_recognition' library [132], [133] was used as the basis for the face recognition algorithm. This library presents a Python wrapper for the C++ dlib library, a toolkit containing machine learning algorithms, including image recognition for face recognition and detection [134]. The model has an accuracy of 99.38% on the Labelled Faces in the Wild [118] benchmark. Notably, Face recognition states that particular user groups will experience a lower accuracy, specifically children and certain ethnic groups (Asian) owning to the training dataset used and states how "the face recognition model is only as good as the training data".

The algorithm can compare two images by extracting faces and encodings and comparing the two encodings. The algorithm returns a dissimilarity score between 0-1, where zero is more similar and one is less similar. By default, the algorithm uses a decision threshold of 0.6.

Two images were analysed from each photo, the original captured from the device and a cropped version cropped to the detected facial region. Doing so was to analyse the background impact of the images. A combination of two open-source face detection algorithms was used to achieve this. Firstly Multitask Cascaded Convolutional Networks (MTCNN) [135], [136], and the second was RetinaFace [137]–[139]. Both algorithms were applied to the data set, MTCNN, and then RetinaFace was applied to the images where MTCNN had failed to extract the facial region. Manual checking was then performed to confirm that facial regions were indeed extracted.

When facial recognition was performed against the original image, the reference (sitting) and probe were the original obtained image. Similarly, when performing the recognition against the cropped version, both the reference and probe were cropped to maintain consistency.

Table 6.2 shows the statistics for the genuine face verification scores achieved from the face recognition algorithm across the scenarios evaluated from the data collection. Unfortunately, due to an error with the development app, the images captured from the Apple iPhone X became inconsistent, and some were lost, meaning the image data collected from the Apple iPhone X is not considered when evaluating the scenarios.

In some instances, and most notably when observing the Factor – Dark scenario, the algorithm failed to detect (FTD) a face within the provided image, causing the number of participants and images to be less than contained within the data set.

*Table 6.2: Scenario Genuine Verification Score Statistics for 'Face'*

| Device | Scenario | Image | Participants | Amount | Mean ($\mu$) | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| Samsung Galaxy S9 | Sitting | Original | 25 | 114 | 0.17 ($\pm$0.08) | 0.15 | 0.06 | 0.40 |
| | | Cropped | 25 | 112 | 0.17 ($\pm$0.07) | 0.15 | 0.05 | 0.38 |
| | Standing | Original | 25 | 123 | 0.25 ($\pm$0.06) | 0.25 | 0.15 | 0.43 |
| | | Cropped | 25 | 116 | 0.26 ($\pm$0.06) | 0.25 | 0.14 | 0.44 |
| | Treadmill | Original | 24 | 116 | 0.31 ($\pm$0.07) | 0.28 | 0.19 | 0.57 |
| | | Cropped | 24 | 107 | 0.31 ($\pm$0.07) | 0.29 | 0.20 | 0.54 |
| | Corridor | Original | 24 | 121 | 0.29 ($\pm$0.06) | 0.29 | 0.21 | 0.56 |
| | | Cropped | 24 | 116 | 0.30 ($\pm$0.05) | 0.30 | 0.21 | 0.42 |
| | Factor – Dark | Original | 14 | 41 | 0.45 ($\pm$0.06) | 0.46 | 0.34 | 0.58 |
| | | Cropped | 5 | 7 | 0.40 ($\pm$0.04) | 0.39 | 0.35 | 0.46 |
| Google Pixel 2 | Sitting | Original | 33 | 100 | 0.16 ($\pm$0.05) | 0.15 | 0.05 | 0.31 |
| | | Cropped | 32 | 94 | 0.16 ($\pm$0.05) | 0.16 | 0.07 | 0.30 |
| | Standing | Original | 33 | 116 | 0.26 ($\pm$0.06) | 0.27 | 0.13 | 0.39 |
| | | Cropped | 32 | 102 | 0.26 ($\pm$0.06) | 0.26 | 0.13 | 0.44 |
| | Treadmill | Original | 35 | 124 | 0.29 ($\pm$0.05) | 0.28 | 0.19 | 0.43 |
| | | Cropped | 33 | 103 | 0.29 ($\pm$0.05) | 0.28 | 0.19 | 0.44 |
| | Corridor | Original | 34 | 141 | 0.30 ($\pm$0.06) | 0.29 | 0.20 | 0.47 |
| | | Cropped | 34 | 133 | 0.30 ($\pm$0.06) | 0.30 | 0.15 | 0.49 |
| | Factor – Dark | Original | 20 | 87 | 0.43 ($\pm$0.06) | 0.42 | 0.29 | 0.59 |
| | | Cropped | 18 | 49 | 0.42 ($\pm$0.06) | 0.41 | 0.32 | 0.54 |
| Apple iPhone 8 | Sitting | Original | 29 | 118 | 0.17 ($\pm$0.06) | 0.15 | 0.07 | 0.35 |
| | | Cropped | 28 | 110 | 0.17 ($\pm$0.06) | 0.17 | 0.08 | 0.37 |
| | Standing | Original | 27 | 135 | 0.27 ($\pm$0.06) | 0.27 | 0.15 | 0.40 |
| | | Cropped | 27 | 123 | 0.28 ($\pm$0.06) | 0.27 | 0.16 | 0.42 |
| | Treadmill | Original | 25 | 122 | 0.32 ($\pm$0.08) | 0.32 | 0.15 | 0.50 |
| | | Cropped | 24 | 88 | 0.31 ($\pm$0.10) | 0.30 | 0.15 | 0.80 |
| | Corridor | Original | 23 | 115 | 0.31 ($\pm$0.07) | 0.30 | 0.17 | 0.46 |
| | | Cropped | 23 | 110 | 0.31 ($\pm$0.07) | 0.31 | 0.15 | 0.48 |
| | Factor – Dark | Original | 16 | 64 | 0.42 ($\pm$0.08) | 0.41 | 0.26 | 0.58 |
| | | Cropped | 14 | 36 | 0.40 ($\pm$0.08) | 0.40 | 0.26 | 0.54 |

One of the first things to consider was to explore if there was any difference between the smartphone hardware when comparing the genuine verification scores. To evaluate this, we took the scores from the baseline sitting scenario and compared the devices using a one-way ANOVA statistical test to assess the significance of the verification scores between devices. The test results in $F(2, 329) = [0.56]$ at $p = 0.57$ $(p > 0.5)$ for the original images and $(2, 313) = [0.62]$ at $p = 0.54$ $(p > 0.5)$ indicating that there is no statistically significant difference between the genuine verification scores across the

smartphone devices used. This evidence can be seen in the box plot showing the sitting scenario scores across devices in Figure 6.1.

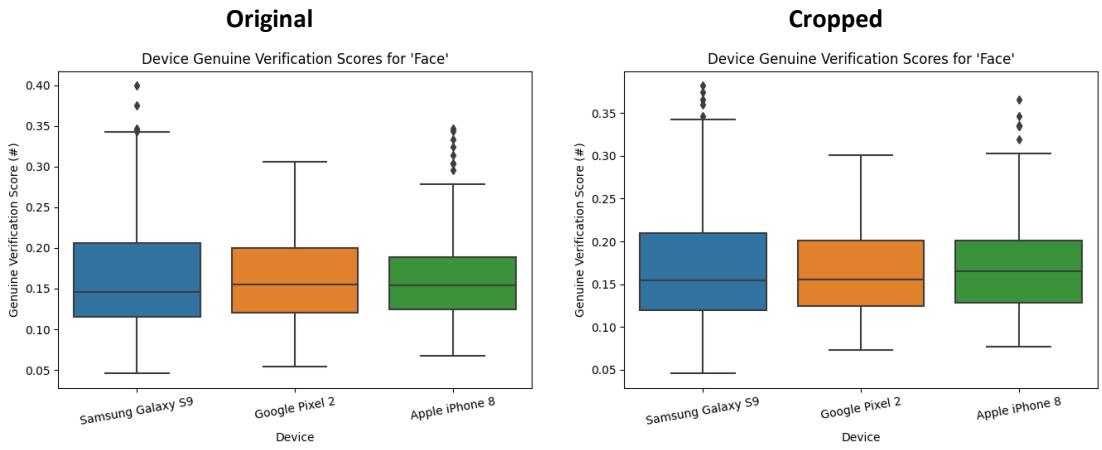



*Figure 6.1: Device Genuine Verification Scores Box Plot for Original and Cropped Image for 'Face'*

Having established that there does not appear to be any significant difference between the individual devices, the genuine verifications for each scenario were analysed for each device to see if the scenarios impacted the genuine verification scores.

Figure 6.2 shows a box plot for genuine verification scores obtained across the scenarios for the data collected from the Samsung Galaxy S9 for both the original and cropped image versions. Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between scenarios with the original images results in $F(4, 510) = [163.65]$ at $p = 5.18 \times 10^{-80}$ $(p < 0.001)$ indicating that there is some statistically significant difference between some scenarios on the genuine verification scores for the original images for the face recognition system. Similarly, when performing an ANOVA statistical test using the cropped images provides the result $F(4, 453) = [94.69]$ at $p = 1.75 \times 10^{-58}$ $(p < 0.001)$, equally indicating statistical significance for the verification scores with the cropped images.

Evidence of some statistical significance between scenarios was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments. Figure 6.2 shows a significance plot indicating where the significant pairs are. The plot highlights that most scenarios show significance against one another, indicating the importance of performing these scenario tests. However, it can also be seen that there is no significance between the two motion-based scenarios, which is potentially to be expected.

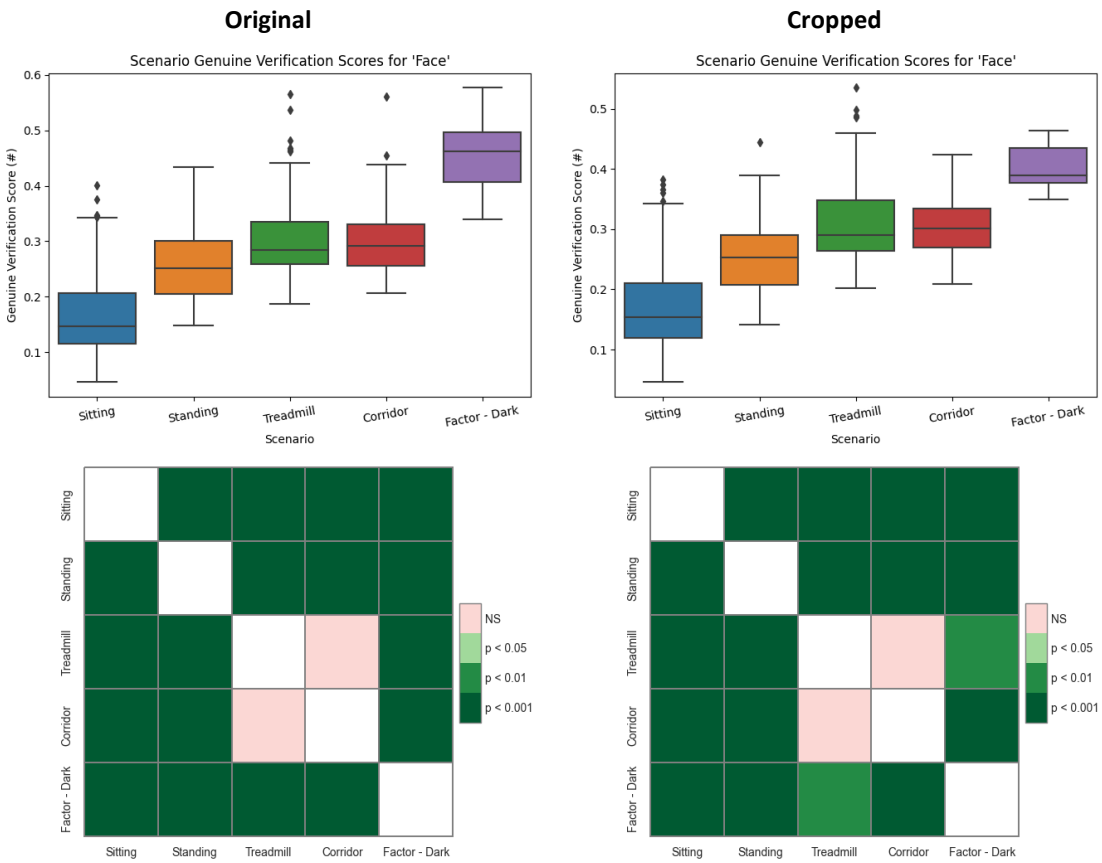



*Figure 6.2: Samsung Galaxy S9 Scenario Genuine Verification Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

This process was then repeated for the Google Pixel 2, and Figure 6.3 shows the box plot for the verification scores for both the original and cropped images. Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between scenarios with the original images results in $F(4, 563) = [270.89]$ at $p = 1.18 \times 10^{-129}$ $(p < 0.001)$ and cropped images results in in $F(4, 476) = [198.60]$ at $p = 5.11 \times 10^{-100}$ $(p < 0.001)$ indicating that there is some statistically significant difference between scenarios for the genuine verification scores.

Following this with a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments provides the significance plots shown in Figure 6.3. The results for the Google Pixel 2 show a similar pattern to the Samsung Galaxy S9, indicating some improbability towards using the scenario approach across devices. The only minor difference is that the cropped images show significance between the two motion-based scenarios that were not apparent with the original images.

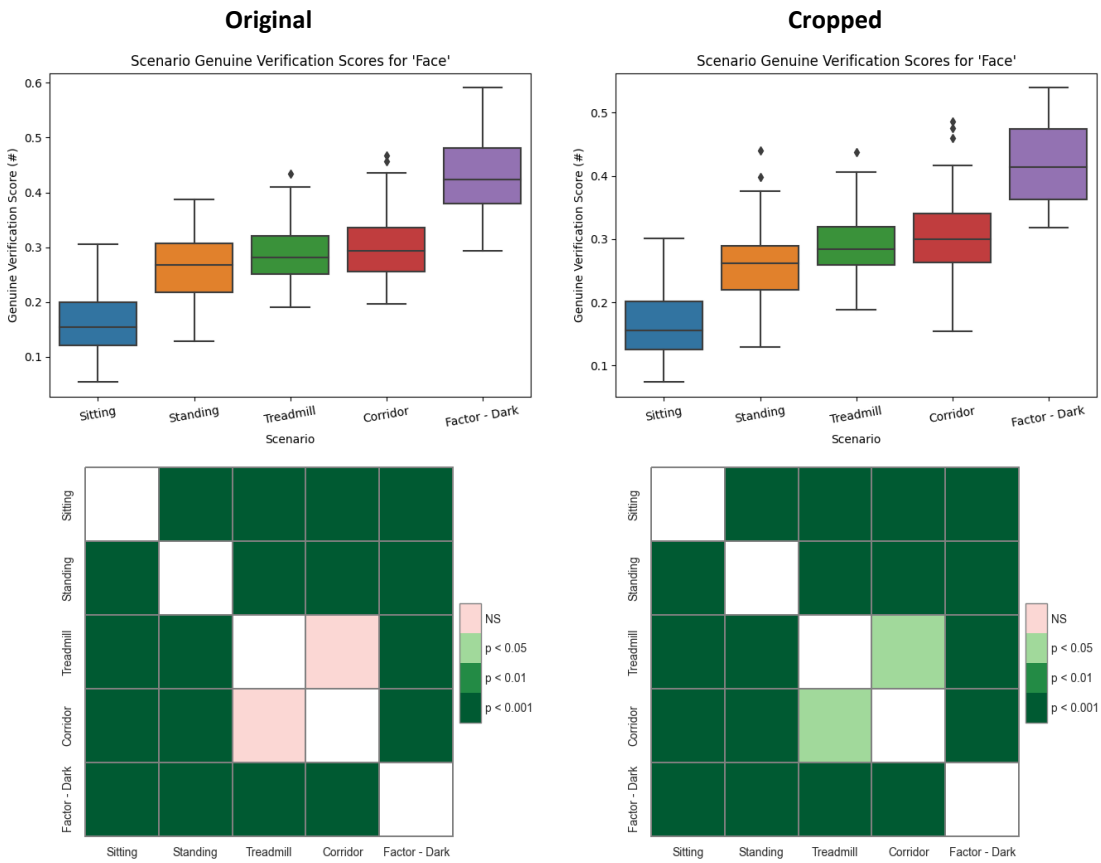



*Figure 6.3: Google Pixel 2 Scenario Genuine Verification Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

Once more, this analysis was performed on the Apple iPhone 8. The box plot for the verification scores across the scenarios for both the original and cropped images is shown in Figure 6.4. A one-way ANOVA statistical test to assess the statistical significance of the verification scores between scenarios for the original images results in $F(4, 549) = [158.28]$ at $p = 5.47 \times 10^{-90}$ $(p < 0.001)$ and for the cropped images results in $F(4, 462) = [93.01]$ at $p = 5.66 \times 10^{-58}$ $(p < 0.001)$ indicating that there is some statistically significant difference between some scenarios on the genuine verification scores for this face recognition system.

Performing a statistical post hoc pairwise t-test provides the significance plot in Figure 6.4. The results support the evaluation of various scenarios as part of the performance framework with all three mobile devices. In each case, the sitting scenario used as a baseline has significance, with the remaining scenarios supporting the concept of using this scenario as the baseline.

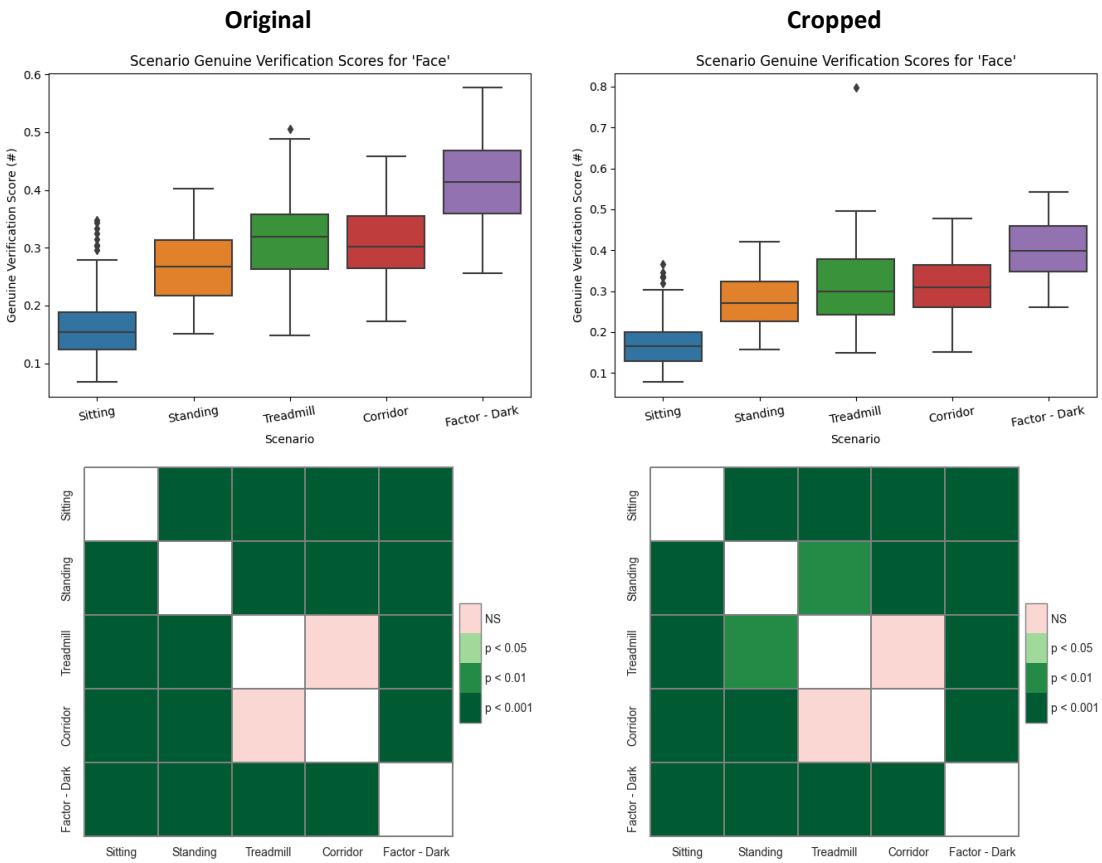



*Figure 6.4: Apple iPhone 8 Scenario Genuine Verification Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

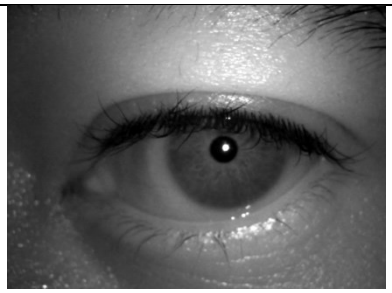
### 6.3.2 Iris (USIT)

For the Iris data, the USIT -- University of Salzburg Iris-Toolkit v3.0.0 [140] was used for the analysis. USIT is an open-source software development kit for iris biometric research to help achieve comparability and reproducibility of research results. There is a three-step process to follow to perform iris recognition with this software:

1. Iris Pre-processing (Segmentation)
2. Feature Extraction
3. Feature Comparison

The approach taken here was to use Contrast-adjusted Hough Transform (CAHT) for segmentation followed by Discrete Cosine Transform (DCT) [141] for feature extraction into an iris code and later comparison as demonstrated. Table 6.3 demonstrates these steps (note that a small extraction section has been zoomed in and shown here for demonstration purposes).

*Table 6.3: Iris Recognition using USIT*

| | |
|---|---|
| **Iris Image** |  |
| **Segmentation (CAHT)** |  |
| **Extraction (DCT)** |  |

Using the USIT library allows further analysis of the collected data with additional insight into potential match scores that were not possible using the commercial devices alone. The results of running the collected iris dataset through the USIT library are shown in Table 6.4. The immediate observation is that there seems to be a minor difference between the genuine scores obtained across the scenarios. Therefore, a statistical test was used to check for any difference between the scenarios. Figure 6.5 shows the box plot for the genuine verification scores across the scenarios using the USIT algorithms (CAHT and DCT).

The algorithm produces a dissimilarity score between a reference and a probe between 0 and 1, where the closer to zero is more similar, and the closer to one is less similar.

*Table 6.4: Scenario Genuine Verification Score Statistics for 'Iris'*

| Device | Scenario | Participants | Amount | Mean ($\mu$) | Median | Min | Max |
|---|---|---|---|---|---|---|---|
| IriTech IriShield | Sitting | 60 | 278 | 0.39 ($\pm$0.09) | 0.41 | 0.00 | 0.46 |
| | Standing | 59 | 297 | 0.39 ($\pm$0.07) | 0.41 | 0.00 | 0.48 |
| | Treadmill | 59 | 291 | 0.39 ($\pm$0.07) | 0.41 | 0.00 | 0.48 |
| | Corridor | 57 | 287 | 0.41 ($\pm$0.06) | 0.42 | 0.00 | 0.47 |
| | Factor – Dark | 60 | 321 | 0.40 ($\pm$0.07) | 0.42 | 0.00 | 0.47 |

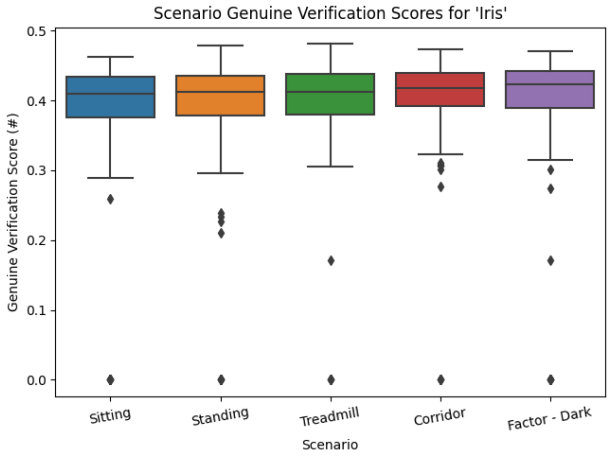



*Figure 6.5: Scenario Genuine Verification Scores Box Plot for 'Iris'*

Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between scenarios results in $F(4, 1469) = [3.46]$ at $p = 8.05 \times 10^{-3}$ $(p < 0.01)$ There is some statistically significant difference between some scenarios on the genuine verification scores for the iris recognition system. Evidence of some statistical significance between scenarios was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

Figure 6.6 shows the results of this post hoc analysis in the form of a significance plot highlighting the significant relationships between groups. The observation made is that there was no significance found between our stationary and motion scenarios and only two groups have significance between them, that being Sitting and Corridor. This result could highlight how using infrared imaging has offset the effect of different scenarios and are a factor (dark). In addition, this result indicates how the hardware can impact performance as one of the core factors.

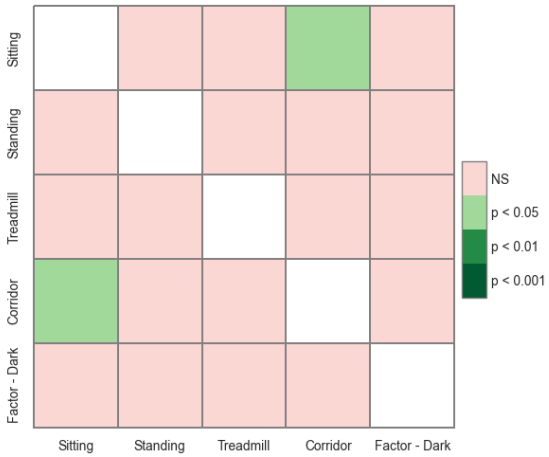



*Figure 6.6: Scenario Genuine Verification Scores P-Value Significance Plot for 'Iris'*

### 6.3.3 Voice (Deep Speaker)

"Deep Speaker" [142] is "a neural speaker embedding system that maps utterances to a hypersphere where speaker similarity is measured by cosine similarity … Deep Speaker reduces the verification equal error rate by 50% (relatively) and improves the identification accuracy by 60% (relatively) on a text-independent dataset". The authors of the GitHub repository note that "models were trained on clean speech data. Remember that the performance will be lower on noisy data" [143]. The analysis presented throughout the "ResCNN Softmax+Triplet trained" model was selected as it currently provides the best performance with an EER of 0.025 when trained against the LibriSpeech [144] dataset.

Deep Speaker produces a similarity score between a reference and a probe using a cosine similarity method. It can produce a score between -1 and one, depending on how similar the samples are. Leaning towards one indicates more similarities, and leaning towards -1 means less similarity. In the experience of using the library, it usually produces results between 0 and 1 and only occasionally drops below 0.

In order for the voice data to be parsed by Deep Speaker, the audio from the devices needed to be converted into ".wav" format from either ".m4a" (iOS) or ".3gp" (Android), the audio was originally recorded in this format as it is the default audio encoding for those devices. The open-source FFmpeg library [145] was used to convert all the audio files into ".wav" format to achieve the conversion for the analysis.

Table 6.5 provides the statistical data when using DeepSpeaker as the biometric algorithm against the collected voice data to obtain genuine verification scores for all the devices and scenarios trialled during the experimental data collection. Using this algorithm also provides the opportunity to compare the devices. Each device has a microphone, but one previously identified core factor was hardware, centred around how different hardware can affect performance. It is possible to compare the device and test if there is a significant difference using the defined optimal scenario (sitting). Figure 6.7 shows the box plot of the genuine verification scores when comparing the devices in the sitting scenario.

Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between devices (hardware) results in $F(3, 485) = [34.25]$ at $p = 4.31 \times 10^{-20}$ $(p < 0.001)$ indicating that there is some statistically significant difference between the devices on the genuine verification scores for the voice recognition system. The evidence of some statistical significance between devices was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

Figure 6.8 shows the results of this post hoc analysis in the form of a significant plot highlighting significant relationships between the devices. The analysis supports that devices are significant (hardware) in biometric performance. It is shown that there is significance between all pairs of devices except for the two iOS devices, potentially supporting a claim that manufacturers will use the same parts,

in this case, microphones, across devices. The same microphone may be used for both the iPhone 8 and the iPhone X. Equally, there is significance, although slightly less, between the two Android devices when comparing Android against iOS.

