Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics

Simon Eberz
University of Oxford
simon.eberz@cs.ox.ac.uk

Kasper B. Rasmussen
University of Oxford
kasper.rasmussen@cs.ox.ac.uk

Vincent Lenders
armasuisse
vincent.lenders@armasuisse.ch

Ivan Martinovic
University of Oxford
kasper.rasmussen@cs.ox.ac.uk

ABSTRACT

In recent years, behavioral biometrics have become a popular approach to support continuous authentication systems. Most generally, a continuous authentication system can make two types of errors: false rejects and false accepts. Based on this, the most commonly reported metrics to evaluate systems are the False Reject Rate (FRR) and False Accept Rate (FAR). However, most papers only report the mean of these measures with little attention paid to their distribution. This is problematic as systematic errors allow attackers to perpetually escape detection while random errors are less severe. Using 16 biometric datasets we show that these systematic errors are very common in the wild. We show that some biometrics (such as eye movements) are particularly prone to systematic errors, while others (such as touchscreen inputs) show more even error distributions. Our results also show that the inclusion of some distinctive features lowers average error rates but significantly increases the prevalence of systematic errors. As such, blind optimization of the mean EER (through feature engineering or selection) can sometimes lead to lower security. Following this result we propose the Gini Coefficient (GC) as an additional metric to accurately capture different error distributions. We demonstrate the usefulness of this measure both to compare different systems and to guide researchers during feature selection. In addition to the selection of features and classifiers, some non-functional machine learning methodologies also affect error rates. The most notable examples of this are the selection of training data and the attacker model used to develop the negative class. 13 out of the 25 papers we analyzed either include imposter data in the negative class or randomly sample training data from the entire dataset, with a further 6 not giving any information on the methodology used. Using real-world data we show that both of these decisions lead to significant underestimation of error rates by 63% and 81%, respectively. This is an alarming result, as it suggests that researchers are either unaware of the magnitude of these effects or might even be purposefully attempting to over-optimize their EER without actually improving the system.

ASIA CCS '17, April 02 - 06, 2017, Abu Dhabi, United Arab Emirates

ISBN 978-1-4503-4944-4/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3052973.3053032

1. INTRODUCTION

Password-based authentication systems only provide login-time authentication, any future change in user identity will go undetected. Continuous authentication is an approach to mitigate this limitation by constantly verifying a user's identity and locking a system once a change in user identity is detected. As such, it is necessary for the system to periodically collect some identifying information about the user. The more frequently such information is collected the faster a potential intruder can be detected. Naturally, approaches that heavily rely on user interaction and cooperation, such as passwords or fingerprints would severely harm user experience. As a result, behavioral biometrics, the use of distinctive user behavior to gain identifying information, has become a popular method to support continuous authentication. Examples include typing behavior (keystroke dynamics), mouse movements, touchscreen inputs and eye movements. These biometrics can be transparently monitored by the authentication system without necessarily requiring any specific input on the user's part.

The extensive body of work on behavioral biometrics calls for reliable ways to compare different systems when faced with the choice of which one to implement. In addition, developers will want to have realistic ideas of what security gains can be expected from using biometric recognition systems. Most papers collect a number of biometric samples from a certain number of users and extract biometric features, with the resulting feature vectors being classified by a machine learning algorithm. Ultimately, this process can result in two types of errors, false rejects and false accepts. Typical metrics reported as a measure of system quality are therefore the (mean) False Accept Rate (FAR), False Reject Rate (FRR) and Equal Error Rate (EER). The EER reflects the error rate at a threshold setting where FAR and FRR are equal. With these metrics being the most common, authors often strive to optimize them, for example by improving classifiers, hyperparameters or feature sets. However, this process of optimizing the mean often overlooks the security implications of different *distributions* of these errors, which may even lead to reduced security. When faced with a continuous authentication system an attacker has to fool the system over a prolonged time, rather than just once (as with a password-based system). Consequently, there is

a big difference between random errors (that will prolong, but not prevent the eventual detection of an attacker) and systematic errors (that can lead to an attacker perpetually escaping detection). Following this intuition we evaluate how prevalent different error distributions are in real-world biometric datasets, with a focus on systematic false negatives (i.e., perpetually undetected intruders). We then propose a number of additional metrics that compactly capture the security implications arising from these types of errors. These metrics can not only be used to compare different systems, but can also guide researchers when evaluating the influence of system design choices (such as feature selection) on error distributions and, ultimately, system security.

