Performance Evaluation Metrics for Biometrics-based Authentication Systems

---

Yuxuan Hong

May 14, 2021
Version: Final Version
Advisor: Rajesh Kumar

Abstract

Research has made many advances in many fields of biometric systems. However, there is a lack of proper research in the performance evaluation field for biometric authentication systems. In this work, we synthesize key ideas from recent publications to show how the landscape for performance evaluation of biometrics-based authentication systems has developed over time. We present the general architecture and fundamentals of an authentication system. We then show why most single-number evaluation metrics have inherent flaws and limited scopes. We synthesize and extract the four significant limitations of commonly used metrics. We also discuss how each can be avoided or addressed by using metrics like Gini Coefficient or the unnormalized frequency counts of scores (FCS) graph. Besides, we explore the limitations of metrics in the context of continuous authentication. We demonstrate how the combination of the Receiver Operating Characteristics (ROC) curve and the FCS graph can help us address the limitations of common metrics and provide a robust performance evaluation approach. From this, we present the state-of-the-art guidelines for evaluating authentication systems. In the end, we identify three critical problems with the current methodologies and propose specific future work plans to address them.

Acknowledgements

I would like to sincerely appreciate my thesis advisor, Professor Rajesh Kumar, for his continuous support, patience, and guidance throughout writing this thesis. I would also like to extend my gratitude to Professor Sorelle Friedler for her guidance and help in completing this thesis.

Contents

1

Introduction

1.1 Background

Identity recognition and management have always played an essential role in human society. Effective identity verification and management systems enable our society to operate by protecting areas, including banking security, database privacy, and various institutional information security.

Traditional knowledge-based and token-based identity verification relies on a password or ID card to represent a user's identity. However, user-chosen passwords can be easily forgotten, guessed, stolen, or shared, which all presents serious security concerns (Jain et al. 2011). Additionally, the traditional authentication system cannot detect multiple enrollments by the same person under different identities. It is becoming more evident that the conventional knowledge-based authentication system alone is not sufficient for reliable user verification and identity management (Jain et al. 2011).

Over the past few decades, with the development of the internet and web-based services, there is an exponential growth in the number of people, institutions, and businesses that store sensitive and essential data online. The proliferation of sensitive data, web-based services, and the development of decentralized services (e.g., online payment companies like PayPal) have led to the increased risk of data and identity theft (Jain et al. 2011). The traditional authentication system can no longer provide the security level required for today's and future's needs (Jain et al. 2011)—all of these call for a better approach to fill the gap. Biometric recognition, or biometrics, offers a more natural, reliable, and user-friendly solution to user authentication.

1.2 Motivation

Biometrics aims to uniquely identify and recognize users through the biological characteristics of the user. A biometric system measures one or more physical or behavioral traits. For instance, physical features include iris, fingerprint, face, and DNA, etc. The behavioral characteristics include gait (walking patterns), signature, or keystroke patterns. These characteristics are often referred to by other terms such as modalities, traits, or identifiers in the field (Jain et al. 2011).

There are usually two types of systems within biometrics: identification and authentication. Biometric identification can be viewed as an N-Classification problem since its goal is to determine the user's identity, where N denotes the number of users stored in the system's database. On the other hand, biometric authentication is a binary classification problem. It aims to verify whether the user is who they claim to be. This work focuses on the performance evaluation aspect of biometrics authentication systems.

Understanding the performance evaluation of biometrics authentication systems is essential for various reasons. First, to be deployed and used in real life, we need accurate and reliable evaluation metrics. Second, although researchers have made many advances in biometrics systems over the past few decades, there is a lack of proper research in the performance evaluation field. So far, there is no consistent approach for evaluating and comparing different biometrics systems nor for reporting metrics. Besides, as this paper will explain in detail later, many widely used single-number metrics used in the field have apparent flaws and limited scopes (Sugrim et al. 2019). Fourth, the landscape of biometrics-based authentication changes rapidly. Issues such as active adversaries and inherent biases, and fairness all call for more research in the field. Therefore, more work is needed in this field to address the challenges mentioned earlier. The robustness, the context of use, the strengths and weakness of important evaluation metrics, and other influential parameters must be discussed and defined. Having a better and clear understanding of performance evaluation is also necessary to facilitate research in the biometrics field.

The contributions of this thesis are as follows:

- We summarize and synthesize key findings from recent publications to show how the performance evaluation landscape for biometrics authentication systems has evolved.

- We present the strengths and weaknesses of all essential evaluation metrics for biometrics authentication systems.

- We demonstrate why commonly used metrics have inherent flaws and limitations and show the state-of-the-art approaches to address the limitations through reporting additional metrics.

- We present three problems of performance evaluation that are still present in the field.

- We propose four future work directions to solve the identified problems to advance research in this field.

The rest of the paper is organized as follows: Section 2 provides the full literature review to show how the landscape of performance evaluation for biometrics authentication system has developed over time. We introduce readers to the fundamentals of an authentication system and familiarize them with commonly used metrics in the field. We then illustrate and analyze the limitations of these metrics in the rest of section 2. By the end of the section, we provide solutions to the constraints and present the current state-of-the-art approach to performance evaluation and reporting. Section 3 identifies three problems that still impact the field and propose future work plans to solve them. We conclude the thesis in section 4.

2

Literature Review

This section will show how the landscape of performance evaluation for biometrics-based authentication systems has evolved. As of today, there are still only a few papers published that directly address this topic. Among them, we select several recent publications that have made significant contributions to the field's development and synthesize their findings to illustrate the development.

The work of Cherifi et al. 2010 is one of the earliest publications that lay the foundation in the field. They propose general guidelines for evaluating biometric systems.

Eberz et al. 2017 directly address the use of performance evaluation metrics in continuous authentication systems. Pointing out flaws in the traditional metrics, the work of Eberz et al. 2017 is one of the earliest papers that raises concerns over the common practice of blind-optimizing metrics without actually improving the security of a system. They point out that commonly reported metrics only capture the errors' mean, not the errors' distributions. They propose using the Gini Coefficient (GC) as an additional metric to effectively capture errors' distribution. Their article is also one of the first papers that quantifies some non-functional machine learning methodologies' impacts on a system's error rates.

Sugrim et al. 2019 further extend the work of Cherifi et al. 2010 and Eberz et al. 2017. They demonstrate that comparing different systems using the traditional single-point metrics such as FAR, FRR, and EER only presents a naive and incomplete comparison because they only focus on one aspect of a system. They also critique the Gini Coefficient's adaption proposed in Eberz et al. 2017 work as GC contains the inherent flaws of other metrics. Sugrim et al. 2019 proposed combining the ROC curve and the unnormalized frequency count of scores (FCS) as a better way to report and measure authentication systems' performance.

Before diving into the field's evolving landscape, we first introduce the essential biometrics and machine learning background information to familiarize readers with critical concepts.

2.1 Biometric Systems and Common Evaluation Metrics

A biometric authentication system commonly consists of two phases - the enrollment and verification phases. In the enrollment phase, a user's biometric data is acquired from a biometric sensor that collects the distinctive physical or behavioral characteristics of the user. The raw biometric measurements will then be preprocessed by filtering, re-centering, and scaling to extract distinguishable and salient features (Jain et al. 2011). Usually, only the extracted features and the corresponding user identity will be stored. At the same time, the raw biometric data will be discarded. In the verification phase, a new version of biometric data is re-collected from the user and compared against the stored template to decide if there is a match (Jain et al. 2011). The scoring operation will then apply an algorithm to the compared data to generate a score that measures the similarity between the two templates (Jain et al. 2011).

A threshold is pre-determined by the implementer to establish the basis for an authorization decision. If the score lies above the threshold, the user will be granted access. If the score is below the threshold, the user will be denied access. By convention, scores from authorized users are higher than scores from unauthorized users. During the training and testing phase of such a system, since the ground truth - the user identity - is known, we can measure an authentication system's performance, such as the proportion of correct decisions vs. the proportion of incorrect decisions. It is important to note that the scores between different authentication systems are not comparable. It is the performance metrics of the systems that allow us to compare. Fig 2.1 shows a diagram for the basic architecture of a biometric authentication system.

The purpose of verification is to determine if two templates being compared come from the same subject. Therefore, a biometric system's essence is a pattern recognition or matching problem (Jain et al. 2011). In particular, the authentication system is essentially a binary classification system. The further away the user's score is above or below the threshold, the more confident the system can be in its classification

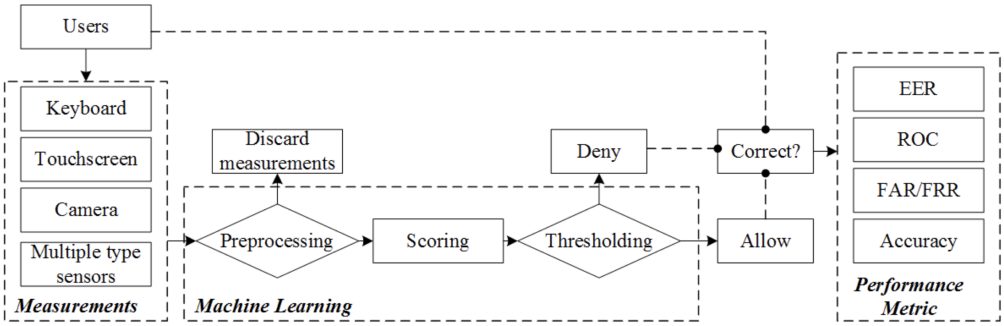

Fig. 2.1: A diagram showing the basic building blocks of a biometric authentication system using machine learning classifiers. The Enrollment phase at measurements collects biometric data from users. A template is initially collected from a user and stored in the database. A fresh sample is re-acquired every time a user attempts to access it. The preprocessing prepare the measurements by extracting key features from the data. The scoring operation applies a function that maps the feature measurements to a numerical score to compare the template of the claimed and true identity. A predefined threshold is then used to decide on the score. Some metrics can then be used to assess the quality of the system (Sugrim et al. 2019).

and decision. The better a system is at separating the unauthorized users' scores from the scores of authorized users, the better is the system's performance. Ideally, a perfect biometric authentication system can separate all unauthorized users' scores from authorized users' scores without any overlap between the two distributions. This is almost always impossible in reality. Realistically, the graph of the distribution for authorized and unauthorized users' scores overlaps (Jain et al. 2011). We use the term genuine attempt to refer to a single attempt made by authorized users and imposter attempt to refer to an attempt made by an imposter or unauthorized user (Dunstone and Yager 2008). Fig 2.2 shows the score distributions of genuine and imposter attempts for a face recognition system. The incorrect decision comes from some poor genuine attempts score lower than some of the serious imposter attempts.