*Table 6.5: Scenario Genuine Verification Score Statistics for 'Voice'*

| Device | Scenario | Participants | Amount | Mean ($\mu$) | Median | Min | Max |
|---|---|---|---|---|---|---|---|
| Samsung Galaxy S9 | Sitting | 24 | 98 | 0.85 ($\pm$0.10) | 0.89 | 0.56 | 0.95 |
| | Standing | 24 | 120 | 0.78 ($\pm$0.10) | 0.79 | 0.51 | 0.93 |
| | Treadmill | 24 | 122 | 0.78 ($\pm$0.09) | 0.79 | 0.52 | 0.93 |
| | Corridor | 24 | 120 | 0.77 ($\pm$0.09) | 0.80 | 0.54 | 0.91 |
| | Factor – Quiet | 23 | 118 | 0.76 ($\pm$0.08) | 0.78 | 0.50 | 0.88 |
| | Factor – Loud | 23 | 116 | 0.68 ($\pm$0.09) | 0.69 | 0.46 | 0.85 |
| Google Pixel 2 | Sitting | 35 | 146 | 0.88 ($\pm$0.07) | 0.90 | 0.51 | 0.97 |
| | Standing | 35 | 177 | 0.83 ($\pm$0.08) | 0.84 | 0.54 | 0.95 |
| | Treadmill | 35 | 178 | 0.82 ($\pm$0.08) | 0.84 | 0.60 | 0.94 |
| | Corridor | 35 | 177 | 0.80 ($\pm$0.08) | 0.80 | 0.48 | 0.94 |
| | Factor – Quiet | 35 | 187 | 0.75 ($\pm$0.08) | 0.77 | 0.40 | 0.90 |
| | Factor – Loud | 35 | 182 | 0.64 ($\pm$0.09) | 0.65 | 0.36 | 0.88 |
| Apple iPhone 8 | Sitting | 33 | 135 | 0.74 ($\pm$0.19) | 0.79 | 0.01 | 0.94 |
| | Standing | 33 | 167 | 0.61 ($\pm$0.19) | 0.65 | -0.02 | 0.94 |
| | Treadmill | 33 | 166 | 0.63 ($\pm$0.17) | 0.65 | 0.02 | 0.95 |
| | Corridor | 33 | 167 | 0.56 ($\pm$0.18) | 0.58 | -0.05 | 0.93 |
| | Factor – Quiet | 33 | 168 | 0.55 ($\pm$0.16) | 0.56 | 0.06 | 0.87 |
| | Factor – Loud | 33 | 167 | 0.43 ($\pm$0.14) | 0.46 | 0.01 | 0.77 |
| Apple iPhone X | Sitting | 27 | 110 | 0.78 ($\pm$0.14) | 0.83 | 0.26 | 0.93 |
| | Standing | 27 | 139 | 0.67 ($\pm$0.11) | 0.68 | 0.25 | 0.83 |
| | Treadmill | 27 | 140 | 0.64 ($\pm$0.14) | 0.67 | 0.20 | 0.83 |
| | Corridor | 27 | 140 | 0.58 ($\pm$0.12) | 0.60 | 0.23 | 0.79 |
| | Factor – Quiet | 27 | 138 | 0.57 ($\pm$0.11) | 0.58 | 0.20 | 0.79 |
| | Factor – Loud | 27 | 140 | 0.44 ($\pm$0.12) | 0.46 | 0.15 | 0.64 |

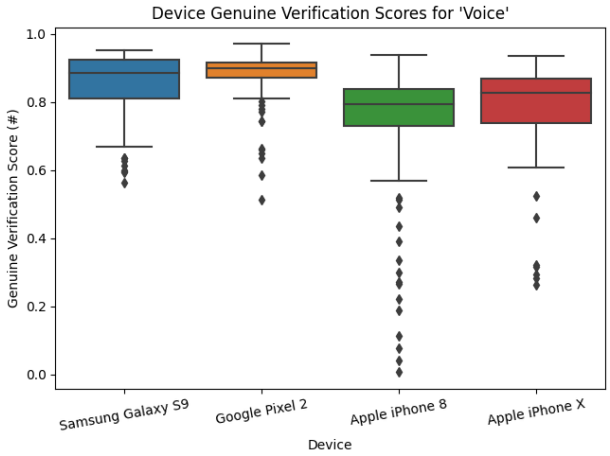



*Figure 6.7: Device Genuine Verification Scores Box Plot for 'Voice'*

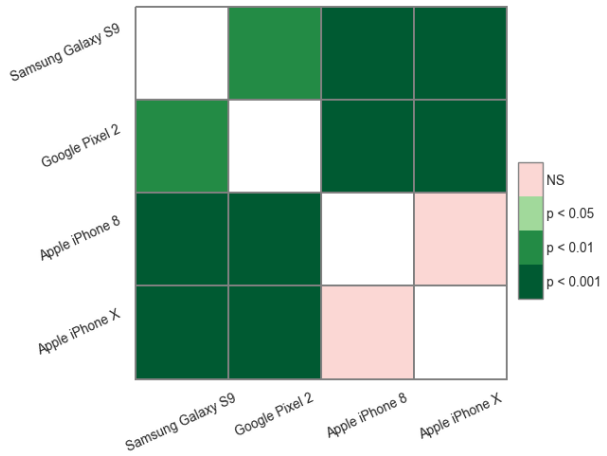



*Figure 6.8: Device Genuine Verification Scores P-Value Significance Plot for 'Voice'*

The next step is to examine the scenario impact of each device, having demonstrated the potential impact of the various devices. Firstly Figure 6.9 shows a box plot for the scenario using the Samsung Galaxy S9. Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between scenarios results in the following for each device:

- Samsung Galaxy S9: $F(5, 688) = [39.86]$ at $p = 5.13 \times 10^{-36}$ $(p < 0.001)$
- Google Pixel 2: $F(5, 1041) = [187.05]$ at $p = 3.66 \times 10^{-142}$ $(p < 0.001)$
- Apple iPhone 8: $F(5, 964) = [51.91]$ at $p = 9.82 \times 10^{-48}$ $(p < 0.001)$
- Apple iPhone X: $F(5, 801) = [103.08]$ at $p = 5.79 \times 10^{-84}$ $(p < 0.001)$

It indicates some statistically significant differences between the scenarios on the genuine verification scores for the voice recognition system. Evidence of some statistical significance between devices was followed by performing a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

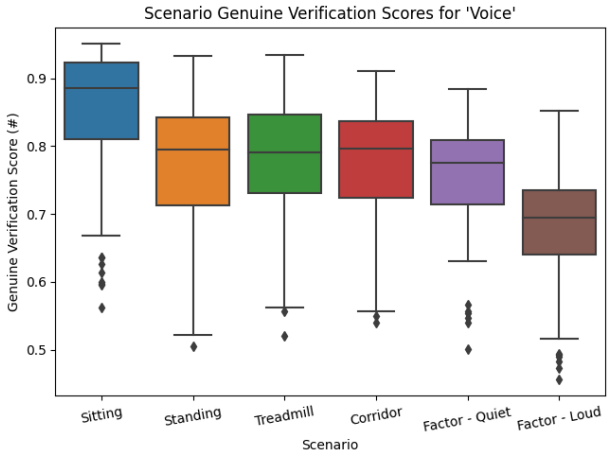



*Figure 6.9: Samsung Galaxy S9 Scenario Genuine Verification Scores Box Plot for 'Voice'*

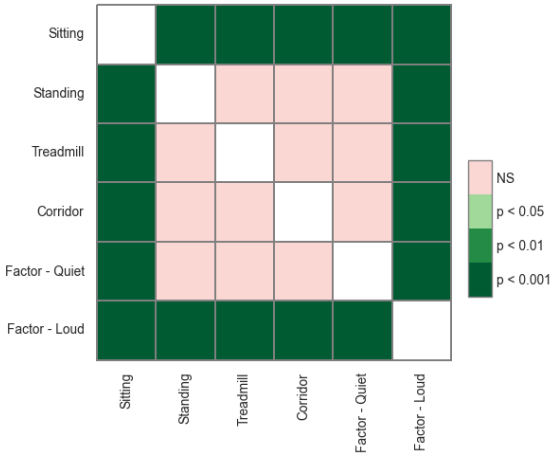



*Figure 6.10: Samsung Galaxy S9 Scenario Genuine Verification Scores P-Value Significance Plot for 'Voice'*

Figure 6.10 shows the significance plot for the verification scores for the Samsung Galaxy S9. The main observation is how the baseline scenario (sitting) shows significance against the remaining scenarios. However, this is likely because it is the same scenario used for the enrolment (reference) sample. Moreover, equally, the introduced influencing factor of the background shows significance when it is loud but not when it is quiet.

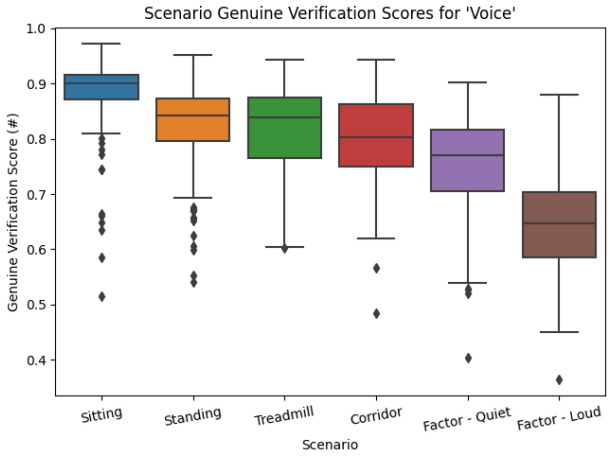



*Figure 6.11: Google Pixel 2 Scenario Genuine Verification Scores Box Plot for 'Voice'*

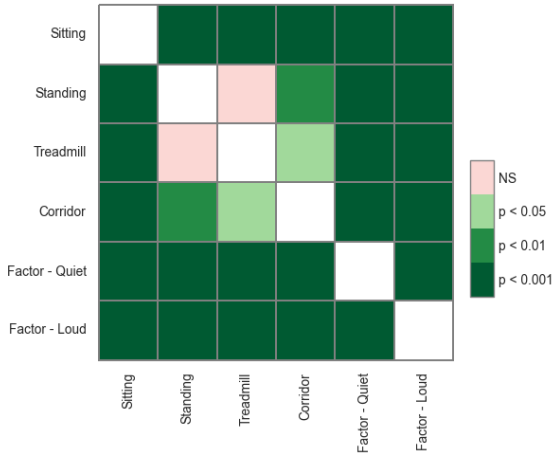



*Figure 6.12: Google Pixel 2 Scenario Genuine Verification Scores P-Value Significance Plot for 'Voice'*

The Google Pixel 2 shows more significance between scenarios than the previous Samsung Galaxy S9, including the baseline scenario and the influencing factors introducing noise into the background.

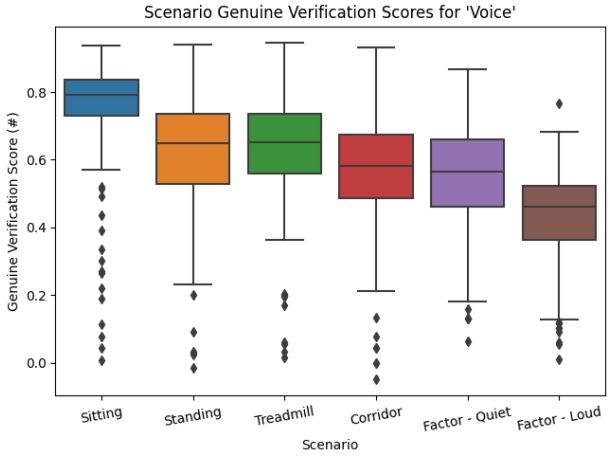



*Figure 6.13: Apple iPhone 8 Scenario Genuine Verification Scores Box Plot for 'Voice'*

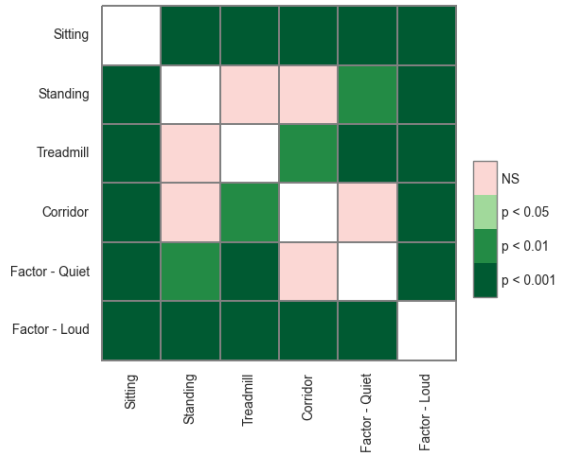



*Figure 6.14: Apple iPhone 8 Scenario Genuine Verification Scores P-Value Significance Plot for 'Voice'*

The two iPhone devices (8 and X) show a similar pattern for the verification scores between scenarios caused by having the same manufacturer (Apple) and identical components. The significance between scenarios is between the baseline, the influenced factor (noise), and the motion-based scenarios.

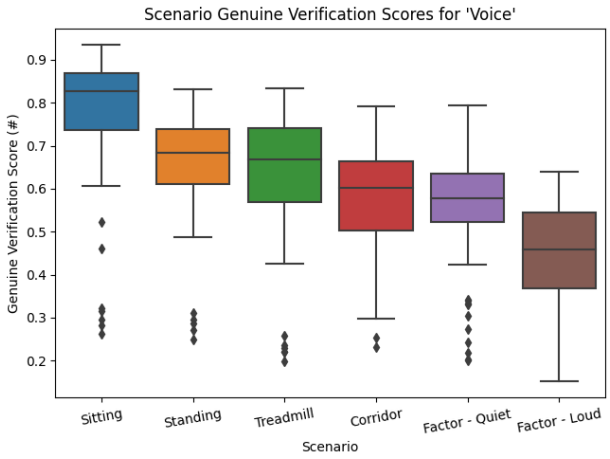



*Figure 6.15: Apple iPhone X Scenario Genuine Verification Scores Box Plot for 'Voice'*

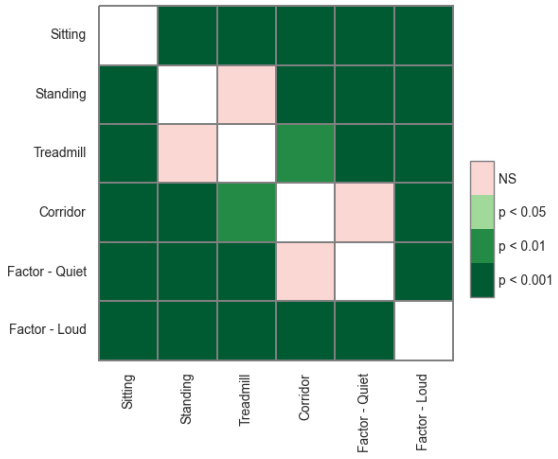



*Figure 6.16: Apple iPhone X Scenario Genuine Verification Scores P-Value Significance Plot for 'Voice'*

The voice data showed the most variation in verification scores between devices compared to the other tested modalities (Face, Iris). The results presented here showcase how scenarios influence the verification scores and, by extension, the performance of the mobile biometric systems supporting their inclusion of specified scenarios within the performance framework.

## 6.4 Quality

A biometric quality score is a "quantitative value of the fitness of a biometric sample to accomplish or fulfil the comparison decision" [5]. Therefore, sample quality is an essential link for analysing the performance of a system, as illustrated by its inclusion within the Human-Biometrics Sensor Interaction (HBSI) model serving as the intersection between the Sensor and Biometric System. "Sample quality is the important link between these two components because the image or sample acquired by the biometric sensor must contain the characteristics or features needed by the biometric system to enrol or match a user in the biometric system" [146].

Yao *et al.* [147] stated that "in a deployed system, the poor acquisition of samples perhaps constitutes the single most important reason for high false reject/accept rates", highlighting how vital the quality score can be to determining performance. There exist some approaches to adopt a standard to create quality assessments for various modalities, including NIST Fingerprint Image Quality (NFIQ) [148] and Face Recognition Vendor Test (FRVT) Quality Assessment [149].

Quality score information can be obtained directly from the manufacturer in the form of an API that allows access to this information or by extracting the sample collected from the capture sensor to analyse offline. The quality score can become more critical to assessing the performance of systems where performance information is limited, such as when the level of access is 'commercial'. In these situations where performance results, such as match score data, are not readily available, it is possible to use quality scores to indicate performance.

Hernandez-Ortega *et al.* developed a face quality assessment tool to "predict the suitability of a specific input image for face recognition purposes" [150]. As part of this, they stated that the "variability [in samples] is associated with the image acquisition conditions … other factors are more related to the properties of the face itself. All these factors influence the quality of the face samples, which is understood as a predictor of the goodness of a given face image for recognition purposes. That is, quality is an estimator of biometric performance".

Overall, quality checks are essential for enrolment, verification, identification, or de-duplication. During image acquisition, it can be used to select the best image from streaming video. It also provides feedback to improve the quality of image capture. During enrolment and identification, it can help reject unqualified images and provide actionable feedback to improve accuracy. A high correlation of quality score with matching accuracy helps reduce error rates [151].

Using the collected data allows for examining the quality of samples for the scenarios and environments explored.

### 6.4.1 Face

For this trial, the facial quality assessment scores were provided by FaceQnet V0 [150] [152], who states that "the results of a computerised system are only as reliable as the data you input. If you input data that is garbage, the result will be unreliable garbage". "FaceQnet can be used as a "black box" that receives a face image and outputs a quality measure between 0 and 1 related to the face recognition accuracy. This quality measure can be understood as proximity between the input image and a hypothetical corresponding [International Civil Aviation Organization] ICAO compliant face image".

Two pre-made models are provided, V0 and V1. For the results presented within this chapter, V0 is used. This decision is partly due to some inconsistent and unexpected behaviour from V1 compared to V0. For example, dark images with hardly visible faces score higher than visible faces. This observation was from manually comparing the results supplied by V0 and V1. For this reason, V0 was used.

Comparing the quality scores between devices using the baseline scenario (sitting) using a one-way ANOVA statistical test results in $F(2, 418) = [9.28]$ at $p = 1.14 \times 10^{-4}$ $(p < 0.001)$ for the original images and $F(2, 414) = [4.50]$ at $p = 1.17 \times 10^{-2}$ $(p < 0.05)$ for the cropped images. Both show statistical significance between the devices used, indicating that the hardware, in this case, the camera embedded into the smartphones, can cause a difference between the obtained quality scores. Figure 6.17 shows the corresponding box and significance plots for the face image quality score device comparison.

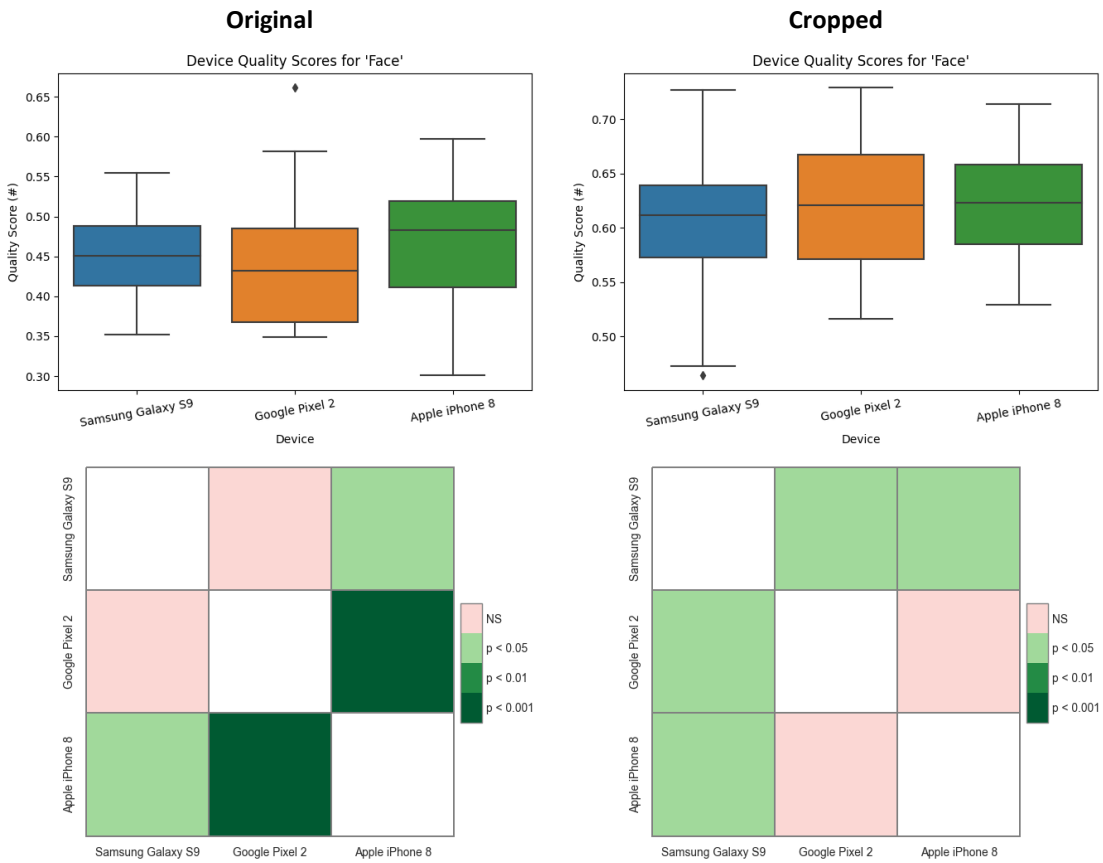



*Figure 6.17: Device Quality Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

The results show how the quality is impacted by the device (hardware) used. However, it is interesting that this did not correlate significantly with the verification scores obtained across devices using the face recognition algorithm. However, it is still worth showing the device impact from a quality score perspective because of the potential impact on the verification performance.

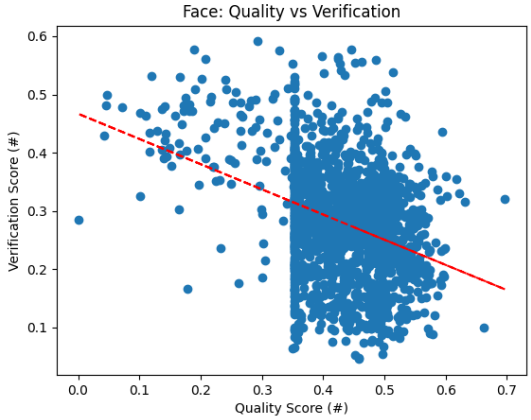



*Figure 6.18: Quality Scores vs Verification Scores for the Original Images for 'Face'*

Figure 6.18 and Figure 6.19 show the comparison between quality and verification for the original and cropped images, respectively. A trendline is also displayed for analysis. For the original images, this trend line has the equation $y = -0.43x + (0.47)$, and for the cropped images, the trendline has the equation $y = -0.52x + (0.58)$. In both cases we see a negative correlation between quality and verification, meaning there is support for the argument that higher quality images result in better verification results, as the verification score for face is a dissimilarity score. This observation becomes more apparent when comparing the cropped images and could be caused by FaceQnet performing better when the images are cropped. However, there are arguments for improving the quality scores for images not cropped to the facial region by applications like FaceQnet.

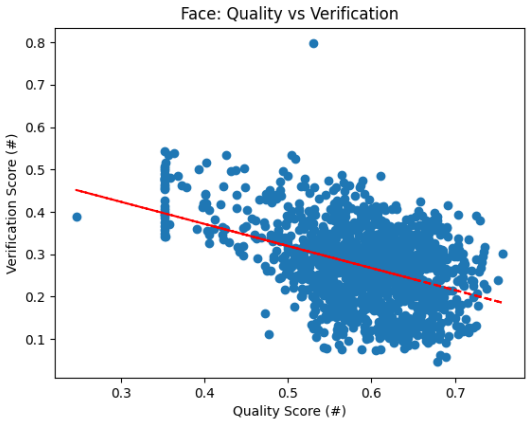



*Figure 6.19: Quality Scores vs Verification Scores for the Cropped Images for 'Face'*

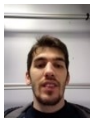
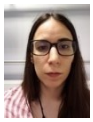
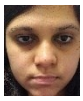
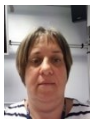
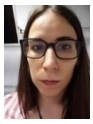
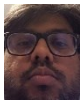
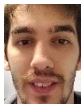
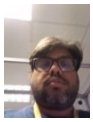
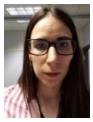
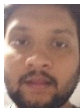
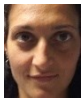
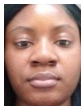
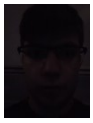
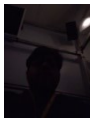
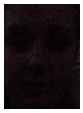
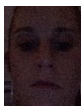
Table 6.6 shows the face quality scores obtained across the scenarios for the cropped and uncropped images for the Samsung Galaxy S9. When considering the scenario quality scores for the original unmodified images, a one-way ANOVA statistical test to assess the statistical significance of the quality scores results in the following:

- Samsung Galaxy S9
  - Original: $F(4, 652) = [342.82]$ at $p = 1.04 \times 10^{-158}$ $(p < 0.001)$
  - Cropped: $F(4, 544) = [267.34]$ at $p = 6.96 \times 10^{-127}$ $(p < 0.001)$
- Google Pixel 2
  - Original: $F(4, 723) = [186.65]$ at $p = 8.00 \times 10^{-110}$ $(p < 0.001)$
  - Cropped: $F(4, 649) = [444.62]$ at $p = 2.94 \times 10^{-184}$ $(p < 0.001)$
- Apple iPhone 8
  - Original: $F(4, 671) = [23.16]$ at $p = 5.98 \times 10^{-18}$ $(p < 0.001)$
  - Cropped: $F(4, 561) = [91.22]$ at $p = 1.03 \times 10^{-59}$ $(p < 0.001)$

All tests indicated some statistically significant difference between the facial image quality scores for both the original and cropped image quality scores across the scenarios. Evidence of some statistical

significance between scenarios was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments for each device for both the original and cropped images. Figure 6.20 shows the box plot of the quality scores and the significance plot between scenarios for the quality scores with the Samsung Galaxy S9.