Besides affecting error distributions, blind optimization of the EER might also lead to unrealistic expectations regarding the system's real-world performance. As systems are usually evaluated on a static dataset, training, operation and the presence of attackers have to be simulated based on this data. There are a number of machine learning methodologies involved with this simulation, including different methods for training data selection and modelling of the attacker class within the classifier. Authors frequently choose to sample training data randomly from the entire set, which would not be possible in actual operation as the training data has to precede the entire testing data. In addition, authors often include some data of the eventual attacker in the (combined) negative class, a decision which is unrealistic outside of some insider threat environments. These disconnects highlight the need to quantify the impact of these different methodologies on error rates in order to accurately compare papers across methodologies. Only an accurate idea of how much each of these decisions impacts error rates will allow researchers to assess whether a papers' low error rates are a result of a better system or merely over-zealous error rate optimization.

The contributions of the paper are as follows: We provide an analysis of the methodology of 25 papers using 5 different biometrics for continuous authentication. We use 13 datasets to quantify the prevalence of systematic errors across 4 biometrics and outline factors influencing these types of errors. We analyse the suitability of different metrics to capture different error distributions and suggest metrics that provide better insights into the system's security. Lastly we quantify the effect of training data selection and attacker models on a system's error rates.

The rest of the paper is organized as follows: Section 2 provides an analysis of the state of the art with regard to metrics and methodologies. We discuss the shortcomings of current state-of-the-practice metrics in Section 3 and propose a number of alternatives to mitigate these problems. In Section 4 we discuss the impact of non-functional design decisions on error rates and conclude the paper in Section 5.

2. ANALYSIS OF COMMON PRACTICES

In this section we present a rigorous analysis of the state of the art, both with regard to metrics reported and the machine learning methodology used to obtain the results. In order to cover a wide cross-section of the field we have analysed 25 systems based on five different biometrics with a focus on recently published work. While these systems differ in experimental design and underlying features, they all provide continuous authentication. As such, we do not consider systems that provide enhanced biometric-based login time

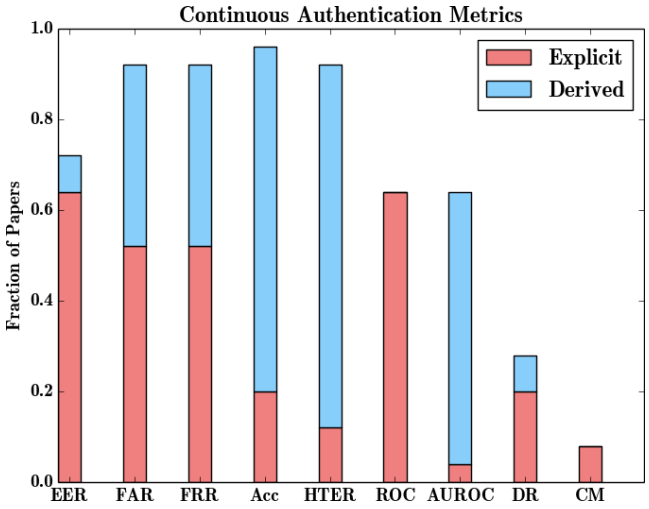


Figure 1: Metrics reported in literature

authentication (such as password hardening or fingerprint scanning).

2.1 Metrics

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

False Accept Rate (FAR) is typically measured as the fraction of intruder samples (rather than intruders) that are incorrectly accepted.