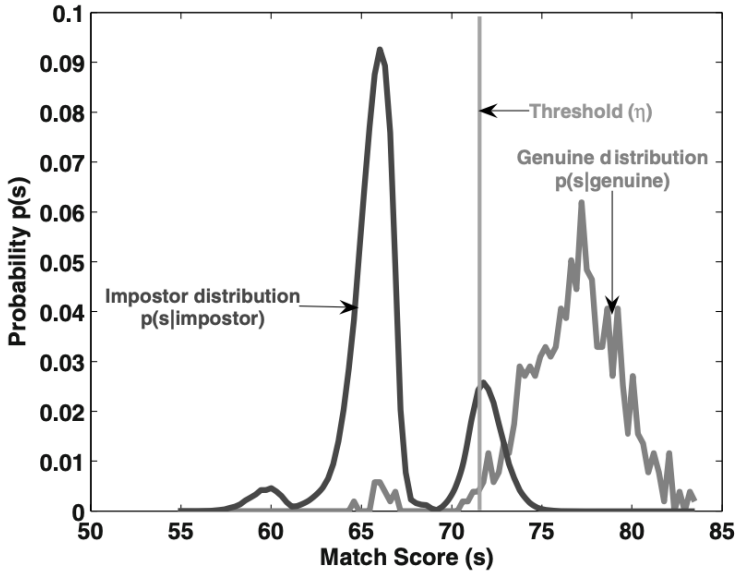

Fig. 2.2: An instance of genuine and imposter score distributions for a face recognition system. The match scores come from a Face-G matcher in the Biometric Score Set Release provided by the National Institute of Standards and Technology (NIST). The vertical line represents the threshold, T, which determines the FAR and FRR of the system. Note that given the distributions, FAR and FRR cannot be reduced simultaneously by adjusting the threshold. Moving the threshold line will increase one at the expense of decreasing the other, and vice versa (Jain et al. 2011).

Formally, we define the authentication as the following binary classification problem: given a claimed identity I, a new input feature set F^N, and a stored template F^S corresponding to I. We need to determine if (I, F^N) belongs to either "genuine" or "imposter" class. Let S be a scoring function that takes in three parameters I, F^N, F^S, and outputs a match or similarity score s. Let T be a pre-determined threshold. Thus, the decision rule can be written as:

$$S(I, F^N, F^S) = s \in \begin{cases} genuine & \text{if s} \geq \text{T} \\ imposter & \text{if s} < \text{T} \end{cases} \tag{2.1}$$

When the identity query is classified as genuine, the user's access to a service or a system is granted. Otherwise, access is denied. The positive class here is the genuine class, and the negative class is the imposter class. Hence, there are four possible outcomes for every decision an authentication system makes (Sugrim et al. 2019):

1. Authorize a legitimate user (true positive, TP)

2. Authorize an imposter (false positive, FP)

3. Deny an imposter (true negative, TN)

4. Deny a legitimate user (false negative, FN)

These decision counts form the Confusion Matrix (CM) for a machine learning-based biometrics authentication system. They are the basis for most other performance evaluation metrics. Fig 2.3 shows instance of such a Confusion Matrix.

	Measurement Source	
	Authorized (Positive)	Unauthorized (Negative)
Grant Access (Positive)	TP	FP
Deny Access (Negative)	FN	TN

Fig. 2.3: A Confusion Matrix in the context of an authentication system (Sugrim et al. 2019).

At its core, a biometric system can make two types of mistakes: false match and false non-match. The False Match Rate (FMR) refers to the proportion of imposter attempts that are falsely declared to match a template of another subject (Jain et al. 2011). The False Non-Match Rate (FNMR) refers to the proportion of genuine attempts falsely reported as non-match for a template of the same subject (Jain et al. 2011). In the context of biometric authentication, FNMR and FMR are generally referred to as False Reject Rate (FRR) and False Accept Rate (FAR), respectively (Jain et al. 2011). However, they are not the same. Jain et al. 2011 note that FMR and FNMR are equivalent to FAR and FRR, respectively, when the system uses one attempt by a user to match their stored template. Thus, in authentication systems, FAR and FRR are often used to replace FMR and FNMR.

The False Positive Rate (FPR) is the probability of authorizing an imposter. The False Negative Rate (FNR) is the probability of denying a legitimate user. FPR is often called the False Accept Rate (FAR), and FNR is called False Reject Rate (FRR) (Sugrim et al. 2019). These two are the fundamental errors an authentication system can make. The pair (FAR, FRR) is the core metric for evaluating an authentication system's performance. Based on the confusion matrix, FAR and FRR can be calculated (Sugrim et al. 2019):

$$FAR = FPR = \frac{FP}{FP + TN}$$

$$FRR = FNR = \frac{FN}{TP + FN}$$

More generally and formally, we let X_0 and X_1 denote the imposter and genuine classes. Note that s here still denotes match score. Let $P(s|X_0)$ and $P(s|X_1)$ be the probability density function of the imposter and genuine classes respectively. As Fig 2.2 shows, given such distribution, the general equations for the FAR and FRR of an authentication system are: (Jain et al. 2011)

$$FAR(T) = P(s \geq T|X_0) = \int_T^{\infty} P(s|X_0)ds$$

$$FRR(T) = P(s \leq T|X_0) = \int_{-\infty}^{T} P(s|X_1)ds$$

We note that FRR and FAR are both functions of the system threshold T. If the threshold T is decreased, FAR will increase, but FRR will decrease. Conversely, if T is increased, FAR will decrease, and FRR will increase (Jain et al. 2011). This shows the inherent trade-off property between FAR and FPR. That's, for a given system, it is impossible to decrease both errors simultaneously. Therefore, the selection of the threshold can be used to tune a biometric system. A lower threshold targets the system at user convenience since fewer genuine attempts will be rejected, at the expense of lower security as more imposter attempts will be authorized (Biometrics 2014). A higher threshold trades convenience for better security because fewer imposter attempts are authorized while simultaneously denying access to some genuine attempts.

There are generally two families of metrics in biometrics: Confusion Matrix derived metrics and ROC curve derived metrics (Sugrim et al. 2019). Confusion Matrix (CM) derived metrics depend on a specific threshold. In contrast, ROC curve derived metrics do not depend on the threshold because the ROC represents many CMs under varying thresholds (Sugrim et al. 2019).

The maximum accuracy (ACC) is an important CM derived metric for authentication systems. It reflects the relative frequency of correct classification of a measurement source (Sugrim et al. 2019). Since accuracy is a function of a threshold, we often see that the reported ACC is the maximum ACC across all thresholds (Sugrim et al. 2019). And the maximum ACC represents the best performance a system's classifier can provide. We note here that since ACC depends on a single threshold, it is incomplete to assess a system's performance by merely looking at the ACC value. We will discuss why ACC and other CM derived metrics can be misleading in section 2.4.2.

In general, the single threshold metrics offer an incomplete picture of a system's performance. It reveals nothing about how changes to threshold might affect the behavior of the metrics. Therefore, some insights into the relationship between the metrics and the threshold are needed.

A Receiver Operating Characteristics Curve (ROC) is a technique for visualizing and selecting classifiers based on their performance (Fawcett 2004). A ROC is computed by varying the authentication threshold from the maximum to the minimum possible values and calculating the TPR and FPR for each threshold (Sugrim et al. 2019). Fig 2.4 shows an example of a ROC curve (Sugrim et al. 2019). A ROC graph is a two-dimensional graph where the true positive rate is plotted on the Y-axis against the false positive rate on the X-axis (Fawcett 2004). A ROC graph depicts the relative trade-offs between the FPR and TPR for a classifier. As the threshold decreases, scores that were initially not high enough to grant access rise above the new threshold and would be granted the access. As the threshold lowers, the number of true positives and false positives will both increase. Each value of the threshold represents a certain trade-off between the benefits (TP) and cost (FP) (Fawcett 2004). The pair (TPR, FPR) is, thus, a parametric function of the threshold. A ROC curve is then plotted as this parametric curve varies across all possible thresholds. In the context of an authentication system, the FAR or FMR is plotted on the X-axis, and the true positive (1 - FRR) or Verification rate (1 - FNMR) is plotted on the Y-axis.