*Table 6.6: Samsung Galaxy S9 Scenario Quality Score Statistics for 'Face'*

| Device | Scenario | Image | Participants | Amount | Average | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| Samsung Galaxy S9 | Sitting | Original | 25 | 139 | 0.45 (±0.05) | 0.45 | 0.35  | 0.55  |
| | | Cropped | 25 | 139 | 0.61 (±0.05) | 0.61 | 0.46  | 0.73  |
| | Standing | Original | 25 | 124 | 0.44 (±0.06) | 0.45 | 0.18  | 0.55  |
| | | Cropped | 25 | 123 | 0.59 (±0.05) | 0.59 | 0.44  | 0.69  |
| | Treadmill | Original | 24 | 121 | 0.43 (±0.05) | 0.43 | 0.35  | 0.53  |
| | | Cropped | 24 | 121 | 0.56 (±0.04) | 0.56 | 0.44  | 0.65  |
| | Corridor | Original | 24 | 122 | 0.46 (±0.06) | 0.47 | 0.33  | 0.58  |
| | | Cropped | 24 | 121 | 0.61 (±0.06) | 0.6` | 0.47  | 0.72  |
| | Factor - Dark | Original | 25 | 151 | 0.21 (±0.10) | 0.20 | 0.00  | 0.45  |
| | | Cropped | 12 | 45 | 0.31 (±0.09) | 0.35 | 0.00  | 0.57  |

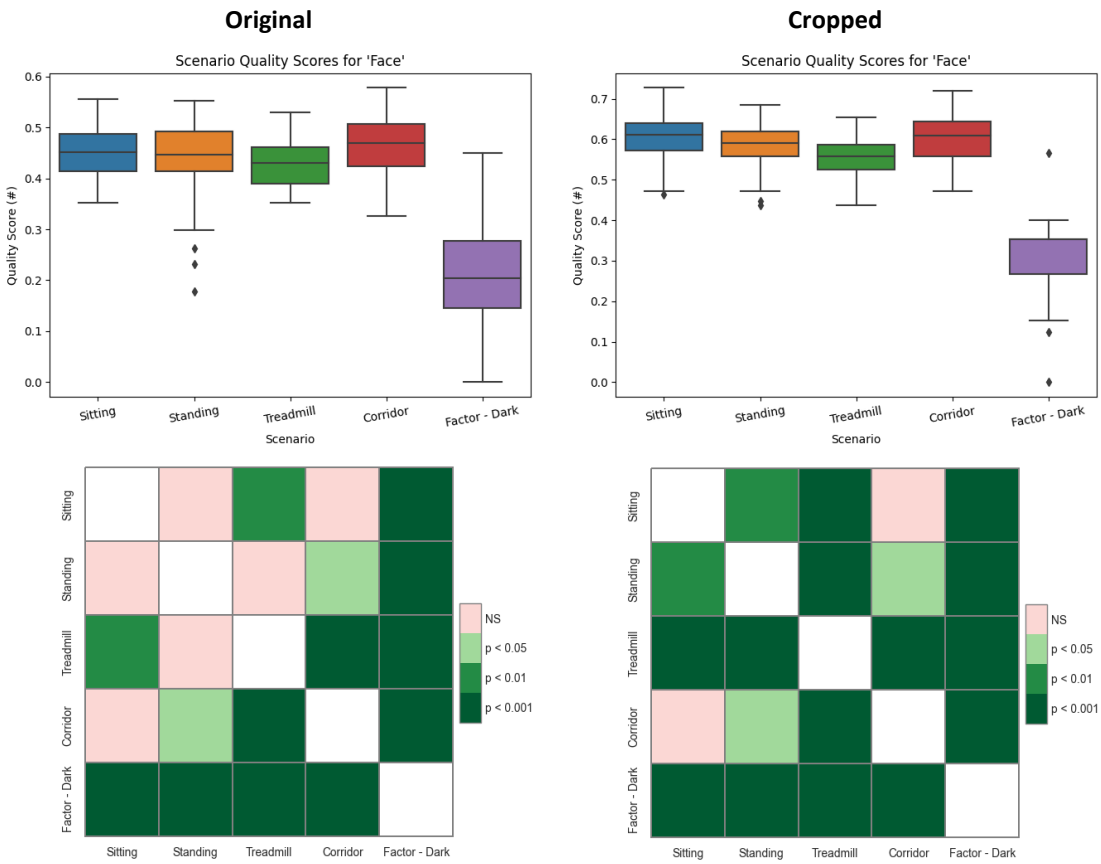
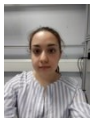
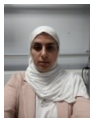
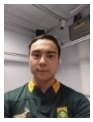
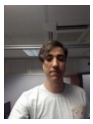
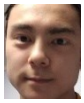
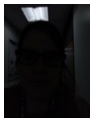
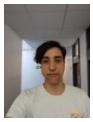
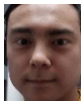
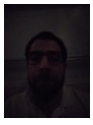
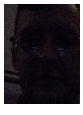



*Figure 6.20: Samsung Galaxy S9 Scenario Quality Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

*Table 6.7: Google Pixel 2 Scenario Quality Score Statistics for 'Face'*

| Device | Scenario | Image | Participants | Amount | Average | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| Google Pixel 2 | Sitting | Original | 35 | 135 | 0.43 (±0.07) | 0.43 | 0.35  | 0.66  |
| | | Cropped | 35 | 131 | 0.62 (±0.05) | 0.62 | 0.52  | 0.73  |
| | Standing | Original | 33 | 116 | 0.44 (±0.07) | 0.42 | 0.31  | 0.58  |
| | | Cropped | 32 | 111 | 0.61 (±0.06) | 0.62 | 0.44  | 0.73  |
| | Treadmill | Original | 35 | 128 | 0.42 (±0.06) | 0.41 | 0.35  | 0.58  |
| | | Cropped | 35 | 127 | 0.58 (±0.06) | 0.58 | 0.48  | 0.73  |
| | Corridor | Original | 34 | 142 | 0.44 (±0.08) | 0.44 | 0.08  | 0.70  |
| | | Cropped | 34 | 142 | 0.62 (±0.07) | 0.63 | 0.16  | 0.75  |
| | Factor - Dark | Original | 35 | 192 | 0.24 (±0.11) | 0.24 | 0.03  | 0.48  |
| | | Cropped | 30 | 143 | 0.34 (±0.09) | 0.35 | 0.00  | 0.50  |

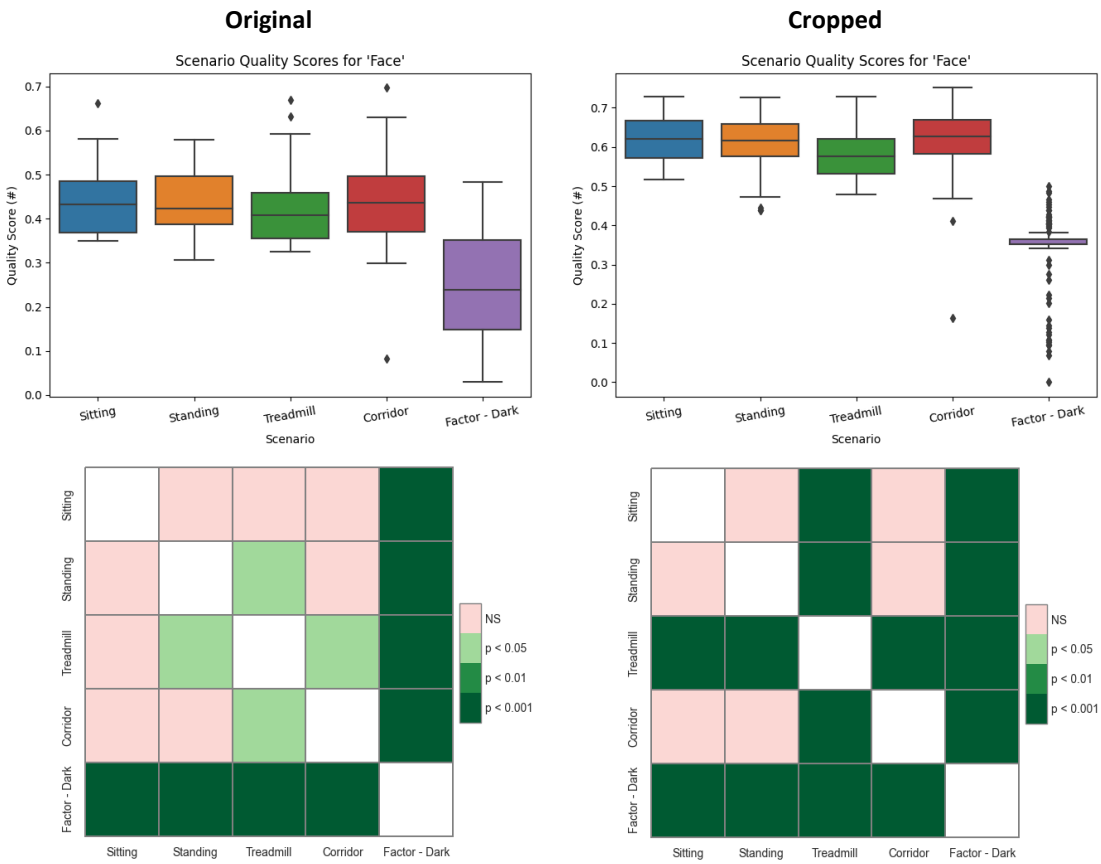
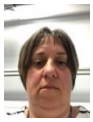
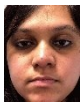
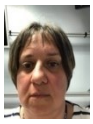
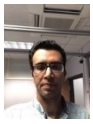
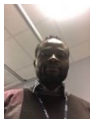
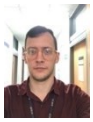
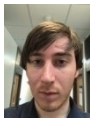
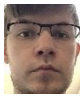
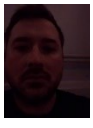



*Figure 6.21: Google Pixel 2 Scenario Quality Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

*Table 6.8: Apple iPhone 8 Scenario Quality Score Statistics for 'Face'*

| Device | Scenario | Image | Participants | Amount | Average | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| iPhone 8 | Sitting | Original | 29 | 147 | 0.47 (±0.07) | 0.48 | 0.30  | 0.60  |
| | | Cropped | 29 | 147 | 0.62 (±0.05) | 0.62 | 0.53  | 0.71  |
| | Standing | Original | 28 | 140 | 0.46 (±0.08) | 0.47 | 0.00  | 0.59  |
| | | Cropped | 28 | 140 | 0.60 (±0.06) | 0.60 | 0.38  | 0.70  |
| | Treadmill | Original | 27 | 136 | 0.44 (±0.06) | 0.44 | 0.35  | 0.63  |
| | | Cropped | 27 | 136 | 0.56 (±0.05) | 0.57 | 0.40  | 0.68  |
| | Corridor | Original | 23 | 115 | 0.47 (±0.07) | 0.47 | 0.35  | 0.60  |
| | | Cropped | 23 | 115 | 0.63 (±0.06) | 0.63 | 0.51  | 0.76  |
| | Factor - Dark | Original | 24 | 119 | 0.39 (±0.07) | 0.38 | 0.08  | 0.53  |
| | | Cropped | 18 | 72 | 0.43 (±0.09) | 0.42 | 0.04  | 0.63  |

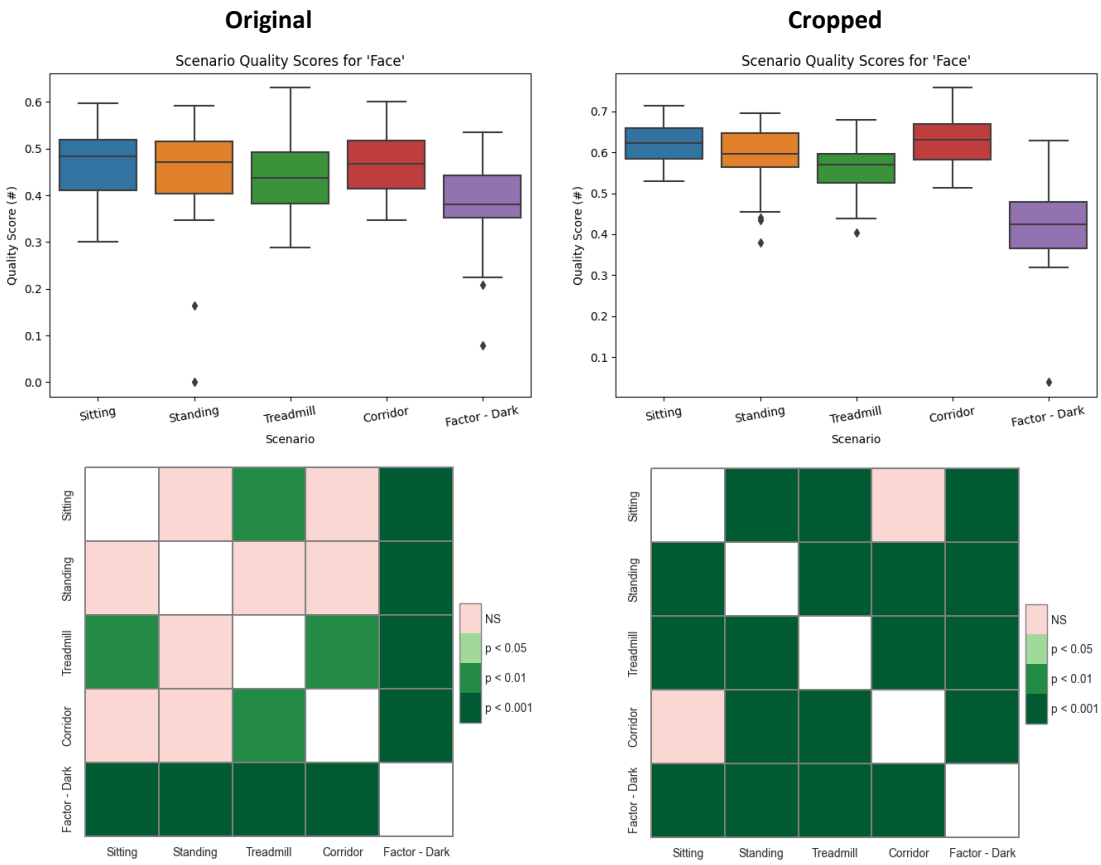



*Figure 6.22: Apple iPhone 8 Scenario Quality Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

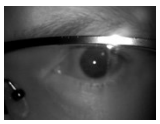
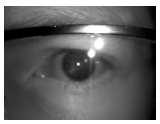
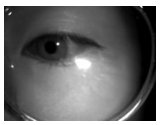
## 6.4.2 Iris

The IriTech IriShield provides an SDK that includes a bespoke algorithm for assessing the quality of an iris image. The quality scores are calculated based on several image quality metrics for iris recognition using IriCore (IriTech Iris SDK) [151]. IriTech's image quality assessment algorithm has been proven as the most accurate one in IREX II. However, IriTech cannot share details of the calculations and algorithms because they are Critical Confidential and Proprietary Information. The metrics used for the calculation include full support of IQCE (IREX II) quality metrics and self-defined metrics:

1. Scalar overall quality
2. Gray level spread
3. Iris radius
4. Pupil iris ratio
5. Usable iris area
6. Iris-sclera contrast
7. Iris-pupil contrast
8. Iris sclera boundary shape
9. Iris pupil boundary shape
10. Margin
11. Sharpness (defocus)
12. Motion blur
13. Signal-to-noise ratio
14. Magnification
15. Head rotation
16. Gaze angle
17. Interlace
18. Vendor-defined metrics

Similarly, to analyse the genuine verification scores for face, the quality scores were examined against the scenario groups to check for any statistical significance. Table 6.9 shows the statistical results for the iris quality scores across the experimented scenario range. Unsurprisingly, the lowest quality scores were found to be for users who were attempting to operate the iris scanner through glasses

*Table 6.9: Scenario Quality Score Statistics for 'Iris'*

| Device | Scenario | Participants | Amount | Average | Median | Max | Min |
|---|---|---|---|---|---|---|---|
| IriTech IriShield | Sitting | 60 | 338 | 71.2 (±29.4) | 83 | 100  | 20  |
| | Standing | 59 | 297 | 77.6 (±27.0) | 90 | 100  | 20.0  |
| | Treadmill | 59 | 291 | 81.3 (±25.7) | 94 | 100  | 21  |
| | Corridor | 57 | 287 | 80.9 (±23.8) | 90 | 100  | 20  |
| | Factor - Dark | 60 | 321 | 76.6 (±25.5) | 86 | 100  | 20  |

When comparing the quality scores against the verification scores, a slight positive correlation, as shown in Figure 6.23, was found. The trendline has the equation $y = 0.00x + (0.34)$. This means that higher iris quality scores had a slight decrease in the verification performance as the verification scores are present as dissimilarity scores. However, as noted from the equation, the gradient is almost flat (0), implying that potentially there is little to no effect between quality and verification observed. This impact could be caused by using infrared imaging to capture and compare the images. It is also worth noting that IriTech IriCore SDK was not used to produce the verification scores due to the closed commercial nature of the SDK.

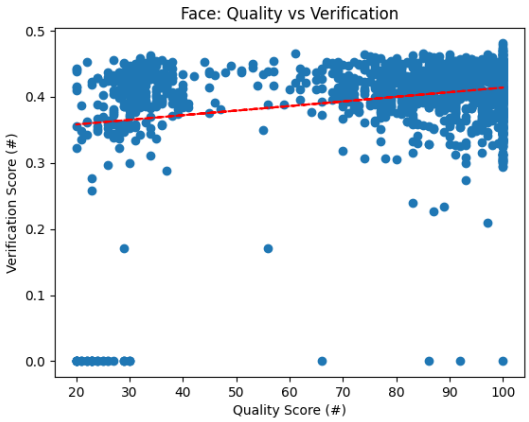



*Figure 6.23: Quality Scores vs Verification Scores for 'Iris'*

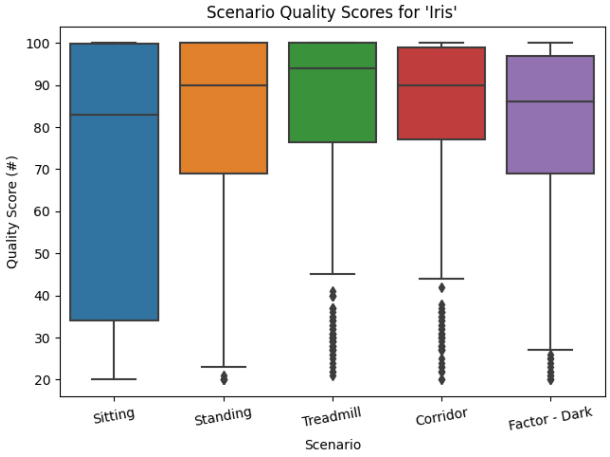



*Figure 6.24: Scenario Quality Scores Box Plot for 'Iris'*

Figure 6.24 shows the box plot results for the quality scores of the iris images between the scenarios. Using a one-way ANOVA statistical test to assess the statistical significance of the quality scores between scenarios results in $F(4, 1529) = [7.51]$ at $p = 5.00 \times 10^{-6}$ $(p < 0.001)$ indicating that the different scenarios have a statistically significant impact on the quality scores for the iris recognition system. Evidence of some statistical significance between scenarios was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

Figure 6.25 shows the results of this post hoc analysis in the form of a significance plot highlighting the significant relationships between groups. The observation is that the significant differences come from comparing the sitting scenario with the other scenarios (except for our challenging condition). The probable cause for this relates to the user's habituation with the device. Sitting was the first scenario that each user experienced for each trialled modality. Therefore, any errors or issues were encountered within this scenario. The results observed here likely reflect this difficulty with users familiarising themselves with the IriTech IriShield for the first time.


*Figure 6.25: Scenario Quality Scores P-Value Significance Plot for 'Iris'*

## 6.5 The Effect of Motion

The effect of motion on mobile biometric performance is a new area of interest, and we aim to begin to analyse the effect of this performance. Previously, the core factors identified 'Diversity of Scenarios' as factors that influence the systems performance and 'motion' and 'stationary' are vital differences within scenarios [45].

Using the collected data, we aim to analyse the effect of motion on biometric performance. The biometric match scores from our face, voice and iris recognition trials were analysed for this purpose. Two scenarios were categorised as stationary (Sitting and Standing), and two were categorised as motion (Treadmill and Corridor).

### 6.5.1 Face

Table 6.10 shows the statistical results of analysing the genuine verification and quality scores for both the original and cropped facial images between the stationary and motion scenarios. For each comparison, Welch's T-Test was performed to test for statistical significance between the stationary and motion scenarios. In all cases, the statistical test indicates that the stationary and motion scenarios are significant, although slightly less so when comparing the quality of the original images. Figure 6.26 and Figure 6.27 show the T-Test histogram density distributions for the verification and quality scores, respectively, for the original and cropped images.

*Table 6.10: Stationary vs Motion Scenario Statistics for 'Face'*

| Variable | Image | Scenario | Amount | Average | Median |
|---|---|---|---|---|---|
| Verification | Original | Stationary | 706 | 0.22 (±0.08) | 0.21 |
| | | Motion | 739 | 0.30 (±0.06) | 0.30 |
| | | **Welch's T-Test:** $t(1348.24) = -22.28, p = 7.18 \times 10^{-94}$ (P < 0.001) | | | |
| | Cropped | Stationary | 657 | 0.22 (±0.08) | 0.21 |
| | | Motion | 657 | 0.31 (±0.07) | 0.30 |
| | | **Welch's T-Test:** $t(1277.31) = -21.76, p = 1.51 \times 10^{-89}$ (P < 0.001) | | | |
| Quality | Original | Stationary | 801 | 0.45 (±0.07) | 0.45 |
| | | Motion | 764 | 0.44 (±0.07) | 0.44 |
| | | **Welch's T-Test:** $t(1562.14) = 1.73, p = 8.30 \times 10^{-2}$ (P < 0.1) | | | |
| | Cropped | Stationary | 791 | 0.61 (±0.05) | 0.61 |
| | | Motion | 762 | 0.59 (±0.06) | 0.59 |
| | | **Welch's T-Test:** $t(1497.35) = 5.17, p = 2.62 \times 10^{-7}$ (P < 0.001) | | | |

**Original** **Cropped**

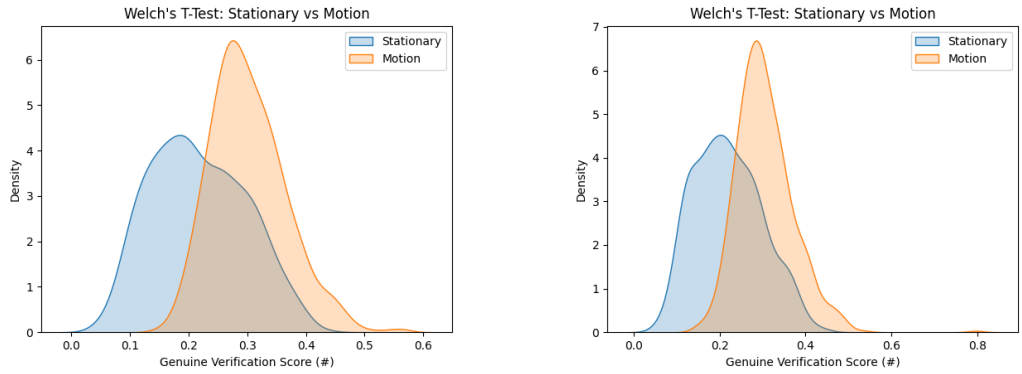



*Figure 6.26: Welch's T-Test Comparing the Genuine Verification Scores for Stationary and Motion Scenarios for Original and Cropped Images for 'Face'*

**Original** **Cropped**

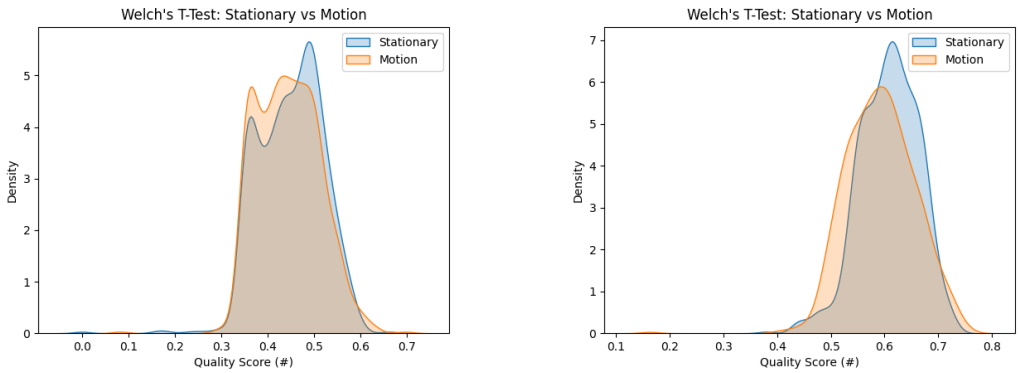



*Figure 6.27: Welch's T-Test Comparing the Quality Scores for Stationary and Motion Scenarios for Original and Cropped Images for 'Face'*

## 6.5.2 Iris

Table 6.11 shows the statistical results of analysing the genuine verification and quality scores for the iris images between the stationary and motion scenarios. For each comparison, Welch's T-Test was performed to test for statistical significance between the stationary and motion scenarios. In all cases, the statistical test indicates that there is significance between the stationary and motion scenarios for both verification and quality, although slightly less so for the verification scores, which, as mentioned previously, is likely down to the capability of the hardware using infrared images over colour RGB images. Figure 6.28 and Figure 6.29 show the T-Test histogram distribution for the verification and quality scores for the stationary and motion scenarios observed for the iris data.

*Table 6.11: Stationary vs Motion Scenario Statistics for 'Iris'*

| Variable | Scenario | Amount | Average | Median |
|---|---|---|---|---|
| Verification | Stationary | 575 | 0.39 (±0.08) | 0.41 |
| | Motion | 578 | 0.40 (±0.07) | 0.42 |
| | **Welch's T-Test:** $t(1086.89) = -2.86, p = 4.38 \times 10^{-3}$ (P < 0.01) | | | |
| Quality | Stationary | 635 | 77.4 (±28.4) | 87.0 |
| | Motion | 578 | 81.1 (±24.8) | 92.0 |
| | **Welch's T-Test:** $t(1208.73) = -4.51, p = 7.26 \times 10^{-6}$ (P < 0.001) | | | |

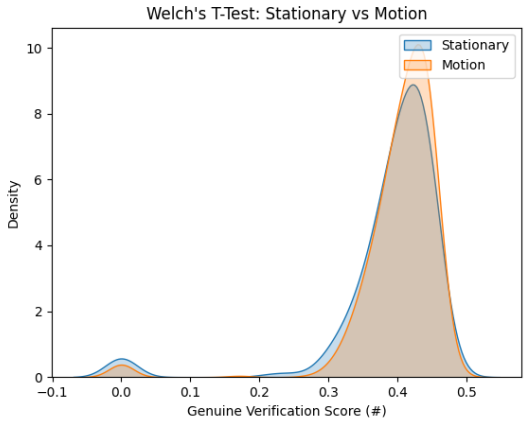



*Figure 6.28: Welch's T-Test Comparing the Genuine Verification Scores for Stationary and Motion Scenarios for 'Iris'*

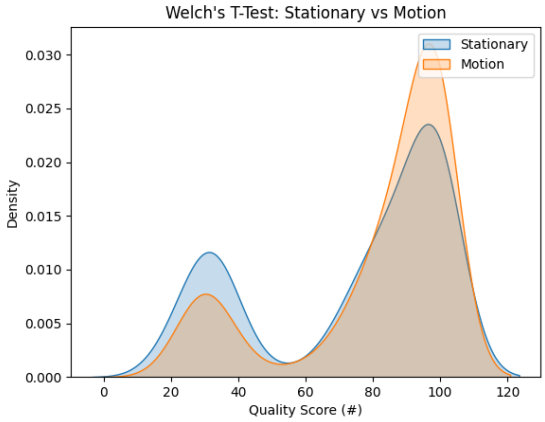



*Figure 6.29: Welch's T-Test Comparing the Quality Scores for Stationary and Motion Scenarios for 'Iris'*

## 6.5.3 Voice

Table 6.12 shows the statistical results of analysing the genuine verification scores for the voice audios between the stationary and motion scenarios. A Welch's T-Test was performed to test for statistical significance, and the statistical test indicates that the stationary and motion scenarios are significant for the verification scores. Figure 6.30 shows the T-Test histogram distribution for the verification scores for the stationary and motion scenarios observed for the voice audio data.

*Table 6.12: Stationary vs Motion Scenario Statistics for 'Voice'*

| Variable | Scenario | Amount | Average | Median |
|---|---|---|---|---|
| Verification | Stationary | 1092 | 0.76 (±0.16) | 0.80 |
| | Motion | 1210 | 0.70 (±0.16) | 0.73 |
| | **Welch's T-Test:** $t(2283.73) = 9.54, p = 3.56 \times 10^{-21}$ (P < 0.001) | | | |

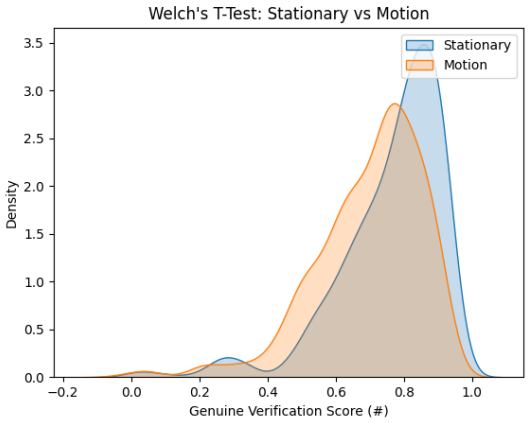



*Figure 6.30: Welch's T-Test Comparing the Genuine Verification Scores for Stationary and Motion Scenarios for 'Voice'*

One of the novel factors that intruded on mobile biometrics is mobility. The concept that the biometric system can be operated while on the move is an area that will require continuous academic research. However, here the results indicate that motion does have a statistically significant for both verification and quality scores for several modalities and supports the inclusion of ensuring that motion-based scenario testing should be included within the performance framework.

## 6.6 The Effect of the Environment

Following on from motion is another concept that will require investigation when considering a mobile biometric system, the environment. Users can authenticate at any time and place, meaning that a performance testing framework should consider the environment when evaluating.

The first data collection session focused on scenarios and challenging conditions in the form of influential factors. The second session focused on the environment by taking the session outdoors in an uncontrolled setting. This trial allows a unique look at how defined (weather, temperature) environmental factors can affect performance and further showcases how we can record operational results within the performance framework.

### 6.6.1 Indoor vs Outdoor

Looking into the performance effects of altering the environment from indoor to outdoor was possible with the inbuilt biometric systems of the tested smartphone devices.

Table 6.13 shows the FNMR achieved by the participants between session one (Indoor) and session two (Outdoor). As noted in Chapter 5, the leading causes of false non-matches were human error and the

user cancelling the authentication rather than a biometric match failure. In addition, the participant enrolled again before starting the session with the baseline sitting scenario.

There are some notable cases where the FNMR increased from the change in environment, most notably the Samsung Galaxy S9 iris, which had the highest FNMR across all the devices and modalities tested. Also, in the case of the two iPhones, the FNMR increased slightly within the outdoor environment than in the indoor environment. However, the remaining modalities saw a fall in the outdoor FNMR. Unfortunately, no firm conclusions can be drawn from this data, partly because of the limited number of participants involved. However, it is still worth highlighting the results achieved directly from the devices as this is typical of the data available from a device offering commercial access to the biometric system.

Going further, using the captured modality samples allows the analysis of the indoor and outdoor environments using third-party libraries. The enrolment reference was the same baseline sitting sample used for all comparisons throughout this chapter. When comparing the data available, the factor scenarios were removed to make the comparison fairer and only compare the indoor and outdoor conditions without potential distractions from other external influences.

*Table 6.13: False Non-Match Rate Results for the In-Built Biometric Systems of the Smartphone Devices Comparting Between Indoor and Outdoor Environments*

| Device | Modality | Session | Participants | Amount | Success | FNMR (%) |
|---|---|---|---|---|---|---|
| Samsung Galaxy S9 | Fingerprint | Indoors | 25 | 548 | 488 | 11 |
| | | Outdoors | 15 | 823 | 746 | 10 |
| | Face | Indoors | 25 | 486 | 448 | 8 |
| | | Outdoors | 15 | 760 | 735 | 4 |
| | Iris | Indoors | 25 | 489 | 368 | 25 |
| | | Outdoors | 15 | 683 | 411 | 40 |
| Google Pixel 2 | Fingerprint | Indoors | 35 | 772 | 694 | 11 |
| | | Outdoors | 14 | 749 | 690 | 8 |
| iPhone 8 | Fingerprint | Indoors | 33 | 653 | 619 | 6 |
| | | Outdoors | 14 | 664 | 615 | 8 |
| iPhone X | Face | Indoors | 27 | 543 | 536 | 2 |
| | | Outdoors | 13 | 658 | 623 | 6 |

#### 6.6.1.1 Face

Table 6.14 shows the statistical results of analysing the genuine verification and quality scores for both the original and cropped facial images between the indoor and outdoor environments. For each comparison, Welch's T-Test was performed to test for statistical significance between the indoor and outdoor environments. In all cases, the statistical test indicates that there is significance between the indoor and outdoor environments, although slightly less so when comparing the quality of the cropped images. Figure 6.31 and Figure 6.32 show the T-Test histogram density distributions for the verification and quality scores, respectively, for the original and cropped images.