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

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

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

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

The *Receiver operating characteristics (ROC) curve* is a plot that shows the dependency between the FAR, FRR and the system's detection threshold. The ROC curve allows to derive a set of pairs (FAR,FRR) at which the system can be run by changing the threshold settings.

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

Detection Rate is a measure of the fraction of attackers that are detected by the system, unlike the FAR it operates on individual users, rather than samples.

The *Confusion Matrix (CM)* plots the fraction of accepted samples for each user pair. As such, it is a representation of raw data, rather than a numeric metric. The CM shows the FRR for each user and the FAR for each user-attacker pair. However, as the number of user pairs scales quadratically

Ref	Biometric	EER	FAR	FRR	Accuracy	HTER	ROC	AUROC	Detection Rate	CM
[16]	Touch	✓[1,4,5]	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[15]		✗	✓	✓	(✓)	(✓)	✗	✗	✗	✗
[35]		✓	(✓)	(✓)	✓	✓	✓	(✓)	✗	✗
[8]		✗	✗	✗	✓	✗	✗	✗	✗	✗
[18]		(✓)	(✓)	(✓)	(✓)	(✓)	✓	✓[2]	✗	✗
[36]		✓	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[12]		✗	✓	✓	(✓)	(✓)	✓[2]	(✓)	✓	✗
[9]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[10]		✓	✓	✓	(✓)	✓[3]	✓	(✓)	✗	✗
[29]		✗	✓	✓	✓[4]	(✓)	✗	✗	✗	✗
[28]		✓[5]	✓[5]	✓[5]	(✓)	(✓)	✗	✗	✗	✗
[31]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[13]	Gaze	✓[5]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✓	✗
[22]		✓	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[14]		✓[5]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✓	✗
[26]	Pulse Response	✓	✓	✓	✓	(✓)	✓	(✓)	✓	✗
[11]	Gait	✓	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[33]		✓	(✓)	(✓)	(✓)	(✓)	✗	✗	✗	✗
[2]		(✓)	✓	✓	✓	✓	✓	(✓)	✗	✗
[27]		✓	✓	✓	(✓)	(✓)	✓	(✓)	✗	✗
[25]	Mouse	✗	✓	✓	(✓)	(✓)	✗	✗	(✓)	✓
[1]		✓	✓	✓	(✓)	(✓)	✓	(✓)	(✓)	✓
[30]		✓[5]	(✓)	(✓)	(✓)	(✓)	✓	(✓)	✗	✗
[37]		✗	✓[5]	✓[5]	(✓)	(✓)	✓	(✓)	✗	✗
[24]		✗	✗	✗	✗	✗	✗	✗	✓	✗

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

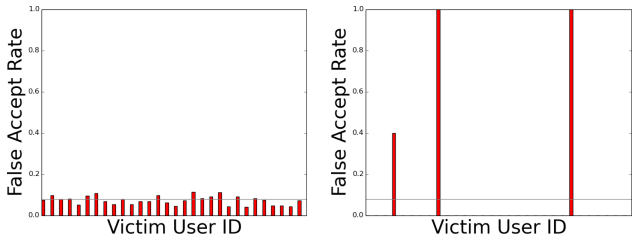
Table 1: Metrics used to evaluate continuous authentication systems. Basic measures such as FAR/FRR/EER are reported by most papers while confusion matrices, which are most informative, are virtually never given.

with the number of users, the space requirements are high for large number of users. In addition the CM is usually given as a plot, which somewhat reduces the space requirement but makes it difficult to obtain more than estimates of the actual numerical results.