There are some areas on a ROC curve that represents particular interests. As Fig 2.4 shows, the upper left corner represents a perfect biometric system without errors (TPR=1, FPR=0), as it classifies all users and imposters correctly. In reality, an ideal biometric system is not achievable due to the inherent uncertainty of biometric authentication (Dunstone and Yager 2008). Nonetheless, all systems strive to obtain curves near this point. The closer a system's curve is to the point, the better is

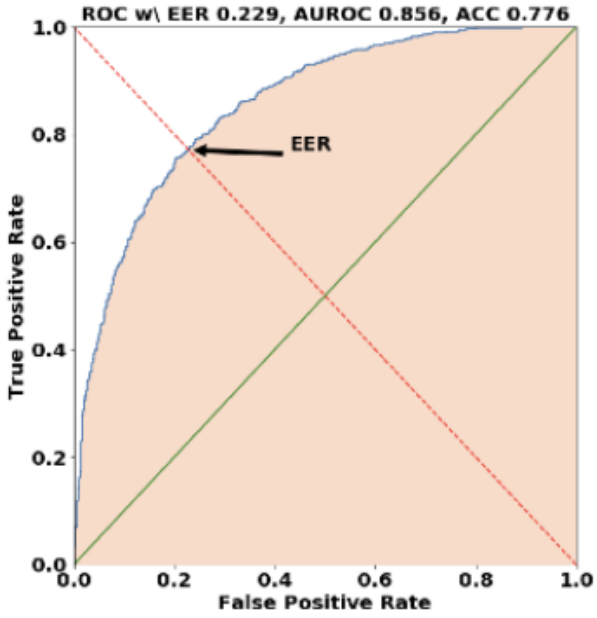

Fig. 2.4: A ROC curve of an authentication system (Sugrim et al. 2019).

the system's performance, which gives us a basis for comparing different systems (Dunstone and Yager 2008).

Another area of interest is the green diagonal line $y = x$, as shown in Fig 2.4. This diagonal line represents the outcome of randomly guessing a class (Fawcett 2004). For instance, if a biometric model randomly guesses the positive class 60% of the time, it is expected to obtain 60% of the positives correct. Still, its FPR will also increase to 60%, resulting in point (0.6,0.6) in the ROC space. Therefore, a random classifier will yield ROC points that slide back and forth on the diagonal line depending on how frequently it selects the positive class (Fawcett 2004). Hence, any good system's ROC curve should stay above the diagonal line.

The intersection of the red diagonal line $y = 1 - x$ and the ROC curve is called the Equal Error Rate (EER) (Dunstone and Yager 2008). The EER reflects the probability of making an incorrect positive or negative decision in equal probability (Sugrim et al. 2019). Only one specific threshold corresponds to an ERR on a ROC curve. The EER is one of the most commonly used metrics in evaluating biometric systems (Sugrim et al. 2019). As mentioned earlier in the architecture section, an authentication system outputs a score for each access. A predefined threshold is then used to classify a score as either imposter or genuine class. Thus, different

thresholds can greatly impact FAR and FRR. And the EER is the threshold where $FAR = FRR$.

One important note here is that EER can only select a threshold and evaluate training errors. The EER cannot be used to assess a system's performance on unknown data (Bengio et al. 2002). That's, EER cannot be used during testing. In testing, a similar metric, the half total error rate (HTER), should be used to evaluate a system's performance.

$$HTER = \frac{FAR + FRR}{2}$$

Another common ROC derived metric is the area under the ROC curve (AUROC). It is the shaded area in the ROC Fig 2.4. The AUROC represents the probability that a random unauthorized user's measurement is scored lower than a random authorized user's measurement (Sugrim et al. 2019). It can be interpreted as a measure of how well a classifier can separate an authorized user's measurements from their unauthorized counterparts (Sugrim et al. 2019).

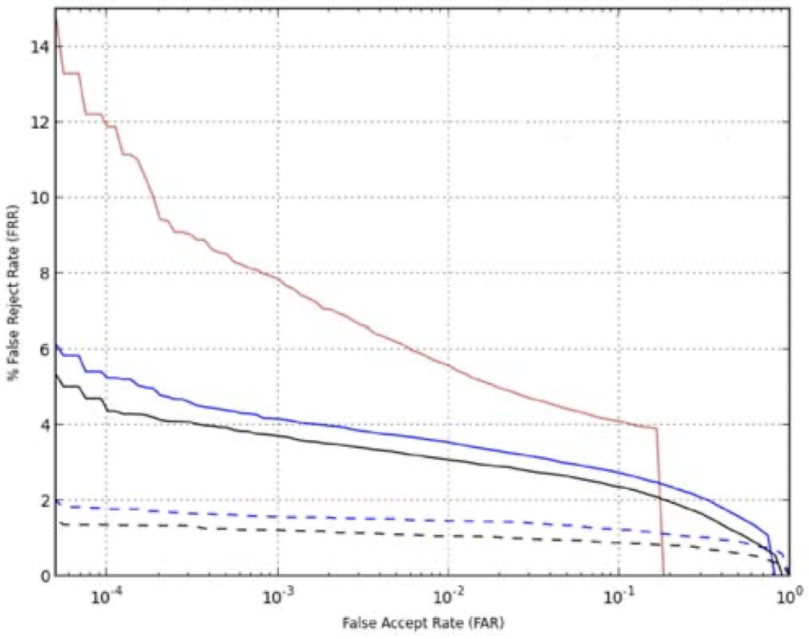

Fig. 2.5: An example of DET curves for a number of different biometric systems. (Biometrics 2014).

In addition to a ROC curve, the Detection Error Trade-off (DET) curve is often used in the field. Fig 2.5 shows an example of a DET graph. The DET curve is a variant of

the ROC curve. The main difference between a ROC and a DET curve is that in a DET curve, the Y-axis is the FNMR (or FRR) instead of TPR in ROC (Dunstone and Yager 2008). The X-axis of a DET curve is still the FMR (or FAR, or FPR), which is the same as that of a ROC curve. Like ROC, DET curves are usually plotted using logarithmic scale (Dunstone and Yager 2008). Since the Y-axis of a DET curve represents the frequency of match errors, the curve closest to the bottom of the plot corresponds to the best performance. We note that this is opposite to a ROC curve since the top left corner of a ROC space is usually the perfect system. As Fig 2.5 shows, the top-most red curve corresponds to the worst performance compared to other curves. The choice between a ROC and a DET curve is subjective and often merely aesthetic because they both capture the same information for biometric systems (Dunstone and Yager 2008).

In retrospect, we have shown the basic architecture of a biometric authentication system. We also offer all the essential and commonly used metrics present in almost all publications that discuss performance evaluations of their authentication systems (Sugrim et al. 2019). The widely used metrics discussed here include FAR (false accept rate), FRR (false reject rate), ACC(maximum accuracy), EER (equal error rate), half total error rate (HTER), ROC curve, DET curve, and AUROC (area under a ROC curve). The information presented in this section summarizes the common metrics mentioned in almost all the papers in the field (Dunstone and Yager 2008). In the following sections, we present the changing landscape of performance evaluation, general guidelines for evaluation, critiques of the common metrics, and many other considerations when evaluating a biometric-based authentication system.

2.2 General Guidelines for Evaluating a Biometric System

Evaluation using statistical metrics is only one aspect of performance evaluation. Cherifi et al. 2010 propose various other important considerations when evaluating biometric systems. The Cherifi et al. 2010 work is one of the earliest papers that propose a general guideline for systems' performance evaluation and comparisons. In biometric systems, the performance evaluation is realized under three contexts: technology, scenario, and operational assessments (Cherifi et al. 2010). The technology evaluation is done using the previously acquired biometric data to determine *a priori* if the developed biometric system meets the requirements. Testing is carried

out using offline processing of saved data, and the results are typically repeatable (Cherifi et al. 2010). The scenario evaluation evaluates an end-to-end system using a simulated environment that models the real-world application of interest. The testing results are only repeatable to the extent that the modeled scenario can be carefully controlled (Cherifi et al. 2010). The operational evaluation is an online test done in real conditions. The result is usually not repeatable due to unknown differences and lurking variables in different operating environments (Cherifi et al. 2010).

When evaluating a biometric system, technology evaluation is usually the most common and feasible method because it is done using acquired data, which renders the results reproducible and the simulation environment relatively easier to construct (Cherifi et al. 2010). The most significant disadvantage of technology evaluation is that it does not accurately reflect the conditions where the system will be deployed (Cherifi et al. 2010). Hence, this is where the quality of data collection plays an important role. The better the acquired samples can mimic the application conditions, the more accurate the evaluation results can be (Cherifi et al. 2010).

A biometric system's performance assessment usually only considers the quality of the input data and the output results. Thus, it is beneficial to treat a biometric system as a black box during an evaluation, ignoring the internal algorithms (Thacker et al. 2008). The black box takes in biometric data and a set of parameters to generate output results. Within this context, Cherifi et al. 2010 proposes the four key considerations when evaluating a biometric system:

1. Quality control of the biometric templates

2. Database related considerations

3. Statistic metrics evaluation

4. Subjective evaluation

As mentioned in the background section, a biometric system is composed of two main components: a sensor that collects biometric templates from the users and some algorithms for the system's enrollment and verification phases. Thus, the quality control of the biometric template is essential to the outcome of a biometric system. Cherifi et al. 2010 presents three main problems that can alter this quality:

1. Problems due to sensors

2. Problems due to users

3. Problems due to the environment

Cherifi et al. 2010 state that researchers must determine strict quality control protocols for the acquired templates before the enrollment and the verification phases. The acquired biometric samples used in technology evaluations are stored in a database. To compare different biometric systems, we need to compute their performance following the same protocols, such as the same acquisition conditions and database. The characteristics of the database can have a huge impact on the outcome of an evaluation. Hence, it is not possible to compare evaluations done using different databases (Cherifi et al. 2010). It follows that when evaluating a biometric system, we must first evaluate the quality of the input templates. When comparing various systems' performance, the systems must be developed using the same database.

In Cherifi et al. 2010 work, their recommendation of statistical metrics are generally equivalent to the common metrics mentioned in Section 2.1. However, they emphasize that statistical metrics should not be the only thing to consider when we appraise a biometric system's performance. Instead, they advocate for a user-centric mindset when assessing a system's usability and quality, stressing the importance of acceptability, usability, convenience, and user confidence. They believe that previous researchers in the field were often too fixated on optimizing the statistic metrics while not paying enough attention to the subjective evaluation aspect (Cherifi et al. 2010).