*Table 6.14: Indoor vs Outdoor Environment Statistics for 'Face'*

| Variable | Image | Environment | Amount | Average | Median |
|---|---|---|---|---|---|
| Verification | Original | Indoor | 1445 | 0.26 (±0.08) | 0.27 |
| | | Outdoor | 1827 | 0.34 (±0.08) | 0.34 |
| | | **Welch's T-Test:** $t(3010.35) = -28.73, p = 1.33 \times 10^{-160}$ (P < 0.001) | | | |
| | Cropped | Indoor | 1314 | 0.26 (±0.08) | 0.27 |
| | | Outdoor | 1739 | 0.34 (±0.07) | 0.33 |
| | | **Welch's T-Test:** $t(2555.11) = -25.79, p = 1.53 \times 10^{-130}$ (P < 0.001) | | | |
| Quality | Original | Indoor | 1565 | 0.45 (±0.07) | 0.45 |
| | | Outdoor | 2017 | 0.46 (±0.07) | 0.46 |
| | | **Welch's T-Test:** $t(3380.91) = -6.64, p = 3.58 \times 10^{-11}$ (P < 0.001) | | | |
| | Cropped | Indoor | 1553 | 0.60 (±0.06) | 0.60 |
| | | Outdoor | 1994 | 0.60 (±0.06) | 0.61 |
| | | **Welch's T-Test:** $t(3417.76) = -0.70, p = 4.82 \times 10^{-1}$ (P < 0.5) | | | |

**Original** **Cropped**

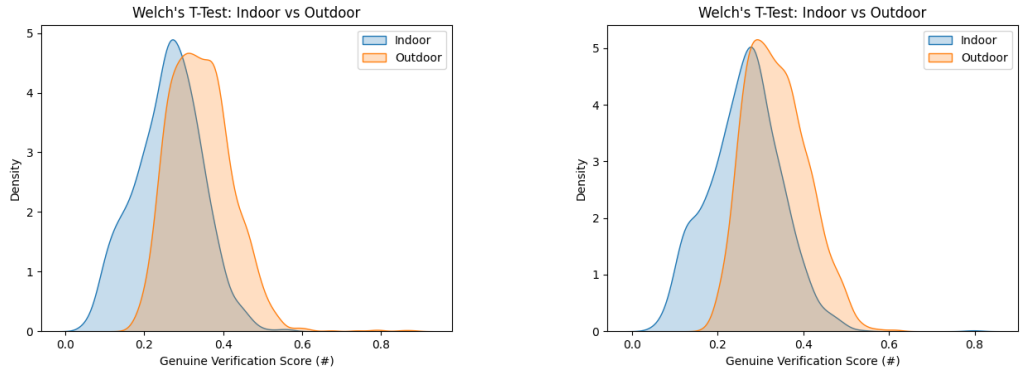



*Figure 6.31: Welch's T-Test Comparing the Genuine Verification Scores for Indoor and Outdoor Environments for Original and Cropped Images for 'Face'*

**Original** **Cropped**

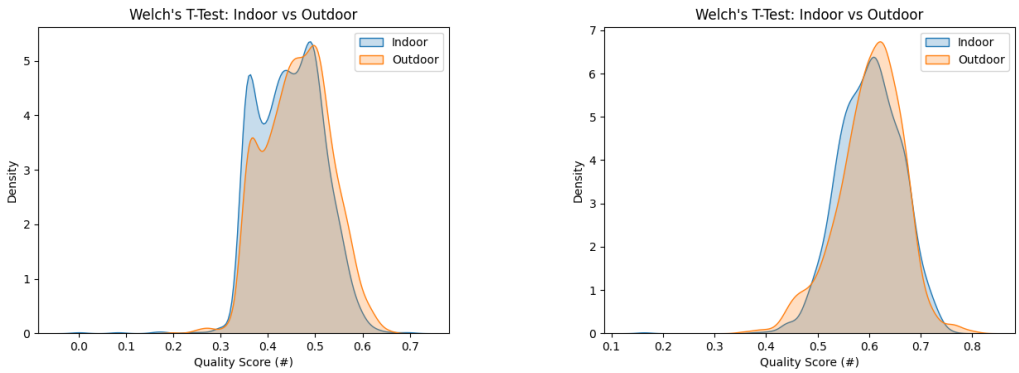



*Figure 6.32: Welch's T-Test Comparing the Quality Scores for Indoor and Outdoor Environments for Original and Cropped Images for 'Face'*

### 6.6.1.2 Iris

Table 6.15 shows the statistical results of analysing the genuine verification and quality scores for the iris images between the indoor and outdoor environments. For each comparison, Welch's T-Test was performed to test for statistical significance between the indoor and outdoor environments. In all cases, the statistical test indicates that there is significance between the indoor and outdoor environments for both verification and quality, although less so for the quality scores. Figure 6.33 and Figure 6.34 show the T-Test histogram distribution for the verification and quality scores for the indoor and outdoor environments, respectively, observed for the iris data.

One observation worth noting when analysing the iris data is that when comparing scenarios, specifically between stationary and motion, the quality scores were impacted more than the verification scores. In contrast, the opposite is true when exploring the environmental impact. The verification impacts outdoor conditions more than the quality scores, which seem to be less affected.

*Table 6.15: Indoor vs Outdoor Environment Statistics for 'Iris'*

| Variable | Environment | Amount | Average | Median |
|---|---|---|---|---|
| Verification | Indoor | 1153 | 0.40 (±0.07) | 0.41 |
| | Outdoor | 1370 | 0.42 (±0.04) | 0.44 |
| | **Welch's T-Test:** $t(1800.29) = -9.20, p = 9.58 \times 10^{-20}$ (P < 0.001) | | | |
| Quality | Indoor | 1213 | 77.5 (±27.0) | 90.0 |
| | Outdoor | 1390 | 76.6 (±23.7) | 86.0 |
| | **Welch's T-Test:** $t(2433.84) = 0.89, p = 3.74 \times 10^{-1}$ (P < 0.5) | | | |

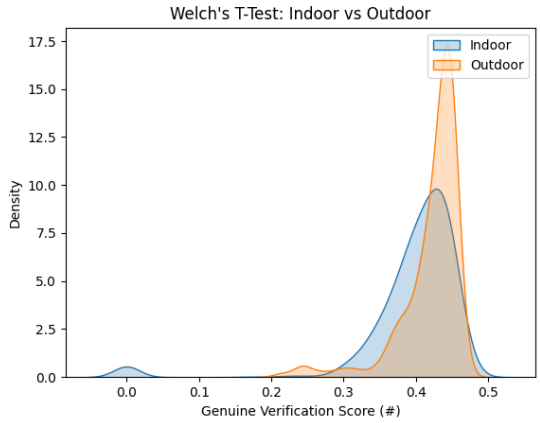



*Figure 6.33: Welch's T-Test Comparing the Genuine Verification Scores for Indoor and Outdoor Environments for 'Iris'*

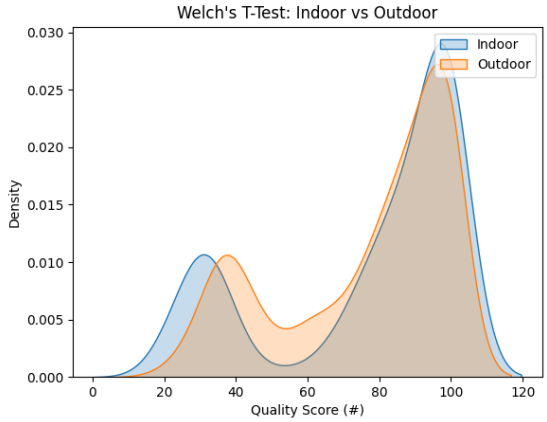



*Figure 6.34: Welch's T-Test Comparing the Quality Scores for Indoor and Outdoor Environments for 'Iris'*

### 6.6.1.3 Voice

Table 6.16 shows the statistical results of analysing the genuine verification scores for the voice audios between the indoor and outdoor environments. A Welch's T-Test was performed to test for statistical significance, and the statistical test indicates that there is significance between the indoor and outdoor scenarios for the verification scores. Figure 6.35 shows the T-Test histogram distribution for the verification scores for the indoor and outdoor environments observed for the voice audio data.

*Table 6.16: Indoor vs Outdoor Environment Statistics for 'Voice'*

| Variable | Environment | Amount | Average | Median |
|---|---|---|---|---|
| Verification | Indoor | 2302 | 0.73 (±0.16) | 0.76 |
| | Outdoor | 2875 | 0.59 (±0.17) | 0.61 |
| | **Welch's T-Test:** $t(5067.91) = 30.07, p = 5.55 \times 10^{-183}$ (P < 0.001) | | | |

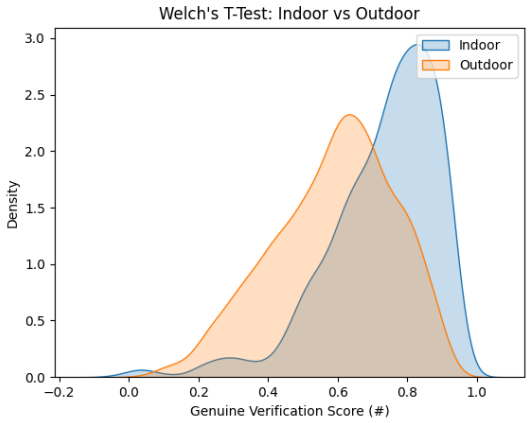



*Figure 6.35: Welch's T-Test Comparing the Genuine Verification Scores for Indoor and Outdoor Environments for 'Voice'*

The comparison between indoor and outdoor environments indicates a significant difference across all tests with verification and quality scores across various modalities. This result helps to strengthen the inclusion of different environments trialled as part of the mobile biometric performance framework with the recommendation that some outdoor activity should be included within the operational testing stage.

### 6.6.2 Weather

Weather analysis collected a snapshot of the weather when the participant started the outdoor trial of the second data collection session. This snapshot was collected from BBC Weather App [153], in association with MeteoGroup [154], which indicates the current weather over the next hour, covering the outdoor trial duration.

Table 6.17 shows the observed weather conditions for the total of the fifty-six participant trials who completed the second session. It is shown that most of the outdoor trials took place in dry, sunny conditions reflecting the time of the year that the trial took place, Summer. It is possible to see if and what the weather conditions have on the biometric performance using these groups, particularly in sample quality and match scores.

*Table 6.17: Summary of Weather Conditions Experienced During Outdoor Trial*

| Wet Conditions | Heavy Rain Showers | Light Rain Showers | Light Rain |
|---|---|---|---|
| Snapshot Sample |  |  |  |
| Participants | 1 | 9 | 4 |
| | | | |
| Dry Conditions | Light Cloud | Sunny Intervals | Sunny |
| |  |  |  |
| Participants | 1 | 25 | 16 |

### 6.6.2.1 Face

Table 6.18 shows the statistical summary of the weather analysis on the genuine verification scores for the face. Figure 6.31 presents the data as a box plot and the significance plot showing a statistical significance between weather conditions for the verification scores.

*Table 6.18: Weather Condition Genuine Verification Score Statistics for 'Face'*

| Weather | Cropped | Participants | Amount | Mean (μ) | Median | Min | Max |
|---|---|---|---|---|---|---|---|
| Heavy Rain Showers | Original | 1 | 48 | 0.34 (± 0.03) | 0.34 | 0.29 | 0.40 |
| | Cropped | 1 | 40 | 0.35 (± 0.03) | 0.35 | 0.29 | 0.41 |
| Light Rain Showers | Original | 5 | 260 | 0.32 (± 0.06) | 0.31 | 0.23 | 0.85 |
| | Cropped | 5 | 253 | 0.31 (± 0.04) | 0.31 | 0.21 | 0.45 |
| Light Rain | Original | 4 | 167 | 0.37 (± 0.06) | 0.37 | 0.26 | 0.59 |
| | Cropped | 4 | 147 | 0.38 (± 0.06) | 0.36 | 0.26 | 0.62 |
| Light Cloud | Original | 1 | 51 | 0.28 (± 0.04) | 0.28 | 0.23 | 0.40 |
| | Cropped | 1 | 43 | 0.29 (± 0.03) | 0.28 | 0.23 | 0.39 |
| Sunny Intervals | Original | 20 | 936 | 0.35 (± 0.08) | 0.35 | 0.17 | 0.88 |
| | Cropped | 20 | 907 | 0.34 (± 0.07) | 0.34 | 0.17 | 0.58 |
| Sunny | Original | 14 | 616 | 0.34 (± 0.09) | 0.33 | 0.19 | 0.85 |
| | Cropped | 14 | 582 | 0.34 (± 0.08) | 0.32 | 0.19 | 0.60 |

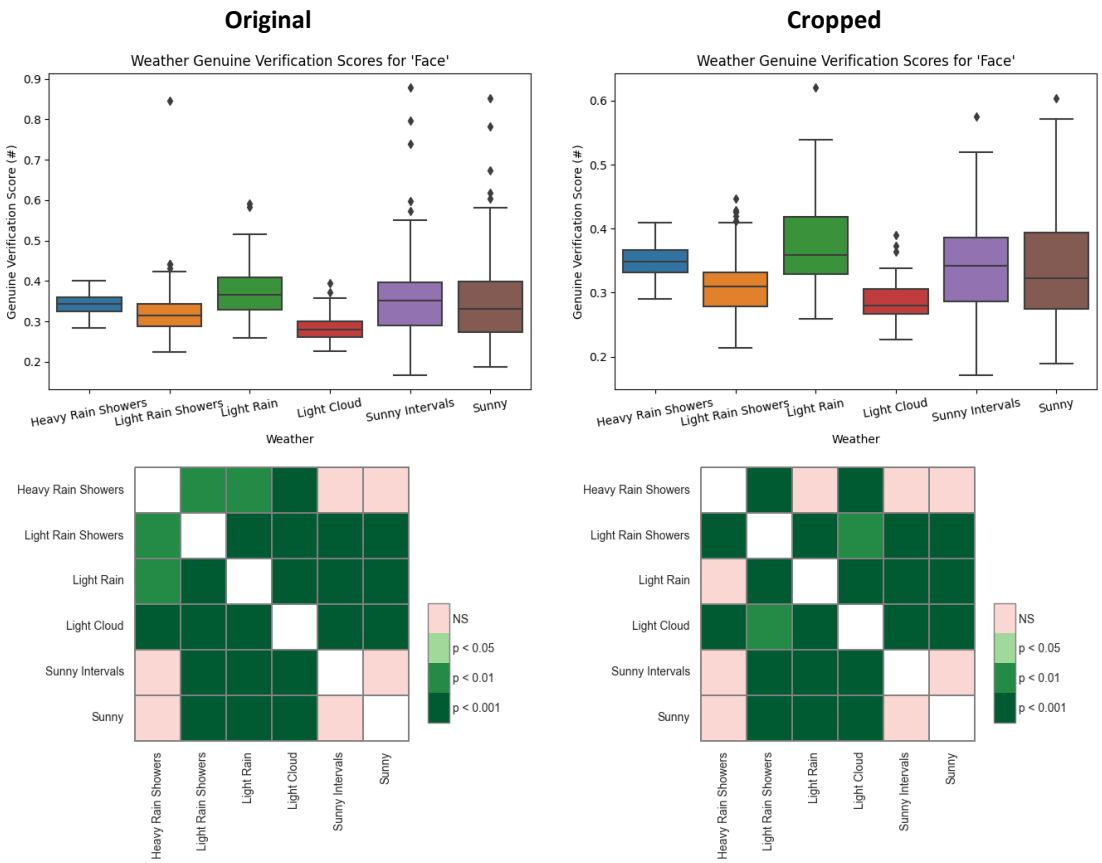



*Figure 6.36: Weather Condition Genuine Verification Scores Box Plot and P-Value Significance Plots for Original and Cropped Images for 'Face'*

Using a one-way ANOVA statistical test to assess the statistical significance of the genuine verification scores between weather conditions results in $F(5, 2072) = [17.45]$ at $p = 5.80 \times 10^{-17}$ $(p < 0.001)$ for the original images and $F(5, 1966) = [23.80]$ at $p = 2.63 \times 10^{-23}$ $(p < 0.001)$ meaning that weather conditions have a statistically significant impact on the genuine verification scores for a face recognition system.

*Table 6.19: Weather Condition Quality Score Statistics for 'Face'*

| Weather | Cropped | Participants | Amount | Mean (μ) | Median | Min | Max |
|---|---|---|---|---|---|---|---|
| Heavy Rain Showers | Original | 1 | 50 | 0.52 (± 0.06) | 0.54 | 0.38 | 0.59 |
| | Cropped | 1 | 45 | 0.64 (± 0.02) | 0.64 | 0.59 | 0.70 |
| Light Rain Showers | Original | 9 | 470 | 0.51 (± 0.07) | 0.49 | 0.37 | 0.71 |
| | Cropped | 9 | 350 | 0.60 (± 0.06) | 0.59 | 0.35 | 0.73 |
| Light Rain | Original | 4 | 187 | 0.41 (± 0.06) | 0.39 | 0.20 | 0.56 |
| | Cropped | 4 | 170 | 0.57 (± 0.07) | 0.55 | 0.37 | 0.71 |
| Light Cloud | Original | 1 | 51 | 0.51 (± 0.03) | 0.51 | 0.43 | 0.56 |
| | Cropped | 1 | 46 | 0.63 (± 0.03) | 0.63 | 0.57 | 0.68 |
| Sunny Intervals | Original | 25 | 1270 | 0.46 (± 0.08) | 0.46 | 0.07 | 0.67 |
| | Cropped | 25 | 1059 | 0.61 (± 0.06) | 0.61 | 0.35 | 0.80 |
| Sunny | Original | 16 | 755 | 0.46 (± 0.07) | 0.46 | 0.20 | 0.65 |
| | Cropped | 16 | 680 | 0.59 (± 0.08) | 0.59 | 0.36 | 0.80 |

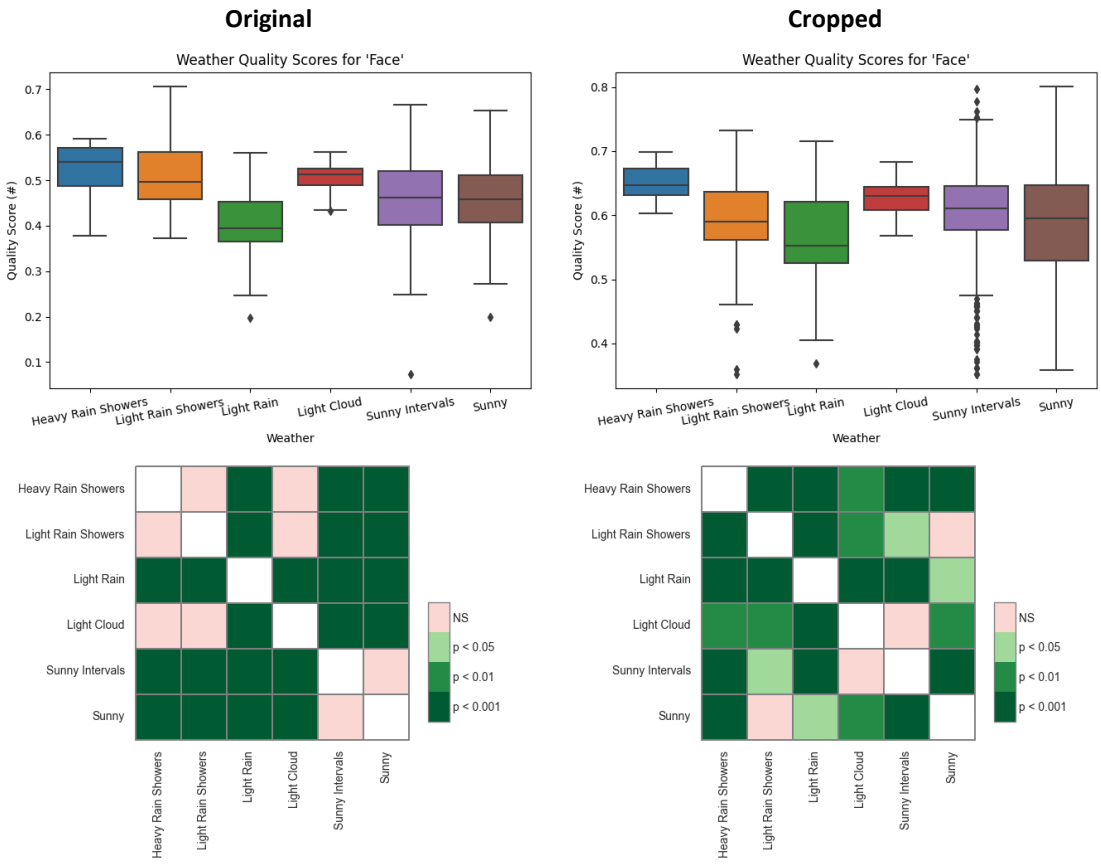



*Figure 6.37: Weather Condition Quality Scores Box Plots and P-Value Significance Plots for Original and Cropped Images for 'Face'*

Using a one-way ANOVA statistical test to assess the statistical significance of the original quality scores between weather conditions results in $F(5, 2777) = [70.81]$ at $p = 6.57 \times 10^{-70}$ $(p < 0.001)$ and the cropped images results in $F(5, 2333) = [20.72]$ at $p = 2.63 \times 10^{-20}$ $(p < 0.001)$ meaning

that weather conditions have a statistically significant impact on the uncropped quality scores for a face recognition system.

#### *6.6.2.2 Iris*

The same approach for analysing the impact of the weather was performed against the obtained iris data. Table 6.20 shows the overall statistics for the genuine verification score data for the inferred iris image captured using the IriTech IriShield. In contrast, outdoors during the second data collection session and Figure 6.38 shows this in a box plot.

Using a one-way ANOVA statistical test to assess the statistical significance of the verification scores between weather conditions results in $F(4, 1365) = [31.26]$ at $p = 6.14 \times 10^{-25}$ $(p < 0.001)$ indicating that the weather conditions statistically impact the genuine verification scores for the iris recognition system. The evidence of some statistical significance between weather conditions was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

Figure 6.39 shows the results of this post hoc analysis in the form of a significance plot highlighting the significant relationships between groups. The significance is present between the 'wet' and 'dry' conditions as demonstrated by the "Light Rain Showers" and "Sunny Intervals" having significance compared to the other trialled weather conditions and more significantly with each other.

***Table 6.20: Weather Condition Genuine Verification Score Statistics for 'Iris'***

| Weather | Participants | Amount | Mean (μ) | Median | Min | Max |
|---|---|---|---|---|---|---|
| Light Rain Showers | 3 | 111 | 0.42 (± 0.02) | 0.43 | 0.36 | 0.46 |
| Light Rain | 2 | 76 | 0.42 (± 0.03) | 0.42 | 0.34 | 0.46 |
| Light Cloud | 1 | 53 | 0.42 (± 0.01) | 0.41 | 0.40 | 0.46 |
| Sunny Intervals | 13 | 635 | 0.41 (± 0.06) | 0.43 | 0.20 | 0.48 |
| Sunny | 10 | 495 | 0.44 (± 0.02) | 0.44 | 0.36 | 0.47 |

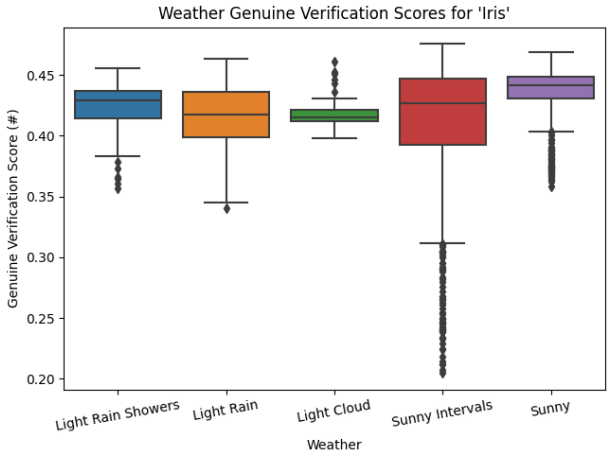



*Figure 6.38: Weather Condition Genuine Verification Scores Box Plot for 'Iris'*

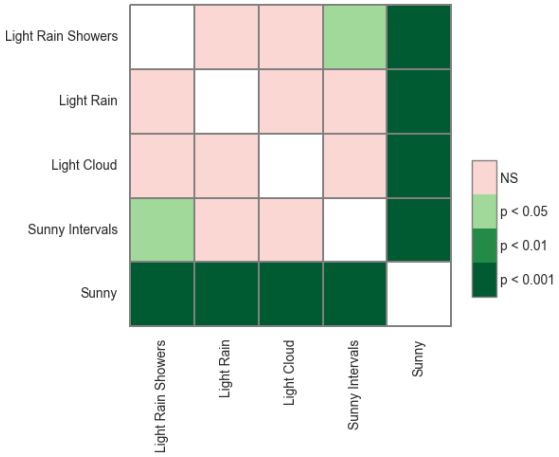



*Figure 6.39: Weather Condition Genuine Verification Scores P-Value Significance Plot for 'Iris'*

Additionally, analysing the iris quality score against the weather conditions produces the statistical data shown in Table 6.21 and the box plot Figure 6.40. Using a one-way ANOVA statistical test to assess the statistical significance of the quality scores between weather conditions results in $F(4, 1385) = [10.42]$ at $p = 2.55 \times 10^{-8}$ $(p < 0.001)$ indicating that the different weather conditions have a statistically significant impact on the quality scores for the iris recognition system. The evidence of some statistical significance between weather conditions was followed by a statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

Figure 6.41 shows the results of this post hoc analysis in the form of a significance plot highlighting the significant relationships between groups. The observation is that significance is present between the 'wet' and 'dry' conditions as demonstrated by the "Light Rain Showers" and "Sunny Intervals" having significance compared to the other trialled weather conditions and more significantly with each other.

*Table 6.21: Weather Condition Quality Score Statistics for 'Iris'*

| Weather | Participants | Amount | Mean (μ) | Median | Min | Max |
|---|---|---|---|---|---|---|
| Light Rain Showers | 3 | 111 | 86 (± 22) | 97 | 23 | 100 |
| Light Rain | 2 | 96 | 79 (± 23) | 88 | 36 | 100 |
| Light Cloud | 1 | 53 | 89 (± 14) | 92 | 39 | 100 |
| Sunny Intervals | 13 | 635 | 75 (± 25) | 85 | 20 | 100 |
| Sunny | 10 | 495 | 74 (± 23) | 80 | 23 | 100 |

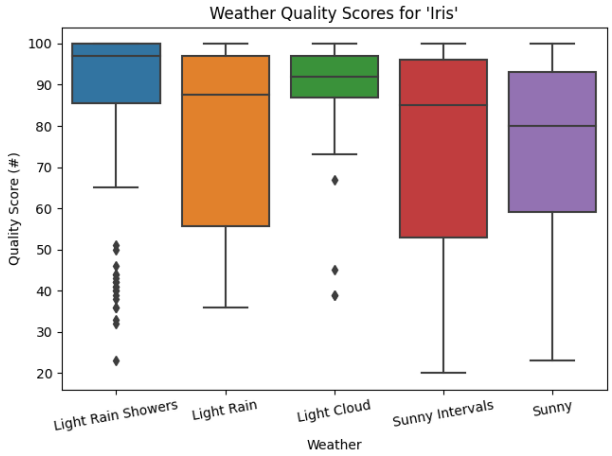



*Figure 6.40: Weather Condition Quality Scores Box Plot for 'Iris'*

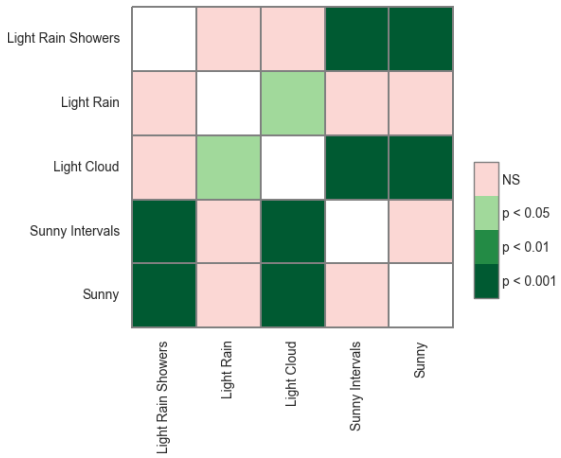



*Figure 6.41: Weather Condition Quality Scores P-Value Significance Plot for 'Iris'*

### 6.6.2.3 Voice

Using a one-way ANOVA statistical test to assess the statistical significance of the genuine verification scores between weather conditions results in $F(5, 2869) = [16.54]$ at $p = 3.95 \times 10^{-16}$ $(p < 0.001)$ indicating that the weather conditions significantly impact the quality scores for the iris recognition system. The evidence of some statistical significance between weather conditions was followed by a

statistical post hoc pairwise t-test for multiple comparisons of independent groups with a step-down method using Bonferroni adjustments.