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

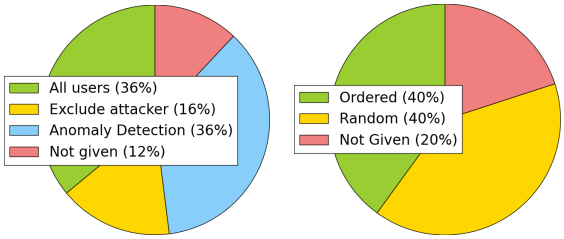
2.1.1 Limitations of common metrics

The EER (as well as the related metrics FAR, FRR, HTER and accuracy) is often used to compare different classifiers, with the assumption being that a lower EER results in attackers being detected more quickly (and more attackers being detected overall) and users being rejected less frequently (i.e., a better system). In the context of one-time (i.e., not continuous) authentication this is a sensible and widely accepted metric. However, continuous authentication provides a unique challenge as errors accumulate over the runtime of the system. Without knowing the exact distribution, an FAR of 10% could signify all attackers being detected 90% of the time (resulting in eventual detection), or 10% of the attackers never being detected while all others are exposed immediately. The second scenario exhibits so-called *systematic false-negatives*. These different scenarios are illustrated in Figure 2. Unlike regular false negatives, which might be randomly distributed across victim-attacker pairs as well as across the time of a session, systematic false negatives are tied to a combination of attacker and victim and are usually more persistent or even permanent as a result of the behavior



(a) Random Errors (b) Systematic Errors

Figure 2: Different distributions of the FAR lead to different security challenges, random errors and eventual detection (left) and systematic false negatives (right). The grey line denotes the (identical) 9% FAR of both samples. Note that this figure shows the success of a single attacker in impersonating multiple victims.



(a) Attacker Models (b) Training Data Selection

Figure 3: A large fraction of papers use random training data selection and inclusion of imposter data in the training set, both of which are likely to underestimate error rates.

of two users being very similar. These types of errors are more problematic from a security perspective, as the undetected attackers can then access the compromised system for a virtually unlimited time. Part of this property is captured through the detection rate, which measures the fraction of attackers with a non-zero FAR. However, the metric does not account for the difference between undetected attackers and those with simply a very high FAR. In practice this might even be determined by a single sample being classified differently. The confusion matrix paints a complete picture, but it is neither compact enough to report for large datasets, nor does it enable readers to easily compare two systems. Most likely, these limitations are the reason it is rarely reported in the literature. The authors of [6] propose to report the number of undetected attackers along with the average number of imposter actions (ANIA), a metric related to the false accept rate. However, they recommend reporting only the ANIA (which is, by definition, an average value), with no regards for its distribution between attackers.

While systematic errors are problematic for the FAR, this type of distribution might be desirable for false rejects. A seemingly low, but non-zero false reject rate for all users might still lead to frequent false alarms due to the base rate fallacy [5] if the system is run continuously throughout the day with a moderate sampling rate. If the false rejects were concentrated on few users they could be authenticated through other means (such as a different biometric) instead, without compromising security for the remaining users. In addition, such a scenario allows the developer of a biometric recognition system to analyze why the system performs poorly for precisely these users.

2.2 Common Evaluation Methodologies

A number of factors affect the distinctive capabilities (and thereby the security and usability) of a biometric system. Prominent examples include the ability of the system to collect high-quality data, the selection of distinctive features and the classifier itself. However, most papers analyze the system on a static dataset, which requires the simulation of training and operation, as well as the modelling of an attacker. In this section we provide a summary of methodologies and present an analysis on their prevalence in related work.

Hyperparameter Tuning. Following the feature extraction and normalization, a suitable classifier has to be chosen. Depending on the classifier, a number of hyperparameters have to be instantiated. Such parameters include the number of datapoints (the value of k) in the k-nearest-neighbors algorithm and the kernel type and soft margin constant C of a Support Vector Machine.