The three critical factors of the subjective evaluation are acceptability, convenience, and user confidence in the system (Cherifi et al. 2010). Acceptability is defined as how acceptable users deem their interaction with a biometric system is (Cherifi et al. 2010). The acceptability factor is highly dependent on the culture of users. For example, European users generally prefer fingerprint recognition to iris recognition. In contrast, Asian users prefer biometric systems with fewer physical contacts. The convenience factor depends on the quality of the sensors and user interface (UI), and the time a system takes for the enrollment and authentication steps (Cherifi et al. 2010). Generally, the easier the UI and the quicker the authentication steps, the more likely users will accept it. User confidence is similar to the acceptability factor as it is also highly dependent on the culture of users. It denotes how reliable users

perceive a system. Users generally have more substantial confidence in physical biometric systems than behavioral biometric systems since the former is less intrusive during use (Cherifi et al. 2010). However, Cherifi et al. 2010 also note that the more efficient a biometric modality is, such as DNA analysis, the more it invades privacy and the less likely users will feel comfortable using it. Thus, this raises another trade-off researchers to need to consider when choosing a biometric modality, which will eventually impact the system's user-perceived quality.

In addition to the subjective evaluation, there are three crucial considerations when evaluating a behavioral biometric system. First, the behavioral biometric template can change with time according to users (Cherifi et al. 2010). For instance, Hocquet et al. 2007 gives the example of keystroke dynamic analysis. They found that users' keystroke patterns change over time as they type more often and learn how to type more efficiently. Moreover, Cherifi et al. 2010 give another example where users' signature changes as they age. The temporal change in behavioral patterns presents two consequences. For one, the number of templates required for a behavioral biometric system at the enrollment phase for a user is significantly higher than the quantity needed for a biological system. And the data collection for a user needs to be spaced out to account for this variability. Additionally, the internal recognition algorithm and the evaluation of such a biometric system must consider the variability of the templates to make a reasonable assessment of their performance (Cherifi et al. 2010).

Second, behavioral biometric systems are sensor dependent, meaning that a user's extracted behavioral data can be different under different sensors (Cherifi et al. 2010). This means that the performance evaluation of behavioral biometric systems must be realized using the same sensors during the enrolment and recognition phases while closely simulating the final targeted application environment.

Third, using behavioral patterns as a biometric characteristic can be different for users depending on various factors, including age, culture, and experiences (Cherifi et al. 2010). Therefore, it is essential to have the evaluation of such a system be realized under a database that encompasses a wide variety of users. Cherifi et al. 2010 also proposes the inclusion of synthetic and artificially generated templates further to test the robustness of a system's performance. Consequently, we see that when evaluating the efficacy of a behavioral biometric system, the quantity sufficiency of user samples and temporal variability of samples must be considered.

To summarize, the four critical considerations of performance evaluation are quality control of the input samples, database quality controls, statistical evaluation, and subjective evaluation. The subjective evaluation aspect is crucial but is often undervalued. The three critical components of subjective evaluation are acceptability, convenience, and user confidence in the system. When evaluating behavioral biometric systems, the quantity sufficiency of samples and samples' temporal variability must be considered. As mentioned earlier, the factors form the basis of general guidelines for assessing and comparing biometric systems.

2.3 Limitations of Common Metrics in Continuous Authentication Systems

Although there has been extensive research on new biometrics systems, there still lacks a reliable way to compare different systems when deciding on which one to choose (Eberz et al. 2017). Furthermore, although the metrics mentioned in the background section are widely used in the field, they have inherent flaws and limited scopes (Sugrim et al. 2019). We start exploring the limitations of common metrics by examining their applications in continuous authentication systems and then in the general authentication context.

So far, we have only mentioned the general notion of an authentication system. This system verifies a user's identity as either genuine or imposter. Nevertheless, based on the frequency of verification, authentication systems can be further divided into one-time authentication and continuous authentication system (Eberz et al. 2017). Password-based and one-time biometric authentication systems only provide login-time authentication. If the user's identity changes afterward, the system will not be able to detect that.

On the other hand, continuous authentication aims to mitigate this limitation by continually verifying the user's identity and locking the user's access once a change in identity is detected (Eberz et al. 2017). Given its nature, a continuous authentication system must periodically collect identifying information from the user. The higher the frequency of information collection, the more secure is the system (Eberz et al. 2017). Hence, it renders the traditional password-based system and other biological biometric systems that continuously require the user's attention and interaction not ideal for continuous authentication. Consequently, behavioral biometrics become

the best non-intrusive biometric modalities for constant authentication systems. Examples of behavioral biometrics include keystroke patterns, mouse movements, touchscreen inputs, eye movements, etc. (Eberz et al. 2017). The advantage of these behavioral biometrics is that these data can be passively collected and transparently monitored without requiring any user's inputs, which results in a secure and user-friendly authentication system.

As stated in Section 2.1, a biometric system can make two types of errors: false rejects and false accepts. Therefore, the two error rates, FRR and FAR, form the basis of errors for any authentication system. We note that the term the single-number metrics often refer to the common metrics presented in section 2.1. The work of Eberz et al. 2017 is one of the first few papers that directly address the limitations of common metrics in the context of continuous authentication systems. In particular, Eberz et al. 2017 believes that the most significant limitation of using common metrics is that commonly reported metrics only capture the errors' means but not errors' distributions.

Through surveying 25 recent publications in the biometrics authentication field, Eberz et al. 2017 found that most papers only report the mean of single-number metrics with little attention paid to their distributions. This is problematic because the magnitude of these metrics' mean can be a misleading indicator of a system's performance. Other underlying factors, such as the metrics' distributions, need to be considered to have a more holistic view of a system's performance.

The most important consideration when evaluating a continuous authentication system is assessing systematic errors. In a continuous authentication system, an attacker has to fool the system over a prolonged time, rather than once in a one-time authentication system. Thus, there is a huge difference between random errors and systematic errors. A continuous system with random errors will prolong but will eventually detect an attacker. In contrast, a continuous system with systematic errors will lead to a few attackers perpetually escaping detection (Eberz et al. 2017). When a publication only reports the error metrics' mean without reporting the errors' distributions, there is no way of knowing whether a low ERR is due to random errors or systematic errors. In the context of single time authentication, Eberz et al. 2017 believes that using EER and other common metrics are reasonable and widely accepted. However, continuous authentication presents a different challenge as errors accumulate over the runtime of a system.

For instance, without knowing the errors' distributions, a FAR of 10% could indicate that either all attackers will be detected 90 % of the time or 10% of attackers can always bypass the systems' detection while others are exposed immediately. The former is the random error distribution scenario, which results in the eventual detection of attackers. The latter is the systemic errors scenario (i.e., systematic false negatives), which results in some attackers never being detected, even though the FAR value seems low on the surface. Fig 2.6 illustrates these scenarios through two graphs.

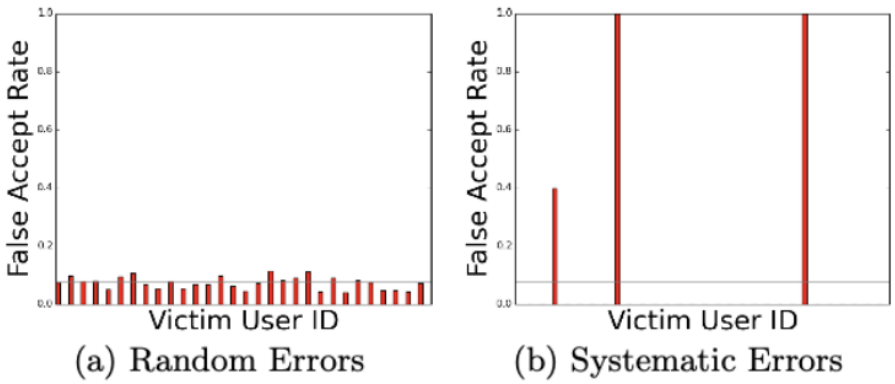

Fig. 2.6: The grey line denotes the same 9% FAR for both samples. However, their FAR distributions are different. On the left, we see random error distribution resulting in eventual detection. On the right, we see the systemic errors. A few attackers can perpetually escape the system's detection by impersonating many victims (Eberz et al. 2017).

Fig 2.6 shows that random errors with moderate and low-variance regular false negatives are often randomly distributed across victim-attacker pairs. The runtime of a session will result in eventual detection by the system and is acceptable for continuous authentication. Systemic errors are more severe from a security perspective because the undetected attackers can access the system perpetually (Eberz et al. 2017).

2.3.1 Gini Coefficient

The confusion matrix can provide a complete picture of the errors (FAR and FRR) distributions. However, it is neither compact enough for large datasets given its high space requirements nor enables readers to compare two systems (Eberz et al.

2017) directly. Additionally, metrics like standard deviation and kurtosis - the fourth standardized moment that measures the skewness of distribution - are also not ideal, because it is difficult to accurately rank biometric systems since preference would be arbitrary (i.e, preferring standard deviation over kurtosis or vice versa) (Eberz et al. 2017). As a result, Eberz et al. 2017 propose the Gini Coefficient(GC) as the best way to capture errors' distributions for continuous authentication systems.

The Gini Coefficient was initially proposed in 1912 as a measure of statistical dispersion to reflect a nation's income distribution (Gini 1912). A GC of 0 indicted absolute equality (i.e., everyone has the same income), while a GC of 1 indicates maximal inequality. In our context, GC - a measure of inequality - can be used to capture the errors' distributions. Specifically, a high GC value indicates more systematic errors. In contrast, a low GC value suggests that errors are relatively evenly spread out through the runtime of a continuous system (Eberz et al. 2017). The Gini Coefficient can be geometrically represented as the area between the Lorenz Curve and the Line of Equality (Eberz et al. 2017). The Lorenz Curve here measures the total error contributed by the bottom X% of users, and the Line of Equality is the Lorenz Curve of a system where all users contribute identical error rates (Eberz et al. 2017). Fig 2.7 shows GC's example of two behavioral biometrics to illustrate GC's usefulness.