*Table 6.22: Weather Condition Genuine Verification Score Statistics for 'Voice'*

| Weather | Participants | Amount | Mean (μ) | Median | Min | Max |
|---|---|---|---|---|---|---|
| Heavy Rain Showers | 1 | 50 | 0.58 (± 0.09) | 0.58 | 0.16 | 0.74 |
| Light Rain Showers | 9 | 468 | 0.53 (± 0.19) | 0.51 | 0.06 | 0.92 |
| Light Rain | 4 | 204 | 0.54 (± 0.18) | 0.60 | 0.10 | 0.80 |
| Light Cloud | 1 | 50 | 0.60 (± 0.06) | 0.60 | 0.48 | 0.73 |
| Sunny Intervals | 25 | 1284 | 0.60 (± 0.16) | 0.61 | 0.02 | 0.93 |
| Sunny | 16 | 819 | 0.61 (± 0.18) | 0.64 | 0.09 | 0.91 |

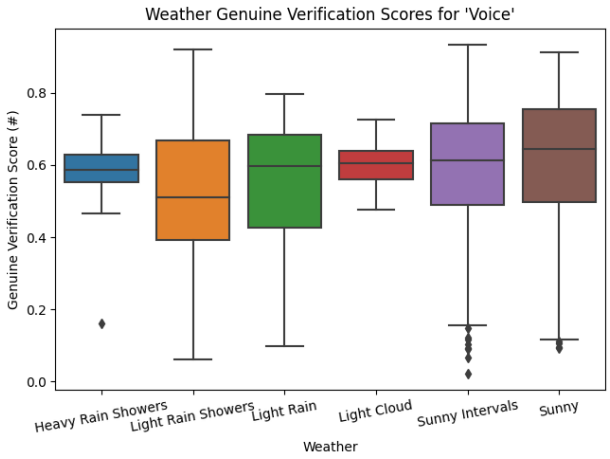



*Figure 6.42: Weather Condition Genuine Verification Scores Box Plot for 'Voice'*

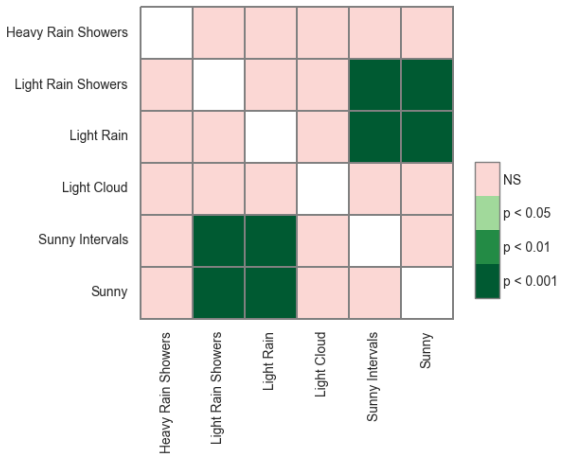



*Figure 6.43: Weather Condition Genuine Verification Scores P-Value Significance Plot for 'Voice'*

## 6.7 Usability

This section will explore usability by examining three key metrics to form an understanding of how users are interacting with the mobile biometric system:

- **Transactional Time** – The total time taken from the biometric prompt being presented to the user asking for the presentation of the biometric sample until a decision is made.
- **Accuracy (1-FNMR)** – The system's accuracy is equivalent to removing the false non-match rate from all the errors that occur, indicating how often the system performs as expected and successfully verifies the user.
- **Satisfaction** – The user's satisfaction is an opinion generated from the survey completed by our participants using a Likert scale ranging from 0 (completely unsatisfied) to 7 (completely satisfied).

From the devices trialled as part of the research, six combinations of device and biometric modality were present. They had a "developer" level of access, allowing the collection of the usability metrics defined above. From the indoor session and scenario, Table 6.23 shows the usability metrics obtained from the study. In addition, the usability metrics presented in Table 6.24 were captured from the outdoor session.

Seeing the transactional time can also reveal some possible habituation effects, as with all the tested modalities, the first scenario presented to the participants (sitting) is also the one which boosts the highest average transactional time, with the presented influencing factors coming in second.

*Table 6.23: Usability Metrics for Smartphone Device Modalities for Session One (Indoor) Scenarios*

| Device | Modality | Scenario | Mean Transactional Time (s) | Accuracy (1-FNMR) (%) | Satisfaction (#) |
|---|---|---|---|---|---|
| Samsung Galaxy S9 | Fingerprint | Sitting | 2.16 | 74 | 4.96 |
| | | Standing | 1.42 | 94 | 4.96 |
| | | Treadmill | 1.32 | 96 | 4.96 |
| | | Corridor | 1.10 | 93 | 4.96 |
| | | Factor - Wet | 3.30 | 23 | 4.96 |
| | Face | Sitting | 3.16 | 92 | 5.08 |
| | | Standing | 1.39 | 90 | 5.08 |
| | | Treadmill | 1.38 | 91 | 5.08 |
| | | Corridor | 1.37 | 93 | 5.08 |
| | | Factor - Dark | 0.75 | 99 | 5.08 |
| | Iris | Sitting | 3.56 | 72 | 3.56 |
| | | Standing | 2.65 | 83 | 3.56 |
| | | Treadmill | 3.42 | 73 | 3.56 |
| | | Corridor | 3.59 | 71 | 3.56 |
| | | Factor - Dark | 3.17 | 75 | 3.56 |
| Google Pixel 2 | Fingerprint | Sitting | 1.55 | 86 | 5.20 |
| | | Standing | 1.02 | 90 | 5.20 |
| | | Treadmill | 1.13 | 87 | 5.20 |
| | | Corridor | 1.08 | 95 | 5.20 |
| | | Factor - Wet | 5.45 | 20 | 5.20 |
| iPhone 8 | Fingerprint | Sitting | 1.63 | 96 | 5.06 |
| | | Standing | 1.16 | 96 | 5.06 |
| | | Treadmill | 1.32 | 93 | 5.06 |
| | | Corridor | 1.19 | 91 | 5.06 |
| | | Factor - Wet | 7.06 | 0 | 5.06 |
| iPhone X | Face | Sitting | 1.13 | 99 | 5.11 |
| | | Standing | 0.81 | 99 | 5.11 |
| | | Treadmill | 0.86 | 98 | 5.11 |
| | | Corridor | 0.86 | 97 | 5.11 |
| | | Factor - Dark | 1.14 | 97 | 5.11 |

*Table 6.24: Usability Metrics for Smartphone Device Modalities for Session Two (Outdoor) Scenarios*

| Device | Modality | Mean Transactional Time (s) | Accuracy (1-FRR) (%) | Satisfaction (#) |
|---|---|---|---|---|
| Samsung Galaxy S9 | Fingerprint | 2.16 | 74 | 4.96 |
| | | 1.42 | 94 | 4.96 |
| | | 1.32 | 96 | 4.96 |
| | | 1.10 | 93 | 4.96 |
| | | 3.30 | 23 | 4.96 |
| | Face | 3.16 | 92 | 5.08 |
| | | 1.39 | 90 | 5.08 |
| | | 1.38 | 91 | 5.08 |
| | | 1.37 | 93 | 5.08 |
| | | 0.75 | 99 | 5.08 |
| | Iris | 3.56 | 72 | 3.56 |
| | | 2.65 | 83 | 3.56 |
| | | 3.42 | 73 | 3.56 |
| | | 3.59 | 71 | 3.56 |
| | | 3.17 | 75 | 3.56 |
| Google Pixel 2 | Fingerprint | 1.55 | 86 | 5.20 |
| | | 1.02 | 90 | 5.20 |
| | | 1.13 | 87 | 5.20 |
| | | 1.08 | 95 | 5.20 |
| | | 5.45 | 20 | 5.20 |
| iPhone 8 | Fingerprint | 1.63 | 96 | 5.06 |
| | | 1.16 | 96 | 5.06 |
| | | 1.32 | 93 | 5.06 |
| | | 1.19 | 91 | 5.06 |
| | | 7.06 | 0 | 5.06 |
| iPhone X | Face | 1.13 | 99 | 5.11 |
| | | 0.81 | 99 | 5.11 |
| | | 0.86 | 98 | 5.11 |
| | | 0.86 | 97 | 5.11 |
| | | 1.14 | 97 | 5.11 |

## 6.8 Tailored Impostors Investigation

To examine the impact of this tailoring approach. The sample tailoring algorithm will try to locate tailors within each group at random. When all tailors within a group are exhausted, they will begin to

select randomly from the group above. As the dataset is small, tailoring is locating tailors from less concentrated groups in the selection process.

The impact of impostors was analysed in the baseline (Sitting) scenario to minimise any interference from external factors. Using voice recording obtained from the Samsung Galaxy S9, we could analyse the impact of the different tailoring processes' Gender' and 'GenderAge' to analyse the performance variations, precisely the false match rate. Figure 6.44 and Table 6.25 shows the results of this analysis.

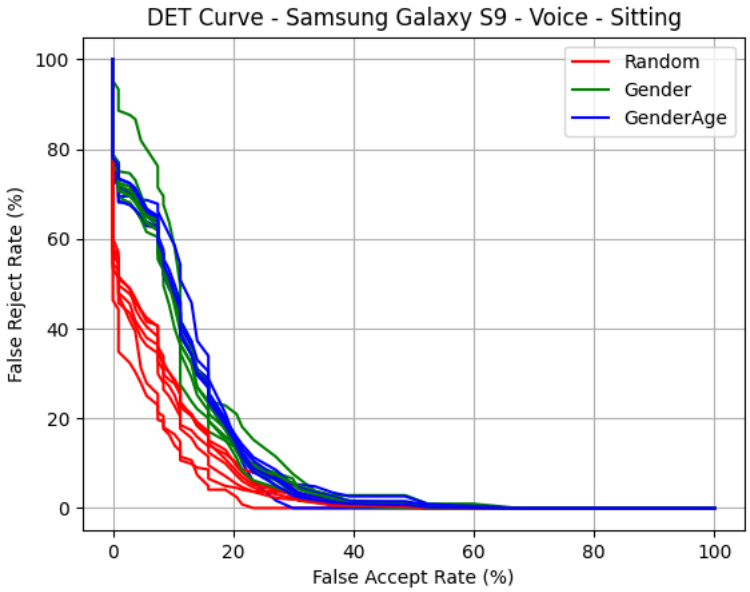



*Figure 6.44: DET Curve Showing Performance Alterations for Varying' Tailoring' Methods*

*Table 6.25: Statistical Tests on Varying Impostor 'Tailoring' Processes*

| | Mean Match Score | | | Independent T-Test | | |
|---|---|---|---|---|---|---|
| **Impostor Amount** | **Random** | **Gender** | **Gender + Age Group** | **(Random, Gender)** | **(Random, Gender + Age Group)** | **(Gender, Gender + Age Group)** |
| 1 | 0.57 (± 0.13) | 0.71 (± 0.09) | 0.68 (± 0.13) | P<0.001 | P<0.001 | P<0.025 |
| 2 | 0.56 (± 0.13) | 0.66 (± 0.11) | 0.66 (± 0.13) | P<0.001 | P<0.001 | P<0.991 |
| 3 | 0.60 (± 0.14) | 0.66 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.494 |
| 4 | 0.59 (± 0.15) | 0.67 (± 0.11) | 0.67 (± 0.13) | P<0.001 | P<0.001 | P<0.519 |
| 5 | 0.59 (± 0.13) | 0.67 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.496 |
| 6 | 0.60 (± 0.14) | 0.66 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.253 |
| 7 | 0.59 (± 0.13) | 0.67 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.815 |
| 8 | 0.59 (± 0.13) | 0.67 (± 0.11) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.683 |
| 9 | 0.59 (± 0.13) | 0.67 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.662 |
| 10 | 0.59 (± 0.13) | 0.67 (± 0.12) | 0.67 (± 0.12) | P<0.001 | P<0.001 | P<0.946 |

The results indicate a statistically significant difference between the average match scores obtained using randomly selected impostors based on gender and between randomly selected impostors and those selected based on gender and age group. However, it is not clear that there is a statistically significant

difference between impostors selected based on gender and those selected based on gender and age. Therefore, it can be stated that the impostor's gender has more of an impact on performance than the age group, but this is due to the sample size and age groups being unbalanced.

The results provided here are beginning to present a picture as to how a tailored impostors approach could work at exploiting the known weakness of the system to bring out an above-average false match rate and thus provide a deeper understanding of the potential securing risk of a worst-case passive impostor attack on the biometric system.

## 6.9 Summary

This chapter has introduced the results of the experimentation work when applied to the core factors and performance framework. In doing so, the argument for including the core factors has been strengthened, and the concepts of the performance framework have been demonstrated.

An illustrative example of the impact of the 'Scenarios' was shown for each tested modality using third-party open-source biometric algorithms against the captured data, including a comparison of stationary and motion scenarios. This chapter has also shown the 'Environment' core factor results by exploring the effect of various weather conditions and temperatures. This observation was achieved by exploring the obtained images' verification and quality scores (Face and Iris).

In providing these results, we have also begun to demonstrate part of the stages of the theoretical framework, including "Stage Two – Target Scenario Evaluation" and "Stage Six – Operational Evaluation", and demonstrated the impact of motion and the environment on the performance by showcasing statistically significant results meaning the impact and performance in these scenarios and conditions should be considered as part of the performance framework.

The next chapter will explore an approach designed to mitigate performance degradation in mobile devices from previously identified factors, focusing on movement patterns using the same collected data. In doing the tailored impostor, the approach was trialled further.

# 7 The Adaptive Threshold Decision

## 7.1 Introduction

Biometric facial recognition is a valuable security tool, allowing authentication with little interaction from the users' perspective since images can be captured from a distance and while in motion, requiring a camera. As a result, technology has advanced in recent years with its incorporation into mobile devices. This chapter develops a proof-of-concept adaptive decision model for mobile devices, which can outperform a static threshold applied to all environments and usage conditions. The motivation for this work was exploring the research question of how any performance deterioration can be mitigated. The performance assessment framework help informs and discover the approach developed in this chapter and serves as a potential solution to help mitigate performance deterioration in mobile biometric systems. The work present in this chapter is an adapted version of previously published work [155].

Section 7.2 introduces related work regarding adaptive methods and the inspiration for the proposed method. Section 7.3 —7.4 introduces the data collection and discuss how movement scenario impact recognition performance. Section 7.5 introduces the theory behind an adaptive framework to better deal with changing movement patterns. Section 7.6 discuss the approach to a detection algorithm for scenarios. Sections 7.7 —7.8 shows the experimental work and results in testing the adaptive threshold algorithm. Finally, Section 7.9 —7.10 provides a concluding discussion and summary and suggests future work.

## 7.2 Adaptive Approaches in Biometrics

Facial recognition has its share of criticism as campaigners claim the current technology is inaccurate, intrusive and infringes on an individual's privacy rights [156]. As a result, several locales have recently implemented or are considering a ban on fixed-system facial recognition technology, including San Francisco [50] and the European Union [156]. Furthermore, on October 05 2021, the European Union recently went as FMR to ban face recognition for mass police surveillance, citing bias and discrimination concerns and an individual's right to privacy [157]. This move is significant because it "sends a strong signal for negotiations of the first-ever EU rules on AI systems" and creates the necessary starting point "to preserve our freedoms and create a human-centric legal framework for AI". Furthermore, the framework helps certify that it fits its purpose of supporting broad technology adoption. One way to achieve this would be to ensure high recognition accuracy across scenarios consistently.

With a camera installed on most smartphone devices, it is increasingly convenient to take a self-portrait image ('selfie') for facial recognition. In addition, service providers are increasingly asking for

users to submit an ID document photo alongside a selfie captured on a mobile device to authenticate their claimed identity as part of Electronic Identity Verification (eIDV) services [158]. Furthermore, smartphones now increasingly incorporate facial technology allowing users to verify themselves and access services and resources within the device and beyond.

Static biometric systems, fixed in position, such as airport eGates, have been used. In these scenarios, the operators have great control over the environment to help optimise recognition performance. However, the same is not valid with mainstream mobile biometrics, where the operator has no control over the operational environment. It, therefore, stands to reason that those mobile biometrics would require a more adaptive approach to handling the authentication system.

The concept of adaptive biometrics systems is not new, as Pisani *et al.* [159] have comprehensively reviewed adaptive biometrics systems. However, most approaches work by updating the biometric reference over time, usually to account for template ageing. Pisani *et al.* note how "there is still a limited number of studies that evaluate adaptive biometric systems on mobile devices" and how researchers "should also acquire data from the sensors on these devices over time". Here the idea is to take a condition-sensitive (and quality index) adaptation criterion approach based on Pisani *et al.* taxonomy.

The method intends not to alter the sample or the probe but to utilise the mobile device's sensor information to determine the operation scenario and set thresholds accordingly. Techniques utilising the sensors embedded into smartphones and combining them with the biometric authentication process are present in the literature, for example, including the creation of behavioural biometric data to assess unique traits to identify individuals either independently or as part of a multimodal system with another physical or behavioural biometric trait to produce accurate biometric systems, commonly for continuous authentication purposes [160]–[162]. Another involvement of smartphone sensor data is liveness detection [163] and defending against presentation attacks. For example, Chen *et al.* [164] demonstrated a presentation attack detection approach using motion sensors to defend against 2D media attacks and virtual camera attacks.

The need for a more adaptive recognition framework is present in the literature as aspects like movement and portability of the device can vary between enrolment and recognition phases [165]. Chapter 3 highlighted the potential factors affecting a mobile biometric system and highlighted 'Scenarios' as one of these factors by categorising them under 'Stationary' and 'Motion'. Gutta *et al.* [166] have filed patents that suggest work and ideas relating to an adaptive biometric threshold, including a light intensity sensor to assist in adjusting the threshold value in a facial recognition system. Similarly, Brumback *et al.* [167] (Fitbit Inc) have also filed patents for continuous authentication on wearable technologies such as smartwatches and fitness trackers. However, they provide no practical examples of the proposals for mobile systems. Castilllo-Guerra *et al.* [168] proposed an adaptive threshold estimation for voice verification systems allowing the threshold to adapt to specific speakers. Similarly, Mhenni *et al.* [169]

proposed using an adaptive strategy specific to each category of users while investigating Doddington's Zoo classification of user keystroke dynamics.

Lunerti *et al.* [170] showed that for face verification in a mobile environment, "it can be possible to ensure good sample quality and high biometric performance by applying an appropriate threshold that will regulate the amplitude on variations of the smartphone movements during facial image capture". This chapter aims to contribute to showing how an adaptive approach can be the answer to having an "appropriate threshold" and begin to explore the gap in mobile biometric adaptive systems by exploring the potential impact of motion scenarios on recognition performance. The aim is to answer the following question: can mobile biometric recognition performance and security be improved by using an adaptive approach to the decision component using knowledge of the operating scenario?

## 7.3 Adaptive Threshold Data Collection

The same data collected from Chapter 5 was used to trial this approach, and this section will briefly identify the data relevant for the remainder of this chapter. This chapter will focus on the results achieved using the Android-based Samsung Galaxy S9 smartphone device. A custom application was developed to collect and capture data from this device to mimic biometric authentication. Using the Samsung Galaxy S9, the collected data included a 'selfie' image taken by the participant in the scenarios and background metadata obtained from the multitude of sensors (including accelerometer, gyro sensor and geomagnetic sensor) within the device. The device features an 8-megapixel (1.22 $\mu$, f/1.7) front-facing camera. However, the default picture size captures images at 5.2 megapixels for the study. Therefore, the resolution of the images was 2640x1980.

Twenty-five participants completed this part of the study during one session visit. The participant was tasked with operating the device in a variety of scenarios, the order of which was:

- Sitting - Participant sat down in a chair.
- Standing - Participant standing.
- Treadmill - Participant walking at a steady speed on a treadmill (speed set by the participant).
- Corridor - Participant walking at a steady speed down a corridor.

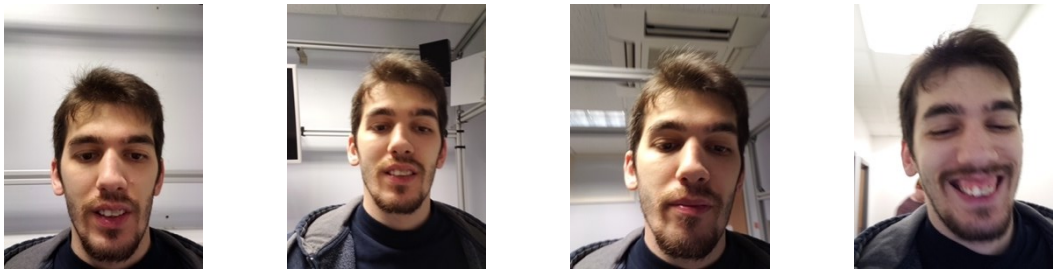
The aim was to mimic likely scenarios for smartphone use, except the treadmill, where the aim was to create a controlled walking scenario. The intention was to ensure the tasks were not too strenuous due to the repetitive nature of the repeat biometric transactions. The theory tests the approach on indoor scenarios in typical biometric authentication environments (room lighting), allowing the work to focus specifically on motion and movement. However, ideally, the approach could be adapted to other scenarios and factors in the future.

In each scenario, the participant held the device with their own hands as they usually would when operating a smartphone device. The participants were pre-enrolled at the start of the session using the device's biometric system while seated. The participants took a 'selfie' image for each scenario. There were no recommendations on positioning the face within the image; the only requirement was that the face was within the image. An additional part of the experiment was to see the impact on the device's facial recognition system.

Once the participant had captured the image, they remained in the same position, including the handling of the device. They were then presented with the device's in-built Android BiometricPrompt [171] to perform an authentication. While this was happening, the device would simultaneously collect the metadata (sensors, including Gyroscope, Linear Acceleration, Magnetic Field, and Orientation) from the moment the device's face authentication started until the process had finished utilising the abilities of Android SensorManager [172].

Because the face recognition authentication can be over within a second, the sensor delay was set to 0.005s to collect as much sensor data as possible. However, the documentation does note that "this is only a hint to the system. Events may be received faster or slower than the specified rate".

Figure 7.1 shows examples of one captured 'selfie' image from each tested scenario, taken by a single participant in the study. Table 7.1 displays the number of images collected from each scenario and how many of those the facial recognition algorithm could detect a face. The work in this paper uses images where the algorithm detected a face. Table 7.2 shows the breakdown of the participants' ages. 76% of the participants who used the Samsung Galaxy S9 were under 30, as a student population was recruited and used for this study. In addition, the participants had a gender split of 52% Female to 48% Male.



(a) Sitting (b) Standing (c) Treadmill (d) Corridor

***Figure 7.1: One example image from each scenario obtained from one participant during the first session***

*Table 7.1: Number of images collected from each scenario*

| Scenario | Images | Face Detected | No Face Detected |
|---|---|---|---|
| Sitting | 139 | 139 | 0 |
| Standing | 124 | 123 | 1 |
| Treadmill | 121 | 116 | 5 |
| Corridor | 122 | 120 | 2 |

*Table 7.2: Participant Age Ranges*

| Age Ranges | Number of Subjects |
|---|---|
| 19-21 | 3 |
| 22-24 | 8 |
| 25-29 | 8 |
| 30-39 | 4 |
| 40-49 | 2 |
| Total | 25 |

## 7.4 Scenario Performance

A prototype was created to test whether the adaptive approach has the potential to outperform a traditional system. Unfortunately, commercial smartphone devices have biometric components tightly locked down for security and privacy concerns. Therefore, an open-source software algorithm was used to help create a prototype of how a potential adaptive system could perform and function. The open-source 'face-recognition' python library (version 1.3.0) by Geitgey [133], [173] was used as the face recognition algorithm for the prototype. This library utilises the machine learning library 'Dlib'.

The first sitting attempt was used as the enrolment reference for each user to act as the base-case scenario. Then, the remaining images from all the scenarios were used as verification probes. 114 verification probes for the sitting scenario, 123 for the standing scenario, 116 for the treadmill scenario, and 120 for the corridor scenario were used. Next, the 'face-recognition' library calculated and returned the dissimilarity distance scores (between 0 and 1) of a given enrolled sample and a new verification probe. Here, a high score indicates that two images are unlikely to be of the same person (no match), and a low score indicates that the two images are likely to be of the same person (match). The library recommends a decision threshold of 0.6, meaning all comparisons that score 0.6 or below are considered the same person, and anything above is different.

Chapter 3 showed how scenarios could impact the false non-match rate of the Samsung Galaxy S9, and Chapter 5 showed the performance results from the device. However, it is also noted how additional factors could have caused this impact. Finally, as an exploratory investigation, the dissimilarity score

information provided by this library was used and investigated if a need existed for having a different threshold for each scenario by examining the performance observed within each. This potential need can be seen by exploring how the dissimilarity scores from genuine transactions vary in each scenario. Table 7.3 shows this information along with the standard deviation and shows that the average dissimilarity score for the stationary scenarios was 0.21 (±0.08).

In contrast, the average score for the in-motion scenarios was 0.30 (±0.06). It indicates a 43% score increase from a user in a stationary scenario to being in a motion scenario. The unpaired two-tailed t-test gives a t-score equal to 13.84 with an associated p-value of less than 0.00001, demonstrating a statistically significant difference between the genuine distance scores in stationary and motion scenarios. Similar statistical tests proved that the difference between the impostor distance scores in this instance was not statistically significant. It can also be seen that the baseline recognition performance varies across scenarios.

Here four impostors for each genuine user were used as discussed in Section 7.7.1, and the largest FMR occurs in the same scenario used for the enrolment. However, this is also the scenario with a mean dissimilarity score significantly lower than the baseline threshold of 0.6, highlighting the problem and effect of using impostor probes taken in the same scenario on the false match rate. Therefore, it is believed that an adaptive threshold could provide greater security by restricting these passive impostor attacks. These findings highlight reasons for the introduction of unique thresholds into biometric algorithms.

*Table 7.3: Performance variations for each tested scenario*

| | Genuine Mean Dissimilarity Score | Baseline Recognition Performance |
|---|---|---|
| Sitting | 0.16 (±0.07) | FNMR: 0.00<br>FMR: 11.30 |
| Standing | 0.25 (±0.06) | FNMR: 0.00<br>FMR: 9.04 |
| Treadmill | 0.31 (±0.07) | FNMR: 0.00<br>FMR: 8.70 |
| Corridor | 0.29 (±0.05) | FNMR: 0.00<br>FMR: 9.41 |

## 7.5 The Adaptive Scenario Threshold

The adaptive threshold approach alters a biometric system's 'Decision' ('Matcher') component. In a traditional static system, this component is relatively straightforward. First, the stored enrolment reference is compared to an additionally provided probe and receives a match score from the system to determine how similar or dissimilar the two are. Then, having received this match score, a pre-defined threshold can allow genuine users to access the system while keeping as many impostors from accessing the system as possible. The aim is to set a threshold to keep the False Non-Match Rate (FNMR), the

percentage of genuine people rejected by the system, and False Match Rate (FMR), the percentage of impostors accepted by the system, as low as possible. The equal error rate (EER) is the value where the FNMR and FMR are identical, with a low equal error rate indicating a high accuracy for the biometric system. Figure 7.2 shows an example of a static system's 'Decision' component.

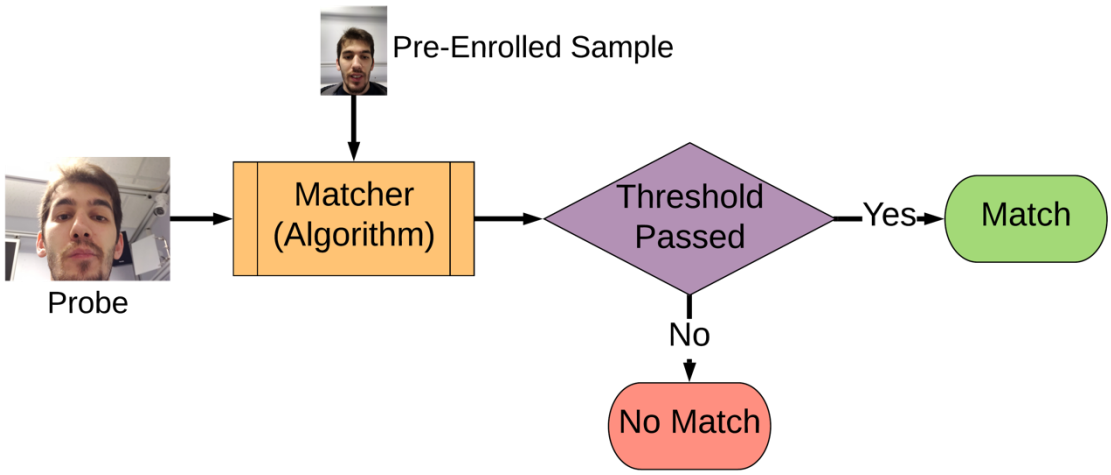



*Figure 7.2: A traditional matcher/decision of a biometric system*

This method addresses whether it can improve overall biometric performance by adapting the threshold based on information from the authentication environment. When using a traditional (static) biometric system, the evaluators can create an appropriate environment and provide directions to users to help ensure optimal usage, giving the best chance of successful authentication. However, with the unpredictability of the environments, scenarios, and conditions in which mobile devices are operated, and hence where the biometric authentication can occur, can the system alter the decision threshold instead to allow optimal performance? The chapter presents how this framework could function in Figure 7.3. Here it is illustrated that instead of having a single threshold to cover the entire spectrum of environments and scenarios, as depicted in Figure 7.2, the system can have a separate limit set for specified situations, such as in this example using 'Stationary' and 'Motion'. This concept is the first work to utilise smartphone sensor data to classify scenarios to create an adaptive biometric system for mobile devices by adjusting the threshold accordingly.