Attacker Model. Most biometrics are evaluated without a committed attacker in mind, this is commonly referred to as the zero-effort threat model. As such, the "attacker" is another user that attempts to access the victim's system without taking action to either circumvent the authentication system or impersonate the legitimate user. Even in this simplified threat model, it is still necessary to test the system's performance in detecting intruders. This is commonly achieved by comparing a user's template against the samples of all other users (i.e., the "attackers"). An important concern is the building of the user model itself. A common choice is to train a binary classifier with one user's samples as the positive class and samples from all other users (including the eventual attacker) as a single combined negative class. The system is then "attacked" individually by each of the users that jointly form the negative class. This approach means that reference data of the attacker is included in the negative class, even though it only forms a fraction of the overall class. In practice, it is impractical to assume that reference data for each potential attacker is available (aside from specific insider threat scenarios, such as [13]) and including this data may lead to overestimating the classifier's performance. A different approach trains a generic attacker model from other users (again, combining them into a negative class), but withholding samples from the actual attacker. The authentication system could then be shipped with this (anonymized) reference data. These two scenarios are also considered in [6] and referred to as external and internal scenarios, respectively. A more straight-forward approach is to perform anomaly detection, which trains a model from a single user's data without the requirement of providing samples for a negative class. New samples are then classified based on how similar they are to the training examples.

Selection of Training Data. An operational authentication system always requires reference data for each legitimate user (training data) in order to classify new observations. In practice, the initial training has to occur before any samples can be classified (although the model can be updated based on new observations). Consequently, a common approach to simulate this setting is to use the first part of the recorded data as training data, and the remaining samples as test

data. Another approach is to randomly sample the training data from the entire dataset, and to use the remaining data for testing. The sampling is often repeated to provide statistical robustness (either by performing several iterations of random sampling or through cross validation). However, this approach violates the requirement that training always has to precede testing (as some training samples may have been recorded after some testing samples).

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

The complete results of our survey can be found in the appendix. One of the most important observations is the (apparent) reluctance of researchers to make their data freely accessible online. However, it should be noted that our survey only accounts for data that is both available online and referenced in the corresponding paper. We have not contacted individual authors and can not make any statement on their willingness to share data on request. The number of papers using and building on this shared data (most notably, the data published as part of Touchalytics [16]) highlights that this is a valuable contribution to the community. In a similar fashion, the code used to generate the results is not usually published. As a number of machine learning steps depend on random numbers, this might make it particularly difficult to reproduce exact results, even if all decisions are clearly stated and raw data is available.

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

The vast majority of papers either do not use aggregation of samples, or don't report on the specifics of their mechanism. If samples are aggregated, this is usually done following classification (i.e., not on a feature vector level).

2.2.1 Limitations of common methodologies

The previous section has shown that a wide variety of methodologies are used to evaluate the static datasets, which suggests that it might not be possible to directly compare papers even if they use similar metrics. This would not necessarily be a problem if the impact of different methodologies on the reported metrics were to be comparatively small. To the best of our knowledge, this effect has not been quantified in the context of continuous authentication. It is, however, well-studied in malware detection. Specifically, Allix et al. have shown that sampling training data randomly from all available data leads to systematic underestimation of error rates [4, 3]. This is problematic, as reference data for future malware helps in the classification, but would not necessarily be available in the real-world (i.e., to classify newly observed malware). One might assume a similar effect for continuous authentication, as random training data selection would make future samples available to help classifying past ones. This allows the classifier to accurately account for short and long term changes in user behavior, which would not be possible when maintaining the temporal integrity of the dataset.