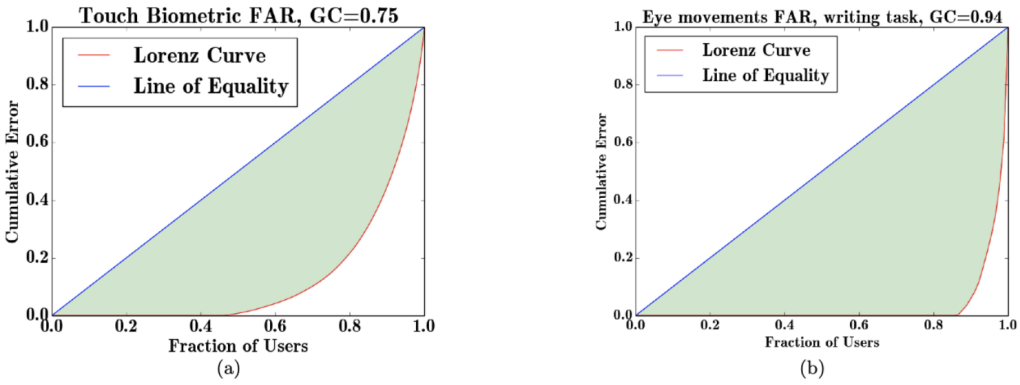

Fig. 2.7: The touch biometric has a comparatively low GC of 0.75 (left), which indicates largely random errors, while the eye movement biometric's higher GC of 0.94 (right) suggesting high systematic errors which will lead to attackers consistently fooling detection (Eberz et al. 2017).

From Fig 2.7, we can see that the touch biometric has a relatively lower GC value of 0.75, which indicates errors are relatively evenly spread out. The eye movement biometric, on the other hand, has a higher GC value, suggesting that the FAR is

caused by a few extremely successful attackers consistently fooling the system. Therefore, for the FAR, systems with lower GC are desirable because they suggest that false accepts relatively evenly spread out across attackers. A system with higher GC indicates that it allows a few attackers to escape detection consistently.

The GC has two essential properties that make it an ideal metric for continuous authentication. First, GC is scale-independent, meaning that it does not depend on the total or average error of a system but only depends on the distribution of values (Eberz et al. 2017). Second, GC is population independent, meaning that it does not depend on the number of samples in a dataset (Eberz et al. 2017). This is crucial since the number of subjects in biometric datasets varies greatly. Using subsets of equal size is not practical given that authors rarely publish their datasets (Eberz et al. 2017). These two properties of GC allow us to compare systems with different error rates and different data sizes, which is not possible using popular common metrics like FAR, FRR, or EER.

Another advantage of the GC is that it allows researchers to select useful features. For example, while testing an eye movement biometric system, Eberz et al. 2017 found that after removing the pupil diameter, one of the most distinctive features of eye movement biometrics, the average error rates increase and the GC decreases. This insinuates that pupil diameter is a key feature contributing to the systematic errors. Due to pupil diameter's relative stability, it can separate most users but consistently confuses users with similar baseline pupil diameters. In other words, this feature helps distinguish users that are already relatively well-separated but does little to reduce the systemic errors if not contributing more. This demonstrates the usefulness of the GC in feature selections. That's, in some cases, adding distinctive features can reduce the security of a system, even if the error rates seem lower (Eberz et al. 2017). Consequently, Eberz et al. 2017 advises researchers not blindly to strive for the lowest average EER. Instead, they advocate for using the GC to examine how changes to features and classifiers affect their system's error distributions.

2.4 Limitations of Metrics in General Authentication Systems

So far, we have examined the limitations of common metrics in the context of continuous authentication systems. In this section, we present the limitations and flaws of common metrics in a more general context.

We compile and distill the flaws and limitations of common single-number metrics into the following four areas:

1. Failure of capturing errors' distributions
2. Susceptible to non-functional methodologies
3. Susceptible to Population Skews
4. Failure of capturing score distributions

The first limitation, described in Section 2.3, is particularly relevant and vital to continuous authentication. The other three limitations affect all authentication systems, generally.

2.4.1 Susceptible to Non-Functional Methodologies

One limitation of single-number metrics is that their numerical outcomes can be easily influenced by some non-functional parameters, which has nothing to do with a system's security. Eberz et al. 2017 quantify the impacts of two popular non-functional machine learning methodologies - the attacker modeling process and the selection of training data - on error rates.

The selection of training data refers to the common practice that majority of the papers randomly select training data from the entire datasets and then use the rest as validation and test sets (Eberz et al. 2017). The attacker modeling process refers to the common practice of merging data from all users to constitute the negative class (attacker) (Eberz et al. 2017). The attacker modeling process is prevalent in the

field and always causes skewed measurement populations (Sugrim et al. 2019). To illustrate, a typical research recruits N participants, take M number of measurements from each, and then select one participant as the authorized user. This results in the skewed ratio (Sugrim et al. 2019): (left is the number of measurements from unauthorized users and the right is that of authorized users)

$$(N-1)*M:M$$

Understanding these methodologies' precise influence is significant as it can tell us whether a lower EER is due to a better system or excessive blind optimization through non-functional design decisions. To show the influence of these factors on errors, Eberz et al. 2017 calculate the EER for several datasets using these two methodologies. The results are shown in Fig 2.8. The graph shows that EER can decrease up to 80% when randomly selecting training data from the same dataset. Including attackers in the training data (as a negative class) can decrease EER up to 63%. From this, selecting training data randomly provides the greatest improvement to the original EER (Eberz et al. 2017).

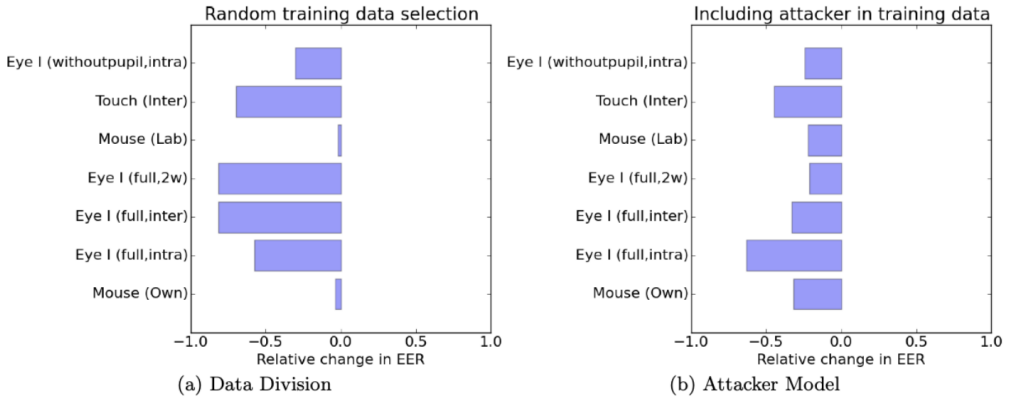

Fig. 2.8: EERs decrease up to 80% when randomly selecting training data. Including the actual attacker in the negative class provides a reduction of up to 63%. (Eberz et al. 2017).

The distribution of errors was almost unaffected by these two factors, indicating that they mainly shift the mean of EER (Eberz et al. 2017). The results suggest that one can significantly reduce or change the EER using different random sampling methods. Further, it shows that only looking at the single-number metric, like the EER of a system, is inadequate. It can be significantly skewed by non-functional

parameters that would not affect the system's actual performance under the real-world application environment. For instance, if the same dataset were evaluated using random and ordered training data selection, one might prefer one based on EER, although their true performance would be identical.

Through analyzing 25 recent publications in the related field, Eberz et al. 2017 found that many authors were not aware of this impact and their reporting methods are implicitly vulnerable to these two methodologies' impacts. This is alarming because it suggests that many researchers in the field are either unaware of the magnitude of these factors' impacts or too fixated on blind optimizing their EERs without actually improving their system performance (Eberz et al. 2017). This limitation of common metrics also reveals that researchers should consider some non-functional design decisions that can influence their system's performance during implementation.

2.4.2 Susceptible to Population Skews

Skews within the measurement population are common and can inflate some confusion matrix derived metrics (Sugrim et al. 2019). Biometric data are often split into training and testing sets. The training dataset is used to develop a model, and the testing set is used to compute performance metrics. Suppose the measurement population (data) itself is skewed. In that case, this skewness of distribution will be present in both training and testing data (Sugrim et al. 2019). Hence, metrics that are susceptible to population skewness can easily give misleading and inaccurate interpretations of a model's performance. For example, the four counts of a confusion matrix mentioned in section 2.1, such as FNR and FPR, are all susceptible to population skews. That's, if a system with a skewed population reports a low FNR, by looking at the FNR alone, we cannot tell if it is due to the system's discriminative capability or skews in population (Sugrim et al. 2019).

Before diving into the section, we first introduce the scores graph's unnormalized frequency counts (FCS). An FCS graph plots each authorized, and unauthorized users' frequency counts on the Y-axis and each corresponding score on the X-axis. Fig 2.9 shows an example of FCS graph with its corresponding ROC curve.