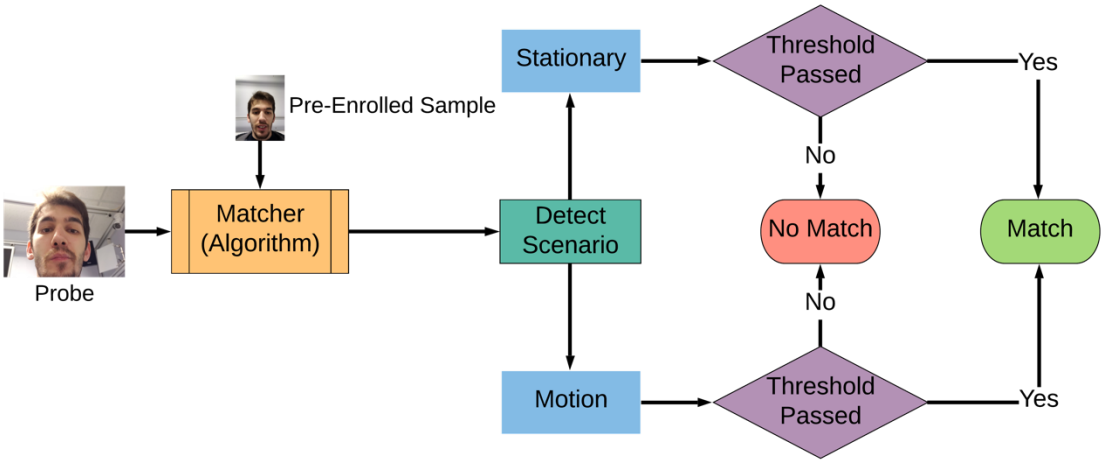



*Figure 7.3: An example framework for a simplified adaptive threshold decision for a biometric system*

In using an adaptive threshold, the expectation is to tailor the authentication experience to better deal with changing movement patterns and allow for enhanced security and user satisfaction. The primary driver is to allow genuine users unobstructed access while keeping out passive impostors. Therefore, using appropriate impostors while designing and testing the framework is vital. The approach to this is discussed further in Section 7.7.1.

## 7.6 Automatic Scenario Detection

A methodology is required to know in what scenario the device user was performing the authentication to achieve this adaptive threshold. The first step is distinguishing between the 'Stationary' and 'Motion' scenarios.

Five features were used for the classifiers, including four in-build mobile sensors, two motion-based sensors, two position-based sensors and a facial image quality assessment. The motion sensors were Gyroscope and Linear Acceleration. The position sensors were Magnetometer (Magnetic Field) and phone Orientation. All these sensors operate on an $x$, $y$, and $z$ axis system, and the data from each channel was collected. The data collection application began collecting the sensor data from the moment the participant started the authentication until the transaction was complete (successful authentication, timeout, attempt limit exceeded). Because the sensor data was collected during the authentication process alone, the entire sample was for analysis. In addition, the participant remained within the scenario when the authentication process began, meaning outliers are not expected in the data from the participants preparing themselves.

The fifth and final feature was the quality assessment of the 'selfie' image. This information came from an open-source library known as 'FaceQnet' and uses a Convolutional Neural Network to "predict the suitability of a specific input image for face recognition purposes" [150]. FaceQnet provides a score for an

input image between 0 and 1, where 0 means the worst quality and 1 means the best quality. In addition, FaceQnet recommends cropping images to the facial region first before assessing them. Using the open-source Multi-task Cascaded Convolutional Neural Networks (MTCNN) library based on the work provided by Zhang *et al.* [135], [174] to achieve this. On the rare occasion that the MTCNN algorithm could not produce a cropped version of the image (usually because the facial region was already over the frames of the images), the original un-cropped image was used instead.

The data in the feature set was processed to achieve reasonable accuracy. The magnitude ($\sqrt{x^2 + y^2 + z^2}$) of the gyroscope, linear acceleration and magnetometer were calculated for each data point obtained for each authentication attempt. Figure 7.4 shows a sample plotted Gyroscope data from one random sitting scenario. The median value from the captured data was used as the feature from each transaction for the orientation.

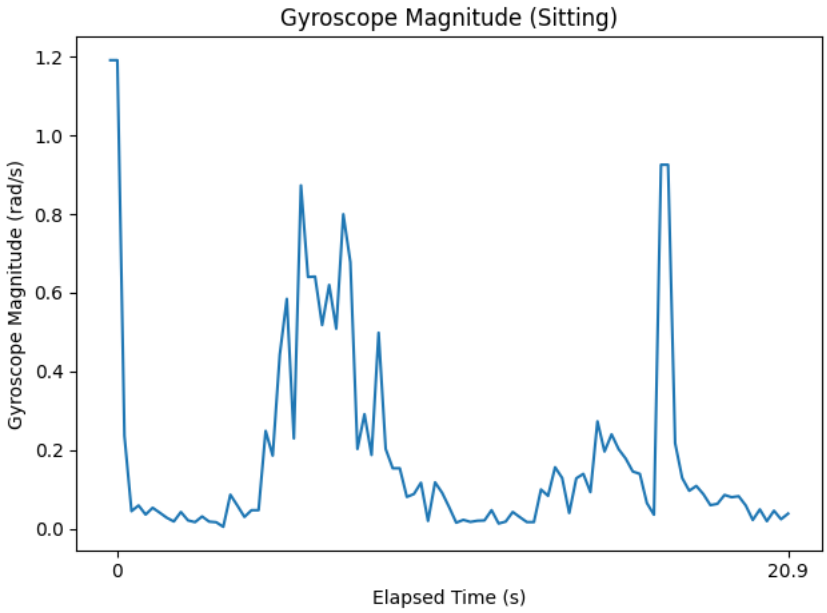



***Figure 7.4: A sample of a gyroscope plot recorded from one transaction during the sitting scenario***

Standard classifier algorithms (SVM, kNN, Naive Bayes, Decision Tree) were tested to see the impact on the performance. The approach started from a 'Stationary' and 'Motion' classifier as it is believed this would provide the most generic form of scenario categories. Next was to create a classifier that could detect the four scenarios explored ('Sitting', 'Standing', 'Treadmill', 'Corridor'). Finally, three classifiers were tested; one to categorise 'Stationary' and 'Motion' and another two to classify into each sub-scenario.

The features were grouped into individual transactions, and a transaction contained multiple rows of features as the sensors continued to release information. Next, half (50%) of the transactions were removed for training and testing. The reason for doing this was to simulate having unseen data for testing

the adaptive framework in its entirety later. This process was repeated five times, selecting a different 50% each time to see the impact of classification accuracy.

Python's Scikit Learn library [175], [176] was used. The features were split into a training (66%) and a testing set (33%). The accuracy was estimated using k-fold cross-validation with a fold value of five and reported the F1-score. The k-nearest neighbour algorithm, with a k-value of three, performed the best with the features. Table 7.4 shows the accuracy results for tested classifiers when classifying the four scenarios.

*Table 7.4: Classification accuracy for standard classifiers*

| Classifier | Cross-Val (F1-score) | Training | Testing |
|---|---|---|---|
| Support Vector Machine | 0.57 (±0.03) | 0.57 | 0.57 |
| Decision Tree | 0.81 (±0.04) | 0.83 | 0.83 |
| Random Forest | 0.80 (±0.02) | 0.79 | 0.81 |
| Naïve Bayes | 0.57 (±0.09) | 0.59 | 0.60 |
| Quadratic Discriminant | 0.58 (±0.09) | 0.60 | 0.62 |

The kNN classifier with a k-value of three could classify all four scenarios with a testing accuracy of **97%**. Table 7.5 shows the accuracy results for each scenario detection classifier using the k-nearest neighbour algorithm for each attempt. The random split of data from attempt five provided the most accurate classifier according to the F1 scores, and this is the one used for the remaining work in this paper. Table 7.6 gives the corresponding confusion matrix for the 'Four Scenarios' classifier when testing with the kNN classifier in attempt five. It is possible to bin most errors under 'Stationary' and 'Motion', where scenarios within each category are misclassified.

*Table 7.5: Scenario Classification Results (kNN)*

| Scenario Classification | Accuracy | Attempt | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Stationary vs Motion | Cross-Val (F1-score) | 0.99 (±0.01) | 0.98 (±0.01) | 0.98 (±0.01) | 0.98 (±0.00) | 0.99 (±0.01) |
| | Training | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 |
| | Testing | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Four Scenarios | Cross-Val (F1-score) | 0.95 (±0.01) | 0.96 (±0.01) | 0.96 (±0.01) | 0.96 (±0.01) | 0.97 (±0.01) |
| | Training | 0.98 | 0.99 | 0.99 | 0.98 | 0.99 |
| | Testing | 0.95 | 0.97 | 0.97 | 0.97 | 0.97 |
| Stationary | Cross-Val (F1-score) | 0.95 (±0.02) | 0.96 (±0.03) | 0.98 (±0.01) | 0.97 (±0.01) | 0.98 (±0.01) |
| | Training | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 |
| | Testing | 0.96 | 0.96 | 0.98 | 0.97 | 0.97 |
| Motion | Cross-Val (F1-score) | 0.96 (±0.02) | 0.98 (±0.01) | 0.97 (±0.01) | 0.98 (±0.02) | 0.97 (±0.01) |
| | Training | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 |
| | Testing | 0.97 | 0.99 | 0.98 | 0.98 | 0.99 |

*Table 7.6: 'Four Scenarios' Confusion Matrix*

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | Sitting | Standing | Treadmill | Corridor |
| True | Sitting | 775 | 14 | 5 | 4 |
| | Standing | 24 | 570 | 3 | 4 |
| | Treadmill | 1 | 2 | 494 | 11 |
| | Corridor | 6 | 5 | 3 | 563 |

Using the kNN classifier to classify all four scenarios provided a **97%** and **99%** testing accuracy when classifying stationary and motion scenarios. In all the classifiers, the achieved testing accuracy was above 90%.

## 7.7 Testing The Adaptive Threshold

Next was to test the approach by using the metadata (features) with the classifier(s) and the 'selfie' image with the 'face-recognition' python library [133], [173]. However, in theory, the approach could be incorporated into commercial devices and work in real-time by integrating it into the biometric authentication process. The approach for this would be like the offline approach, the main difference being the real-time data collection. The device would collect the sensor information about motion and position as the biometric process was happening and turn this information into a feature set like ours.

This feature set would be continuously passed to a classifier to assign a scenario. A majority vote method was then used to assign the overall scenario classification. In other words, the operational scenario with the highest number of occurrences from the scenario detection algorithm was selected. Finally, the overall scenario classification will assign an adaptive decision threshold. The same approach was tested offline for the prototype by using the pre-collected.

A custom Python program that works by excepting two facial images were used to achieve this. All the image files had unique names to locate the data associated with each one. The first image was an enrolment template (taken as the first sitting attempt for each user), and the second was the verification probe. First, the authentication sensor data for the supplied probe image was retrieved. Next, the classifier processed each feature row of data to predict a scenario—a majority vote method assigned the final scenario classification. Once the predicted scenario was known, the program set the decision threshold appropriately. Based on the dissimilarity score and the set threshold, the program then marks the probe image as either a 'match' or a 'no match' decision. It was possible to begin to validate the approach using this information to produce the performance results.

### 7.7.1 Choosing the Impostors

It was necessary to test the potential to keep out passive impostors and evaluate the false match rate of the system to assess the effectiveness of the proposed adaptive framework. In addition, the aim was to find the most suitable (tailored) participants to act as impostors for each enrolled participant. The set theory below represents the algorithm for achieving this.

- Number of Impostors Required: $x$
- Current user: $c$
- Set of Users: $U$ where $c \in U$
- Set of Impostors: $I \subset U = c \notin U$
- An Impostor: $i$ where $i \in I$
- Gender Subset: $G \subseteq I \; \forall \; c.Gender == i.Gender$
- Age Group Subset: $A \subseteq I \; \forall \; c.AgeGroup == i.AgeGroup$
- Nationality Subset: $N \subseteq I \; \forall \; c.Nationality == i.Nationality$
- A subset of Tailored Impostors: $T = G \cap A \cap E \cap N \subseteq I$
- If $|T| \geq x$ {Randomly select $x$ elements from set}
- while $|T| < x$
    - Randomly select from $G \cap A \cap N$ until$|T| == x$ is reached
    - if $G \cap A \cap N$ becomes $\emptyset$ {Randomly select from $G \cap A$ until $|T| == x$ is reached}
    - if $G \cap A$ becomes $\emptyset$ {Randomly select from $G$ until $|T| == x$ is reached}

The 'AgeGroup' was specified in ranges, as shown in Table 7.2. This algorithm should result in a set of $x$ tailored impostors for each participant who resembles the participants. The outcome was to expect the impostor set to provide the most likely cases to cause a false match. An experiment was performed by adjusting the number of tailored impostors to provide meaningful results. When using the algorithm, the more impostors added, the less tailored they will be, diluting the results. For the data, using four impostors per genuine user (2015 impostor comparisons) seemed to provide a fair balance before the impostors became less tailored. This idea is discussed further in Section 7.8 and Figure 7.5.

### 7.7.2 Examining the Threshold

The recommended threshold from the python 'face-recognition' library [133], [173] is 0.6. Using the data gives a false non-match rate of 0.00%, a false match rate of 10.22%, and an equal error rate of approximately 0.64%. It seems the library is recommending a practical threshold value for most cases. However, the concept was to devise a scenario whereby security is of great concern to test the adaptive theory. Therefore, a low (<1%) false match rate is required by setting tighter, more restrictive thresholds.

Section 7.4 identified that the match score varies across scenarios and that the system should set other thresholds for each. Several approaches were trialled to set appropriate threshold values, and in this case, trialling multiple thresholds for the scenarios. The trials allowed the biometric community to see how varying thresholds could affect overall system performance. For example, the maximum distance score obtained from the data could be used. The $95^{th}$ percentile, maximum distance, and the EER threshold value from the scenario data were used as the threshold values to explore the approach. The theory is that this will allow for most genuine cases without causing extremes and outliers in the data to be accepted.

Similarly to how the creation of the scenario classifier was handled, a random 75% sample from the dissimilarity score data (75% from genuine and 75% from impostors) created the thresholds. The impostor scores used for this were created using the tailored impostors. This process was repeated five times, picking a new random set to see the impact, as shown in Table 7.7.

## 7.8 Results

The classifiers produced as discussed ( Section 7.6 ), along with the thresholds found in Section 7.7.2 and chosen tailored impostors based on Section 7.7.1, bring the framework together. Finally, it is possible to examine how the adaptive framework could perform using the open-source 'face-recognition' library [133], [173] and the pre-collected data.

Figure 7.5 shows how the false match rate changes as the algorithm are used to alter the number of impostors used (the algorithm was rerun for each iteration). A randomly selected 450 comparisons were used to provide a reasonable sample from the impostor comparisons pool with $x$ number of impostors per genuine user. This process was repeated three times, and an average was taken to produce the graph. The baseline's FMR declines as the impostors become less tailored; however, the adaptive approach outperforms the baseline with the most tailored impostors and continues to do so even when less tailored impostors are included.

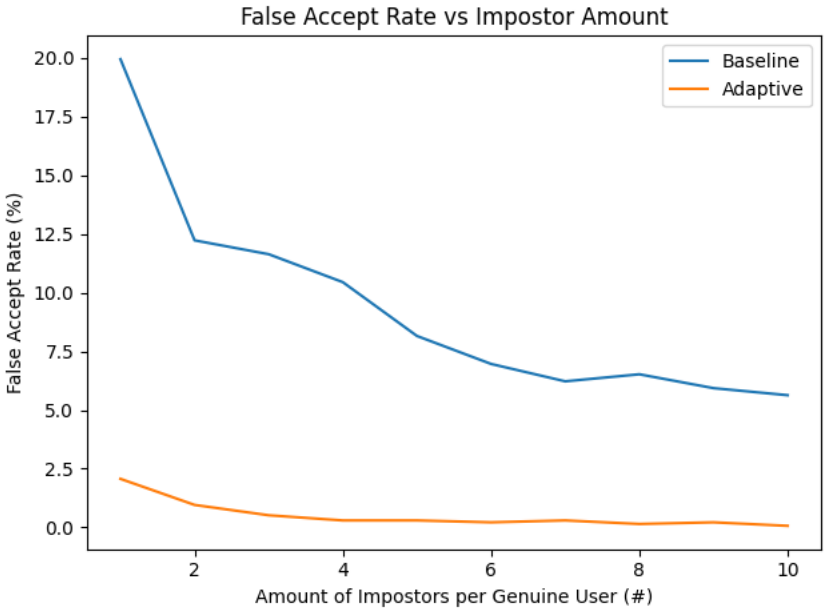



*Figure 7.5: Changes to false match rate with varying impostor amounts*

Firstly, the more generic classifier was tested to classify the authentication metadata into a 'Stationary' and 'Motion' category, followed by a test of the classier that could distinguish between the four scenarios experimenting with: 'Sitting', 'Standing', 'Treadmill', 'Corridor'. Finally, a combination of the two classifiers. The data would first classify into 'Stationary' and 'Motion' and then into separate classifiers for the scenario that belonged to either category. The approach can achieve recognition results by trialling both using '95th', 'Max' and 'EER' thresholds, as shown in Table 7.2.

*Table 7.7: Recognition performance results when trialling the adaptive threshold*

| Classifier | Threshold | Attempt | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Stationary vs Motion | 95th | FNMR: 5.29 FMR: 0.00 | FNMR: 4.65 FMR: 0.00 | FNMR: 6.98 FMR: 0.00 | FNMR: 5.07 FMR: 0.00 | FNMR: 6.13 FMR: 0.00 |
| | Max | FNMR: 0.00 FMR: 2.53 | FNMR: 0.00 FMR: 2.53 | FNMR: 0.00 FMR: 2.53 | FNMR: 0.00 FMR: 2.53 | FNMR: 0.63 FMR: 0.35 |
| | EER | FNMR: 0.42 FMR: 0.35 | FNMR: 0.42 FMR: 0.69 | FNMR: 0.42 FMR: 0.50 | FNMR: 0.42 FMR: 0.50 | FNMR: 0.85 FMR: 0.25 |
| Four Scenarios | 95th | FNMR: 6.13 FMR: 0.05 | FNMR: 7.19 FMR: 0.10 | FNMR: 7.40 FMR: 0.00 | FNMR: 5.92 FMR: 0.10 | FNMR: 6.98 FMR: 0.00 |
| | Max | FNMR: 0.85 FMR: 1.39 | FNMR: 1.06 FMR: 1.39 | FNMR: 0.00 FMR: 2.53 | FNMR: 0.00 FMR: 2.53 | FNMR: 1.48 FMR: 0.25 |
| | EER | FNMR: 0.85 FMR: 1.39 | FNMR: 1.27 FMR: 0.89 | FNMR: 1.06 FMR: 0.50 | FNMR: 1.06 FMR: 0.55 | FNMR: 1.48 FMR: 0.25 |
| Stationary/Motion + Scenarios | 95th | FNMR: 6.13 FMR: 0.05 | FNMR: 7.19 FMR: 0.10 | FNMR: 7.40 FMR: 0.00 | FNMR: 5.92 FMR: 0.10 | FNMR: 6.98 FMR: 0.00 |
| | Max | FNMR: 0.85 FMR: 1.39 | FNMR: 1.06 FMR: 1.39 | FNMR: 0.85 FMR: 1.39 | FNMR: 0.00 FMR: 2.53 | FNMR: 1.48 FMR: 0.25 |
| | EER | FNMR: 1.06 FMR: 0.55 | FNMR: 1.27 FMR: 0.89 | FNMR: 1.06 FMR: 0.50 | FNMR: 1.06 FMR: 0.55 | FNMR: 1.48 FMR: 0.25 |

It is known from the classifier accuracy that the classifier is not classifying all the scenarios correctly every time. As a result, there is a risk of incorrect classification of a scenario with an alternative acceptance threshold. This misclassification poses a risk for impostors to be accepted by the system. Further work to improve the classifier accuracy will result in improved recognition performance.

The results show that an adaptive approach can produce reliable recognition accuracy, notably by maintaining and improving a low false match rate above a traditional fixed value. Table 7.8 highlights this comparison using the four-scenario classifier and EER (number 3 in Table 7.7 thresholds to the baseline and a perfect classifier. A perfect classifier would be able to accurately categorise the scenarios 100% of the time. The success of the most effective adaptive approach has reduced the false match rate by approximately 95% from baseline performance.

*Table 7.8: Comparing Recommended Baseline Performance to the Adaptive Approach*

| | Recognition Performance |
|---|---|
| **Baseline** | FNMR: 0.00<br>FMR: 10.22 |
| **Adaptive Threshold** | FNMR: 1.06<br>FMR: 0.50 |
| **Perfect Scenario Classifier** | FNMR: 0.42<br>FMR: 0.60 |

### 7.8.1 Verification

Having tested the scenario adaptive threshold method on the Samsung Galaxy S9, the same concept was tested on another device to see if the approach was interoperable. Therefore, another Android-based device, the Google Pixel 2, was tested. However, Google Pixel 2 does not allow developers access to its 'Trusted Face' feature, meaning it is impossible to collect background sensor data during the authentication process. The device collected the sensor features while the participant was operating the in-built camera and taking a 'selfie' to simulate the authentication process to counter this. Unfortunately, the side effect means there was a lot more sensor data collected. On average, operating and using the camera takes more time than the usual biometric authentication prompt to complete.

The data were collected under the same scenario conditions. Resulting in an additional 100 genuine 'sitting' transactions, 116 'standing' transactions, 124' treadmill' and 141 'corridor' from around 30 individuals of a similar student demographic that operated the Samsung Galaxy S9. Furthermore, when running the 'selfie' data collected from the Google Pixel 2 through the 'face-recognition' Python library with a baseline threshold (0.6), the performance results were FNMR: 0.00% and FMR: 9.50%.

The same approach was used by removing half of the transactions before classifying them. As devices are unique with different sensors, it is impossible to rely on using the same classifier as before, and unique ones will need to be produced for each device/model. The classifier evaluation results were promising, with the 'four scenarios' classifier reporting cross-validation accuracy as 1.00 ($\pm$0.00) and the training and testing accuracy being 1.00%. Again 75% of the dissimilarity score data set appropriate thresholds. Three sets of EER thresholds were generated as a trial by altering 75% of the data used. For the three trials, the results were again showing improvements over the baseline case and beginning to prove that the adaptive threshold is interoperable:

- FNMR: 1.25% and FMR: 0.00%
- FNMR: 2.91% and FMR: 0.00%
- FNMR: 1.66% and FMR: 0.00%

## 7.9 Discussion

The approach utilised the sensors readily available on most modern smartphones (and wearables) with developer access. It showed the transformation into potential features for a classifier that could recognise simple scenario categories. The classifier for detecting the four simple scenarios had a testing accuracy of **97%**. The adaptive threshold relies on having the ability to identify the scenario reliably, and the results suggest that this ability is a significant factor in the overall function of the adaptive framework to perform optimally. The chapter focuses on using an adaptive approach for face recognition, but no significant obstacles are foreseen for using the same technique for other physical and behavioural biometric modalities.

The process was demonstrated using collected data from a commercial device and an open-source face recognition algorithm showing that this method has potential merit by imagining a scenario where security and privacy are of grave concern. Hence, a low false match rate may be more important than a false non-match rate. The method was tested against a static, fixed threshold. An algorithm was produced to help identify the best impostors for each participant to help stress-test the approach. The impact of tailoring was demonstrated in Figure 7.5, which proved that the algorithm was working.

Offline testing was then performed using Python's 'face recognition' library [133], [173], which recommends a threshold of around 0.6. This threshold gave a false non-match rate of 0.00% and a false match rate of 10.22% using the previously collected data. One of the best methods found by taking the adaptive decision approach was using the classifier to detect between stationary and motion scenarios and the EER threshold value. The approach achieved a result that gave a false non-match rate of **0.42%** and a false match rate of **0.35%**, a reduction of 95% over the algorithm's baseline threshold. The interoperability of the approach was shown by replicating it using another device with similarly successful results. This relatively simple adaptive method was able to produce an improvement in recognition performance, which could outperform an algorithm using a single static threshold value.

## 7.10 Summary

This chapter presented a novel adaptive approach to biometric authentication for a mobile device, an area of research currently lacking in the literature, as noted by Pisani *et al.* [159]. The proposal created an extendable 'Adaptive Decision Threshold', whereby a unique threshold value is set for specified scenarios. The theoretical advantage of this approach is to allow for stricter control over access by not having to specify a one-off static threshold value to account for the vast number of conditions where a biometric authentication may occur.

This relatively simple example demonstrates the practicalities and proof-of-concept of using this adaptive threshold approach. Further testing will be required to prove the competency of the method

thoroughly, including a greater variety of scenarios and environmental lighting and weather conditions. In addition, testing should include adapting presentation attack detection (PAD) methods for individual scenarios and mitigating malicious actors in exploiting weaknesses in the adaptive approach. The hope is that others will take the work started here to produce and further investigate the method's effectiveness. As well as allow developers and manufacturers to incorporate a scenario-based threshold adaptive approach into future algorithms in mobile biometric systems to allow for higher security without jeopardising performance.

The next chapter will present the conclusion for the thesis, providing what has been addressed and considerations for future work.

# 8 Conclusions

## 8.1 Introduction

This thesis provides the first foundational step towards a novel performance assessment framework for mobile biometrics. However, producing a comprehensive and extensive framework that can cover and capture the entire breadth and depth of performance is more than a single PhD project. To fully achieve this goal will require collaboration on a global scale using the expertise of the ISO SC 37 WG5 committees and the FIDO Alliance and incorporating private companies.

The novelty of the work mainly comes from the focus on mobile devices and analysing data captured entirely from mobile devices. The proposed framework also considers the level of access the evaluator will have to the device. The current literature assumes that a certain level of access to evaluate a system will be available.

The contribution is to allow this thesis and the recommendations to serve as a starting point for further exploration with the potential to incorporate some of the identified areas into future standards and documents within the scope of mobile biometric performance testing. Furthermore, academia can use this comprehensive guide to provide a comparable guide for performance results with a specified captured scenario.

Although the term 'mobile' includes a range of devices, the core focus of the thesis was considering a smartphone scenario. However, the proposal should be compatible with a range of mobile devices. A handheld iris scanner that cooperated with a smartphone was used to indicate that the framework could be interoperable between mobile devices.

This concluding chapter will provide an overview of the work presented throughout this thesis, including lessons learnt and contributions. Section 8.2 acknowledges lessons learnt and limitations mainly regarding the collected dataset. Section 8.3 provides a brief illustrative walkthrough through the assessment framework concerning previous chapter results. Finally, section 8.4 discusses thesis contributions, and Section 8.5 and 8.6 provides a reflection and closing summary.

## 8.2 Limitations and Lessons Learnt

The project was not without its share of problems, particularly regarding running a biometric data collection exercise. Collecting demographic information was not done as well as it should have been, and important information for later analysis was lacking. For example, ethnicity information and age values

(as opposed to age ranges) should have been used to evaluate better are tailored impostor theory. Initially, the decision not to include this information was made to ensure the experiment was run appropriately and maintain a high level of ethical standards to gain ethical approval from the university.

The data collection was produced using commercial smartphones, which means the same issues resulting in the lack of access to the biometric system were present. However, this ensured that the proposed framework could handle a range of device access considerations. However, it did mean certain aspects of the framework could not be fully explored, such as security and privacy.

At the start of the project, it was decided that PAD should not be a focus of the presented work. However, given PAD prevalence in existing requirements and analysis of biometric system performance, it was later deemed to be an essential component for a full-scale mobile biometric performance evaluation, and it was included; however, due to its late-stage inclusion, practical work was not conducted, and therefore the inclusion in the framework remains theoretical at this point.

## 8.3 The Performance Assessment Framework

So far, this thesis has analysed the effects of various core factors previously identified as crucial to understanding the performance of mobile biometric systems. The proposed framework utilised these factors in its formation, and pulling in the results obtained will demonstrate an illustrative theoretical walkthrough of the framework using the experimental data collection results. One of the key differences regarding the proposed framework is working with multiple access levels.

Not every element of the proposed framework was explored for this project owing to time constraints and the ongoing COVID-19 pandemic, including PAD. However, walking through each stage of the framework produces the following outcome. First, the experimental data collection involved collecting data from commercial devices, meaning obtaining significant performance results is limited. However, using ‘developer’ access and our custom application allows for more usability-related performance metrics. Equally, taking the obtained biometric sample data captured from the mobile devices and applying third-party open-source libraries to act as the algorithm can allow the simulated evaluation of the proposed framework as if a higher level of access (e.g., ‘tester’) was available.

### 8.3.1 Stage One: Determine Evaluation Parameters

Stage one required defining three core parameters: modality, access level, and desired security level. This exercise examined fingerprint, Face, Iris, and voice as unimodal (not multimodal) modalities. For the level of access, the intent is to illustrate several outcomes from ‘developer’ through to ‘open’; however, only one level of access is assumed when using the performance framework in practice. No samples were collected and operated only at the ‘developer’ level of access for fingerprint. The desired security level is

of less concern for this illustration, as this parameter dictates the extent of testing and the necessary performance results that need to be achieved to consider the system a 'pass'.