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

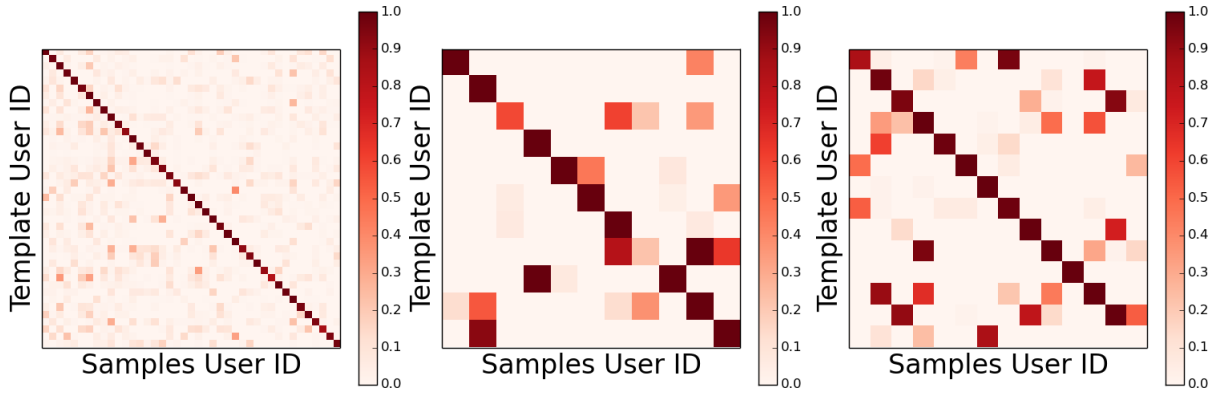
Out of the 25 papers we analyzed (see Table 3 in the appendix), 13 use at least one of these methodologies and a further 6 don't report the methodology used. As such, it is crucial to quantify the precise impact of these choices and adapt the state of the practice if necessary.

3. EFFECTS OF ERROR DISTRIBUTIONS

In order to evaluate the impact of the limitations outlined in the previous section we require a number of diverse biometric datasets, all of which have to be suitable for continuous authentication. Some differences in error distributions might be due to the biometric, while some can be attributed to specifics of a dataset. As such, we require datasets covering multiple biometrics and ideally several datasets per biometric. For this analysis we use 13 datasets collected by related work and 3 datasets collected for this study. Details of the datasets can be found in the appendix. In this section we investigate the previously described sixteen datasets with regard to the distribution of their errors. Based on the insights gained from this analysis we will discuss a number of novel metrics with regard to how well they capture these distributions.

3.1 Systematic Errors in the Wild

The most complete way to visualize the exact distribution of errors (both FAR and FRR) is a confusion matrix. A confusion matrix shows the fraction of accepted samples for each combination of template and samples (see Figure 4 for an example). As such, the TPR (i.e., 1-FRR) is shown on the diagonal and the remaining fields show the FAR for each combination of attacker and victim. The confusion matrix of an ideal system would be 1 on the diagonal and 0 otherwise. As discussed in Section 2, systematic false negatives (i.e., attackers that consistently remain undetected) are a more severe problem than a moderate, low-variance FAR for all attackers. This is due to the nature of continuous authentication, which requires an attacker to consistently fool the authentication system, rather than only succeed once.



(a) Touch Dataset (b) Mouse (own machine) (c) Gait II

Figure 4: Fraction of accepted samples for different user combinations, the values outside the diagonal reflect the FAR. The touch dataset shows an even distribution of the FAR, resulting in a low standard deviation. Both the mouse and gait datasets show more systematic errors, as indicated by few dark spots and a high standard deviation and kurtosis.

In our datasets we actually observe both of these scenarios, leading to a need to accurately distinguish them without the need of manually examining confusion matrix. Figure 4 suggests that the mouse movement and gait biometrics show a high number of extreme outliers for the FAR (as indicated by the dark spots off the diagonal). Conversely, the false accepts seem to be more evenly distributed between attackers for the touch input biometric, suggesting it would be better suited for continuous authentication from a pure security perspective.