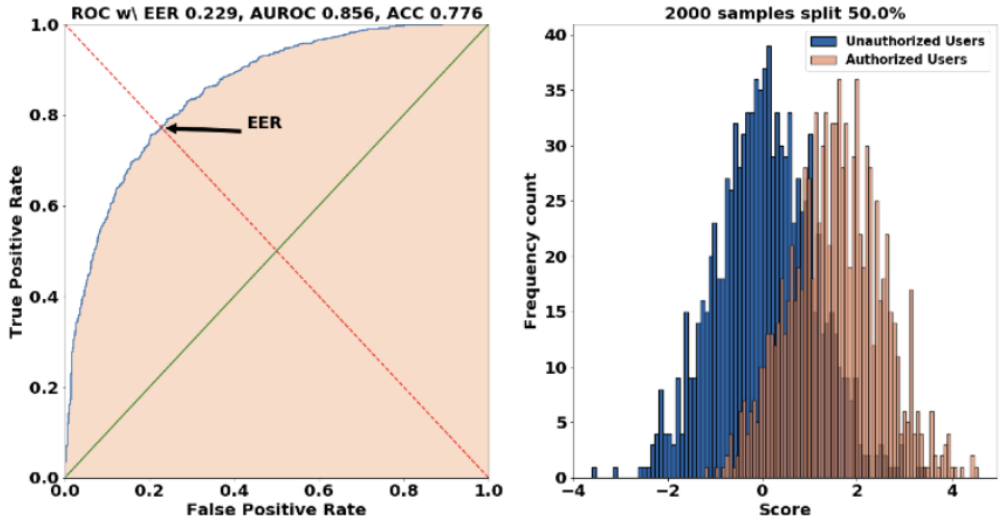

Fig. 2.9: A graph of ROC (left) and FCS (right) (Sugrim et al. 2019).

The FCS is created by identifying the maximum and minimum scores across all measurements and choosing a common bin width over the range (Sugrim et al. 2019). Scores are separated by their class labels (genuine or imposter) and plotted as histograms over that common bin width. The bin width is a free parameter that the implementer decides to represent the score ranges and variability (Sugrim et al. 2019). One important note here is that when we refer to FCS in this work, we only refer to the "un-normalized" version of FCS, not the "normalized" version. Besides, when comparing systems using unnormalized FCS graphs, it is imperative to make sure that FCS graphs are plotted using the same bin width. Otherwise, the overlapping region will seem different. In this and the following section, we use FCS to present metrics' limitations and show how FCS can be used to diagnose these limitations. We will discuss in details FCS's advantages and usage in Section 2.4.4.

The maximum accuracy (ACC) is another essential common metric that is vulnerable to population skews (Sugrim et al. 2019). If we only use ACC to compare two systems, we would naturally assume that the system with a higher ACC value is better. This seems to be reasonable on the surface since ACC estimates the probability of correct classification. However, ACC can also be misleading if that's the only metric we consider as Fig 2.10 shows.

When a classifier has a low ACC value, the genuine and imposter classes' scores will have a large amount of overlap. Suppose we make one measurement population

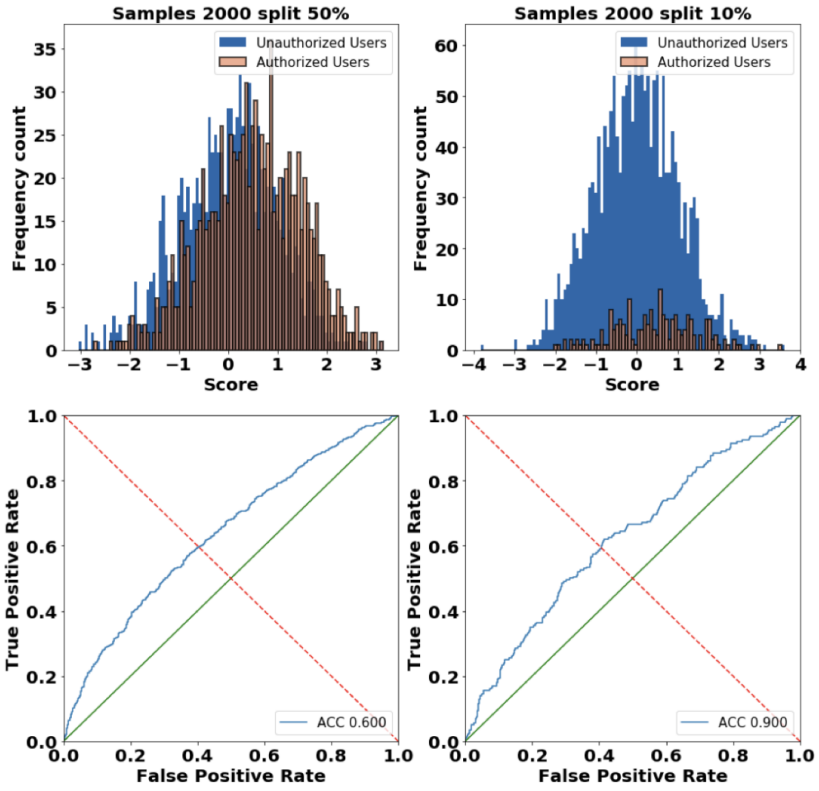

Fig. 2.10: A demonstration of how population skews can cause metrics like ACC to be misleading. Here are two FCS and ROC graphs for two measurements population. The right side has even split between unauthorized and authorized users. The left has skewed distributions. The same scoring function is applied to both. By adjusting the threshold to have a system just make negative decisions most of the time, we can get the right-side system to have 90% accuracy. The system on the left only has 60% accuracy. However, the ROC curves show that they have the same performance. (Sugrim et al. 2019).

to have predominantly unauthorized users (imposters) and make the proportion of authorized users significantly less, as shown on the right of the Fig 2.10. In that case, the population will be skewed toward having much more imposter (negative) class. In this case, the system's classifier will learn quickly to make mostly negative decisions (deny access) because most classifiers learn to optimize by minimizing their training data errors. Since the test data comes from the same population distribution, the test data will contain the same population skew pattern, making the system output an inflating high ACC value, rendering the system more accurate than it is (Sugrim et al. 2019).

Fig 2.10 shows ROC and FCS graphs for two samples of measurements. On the left side, the 2000 measurements are evenly split between unauthorized and authorized

27

users. On the right side, only 10% are authorized users while the rest 90% are unauthorized. By looking at the FCS graphs on the top, we can see their population distribution. For the even split side, the maximum accuracy ACC is roughly 60%, while the ACC for the skewed case is 90%, which is achieved by choosing a threshold that results in mostly negative decisions (Sugrim et al. 2019). By looking at the ACC values alone without knowing their respective data distribution, we would conclude that the right model is more accurate. This is misleading because they both use the same scoring function, and the ROC curves in Fig 2.10 show that they have the same performance. In general, any metrics that depend on both negative counts (N) and positive counts (P) are susceptible to this limitation (Sugrim et al. 2019). Recall that:

$$N = TN + FN$$

$$P = TP + FP$$

This example demonstrates that reporting based only on single summary metrics is incomplete (Sugrim et al. 2019). Some of the common metrics, especially the Confusion Matrix derived metrics, are susceptible to population skews. If a system was trained on mostly unauthorized data, it would learn to recognize unauthorized users well. And it will generate seemingly high metrics value during testing since the testing data is skewed the same way. However, suppose the system is deployed in real-life applications. In that case, it might not perform as expected because it may not recognize authorized users very well since it wouldn't learn a good model for authorized users (Sugrim et al. 2019). The same reasoning applies if the measurement population is skewed the other way. Therefore, we can now see why researchers must have this awareness and report the frequency of both authorized and unauthorized users since common metrics like ACC provide little insight into the population distribution.

In Fig 2.10, we see how FCS and ROC graphs can help us identify this weakness. Fig 2.10 shows that ROC curves are mostly unaffected by the population skews. The unnormalized counts in FCS graphs reveal the population skews directly and visually. Population skews allow some single-summary metrics to mask the poor score separation of a system by having seemingly high metric values (Sugrim et al. 2019). The visualization from FCS and ROC helps us quickly examine the distributional skews and justify our metrics.

28

2.4.3 Failure of Capturing Score Distributions

The previous section shows how population skews can lead to misleading interpretations of some Confusion Matrix derived metrics. In this section, we present another limitation of standard metrics: failure of capturing score distributions. The score distribution of a system contains important information regarding the system's performance. Single-summary metrics fail to capture the score distributions. Thus, when comparing systems using common metrics, it will lead to misleading and inaccurate interpretations (Sugrim et al. 2019). We also show how some ROC derived metrics can exhibit this limitation.

When reporting a single performance metric as a summary of a system, the metric's value reveals little information regarding a system's classifier's uniqueness and the system's performance. A common practice in the field is to compare different systems using EER (Sugrim et al. 2019). In Fig 2.11 and the following analyses, we show how comparing systems using EER can be misleading.

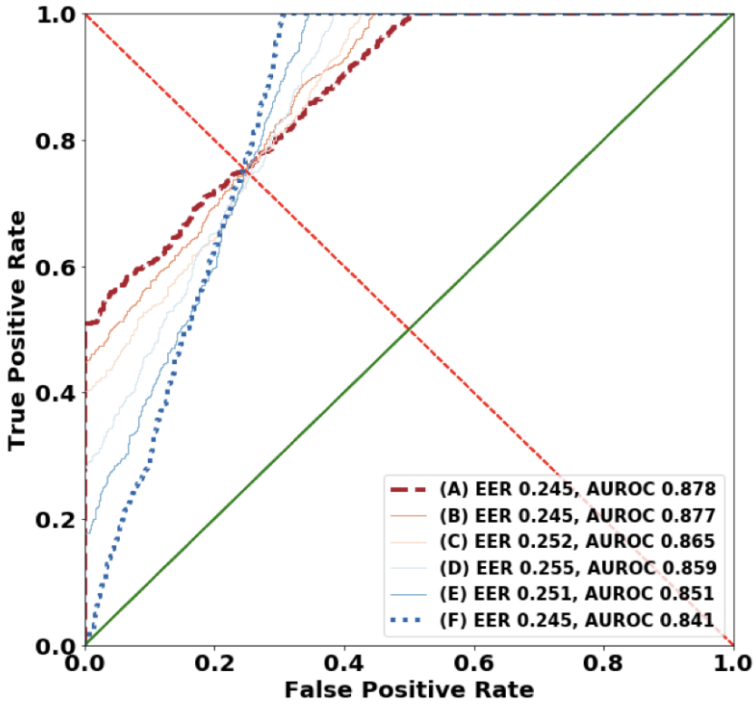

Fig. 2.11: An example of why an EER value does not represent a unique ROC curve. Here we have six different ROC curves corresponding to six different systems with similar EER and AUROC values. The linear portions of the ROC curves have different slopes, representing different sensitivities to changes in the threshold. (Sugrim et al. 2019).