### 8.3.2 Stage Two: Algorithmic Evaluation

Stage two provides the algorithmic test for evaluating the algorithm(s), usually involving existing frameworks. Looking at each level of access:

- Closed, User and, Developer – Algorithmic evaluation is considered impossible due to the inability to operate the algorithm offline and allow data injection.
- Tester and Open–Algorithmic evaluation is considered possible as the ability for offline access and the injection of data is possible.

For this illustrative example, the reported results of the open-source algorithms for face and voice are displayed:

- Face Recognition - The model has an accuracy of 99.38% on the Labelled Faces in the Wild benchmark
- Deep Speaker - Deep Speaker reduces the verification equal error rate by 50% (relatively) and improves the identification accuracy by 60% (relatively) on a text-independent dataset.

### 8.3.3 Perform Baseline Evaluation

Stage three requires the results of a baseline evaluation in 'optimal' conditions. The condition defined for this is indoors with no noise, with the device handheld by the user while seated. The experimental data collection corresponds to the data obtained from the first scenario of session one, where the user is operating the device in a seated position. It should be noted that no habituation transactions occurred, something added to the framework from this experience.

The first results shown in Chapter 5 form part of the baseline evaluation from a 'developer' level of access and, therefore, can be attributed to the device used. In practice, the baseline evaluation should include false match rates; however, this study omitted to perform offline FMR due to time constraints.

### 8.3.4 Stage Four: Targeted Scenario Evaluation

Stage four requires defining scenarios to conduct a performance test within. The main requirement is to involve a motion-based scenario and a challenging known influencing factor scenario. Again, with the 'developer' level of access, this comes directly from the Boolean result of the verification.

This approach allows for more in-depth analysis and explanation of the performance obtained. Specifically, we can group the scenarios into stationary and motion (not the influencing factor scenario). The results for this are shown within Chapter 5, where a breakdown of the scenario results achieved and a further breakdown showcasing what caused the false non-match to occur.

### 8.3.5 Stage Five: Presentation Attack Detection and Architectural Security

Stage five of the framework is concerned with security regarding the circumvention of the system. This stage was not explicitly evaluated as part of the experimental data collection. A high-level investigation was conducted to see if anything was obtainable from the logs regarding architectural security; however, no evidence of security or privacy concerns was found.

### 8.3.6 Stage Six: Operational Evaluation

Stage six explores the operational performance of the device, and the requirements for this are device-specific. The recommendation, where possible, is to trial the device in various outdoor conditions. How the results are presented can differ based on the evaluation performed. The experimental data collected contained information about the conditions (weather) at the time (for the current hour) of collection. The results for this were included in Chapter 6.

### 8.3.7 Stage Seven: (Final) Reporting

Stage seven is the reporting stage, although the recommendation, as demonstrated here, is to produce a report of results as the evaluators move through the stages of the performance framework. The fundamental reporting criteria are the false non-match rate (FNMR), the false match rate (FMR), and the defined scenario in which the results were achieved. Ideally, the results should include other usability quality scores (satisfaction, timings). This approach should showcase the baseline (optimal) results and the targeted scenario evaluation composed of, for example, the defined stationary scenarios, the defined motion scenarios, and the defined factor scenarios.

## 8.4 Thesis Contributions

Building on and providing an overview of best practices for biometric assessment, including the 19795 series and the FIDO work, the thesis defined a range of performance metrics relating to biometric system accuracy, timing, environmental considerations, usability errors and security levels. In doing so, an initial performance framework for reporting and exploring system performance was designed with the intention of an extensible design that will have direct generic applicability to other existing and emerging

modalities, with specifics where required. Furthermore, seven core factors relating to mobile biometric performance were identified and utilised in discussions around performance, serving as the foundation for all future researchers looking into presented performance-related results.

The thesis assessed mobile performance in a controlled environment (lighting and noise) and using standard smartphone in-built biometric sensors in scenarios including holding in hand (while sitting and standing) and moving on a treadmill device using the performance framework. A range of influencing factors was also introduced, including varying the controlled illumination within the test room to simulate dark lighting conditions and background noise. Data were collected from 60 participants across mainly a student demographic.

The performance assessment of both the black-box OS biometric systems, from commercial smartphone devices and open-source algorithms that can return more performance data, were explored to assess the framework—first, assessing the baseline performance, allowing the exploration of performance alterations. The performance was assessed according to verification rates, sample quality, timings, usability assessments, and offline questionnaires. Local Ethics Committee Approval was obtained for the data collection.

The performance of mobile devices across various environments was explored using the framework. Subjects used the same devices to interact with biometric sensors whilst on the move. The environment involved data collection both indoors and outside and whilst walking. In addition, the experiment involved a 'free-style' capture session whilst moving around the university campus. The thesis has explored the environmental impact, how performance characteristics are affected by environmental data and scenario impact, and how performance characteristics are affected by scenario usage.

An approach to overcome performance deficit through an introductory study regarding a novel authentication approach for mobile devices using embedded sensors to classify scenarios to improve recognition performance. In doing so, further analysis of the tailored impostor approach was presented.

The thesis has started to provide a robust assessment of mobile biometric platforms by establishing a theoretical framework for mobile performance assessment. The framework considers the following:

- Use of mobile devices in various environments, including indoors and outdoors, using whilst stationary, walking
- The user base and deployment platform
- Capture requirements at enrolment to enhance the performance of mobile biometrics
- The nature of the transaction utilising the biometric authentication
- System performance, accuracy, and timings
- The security level needed – a three-way balance between the 'quantity' of a transaction, ease of use and susceptibility to presentation attack scenarios

- Metrics for analysis and the presentation of understandable/interpretable performance results.

Biometric performance testing is complex, time-consuming and expensive, with large-scale testing often restrictive for non-large-scale multinational corporations, resulting in requirements such as the rule of 3 and rule of 30 test sizes becoming ignored [27]. The approach taken by the proposed performance assessment framework presented within this thesis attempted to address some of these concerns for mobile platforms. The concept presented was to target a worse-case approach to all aspects of the testing process, which is generally not considered within the literature and current testing standards that aim to show the system positively. Manufacturers will be resistant to showcasing results that showcase their product negatively.

By introducing a tailored impostor's approach to non-mated comparisons, we aim to exploit this worst-case analysis using more challenging populations, i.e., those that are more similar. The hypothesis is that the performance will increase the false match rate, and since that rate dominates the sample size calculations, the required sample sizes are likely to be smaller. However, there is still a potential issue regarding how those results generalise to a population that is not as similar if the effort required to gather a more specific sample will outweigh the potential benefits of running the evaluation.

The outcome is the foundation for a robust performance assessment of mobile biometrics concerning environmental and usage variation. In addition, the performance framework will allow for a deeper academic and security-focused understanding of anticipated performance levels, with the possibility to develop mitigation techniques and support tailored to specific end-use scenarios.

## 8.5 Reflection

The PhD and, by extension, this thesis had been a journey, and like all large-scale projects, there were twists and turns, and not everything planned at the start could be completed in detail by the end. One such element was an exploration into automated approaches to biometric testing using statistical methods to predict performance based on certain factors, mainly the core factors presented in Chapter 3. This idea is currently exploring an ongoing ISO standard [95]. The second such element and an ongoing area of research within the biometric community is the black-box nature of biometric algorithms that rely more on artificial intelligence and deep learning techniques. Initially, the intent was to explore and uncover the black box (explainable AI). However, after initial scoping, it was revealed that this work could form its PhD project.

Regardless of some elements that could not be fully explored in work presented in this thesis, the thesis has extensively explored the world of mobile biometric performance and produced a performance

evaluation framework. Furthermore, looking back at the research questions ( Section 1.6 ), the thesis has made progress in answering them all.

- **How can mobile biometric performance be measured?**

Arguably, this thesis's main aim is to explore and produce a definitive standard for measuring mobile biometric performance. Has this been achieved? The thesis provides a strong foundation for measuring mobile biometric performance by utilising and amalgamating existing work in this area and adding necessary recommendations for moving this area of research forward with the inclusion of core factors. However, changing current practices and making ground-breaking additions to the current standards will take time. However, hopefully, this work will form a reference for updating and improving as research and industry move forward.

- **How do environment and motion affect biometric performance and security?**
- **How does the quality/performance of <modality> change when placed in diverse scenarios and Environments?**

It was acknowledged that operating within various environments and scenarios is one of the principal elements separating mobile biometric systems from fixed systems. Therefore, including recommendations to make sure this was fully addressed within the performance framework was paramount. This framework was followed up by a test study to explore the impact of these factors and consolidate the inclusion within the framework, the results of which were presented in Chapter 6, showing the potential impact of the environmental conditions and scenarios. Quality scores were used to gauge further information regarding performance and usability

- **How can any performance deterioration be mitigated?**

The next step was to explore if it was possible to mitigate against the established core factors, having established the performance framework and then proceeding to justify it. For this, a novel authentication approach was designed, explored and tested by utilising the typical sensors available on mobile and wearable devices to predict the scenario that the user was performing the authentication within and adjust the decision threshold dynamically and accordingly. This work was presented within Chapter 7 and shows the potential of this method as a successful approach to improve performance deterioration without jeopardising security.

- **What is the current smartphone security (locking) habits among users?**

This question was provided as an insight into the current user habits of mobile biometrics and to help target the framework accordingly. It was interesting to see where users placed their security habits regarding phone unlock with a preference for fingerprint. However, it would be expected that if the same questions were repeated today (2022), a preference and use towards face unlock would be seen due to the industry and technology movement in this direction, with most smartphones today incorporating the technology. The overview of these results is showcased in Chapter 5.

- **Do they primarily use the biometric authentication method available on their device?**

This question is built on the previous one. It was seen that biometrics did form the primary method of authentication for users wishing to unlock their devices and is continuing to see rises in use across other sectors as a secure yet usable method of authentication, with trends showing that biometrics will likely authenticate the vast majority of banking transactions [177].

- **Is there an overwhelming preference towards a particular modality?**

Perhaps unsurprisingly, there was a strong preference towards fingerprint as the biometric modality for authentication. However, as this was also the most popular modality in use at the time among the participants, it supports the usability and habituation theories that the more use and exposure users have, the more comfortable and greater preference they are likely to show. This trend will change over time, and it is suspected that if a repeat of the question was issued today, a higher, if not majority, preference for the face modality is likely to be observed.

- **Is it possible to achieve reliable performance metrics with a commercial (off-the-shelf) device with limited knowledge of its internal workings?**

This question proved to be an exciting exploration of the work presented in this thesis. The limited access to the workings of the biometric system of commercial devices is troublesome for researchers but utterly understandable from a privacy and security perspective. Whether it is possible to achieve reliable performance metrics is debatable, and evaluators are certainly limited in what is achievable. However, it is still possible to perform a biometric performance evaluation with careful consideration. One of the essential requirements for the proposed performance framework is the ability to consider the level of access an evaluator has and adjust the flow accordingly so that meaningful information can be obtained without exhaustive testing, but this does come at the expense of scaling down the evaluation. The introduction of tailored impostors to perform impostor testing with fewer participants but with the most significant impact was designed to overcome some issues.

- **What effect does usability have on the performance of biometrics on a mobile platform?**

Regarding usability, the habituation effect was discussed and demonstrated within the results leading to the alteration of the performance framework to include some trial or habituation transactions before the principal transactions take place to avoid potential bias. Alongside this transactional time, satisfaction scores and error rates showed some of the usability impacts of the devices. It is noted that usability was not explored in-depth as some of the other factors, partly due to the sheer scope of analysing as many aspects as possible that comprise mobile biometric performance. However, it was discovered as part of the scoping of the core factors that the users play a central role in understanding the performance and the consideration to make sure that usability metrics were included as part of the performance framework and analyse was added for this reason. Furthermore, it was observed that specific metrics, like satisfaction scores, often related to the error rates achieved.

- **Is it possible to produce a definitive score or ranking regarding the device's performance?**

It turns out that this question may not be as simple to answer as it first appears. Biometric performance can be broken down into several areas (recognition performance, usability, security) and trying to provide a singular score to combine all this information may not be ideal as, depending on use cases, some areas of performance may be more necessary than others which is why the results section of the framework breaks these respective areas and suggests reporting the information from each allowing a more informed decision to be reached.

Overall, the thesis has looked into the three core questions it hoped to achieve and gone further to look into some sub-questions that helped answer the core questions, including current trends and habits among mobile users and how any performance deterioration could be mitigated once measured. As stated at the start of this section, not everything was envisaged, and the start of the project was achievable by the end, leaving future scope for further work to help complete some of these consolidating areas. Continuing and incorporating the findings of this work will require global standard cooperation, and discussions are likely still required on how biometric performance should be scored and reported and whether it is possible to achieve this with a definitive score or rank. Hopefully, this thesis will play its part in aiding future researchers, standard makers and industry stakeholders in the performance assessment of mobile biometric systems.

## 8.6 Summary

This chapter has wrapped up the work presented within this thesis and performed an initial walkthrough of the proposed performance evaluation framework stating some current limitations with the work presented and noting possible improvements and future work. As stated at the start of this chapter, creating a fully-fledged performance assessment framework is more than any PhD project. However, the intention is to serve as the start of something for academia, industry, governments, and future standardisation work, allowing for further exploration and research into the proposals presented in this thesis.

Regarding future work ideally, it would be ideal to evaluate if performance trends can be predicted from existing results, something that the proposed framework will be able to handle more effectively due to the inclusion of defined scenarios to answer the question: can we identify trends in performance that are common between devices/modalities that can predict performance for an unknown device?

Further analysis and investigations into using the approaches discussed within the proposed framework are required to explore all the approaches it can offer. However, certain aspects will hopefully be helpful for future mobile biometric performance assessments. The thesis aimed to present a comprehensive testing framework for mobile biometrics to provide users and businesses with the assurance they need to satisfy their requirements.

The framework utilised existing standards and approaches to amalgamate current industry standards and recommendations, including the likes of ISO, FIDO Alliance, and Google. The intention was also to go further and improve upon the existing standards in certain aspects by incorporating more performance areas into our framework, most notably usability. Hopefully, evaluators will see the benefit of some of the ideas present and will work towards incorporating and analysing the benefits further:

- Separate FTA from FRR
- Consider FNMR and FMR (attempt-based) (for usability purposes)
- Considers options for when the 'common test harness' is not available
- Use targeted impostors (the current suggestion is to use random for attempt-based)
- Defined enrolment scenario
- Defined multiple verification scenarios (including motion) + operational
- Different levels of security testing

# *References*


[1] 'Biometrics Institute Industry Survey 2020', Biometrics Institute, Jul. 2020. Accessed: Mar. 16, 2020. [Online]. Available: https://www.biometricsinstitute.org/industry-survey-2020-summary/

[2] 'Biometrics Institute Industry Survey 2021', Biometrics Institute, Jul. 2021. Accessed: Jul. 18, 2021. [Online]. Available: https://www.biometricsinstitute.org/industry-survey-2021-summary-2/

[3] FIDO Alliance, 'Biometric Component Certification', *FIDO Alliance*. https://fidoalliance.org/certification/biometric-component-certification/ (accessed Apr. 22, 2021).

[4] M. Farik, N. Lal, and S. Prasad, 'A Review Of Authentication Methods', *International Journal of Scientific & Technology Research*, vol. 5, pp. 246–249, Nov. 2016.

[5] 'Information technology — Vocabulary — Part 37: Biometrics', International Organization for Standardization, Geneva, Standard ISO/IEC 2382-37:2017, Feb. 2017.

[6] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*. Springer Science & Business Media, 1999.

[7] 'Information technology — Biometric performance testing and reporting — Part 1: Principles and framework', International Organization for Standardization, Geneva, Standard ISO/IEC 19795-1:2021, May 2021.

[8] 'Intro-to-FIDO.pdf'. Accessed: Mar. 19, 2022. [Online]. Available: https://fidoalliance.eventstreamvm.com/wp-content/uploads/2020/10/Intro-to-FIDO.pdf

[9] National Cyber Security Centre, 'Biometric Recognition and Authentication Systems - Measuring Performance', *Measuring Performance - NCSC.GOV.UK*, Jan. 24, 2019. https://www.ncsc.gov.uk/collection/biometrics (accessed Mar. 16, 2021).

[10] T. Mansfield and G. Kelly, 'Measuring and comparing the performance of biometric systems', *Information Security Technical Report*, vol. 3, no. 1, pp. 70–76, Jan. 1998, doi: 10.1016/S1363-4127(98)80021-6.

[11] 'About Face ID advanced technology', *Apple Support*. https://support.apple.com/en-gb/HT208108 (accessed Mar. 02, 2021).

[12] | Chris Burt, 'Survey shows biometrics not trusted or understood by majority of consumers | Biometric Update', Jun. 09, 2019. https://www.biometricupdate.com/201906/survey-shows-biometrics-not-trusted-or-understood-by-majority-of-consumers (accessed Nov. 08, 2021).

[13] A. Wojciechowska, M. Choraś, and R. Kozik, 'The overview of trends and challenges in mobile biometrics', *Journal of Applied Mathematics and Computational Mechanics*, vol. Vol. 16, no. nr 2, 2017, doi: 10.17512/jamcm.2017.2.14.

[14] R. R. Jillela and A. Ross, 'Segmenting iris images in the visible spectrum with applications in mobile biometrics', *Pattern Recognition Letters*, vol. 57, pp. 4–16, May 2015, doi: 10.1016/j.patrec.2014.09.014.

[15] R. Giot, C. Rosenberger, and B. Dorizzi, 'Performance Evaluation of Biometric Template Update', *arXiv:1203.1502 [cs]*, Feb. 2012, Accessed: Mar. 13, 2022. [Online]. Available: http://arxiv.org/abs/1203.1502

[16] A. Buriro, Z. Akhtar, B. Crispo, and S. Gupta, 'Mobile Biometrics: Towards A Comprehensive Evaluation Methodology', in *2017 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2017, pp. 1–6. doi: 10.1109/CCST.2017.8167859.

[17] R. Blanco-Gonzalo, R. Sanchez-Reillo, O. Miguel-Hurtado, and J. Liu-Jimenez, 'Usability analysis of dynamic signature verification in mobile environments', in *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*, Sep. 2013, pp. 1–9.

[18] Z. Syed, J. Helmick, S. Banerjee, and B. Cukic, 'Effect of User Posture and Device Size on the Performance of Touch-Based Authentication Systems', in *2015 IEEE 16th International Symposium on High Assurance Systems Engineering*, Jan. 2015, pp. 10–17. doi: 10.1109/HASE.2015.10.

[19] B. Fernandez-Saavedra, R. Sanchez-Reillo, R. Ros-Gomez, and J. Liu-Jimenez, 'Small fingerprint scanners used in mobile devices: the impact on biometric performance', *IET Biometrics*, vol. 5, no. 1, pp. 28–36, Mar. 2016, doi: 10.1049/iet-bmt.2015.0018.

[20] P. Biometrics, 'Understanding biometric performance evaluation', *URL: https://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation.pdf (pristupljeno: srpanj 2018.)[10]*, 2014.

[21] B. Fernandez-Saavedra, R. Sanchez-Reillo, C. Sanchez-Redondo, and R. Blanco-Gonzalo, 'Testing of Biometric Systems Integrated in Mobile Devices', in *2015 International Carnahan Conference on Security Technology (ICCST)*, Sep. 2015, pp. 321–326. doi: 10.1109/CCST.2015.7389704.

[22] 'Information technology — Biometric performance testing and reporting — Part 1: Principles and framework', International Organization for Standardization, Geneva, Standard ISO/IEC 19795-1:2006, Apr. 2006.

[23] E. Ellavarason, R. Guest, and F. Deravi, 'A Framework for Assessing Factors Influencing User Interaction for Touch-based Biometrics', in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 553–557. doi: 10.23919/EUSIPCO.2018.8553537.

[24] E. Ellavarason, R. Guest, and F. Deravi, 'Evaluation of stability of swipe gesture authentication across usage scenarios of mobile device', *EURASIP J. on Info. Security*, vol. 2020, no. 1, p. 4, Dec. 2020, doi: 10.1186/s13635-020-00103-0.

[25] E. Ellavarason, R. Guest, F. Deravi, R. Sanchez-Riello, and B. Corsetti, 'Touch-dynamics based Behavioural Biometrics on Mobile Devices – A Review from a Usability and Performance Perspective', *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–36, Feb. 2021, doi: 10.1145/3394713.

[26] C. Bhagavatula, B. Ur, K. Iacovino, S. M. Kywe, L. F. Cranor, and M. Savvides, 'Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption', presented at the Workshop on Usable Security, San Diego, CA, 2015. doi: 10.14722/usec.2015.23003.

[27] T. Eglitis, E. Maiorana, and P. Campisi, 'Influence of Test Protocols on Biometric Recognition Performance Estimation', in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2021, pp. 1–5. doi: 10.1109/BIOSIG52210.2021.9548315.

[28] A. Dube, D. Singh, R. K. Asthana, and G. Singh Walia, 'A Framework for Evaluation of Biometric Based Authentication System', in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020, pp. 925–932. doi: 10.1109/ICISS49785.2020.9315933.

[29] Gitprakhar13, 'Windows Hello biometrics in the enterprise (Windows) - Windows security'. https://docs.microsoft.com/en-us/windows/security/identity-protection/hello-for-business/hello-biometrics-in-enterprise (accessed Mar. 14, 2022).

[30] windows-driver-content, 'Windows Hello biometric requirements'. https://docs.microsoft.com/en-us/windows-hardware/design/device-experiences/windows-hello-biometric-requirements (accessed Mar. 14, 2022).

[31] 'Measuring Biometric Unlock Security', *Android Open Source Project*. https://source.android.com/security/biometric/measure (accessed Feb. 08, 2022).

[32] 'Apple publishes whitepaper on iOS security, details Touch ID fingerprint sensor specs, functionality | Biometric Update', Feb. 28, 2014. https://www.biometricupdate.com/201402/apple-publishes-whitepaper-on-ios-security-details-touch-id-fingerprint-sensor-specs-functionality (accessed Mar. 16, 2022).

[33] 'About Face ID advanced technology', *Apple Support*. https://support.apple.com/en-gb/HT208108 (accessed Mar. 16, 2022).

[34] 'Apple buys mobile security firm AuthenTec for $356 million', *Reuters*, Jul. 27, 2012. Accessed: Mar. 16, 2022. [Online]. Available: https://www.reuters.com/article/us-authentec-acquisition-apple-idUSBRE86Q0KD20120727

[35] A. J. Mansfield and J. L. Wayman, 'Best practices in Testing and Reporting Performance of Biometric Devices', National Physical Laboratory, Teddington, Report/Guide (NPL Report) CMSC 14/02, Aug. 2002. Accessed: Feb. 28, 2021. [Online]. Available: https://eprintspublications.npl.co.uk/2460/

[36] 'Information technology — Biometric performance testing and reporting — Part 2: Testing methodologies for technology and scenario evaluation', International Organization for Standardization, Geneva, Standard ISO/IEC 19795-2:2007, Feb. 2007.

[37] 'Information technology — Biometric performance testing and reporting — Part 9: Testing on mobile devices', International Organization for Standardization, Geneva, Standard ISO/IEC TS 19795-9:2019, Dec. 2019.

[38] 'Information technology — Biometrics used with mobile devices', International Organization for Standardization, Geneva, Standard ISO/IEC TR 30125:2016, 2016.

[39] S. Schuckers, G. Cannon, and N. Tekampe, Eds., 'FIDO Biometrics Requirements'. FIDO Alliance, Dec. 06, 2021. Accessed: Feb. 08, 2022. [Online]. Available: https://fidoalliance.org/specs/biometric/requirements/

[40] FIDO Alliance, 'FIDO Alliance Overview - Changing the Nature of Authentication', *FIDO Alliance*. https://fidoalliance.org/overview/ (accessed Apr. 22, 2021).

[41] 'Information technology — Biometric presentation attack detection — Part 3: Testing and reporting', International Organization for Standardization, Geneva, Standard ISO/IEC 30107-3:2017, Sep. 2017.

[42] N. Whiskerd, J. Dittmann, and C. Vielhauer, 'A Requirement Analysis for Privacy Preserving Biometrics in View of Universal Human Rights and Data Protection Regulation', in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 548–552. doi: 10.23919/EUSIPCO.2018.8553045.

[43] V. M. Patel, N. K. Ratha, and R. Chellappa, 'Cancelable Biometrics: A review', *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 54–65, Sep. 2015, doi: 10.1109/MSP.2015.2434151.

[44] A. K. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*. Springer Science & Business Media, 2006.

[45] M. Boakes, R. Guest, F. Deravi, and B. Corsetti, 'Exploring Mobile Biometric Performance Through Identification of Core Factors and Relationships', *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 4, pp. 278–291, Oct. 2019, doi: 10.1109/TBIOM.2019.2941728.

[46] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn, 'Pupil dilation degrades iris biometric performance', *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 150–157, 2009, doi: 10.1016/j.cviu.2008.08.001.

[47] Y. Adini, Y. Moses, and S. Ullman, 'Face recognition: The problem of compensating for changes in illumination direction', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, 1997, doi: 10.1109/34.598229.

[48] R. Cappelli, M. Ferrara, and D. Maltoni, 'The Quality of Fingerprint Scanners and Its Impact on the Accuracy of Fingerprint Recognition Algorithms', in *Multimedia Content Representation, Classification and Security*, Berlin, Heidelberg, 2006, pp. 10–16. doi: 10.1007/11848035_3.

[49] 'FACTOR | Definition of FACTOR by Oxford Dictionary on Lexico.com also meaning of FACTOR', *Lexico Dictionaries | English*. https://www.lexico.com/definition/factor (accessed Jun. 14, 2021).

[50] 'San Francisco is first US city to ban facial recognition', *BBC News*, May 14, 2019. Accessed: Jun. 14, 2021. [Online]. Available: https://www.bbc.com/news/technology-48276660

[51] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, 'Quality Measures in Biometric Systems', *IEEE Security Privacy*, vol. 10, no. 6, pp. 52–62, Nov. 2012, doi: 10.1109/MSP.2011.178.

[52] A. Jain, K. Nandakumar, and A. Ross, 'Score normalization in multimodal biometric systems', *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005, doi: 10.1016/j.patcog.2005.01.012.

[53] M. He *et al.*, 'Performance evaluation of score level fusion in multimodal biometric systems', *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010, doi: 10.1016/j.patcog.2009.11.018.

[54] D. Gafurov, 'A survey of biometric gait recognition: Approaches, security and challenges', in *Annual Norwegian computer science conference*, 2007, pp. 19–21.

[55] K. Ito and T. Aoki, 'Recent Advances in Biometric Systems', vol. 6, no. 1, pp. 64–80, 2018, doi: 10.3169/mta.6.64.

[56] P. Yan and K. W. Bowyer, 'Biometric Recognition Using 3D Ear Shape', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1297–1308, Aug. 2007, doi: 10.1109/TPAMI.2007.1067.

[57] R. Blanco-Gonzalo, R. Sanchez-Reillo, L. Martinez-Normand, B. Fernandez-Saavedra, and J. Liu-Jimenez, 'Accessible Mobile Biometrics for Elderly', in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15*, Lisbon, Portugal, 2015, pp. 419–420. doi: 10.1145/2700648.2811332.

[58] R. Sanchez-Reillo, R. Blanco-Gonzalo, J. Liu-Jimenez, M. Lopez, and E. Canto, 'Universal access through biometrics in mobile scenarios', in *2013 47th International Carnahan Conference on Security Technology (ICCST)*, Oct. 2013, pp. 1–6. doi: 10.1109/CCST.2013.6922051.

[59] S. J. Elliott, E. P. Kukula, and N. C. Sickler, 'The Challenges of the Environment and the Human / Biometric Device Interaction on Biometric System Performance', 2004. doi: 10.1.1.133.910.

[60] C. Lunerti, R. M. Guest, R. Blanco-Gonzalo, R. Sanchez-Reillo, and J. Baker, 'Environmental effects on face recognition in smartphones', in *2017 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2017, pp. 1–6. doi: 10.1109/CCST.2017.8167825.

[61] Y. Gong, 'Speech recognition in noisy environments: A survey', *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995, doi: 10.1016/0167-6393(94)00059-J.

[62] T. Yamada, M. Kumakura, and N. Kitawaki, 'Performance Estimation of Speech Recognition System Under Noise Conditions Using Objective Quality Measures and Artificial Voice', *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006, doi: 10.1109/TASL.2006.883254.

[63] R. Guest, M. Brockly, S. Elliott, and J. Scott, 'An assessment of the usability of biometric signature systems using the human-biometric sensor interaction model', *IJCAT*, vol. 53, no. 4, pp. 336–347, 2016, doi: 10.1504/IJCAT.2016.076810.

[64] S. Elliott, M. Mershon, V. Chandrasekaran, and S. Gupta, 'The evolution of the HBSI model with the convergence of performance methodologies', in *2011 Carnahan Conference on Security Technology*, Oct. 2011, pp. 1–4. doi: 10.1109/CCST.2011.6095938.