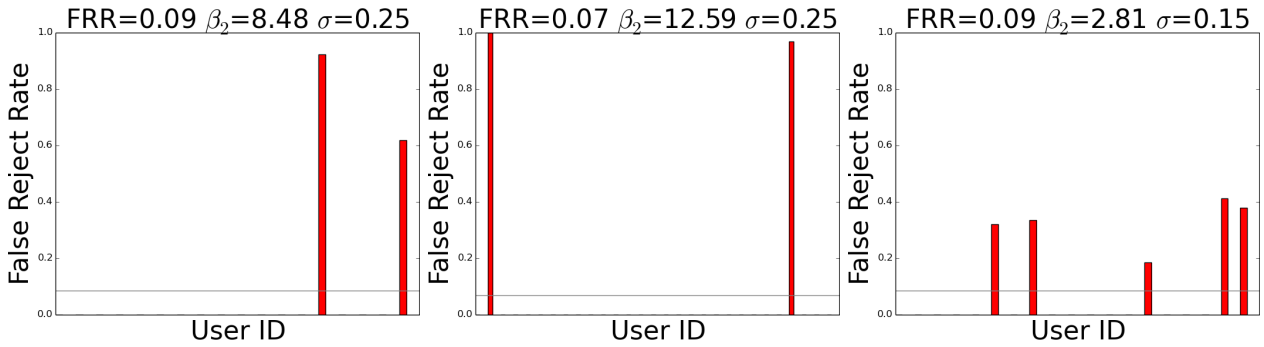
For the FRR we observe similar differences in distributions, although the consequences are different. Systematic rejections of individual users might indicate erratic behavior (such as excessive head movements or poor calibration for the eye movement biometric), while even distributions of errors suggest a lower distinctiveness of features in general. The former could be mitigated by examining the root cause of error for the affected users and, if these can not be fixed, authenticating users through a different mechanism. Multimodal authentication systems are particularly well-suited for this, as they can dynamically choose biometrics that work well for this specific user. As such, biometrics where the FRR is focused on few users might be easier to use in practice. Figure 5 shows the distribution of the FRR for different over-time datasets for the eye movement biometric. Errors are focused on few users given a short time-distance and start to evenly affect more users over two weeks.

3.2 Metrics to Quantify Systematic Errors

In this section we will discuss a number of statistical measures to better capture systematic errors and analyse how well they perform on our real-world data.