29

Fig 2.11 presents six different ROC curves corresponding to six other systems with similar EER and AUROC values. The linear portions of the ROC curves have different slopes, representing different sensitivities to changes in threshold due to different score distributions (Sugrim et al. 2019). For example, a shift in threshold has more impacts on ROC (F) than ROC (A) since ROC (F) has a steeper slope (Sugrim et al. 2019). Hence, if we only compare the two systems using similar EER values and pick one to implement for our application, the implementation can fail unexpectedly since we are unaware of the sensitivity to the threshold.

This reveals that the direct comparison of systems using EER is inappropriate because different system's ROCs can have similar EER values. When implementing a biometric authentication system in real life, the implementer usually has a target application in mind that requires certain levels of FAR and FRR (Sugrim et al. 2019). When we are only given EERs to evaluate a system, and we have a specific requirement for FAR or FRR, we can not determine if the system can satisfy the need from EER alone (Sugrim et al. 2019).

From section 2.1, we know that EER is the point on a ROC curve where FAR equals FRR. In other words, EER represents the equal probability of making incorrect deny access and incorrect granting access decisions. Therefore, EER represents the overlapping region between unauthorized and authorized users' scores distributions (i.e., proportional to overlap width in an FCS graph). Based on this, the corresponding FCS for each ROC in Fig 2.11 is presented in Fig 2.12. We note that there is no population skews in any of the graphs in Fig 2.12; rather, only the score distributions change (i.e., the authorized range shrinks). Each FCS graph in Fig 2.12 corresponds to a system with a unique score distribution.

As we can see, each score distribution has the same width of the overlap region. The overlapping area moves rightward as we read the graph from left to right and top to bottom. The range of authorized users' scores region shrinks as the overlapping areas move to the right. The unauthorized users' region grows to maintain the same width of overlap (Sugrim et al. 2019). All systems' classifiers can be tuned by choosing a specific threshold to achieve a similar EER (Sugrim et al. 2019). As the unauthorized region takes over the authorized area, we can see fewer distinct scores from the authorized class and, hence, fewer ways to obtain true class declaration. This explains why the range of potential trade-offs is worse for F than A, as reflected in the more obscured overlapping regions in the FCS graph and also reflected in the different slopes of their ROC curves in Fig 2.11 (Sugrim et al. 2019). Furthermore, the unnormalized FCS graphs in Fig 2.12 show that the probability

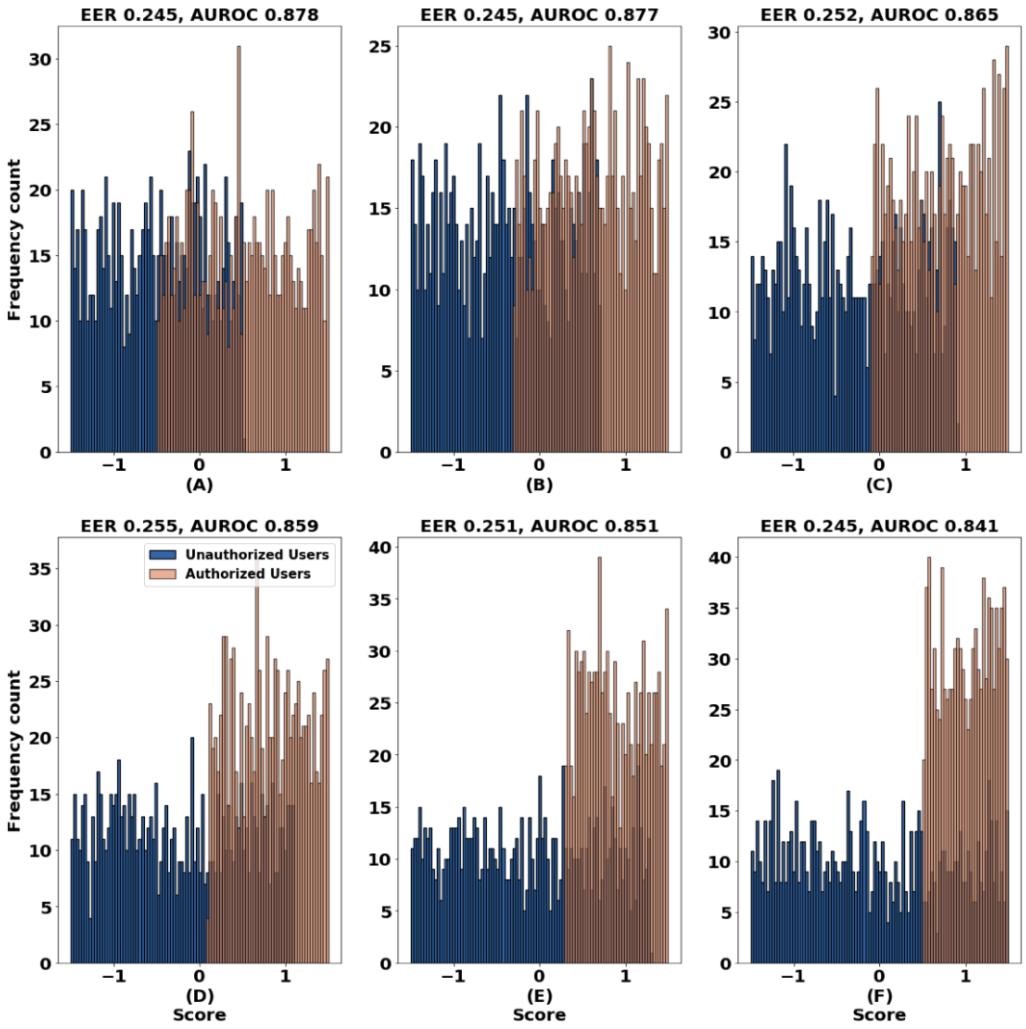

Fig. 2.12: The corresponding FCS graphs to the ROC curves shown in Fig 2.11. There is no population skews in any of the graphs. The difference displayed reflects different scores of distributions over the same overlap region. Each of the FCS graphs has the same width of overlap. The overlapping region moves rightward as we read the graph from left to right and top to bottom. The range of authorized users' scores ranges shrinks as the overlapping regions move to the right. The unauthorized users' region grows to maintain the same width of overlap. (Sugrim et al. 2019).

mass for authorized scores is re-distributed over a smaller region, which renders the distributions of the scores more asymmetric (Sugrim et al. 2019). This indicates that the system's classifier becomes more biased as the probability of observing an authorized score reduces.

We can conclude that if the distribution shape changes but the overlapping region remain the same, the EER value will remain unchanged. At the same time, the ROC

31

curve becomes more curved (Sugrim et al. 2019). Thus, if we compare systems only based on EER, we would assume that different systems with similar EER have similar performances without knowing that they can exhibit very different applications.

From this, we can see that the major limitation of EER is that it only concentrates on the overlap between the distribution of two classes and ignores the region that lies outside the overlap (Sugrim et al. 2019). The proportion of the scores outside the overlap is as significant as the overlap scores because it reflects the probability of observing an easily confused score (Sugrim et al. 2019). Thus, we see why EER fails to account for the asymmetries in score distributions and the regions outside of overlap. In contrast, we can see how the FCS graphs in Fig 2.12 helps us visually observe the asymmetries in score distributions and all other important information mentioned above.

Following similar reasoning, Sugrim et al. 2019 state that we can see similar flaws in other ROC derived metrics, such as AUROC and GC. Recall from section 2.1 that AUROC denotes the area under a ROC curve. AUROC essentially represents the probability that scores from different measurement classes separate well (Sugrim et al. 2019). This probability is proportional to the width of overlapping score regions. As a result, AUROC has the same inherent flaws as EER does.

Although Eberz et al. 2017 appraise the GC in fixing the first limitation of capturing errors' distributions in a continuous authentication system, Sugrim et al. 2019 believes that GC fails to capture the score distributions. Functionally, GC can be derived from AUROC as (Sugrim et al. 2019):

$$GC = 2 * AUROC - 1$$

Similar to AUROC and EER, GC is also a measure of separation between different measurement sources. Therefore, GC contains the same flaw as AUROC and ERR.

2.4.4 Robust Approach to Evaluation and Reporting

From the previous four sections, we have seen the critical limitations of common metrics. From the last two sections, we have also seen how the combination of ROC curves and FCS graphs can visually diagnose the problems in population

skews and asymmetries of score distributions. In this section, we summarize the advantages of using the FCS graph and explain why the combination of ROC and FCS graphs, proposed by Sugrim et al. 2019, is currently a state-of-the-art approach to performance reporting in the field. In the end, we also present the new reporting guidelines proposed by Sugrim et al. 2019 as the new basis of performance evaluation.

In short, the advantages of using FCS are that it can effectively detect measurement population skew, asymmetries in the scoring distribution, and sensitivity to threshold changes. From the last two sections, we have seen how FCS allows us to visually diagnose and justify the achieved performance reported in the ROC and other metrics. We note again that FCS should not be normalized because the unnormalized frequency counts allow us to visually observe population skews, score distributions imbalances, and score overlap regions (Sugrim et al. 2019). These visual analyses were previously not possible to perform through common ROC and CM derived metrics. It allows researchers to check if their metrics exhibit any vulnerability to population skews or score distribution imbalances.