[65] M. Brockly, R. Guest, S. Elliott, and J. Scott, 'Dynamic signature verification and the human biometric sensor interaction model', in *2011 Carnahan Conference on Security Technology*, Oct. 2011, pp. 1–6. doi: 10.1109/CCST.2011.6095937.

[66] Z. Sitová *et al.*, 'HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users', *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 877–892, May 2016, doi: 10.1109/TIFS.2015.2506542.

[67] K. Nguyen, C. Fookes, R. Jillela, S. Sridharan, and A. Ross, 'Long range iris recognition: A survey', *Pattern Recognition*, vol. 72, pp. 123–143, Dec. 2017, doi: 10.1016/j.patcog.2017.05.021.

[68] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, 'Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation', National Inst of Standards and Technology Gaithersburg Md, 1998.

[69] N. Yager and T. Dunstone, 'Worms, Chameleons, Phantoms and Doves: New Additions to the Biometric Menagerie', in *2007 IEEE Workshop on Automatic Identification Advanced Technologies*, Jun. 2007, pp. 1–6. doi: 10.1109/AUTOID.2007.380583.

[70] N. Popescu-Bodorin, V. E. Balas, and I. M. Motoc, 'The Biometric Menagerie – A Fuzzy and Inconsistent Concept', in *Soft Computing Applications*, vol. 195, V. E. Balas, J. Fodor, A. R. Várkonyi-Kóczy, J. Dombi, and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 27–43. doi: 10.1007/978-3-642-33941-7_6.

[71] R. Blanco-Gonzalo, N. Poh, R. Wong, and R. Sanchez-Reillo, 'Time evolution of face recognition in accessible scenarios', *Hum. Cent. Comput. Inf. Sci.*, vol. 5, no. 1, p. 24, Dec. 2015, doi: 10.1186/s13673-015-0043-0.

[72] M. El-Abed, R. Giot, B. Hemery, and C. Rosenberger, 'A study of users' acceptance and satisfaction of biometric systems', in *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, Oct. 2010, pp. 170–178. doi: 10.1109/CCST.2010.5678678.

[73] C. Riley, K. Buckner, G. Johnson, and D. Benyon, 'Culture & biometrics: regional differences in the perception of biometric authentication technologies', *AI & Soc*, vol. 24, no. 3, pp. 295–306, Oct. 2009, doi: 10.1007/s00146-009-0218-1.

[74] X. Qu, D. Zhang, G. Lu, and Z. Guo, 'Door Knob Hand Recognition System', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 11, pp. 2870–2881, Nov. 2017, doi: 10.1109/TSMC.2016.2531675.

[75] O. Miguel-Hurtado, R. Blanco-Gonzalo, R. Guest, and C. Lunerti, 'Interaction evaluation of a mobile voice authentication system', in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, Oct. 2016, pp. 1–8. doi: 10.1109/CCST.2016.7815697.

[76] M. A. Sasse, 'Red-Eye Blink, Bendy Shuffle, and the Yuck Factor: A User Experience of Biometric Airport Systems', *IEEE Security Privacy*, vol. 5, no. 3, pp. 78–81, May 2007, doi: 10.1109/MSP.2007.69.

[77] T. Feng *et al.*, 'Continuous mobile authentication using touchscreen gestures', in *2012 IEEE Conference on Technologies for Homeland Security (HST)*, Nov. 2012, pp. 451–456. doi: 10.1109/THS.2012.6459891.

[78] D. Buschek, A. De Luca, and F. 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*, Seoul Republic of Korea, Apr. 2015, pp. 1393–1402. doi: 10.1145/2702123.2702252.

[79] L. Jain, J. V. Monaco, M. J. Coakley, and C. C. Tappert, 'Passcode keystroke biometric performance on smartphone touchscreens is superior to that on hardware keyboards', *International Journal of Research in Computer Applications & Information Technology*, vol. 2, no. 4, pp. 29–33, 2014.

[80] J. Chakrabarthy, 'Fingerprint Scanner On Phones: History & Evolution | IGadgetsworld', Apr. 17, 2016. https://www.igadgetsworld.com/fingerprint-scanner-history-evolution-but-do-we-really-need-that/ (accessed Jul. 10, 2021).

[81] Apple, 'About Touch ID advanced security technology', *Apple Support*, Sep. 11, 2017. https://support.apple.com/en-gb/HT204587 (accessed Jul. 10, 2021).

[82] D. D. Hwang and I. Verbauwhede, 'Design of portable biometric authenticators - energy, performance, and security tradeoffs', *IEEE Transactions on Consumer Electronics*, vol. 50, no. 4, pp. 1222–1231, Nov. 2004, doi: 10.1109/TCE.2004.1362523.

[83] R. R. Schaller, 'Moore's law: past, present and future', *IEEE Spectrum*, vol. 34, no. 6, pp. 52–59, Jun. 1997, doi: 10.1109/6.591665.

[84] E. Cantó-Navarro, M. López-García, and R. Ramos-Lara, 'Floating-point accelerator for biometric recognition on FPGA embedded systems', *Journal of Parallel and Distributed Computing*, vol. 112, pp. 20–34, Feb. 2018, doi: 10.1016/j.jpdc.2017.09.010.

[85] P. Hu, H. Ning, T. Qiu, Y. Xu, X. Luo, and A. K. Sangaiah, 'A unified face identification and resolution scheme using cloud computing in Internet of Things', *Future Generation Computer Systems*, vol. 81, pp. 582–592, Apr. 2018, doi: 10.1016/j.future.2017.03.030.

[86] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification', in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220.

[87] N. Ortiz, R. D. Hernandez, R. Jimenez, M. Mauledeoux, and O. Aviles, 'Survey of biometric pattern recognition via machine learning techniques', *ces*, vol. 11, no. 34, pp. 1677–1694, 2018, doi: 10.12988/ces.2018.84166.

[88] 'Definitions, Meanings, Synonyms, and Grammar by Oxford Dictionary on Lexico.com', *Lexico Dictionaries | English*. https://www.lexico.com/ (accessed Jul. 10, 2021).

[89] 'Information technology — Biometric sample quality — Part 1: Framework', International Organization for Standardization, Geneva, Standard ISO/IEC 29794-1:2016, Jan. 2016.

[90] C. G. Allen and S. Komandur, 'The Relationship Between Usability and Biometric Authentication in Mobile Phones', in *HCI International 2019 - Posters*, C. Stephanidis, Ed. Cham: Springer International Publishing, 2019, pp. 183–189.

[91] P. Jorgensen, *Software testing: a craftsman's approach*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2014.

[92] T. Wescott, 'The Discipline of System Design', in *Developing and Managing Embedded Systems and Products*, Elsevier, 2015, pp. 235–328. doi: 10.1016/B978-0-12-405879-8.00008-8.

[93] Google, 'Measuring Biometric Unlock Security', *Android Open Source Project*, Mar. 05, 2021. https://source.android.com/security/biometric/measure (accessed Apr. 26, 2021).

[94] W. An *et al.*, 'Performance Evaluation of Model-Based Gait on Multi-View Very Large Population Database With Pose Sequences', *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 421–430, Oct. 2020, doi: 10.1109/TBIOM.2020.3008862.

[95] 'Biometric performance estimation methodologies using statistical model', International Organization for Standardization, Geneva, Standard ISO/IEC WD 5152.

[96] S. K. Modi and S. J. Elliott, 'Impact of image quality on performance: Comparison of young and elderly fingerprints', in *The 6th International Conference on Recent Advances in Soft Computing (rasc*, 2006, pp. 10–12.

[97] 'Information technology — Biometric performance testing and reporting — Part 3: Modality-specific testing', International Organization for Standardization, Geneva, Standard ISO/IEC TR 19795-3:2007, Dec. 2007.

[98] M. Boakes, R. Guest, F. Deravi, and B. Corsetti, 'Exploring Mobile Biometric Performance Through Identification of Core Factors and Relationships', *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 1, no. 4, pp. 278–291, Oct. 2019, doi: 10.1109/TBIOM.2019.2941728.

[99] S. Schuckers, G. Cannon, N. Tekampe, E. Tabassi, M. Karlsson, and E. Newton, 'FIDO Biometrics Requirements', FIDO Alliance, Document, Oct. 2020. [Online]. Available: https://fidoalliance.org/specs/biometric/requirements/

[100] M. Gomez-Barrero and J. Galbally, 'Reversing the irreversible: A survey on inverse biometrics', *Computers & Security*, vol. 90, p. 101700, Mar. 2020, doi: 10.1016/j.cose.2019.101700.

[101] Manisha and N. Kumar, 'Cancelable Biometrics: a comprehensive survey', *Artif Intell Rev*, vol. 53, no. 5, pp. 3403–3446, Jun. 2020, doi: 10.1007/s10462-019-09767-8.

[102] A. K. Singh, P. Joshi, and G. C. Nandi, 'Face recognition with liveness detection using eye and mouth movement', in *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014)*, Jul. 2014, pp. 592–597. doi: 10.1109/ICSPCT.2014.6884911.

[103] European Union, 'General Data Protection Regulation', *Official Journal of the European Union*, vol. 59, no. L 119, pp. 1–88, May 2016.

[104] O. Radley-Gardner, H. Beale, and R. Zimmermann, Eds., *Fundamental Texts On European Private Law*. Hart Publishing, 2016. doi: 10.5040/9781782258674.

[105] A. Cavoukian and A. Stoianov, 'Biometric Encryption', in *Encyclopedia of Cryptography and Security*, H. C. A. van Tilborg and S. Jajodia, Eds. Boston, MA: Springer US, 2011, pp. 90–98. doi: 10.1007/978-1-4419-5906-5_880.

[106] 'Information technology — Security techniques — Biometric information protection', International Organization for Standardization, Geneva, Standard ISO/IEC 24745:2011, Jun. 2011.

[107] 'Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts', International Organization for Standardization, Geneva, Standard ISO 9241-11:2018, Mar. 2018.

[108] M. Brockly, S. Elliott, R. Guest, and R. B. Gonzalo, 'Human-Biometric Sensor Interaction', in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Boston, MA: Springer US, 2014, pp. 1–9. doi: 10.1007/978-3-642-27733-7_2261-3.

[109] S. J. Elliott and E. P. Kukula, 'A definitional framework for the human/biometric sensor interaction model', in *Biometric Technology for Human Identification VII*, Apr. 2010, vol. 7667, p. 76670H. doi: 10.1117/12.850595.

[110] J. J. Howard and D. Etter, 'The effect of ethnicity, gender, eye color and wavelength on the biometric menagerie', in *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, Nov. 2013, pp. 627–632. doi: 10.1109/THS.2013.6699077.

[111] H. E. Khiyari and H. Wechsler, 'Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning', *J Biom Biostat*, vol. 07, no. 04, 2016, doi: 10.4172/2155-6180.1000323.

[112] C. Rathgeb *et al.*, 'Impact of Doppelgängers on Face Recognition: Database and Evaluation', in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2021, pp. 1–4. doi: 10.1109/BIOSIG52210.2021.9548306.

[113] A. Popescu, L.-D. Ștefan, J. Deshayes-Chossart, and B. Ionescu, 'Face Verification With Challenging Imposters and Diversified Demographics', 2022, pp. 3357–3366. Accessed: Mar. 19, 2022. [Online]. Available: https://openaccess.thecvf.com/content/WACV2022/html/Popescu_Face_Verification_With_Challenging_Imposters_and_Diversified_Demographics_WACV_2022_paper.html

[114] ‘Information technology — A study of the differential impact of demographic factors in biometric recognition system performance’, International Organization for Standardization, Geneva, Standard ISO/IEC TR 22116:2021, Jun. 2021.

[115] ‘Information technology — Guidance for specifying performance requirements to meet security and usability needs in applications using biometrics’, International Organization for Standardization, Geneva, Standard ISO/IEC TR 29156:2015, Nov. 2015.

[116] ‘Information technology — Evaluation methodology for environmental influence in biometric system performance’, International Organization for Standardization, Geneva, Standard ISO/IEC 29197:2015, Apr. 2015.

[117] H.-W. Ng and S. Winkler, ‘A data-driven approach to cleaning large face datasets’, in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 343–347. doi: 10.1109/ICIP.2014.7025068.

[118] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, ‘Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments’, presented at the Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition, Oct. 2008. Accessed: Oct. 25, 2021. [Online]. Available: https://hal.inria.fr/inria-00321923

[119] M. M. Kalayeh, M. Seifu, W. LaLanne, and M. Shah, ‘How to Take a Good Selfie?’, in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane Australia, Oct. 2015, pp. 923–926. doi: 10.1145/2733373.2806365.

[120] StatCounter, ‘StatCounter’, *Market share of leading mobile device vendors in the United Kingdom (UK) from 2010 to 2020. Statista. Statista Inc.*, 2021. https://www.statista.com/statistics/487780/market-share-of-mobile-device-vendors-uk/ (accessed Sep. 02, 2021).

[121] ‘IriShield™ Series’, *Iris Scanner | Iris Biometrics Technology | Iris Recognition*. https://www.iritech.com/products/hardware/irishield%E2%84%A2-series (accessed Oct. 12, 2021).

[122] ‘BiometricPrompt | Android Developers’. https://developer.android.com/reference/androidx/biometric/BiometricPrompt (accessed Sep. 14, 2021).

[123] ‘LAPolicy.deviceOwnerAuthenticationWithBiometrics | Apple Developer Documentation’. https://developer.apple.com/documentation/localauthentication/lapolicy/deviceownerauthenticationwithbiometrics (accessed Sep. 14, 2021).

[124] ‘Motion sensors | Android Developers’. https://developer.android.com/guide/topics/sensors/sensors_motion (accessed Oct. 05, 2021).

[125] ‘Position sensors’, *Android Developers*. https://developer.android.com/guide/topics/sensors/sensors_position (accessed Oct. 05, 2021).

[126] ‘Environment sensors | Android Developers’. https://developer.android.com/guide/topics/sensors/sensors_environment (accessed Oct. 05, 2021).

[127] ‘LAError.Code | Apple Developer Documentation’. https://developer.apple.com/documentation/localauthentication/laerror/code (accessed Jan. 24, 2022).

[128] M. Terpilowski, ‘scikit-posthocs: Pairwise multiple comparison tests in Python’, *JOSS*, vol. 4, no. 36, p. 1169, Apr. 2019, doi: 10.21105/joss.01169.

[129] M. Theofanos, B. Stanton, R. Micheals, and S. Orandi, ‘Biometric Systematic Uncertainty and the User’, in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, Sep. 2007, pp. 1–6. doi: 10.1109/BTAS.2007.4401918.

[130] E. P. Kukula, S. J. Elliott, and V. G. Duffy, ‘The Effects of Human Interaction on Biometric System Performance’, in *Digital Human Modeling*, Berlin, Heidelberg, 2007, pp. 904–914. doi: 10.1007/978-3-540-73321-8_102.

[131] C. Lunerti, ‘Facial Biometrics on Mobile Devices: Interaction and Quality Assessment’, phd, University of Kent, 2019. Accessed: Mar. 01, 2022. [Online]. Available: https://kar.kent.ac.uk/77501/

[132] A. Geitgey, ‘Face Recognition’. Mar. 01, 2022. Accessed: Mar. 01, 2022. [Online]. Available: https://github.com/ageitgey/face_recognition

[133] A. Geitgey, ‘Machine Learning is Fun! Part 4: Modern Face Recognition with Deep Learning’, *Medium*, Sep. 24, 2020. https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78 (accessed Jan. 17, 2022).

[134] D. E. King, ‘Dlib-ml: A Machine Learning Toolkit’, *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[135] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, ‘Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks’, *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[136] I. de P. Centeno, ‘ipazc/mtcnn’. Feb. 28, 2022. Accessed: Feb. 28, 2022. [Online]. Available: https://github.com/ipazc/mtcnn

[137] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, ‘RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild’, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 5202–5211. doi: 10.1109/CVPR42600.2020.00525.

[138] S. I. Serengil and A. Ozpinar, ‘HyperExtended LightFace: A Facial Attribute Analysis Framework’, in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, Oct. 2021, pp. 1–4. doi: 10.1109/ICEET53442.2021.9659697.

[139] S. I. Serengil, ‘RetinaFace’. Feb. 28, 2022. Accessed: Feb. 28, 2022. [Online]. Available: https://github.com/serengil/retinaface

[140] C. Rathgeb, A. Uhl, P. Wild, and H. Hofbauer, ‘Design Decisions for an Iris Recognition SDK’, in *Handbook of Iris Recognition*, Second edition., K. Bowyer and M. J. Burge, Eds. Springer, 2016.

[141] D. M. Monro, S. Rakshit, and D. Zhang, ‘DCT-Based Iris Recognition’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 586–595, Apr. 2007, doi: 10.1109/TPAMI.2007.1002.

[142] C. Li *et al.*, ‘Deep Speaker: an End-to-End Neural Speaker Embedding System’, *arXiv:1705.02304 [cs]*, May 2017, Accessed: Nov. 30, 2021. [Online]. Available: http://arxiv.org/abs/1705.02304

[143] P. Rémy, 'philipperemy/deep-speaker'. Feb. 10, 2022. Accessed: Feb. 13, 2022. [Online]. Available: https://github.com/philipperemy/deep-speaker

[144] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 'Librispeech: An ASR corpus based on public domain audio books', in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[145] 'FFmpeg'. https://ffmpeg.org/ (accessed Mar. 02, 2022).

[146] E. P. Kukula, M. J. Sutton, and S. J. Elliott, 'The Human–Biometric-Sensor Interaction Evaluation Method: Biometric Performance and Usability Measurements', *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 784–791, Apr. 2010, doi: 10.1109/TIM.2009.2037878.

[147] M. Y.-S. Yao, S. Pankanti, and N. Haas, 'Fingerprint Quality Assessment', in *Automatic Fingerprint Recognition Systems*, N. Ratha and R. Bolle, Eds. New York: Springer-Verlag, 2004, pp. 55–66. doi: 10.1007/0-387-21685-5_3.

[148] E. Tabassi *et al.*, 'NFIQ 2 NIST Fingerprint Image Quality', National Institute of Standards and Technology, Jul. 2021. doi: 10.6028/NIST.IR.8382.

[149] P. Grother, M. Ngan, and K. Hanaoka, 'Face Recognition Vendor Test - General Evaluation Specifications', National Institute of Standards and Technology, Sep. 2020.

[150] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, 'FaceQnet: Quality Assessment for Face Recognition based on Deep Learning', in *2019 International Conference on Biometrics (ICB)*, Jun. 2019, pp. 1–8. doi: 10.1109/ICB45273.2019.8987255.

[151] 'IriCore', *Iris Scanner | Iris Biometrics Technology | Iris Recognition*. https://www.iritech.com/products/software/iricore-eye-recognition-software (accessed Mar. 02, 2022).

[152] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, 'Biometric Quality: Review and Application to Face Recognition with FaceQnet', *arXiv:2006.03298 [cs]*, Feb. 2021, Accessed: Jul. 20, 2021. [Online]. Available: http://arxiv.org/abs/2006.03298

[153] BBC, 'BBC Weather', *BBC Weather*. https://www.bbc.co.uk/weather/ (accessed Aug. 08, 2021).

[154] MeteoGroup, 'Weather', *DTN*. https://www.dtn.com/weather/ (accessed Aug. 08, 2021).

[155] M. Boakes, R. Guest, and F. Deravi, 'Adapting to Movement Patterns for Face Recognition on Mobile Devices', in *Pattern Recognition. ICPR International Workshops and Challenges*, vol. 12668, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 209–228. doi: 10.1007/978-3-030-68793-9_15.

[156] 'Facial recognition: EU considers ban of up to five years', *BBC News*, Jan. 17, 2020. Accessed: Oct. 19, 2021. [Online]. Available: https://www.bbc.com/news/technology-51148501

[157] 'MEPs back EU facial-recognition ban for police', *EUobserver*. https://euobserver.com/democracy/153135 (accessed Oct. 19, 2021).

[158] A. Goode, 'Digital identity: solving the problem of trust', *Biometric Technology Today*, vol. 2019, no. 10, pp. 5–8, Nov. 2019, doi: 10.1016/S0969-4765(19)30142-0.

[159] P. H. Pisani *et al.*, 'Adaptive Biometric Systems: Review and Perspectives', *ACM Comput. Surv.*, vol. 52, no. 5, p. 102:1-102:38, Sep. 2019, doi: 10.1145/3344255.

[160] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, 'Continuous User Authentication on Mobile Devices: Recent progress and remaining challenges', *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 49–61, Jul. 2016, doi: 10.1109/MSP.2016.2555335.

[161] E. Vasiete *et al.*, 'Toward a non-intrusive, physio- behavioral biometric for smartphones', in *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, New York, NY, USA, Sep. 2014, pp. 501–506. doi: 10.1145/2628363.2634223.

[162] R. Kumar, V. V. Phoha, and A. Serwadda, 'Continuous authentication of smartphone users by fusing typing, swiping, and phone movement patterns', in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sep. 2016, pp. 1–8. doi: 10.1109/BTAS.2016.7791164.

[163] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng, 'Seeing Your Face Is Not Enough: An Inertial Sensor-Based Liveness Detection for Face Authentication', in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, Oct. 2015, pp. 1558–1569. doi: 10.1145/2810103.2813612.

[164] S. Chen, A. Pande, and P. Mohapatra, 'Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones', in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, New York, NY, USA, Jun. 2014, pp. 109–122. doi: 10.1145/2594368.2594373.

[165] N. Poh, R. Wong, J. Kittler, and F. Roli, 'Challenges and Research Directions for Adaptive Biometric Recognition Systems', in *Advances in Biometrics*, Berlin, Heidelberg, 2009, pp. 753–764. doi: 10.1007/978-3-642-01793-3_77.

[166] S. Gutta, M. Trajkovic, and V. Philomin, 'System and method for adaptively setting biometric measurement thresholds', US20070003110A1, Jan. 04, 2007 Accessed: Jan. 17, 2022. [Online]. Available: https://patents.google.com/patent/US20070003110A1/en

[167] C. B. Brumback, D. W. Knight, J. D. M. Messenger, and J. O. Hong, 'Biometric sensing device having adaptive data threshold, a performance goal, and a goal celebration display', US8734296B1, May 27, 2014 Accessed: Jan. 17, 2022. [Online]. Available: https://patents.google.com/patent/US8734296B1/en

[168] E. Castilllo-Guerra, R. Díaz-Amador, and C. L. Julian, 'Adaptive threshold estimation for speaker verification systems', *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3877–3877, May 2008, doi: 10.1121/1.2935778.

[169] A. Mhenni, E. Cherrier, C. Rosenberger, and N. E. B. Amara, 'Adaptive Biometric Strategy using Doddington Zoo Classification of User's Keystroke Dynamics', in *2018 14th International Wireless Communications Mobile Computing Conference (IWCMC)*, Jun. 2018, pp. 488–493. doi: 10.1109/IWCMC.2018.8450401.

[170] C. Lunerti, R. Guest, J. Baker, P. Fernandez-Lopez, and R. Sanchez-Reillo, 'Sensing Movement on Smartphone Devices to Assess User Interaction for Face Verification', in *2018 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2018, pp. 1–5. doi: 10.1109/CCST.2018.8585547.

[171] 'BiometricPrompt', *Android Developers*. https://developer.android.com/reference/androidx/biometric/BiometricPrompt (accessed Jan. 17, 2022).

[172] 'SensorManager', *Android Developers*. https://developer.android.com/reference/android/hardware/SensorManager (accessed Jan. 17, 2022).

[173] A. Geitgey, *face-recognition: Recognize faces from Python or from the command line*. Accessed: Jan. 17, 2022. [Online]. Available: https://github.com/ageitgey/face_recognition

[174] I. de P. Centeno, 'mtcnn: Multi-task Cascaded Convolutional Neural Networks for Face Detection, based on TensorFlow'. Accessed: Jan. 17, 2022. [Online]. Available: http://github.com/ipazc/mtcnn

[175] 'scikit-learn: A set of python modules for machine learning and data mining'. Accessed: Jan. 17, 2022. [MacOS, Microsoft :: Windows, POSIX, Unix]. Available: http://scikit-learn.org

[176] F. Pedregosa *et al.*, 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[177] 'The Future of Finance: Mobile Biometric Banking - FindBiometrics'. https://findbiometrics.com/the-future-of-finance-mobile-biometric-banking-906119/ (accessed Jul. 26, 2020).

# Appendix A: Questionnaire

**A Performance Assessment Framework for Mobile Biometrics**

**Pre-Experiment Questionnaire**

| **Participant ID:** | (Research Use Only) |
|---|---|

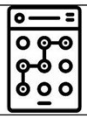
**Personal Information**

| How old are you? | | | | | | | |
|---|---|---|---|---|---|---|---|
| 16-18 | 19-21 | 22-24 | 25-29 | 30-39 | 40-49 | 50-64 | 65+ |
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| What is your gender | Male: ☐ | Female: ☐ | Other: ☐ |
|---|---|---|---|
| What is your occupation? | | | |
| What is your nationality? | | | |
| Are you left or right handed? | Left: ☐ | Right: ☐ | |
| Do you primarily have to use glasses or contact lenses to view smartphone screens (leave blank if not)? |  Glasses: ☐ |  Contact: ☐ | |

**Device Information**

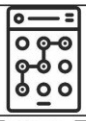
| Do you currently own a mobile phone? | Yes: ☐ | No: ☐ |
|---|---|---|
| What mobile phone do you currently own? | | |
| Do you currently operate a screen lock on your mobile phone? | Yes: ☐ | No: ☐ |

| What type of screen lock do you use primarily? | | | |
|---|---|---|---|
|  |  |  |  |
| Biometric: ☐ | Pattern: ☐ | Pin: ☐ | Other: ☐ |

If you use a biometric screen lock, please specify which you use?

|  |  |  |  |  |
|---|---|---|---|---|
| Fingerprint: ☐ | Face: ☐ | Voice: ☐ | Iris: ☐ | Other: ☐ |

If you operate a secondary (back-up) screen lock, what type do you use?

|  |  |  |  |
|---|---|---|---|
| Password: ☐ | Pattern: ☐ | Pin: ☐ | Other: ☐ |

How satisfied are you with the ease of use of your current phone lock?

| Very Dissatisfied | <= | == | == | == | => | Very Satisfied |
|---|---|---|---|---|---|---|
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

Observe a definition of 'reliable' for a biometric scenario and answer the following question.

*Reliable: The ability to accurately identify you promptly on a consistent basis.*

How reliable do you find biometrics as a form of authentication on smartphones?

| Very Unreliable | <= | == | == | == | => | Very Reliable |
|---|---|---|---|---|---|---|
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

**Post-Experiment Questionnaire**

**Experience**

Having now experienced a variety of different biometrics on smartphones please answer the following questions.

| How satisfied are you with the ease of use of **fingerprint** authentication on smartphones? | | | | | | |
|---|---|---|---|---|---|---|
| Very Dissatisfied | <= | == | == | == | => | Very Satisfied |
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

| How satisfied are you with the ease of use of **face** authentication on smartphones? | | | | | | |
|---|---|---|---|---|---|---|
| Very Dissatisfied | <= | == | == | == | => | Very Satisfied |
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

| How satisfied are you with the ease of use of **voice** authentication on smartphones? | | | | | | |
|---|---|---|---|---|---|---|
| Very Dissatisfied | <= | == | == | == | => | Very Satisfied |
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

| How satisfied are you with the ease of use of **iris** authentication on smartphones? | | | | | | |
|---|---|---|---|---|---|---|
| Very Dissatisfied | <= | == | == | == | => | Very Satisfied |
| 1: ○ | 2: ○ | 3: ○ | 4: ○ | 5: ○ | 6: ○ | 7: ○ |

| Which would be your preferred modality for authentication on a smartphone? | | | |
|---|---|---|---|
|  |  |  |  |
| Fingerprint: ☐ | Face: ☐ | Voice: ☐ | Iris: ☐ |

| If you were not previously using a biometric screen lock are you now considering using it? | Yes: ☐ | No: ☐ |
|---|---|---|

Observe a definition of 'reliable' for a biometric scenario and answer the following question.

*Reliable: The ability to accurately identify you promptly on a consistent basis.*

| How reliable do you find biometrics as a form of authentication on smartphones? | | | | | | |
|---|---|---|---|---|---|---|
| Very Unreliable | <= | == | == | == | => | Very Reliable |
| 1: O | 2: O | 3: O | 4: O | 5: O | 6: O | 7: O |

**Behavioural Biometrics**

| Are you familiar with continuous authentication? (*Definition - It utilises a user's behaviour to continuously verify identity throughout a session, not just at the entry login point. Such as swipe behaviour or movement.*). | | | | | | |
|---|---|---|---|---|---|---|
| Never Heard of It | <= | == | == | == | => | Very Familiar |
| 1: O | 2: O | 3: O | 4: O | 5: O | 6: O | 7: O |

| Do you feel that continuous authentication would be an invasion of your privacy? | | | | | | |
|---|---|---|---|---|---|---|
| Very Invasive | <= | == | == | == | => | Very Non-Invasive |
| 1: O | 2: O | 3: O | 4: O | 5: O | 6: O | 7: O |

# Appendix B: Data Collection Apps

Andoird App: https://bitbucket.org/MattyB95/biometric-data-collection-android/

iOS App: https://bitbucket.org/MattyB95/biometric-data-collection-ios/