By examining the distribution skews and overlap of score frequencies, we can tell if a system tends to show bias toward either a positive or negative decision (Sugrim et al. 2019). By looking at how well scores distributions are separated, we can justify the performance observed in ROC. Moreover, a system's sensitivity to threshold changes can be understood by looking at scores that are separated relative to each class (Sugrim et al. 2019). Additionally, FCS is derived from neither ROC nor CM, which means reporting FCS brings additional insights not offered by ROC and Confusion Matrix derived metrics. Therefore, we can see that the unnormalized FCS graph complements ROC and common metrics well. That's why Sugrim et al. 2019 propose reporting ROC, FCS, and other necessary single-number metrics as a more holistic way of evaluating systems' performance.

From here, Sugrim et al. 2019 propose the new guidelines for performance reporting for the field. The first suggestion is to report as many metrics as possible, including both the ROC and FCS. In cases where the FCS cannot be reported, Sugrim et al. 2019 suggest reporting the ROC curve to allow others to decide if the system has a threshold that satisfies the FAR or FRR targets of their applications. Last, suppose the ROC cannot be reported. In that case, they suggest multiple summary metrics that not functionally dependent and present justifications for the metrics' validity (Sugrim et al. 2019). This forms the basis of the state-of-the-art approach of performance evaluation and reporting for biometrics-based authentication systems.

33

3

Problems and Future Work Proposal

3.1 Statement of Problems

Although the researcher's understanding of performance evaluation for biometric systems has come a long way over the past decade, several issues still exist with the current methodologies. We identify the following problems:

1. Lack of proper research in distribution level metrics

2. Lack of proper research in quantifying the impacts of some basic metrics on systems' performance

3. Lack of standard evaluation metrics package in the field

We explain each problem in more detail below.

The first problem is that the field currently lacks proper research in distribution level metrics, which are metrics that quantify the overlap regions of scores and areas outside the overlap. We now understand that an authentication system's essence is the study of two classes' (imposter and genuine) distributions and how well they separate. The traditional single-number metrics, mentioned in section 2.1, often focus on a single aspect of a system and captures limited information about classes' distributions. And we have seen why common metrics alone are inadequate in achieving that goal. However, some lesser-known distribution level metrics, such as the KL divergence, modified Bhattacharyya distance, and Mahalanobis distance. These metrics are lesser studied in the field. And they could potentially provide

additional insights into understanding a system's performance by capturing more information regarding the distributional overlap (Dunstone and Yager 2008).

Second, some basic metrics, such as the sufficiency of samples and confidence interval, are often left out in the final reported results. Some researchers have taken them for granted without checking their validity (Dunstone and Yager 2008). The sufficiency of samples refers to the sufficiency of data or sample size, which can be viewed as the minimum level of sample size needed to validate a claim (Jain et al. 2011). Sometimes, researchers in the field drew overly-promising results based on systems trained on limited sample sizes (Dunstone and Yager 2008). Moreover, the impacts of the sufficiency of samples are not quantified in the field.

Third, there is a lack of standard evaluation metrics packages in the field. Researchers often have to reinvent the wheels and spend a lot of time on writing metrics functions. If an evaluation metrics package can be created, standardized, and open-sourced to the public, like other machine learning communities, this could significantly improve research efficiency and facilitate progress in the field.

3.2 Future Work

Based on the field's current landscape and the problems mentioned earlier, we propose the following future work directions to contribute.

We first plan to make a list of all distribution level metrics that could be useful to authentication systems. So far, the list includes modified Bhattacharyya distance, KL Divergence, Mahalanobis distance, and other variants of histogram overlap graphs and metrics.

We then plan to define these metrics first from a mathematical perspective and then find relevant literature, which could be in other fields of machine learning or statistics, on these metrics. Afterward, we plan to do an in-depth literature review of how these metrics were used in other fields. Moreover, we plan to add the sufficiency of samples and confidence scores to a list of investigating metrics. We will propose methods to compute the adequacy of samples and confidence scores in the reported metrics. Then, we plan to use both synthetic and real biometric scores to create a testing database. We then plan to investigate the impacts of the

metrics as mentioned earlier on the database. Specifically, we will examine if these new metrics can present any additional useful insights or information, which are not reflected through the traditional metrics. We also aim to quantify these metrics' impacts and define strengths and weaknesses for each of them.

Furthermore, we plan to use behavioral biometrics such as gait or keystroke patterns to create a new dataset for testing continuous authentication systems. We then repeat a similar process to explore novel evaluation metrics that are suitable for continuous authentication. We will also test the aforementioned new metrics to see if they are ideal for continuous authentication.

Through the process, we will have a list of useful novel metrics. We then plan to implement all the essential metrics, including common and novel metrics, in a comprehensive Python package. We plan to streamline function usage in the package so that researchers in the field can use them off-the-shelf to evaluate their systems right away. Our goal is to make it a comprehensive open-source package to make the performance evaluation of biometrics-based authentication systems more straightforward and convenient.

4

Conclusion

We have synthesized key ideas from recent publications to show how the landscape for performance evaluation of biometrics-based authentication systems has developed over time. We present the general architecture and fundamentals of an authentication system. We have defined every popular metric commonly used in literature and illuminate the strengths and weaknesses for each. We also present the four old guidelines for evaluation: quality control of input templates, database-related considerations, statistical evaluation, and subjective evaluation.

Besides, we have shown the four major limitations of commonly used metrics. They are the failure to capture errors' distributions, susceptible to non-functional parameters, susceptible to population skews, and inability to capture score distributions. We discuss why the limitations of commonly used metrics are significant to continuous authentication and how the Gini Coefficient can capture systemic errors. Furthermore, we use examples from Eberz et al. 2017 work to illustrate how GC can also be used in feature selections. Afterward, we present how adopting two non-functional parameters - training data selection and attacker modeling - can reduce error rates without improving the system. This reveals another non-functional design decision that researchers need to consider when designing their systems. Next, we show how population skews can artificially inflate some Confusion Matrix derived metrics, making a system seem more accurate than it is. We demonstrate how we can use FCS graphs to check for population skews and use ROC curves to compare systems better. Last, we show why single-summary metrics like EER, AUROC, and GC can not be used as the sole factor while comparing systems because they fail to capture score distributions for authorized and unauthorized classes. Different authentication systems can achieve similar EER and AUROC values through tuning parameters and selecting specific thresholds. We then show how we can use FCS graphs to detect score distribution imbalances and understand a system's sensitivity to changes in threshold by looking at how scores are relatively separated.

We then summarize the advantages of using FCS: it can detect measurement population skews, asymmetries in the scoring distribution, and sensitivity to threshold changes. And we explain why the combination of ROC and FCS provides a more robust evaluation approach. Finally, we present the new evaluation guidelines, proposed by Sugrim et al. 2019, which forms the basis of the state-of-the-art approach to evaluating authentication systems.

In the end, we identify three important problems with the current methodologies and propose specific future work plans to address them. Although research in other biometrics fields has come a long way over the past few decades, there has been a lack of proper research in the performance evaluation field. We hope this work can help the community raise awareness of the importance of appropriate performance evaluation. We also hope this work can help our community to see the significance of adopting more robust and comprehensive evaluation methodologies.

References

Bengio, Samy, Christine Marcel, Sébastien Marcel, and Johnny Mariéthoz (2002). "Confidence measures for multimodal identity verification". In: *Information Fusion*. DOI: https://doi.org/10.1016/S1566-2535(02)00089-1. URL: http://www.sciencedirect.com/science/article/pii/S1566253502000891 (cit. on p. 13).

Biometrics, PRECISE (2014). *Understanding biometric performance evaluation*. https://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation-QR.pdf (cit. on pp. 10, 13).

Cherifi, Fouad, Baptiste Hemery, Romain Giot, Marc Pasquet, and Christophe Rosenberger (2010). "Performance evaluation of behavioral biometric systems". In: *Behavioral Biometrics for Human Identification: Intelligent Applications*. IGI Global, pp. 57–74 (cit. on pp. 5, 14–17).

Dunstone, T. and N. Yager (2008). *Biometric System and Data Analysis: Design, Evaluation, and Data Mining*. Biometric System and Data Analysis: Design, Evaluation, and Data Mining. Springer US. ISBN: 9780387776279. URL: https://books.google.com/books?id=HXtagjFVEyIC (cit. on pp. 7, 11, 12, 14, 35).

Eberz, Simon, Kasper B Rasmussen, Vincent Lenders, and Ivan Martinovic (2017). "Evaluating behavioral biometrics for continuous authentication: Challenges and metrics". In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 386–399 (cit. on pp. 5, 18–25, 32, 37).

Fawcett, Tom (2004). "ROC graphs: Notes and practical considerations for researchers". In: *Machine learning* 31.1, pp. 1–38 (cit. on pp. 11, 12).

Gini, C. (1912). "Variabilità e mutabilità". In: *Reprinted in Memorie di metodologica statistica*. Ed. by T Pizetti E Salvemini (cit. on p. 21).

Hocquet, Sylvain, Jean-Yves Ramel, and Hubert Cardot (2007). "User Classification for Keystroke Dynamics Authentication". In: *Advances in Biometrics*. Ed. by Seong-Whan Lee and Stan Z. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 531–539. ISBN: 978-3-540-74549-5 (cit. on p. 17).

Jain, Anil K, Arun A Ross, and Karthik Nandakumar (2011). *Introduction to biometrics*. Springer Science & Business Media (cit. on pp. 2, 3, 6–10, 35).

Sugrim, Shridatt, Can Liu, Meghan McLean, and Janne Lindqvist (2019). "Robust Performance Metrics for Authentication Systems." In: *The Network and Distributed System Security Symposium (NDSS)* (cit. on pp. 3, 5, 7–14, 18, 24–33, 38).

Thacker, Neil A., Adrian F. Clark, John L. Barron, J. Ross Beveridge, Patrick Courtney, William Crum, Visvanathan Ramesh, and Christine Clark (Mar. 2008). "Performance characterization in computer vision: A guide to best practices". English. In: *Computer Vision and Image Understanding* 109.3, pp. 305–334. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.04.006 (cit. on p. 15).