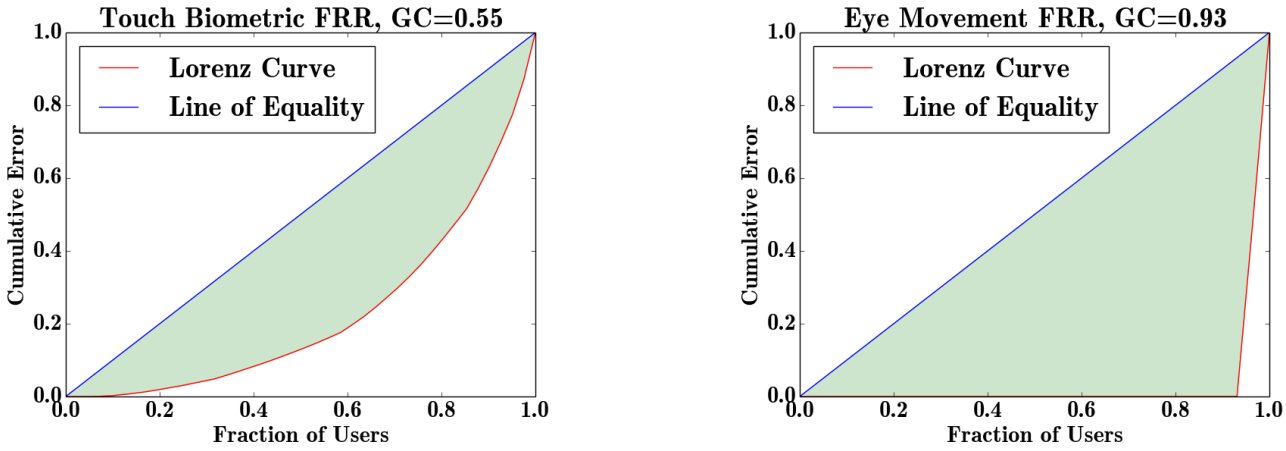
3.2.1 False Accept Rate

As discussed above, the false accept rate should ideally spread out evenly across attackers and therefore minimize systematic errors. In order to reflect systematic false negatives it might be an obvious choice to report the maximal FAR observed, this would then allow to give estimates of the maximal time it takes to find an attacker. However, Table 2 shows that this measure is 1 for the vast majority of datasets, suggesting at least some degree of systematic errors for most biometrics. In addition, it would unfairly penalize larger datasets, as the probability of the set including two very similar users increases with the sample size. This could be mitigated by reporting the fraction of undetected attackers (i.e., the fraction of user-attacker pairs with an FAR of 1, given as "1's" in Table 2). However, given the relatively small number of samples per user for each dataset, there might not be a statistical difference between an FAR of 1, and one very close to 1, suggesting that this feature would also be overly sensitive. Another candidate metric is the standard deviation of the sample. Table 2 shows that the standard deviation varies between 0.05 and 0.37. However, the standard deviation quantifies the variation in a dataset, but does not reveal whether this variation is due to a few extreme outliers (which would be problematic) or a high number of moderate outliers (which would be a less severe problem). This limitation can be mitigated by also taking into account the kurtosis of the sample. Kurtosis is the fourth standardized moment and is a measure of the tailedness of a distribution. As such, a high kurtosis indicates that the distribution tends to produce more extreme outliers. Combining standard deviation and kurtosis (i.e., an ideal distribution being low standard deviation and low kurtosis) seems to fit our required profile. Figure 7 shows datasets with similar standard deviation but different kurtosis. The first gait dataset shows systematic errors, indicated by a high kurtosis of 11.53 while the second one exhibits more random errors, leading to a lower value of 2.16. Despite this combination seeming fit for purpose, it would be difficult to use to accurately rank biometrics as any total ordering (i.e., preferring kurtosis over standard deviation or vice-versa) would be somewhat arbitrary. The Gini Coefficient (GC) has been proposed in 1912 as a measure of statistical dispersion to reflect the income distribution of a nation's residents [19]. A GC of 0 indicates a maximal equality of values (i.e., every resident having the same income), while a value close to 1 represents maximal inequality (i.e., one resident earning all the income). As a measure of inequality the GC is also intuitively applicable to capture types of error distributions, with a high GC reflecting more systematic errors. An intuitive geometric representation of



(a) Intra-Session (b) Inter-Session (c) Over Two Weeks

Figure 5: Distribution of the FRR between users for three different datasets based on the eye movement biometric using all features. While the average FRR is similar for all datasets the distributions are not. The two-weeks dataset shows moderate error rates for many users while the errors are concentrated on few users for the other two. This property is modelled by the kurtosis and to a lesser degree by the standard deviation.

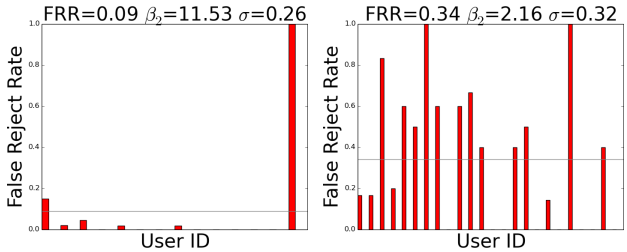


(a) (b)

Figure 6: False rejects are spread evenly for the touch input biometric and are focused on very few users for the eye movement biometric. This is reflected in the difference in Gini coefficients (0.55 vs 0.93).

the Gini Coefficient is the area between the Lorenz Curve (which, in our scenario, measures the total error contributed by the bottom x % of users) and the Line of Equality (which is the Lorenz curve of a system where all users contribute identical error rates). The GC is shown as the shaded area in Figure 6. The GC has two important properties that makes it a suitable metric: Its scale idependence means that it does not depend on the total or average error of a system, only the distribution of values. As such, it can be used to compare systems with different error rates. Conveniently, the GC always lies between 0 and 1, unlike standard deviation and kurtosis, which can take arbitrarily high values. In addition, it is population independent and does not depend on the number of samples in the dataset. This is of crucial importance, as the number of subjects in biometric datasets varies greatly and using only subsets of equal size seems infeasible due to authors rarely publishing their raw data.

Figure 8 shows the Gini Coefficient for the two most extreme cases we observe in our datasets. For the touch input biometric many attackers contribute to the overall FAR, while the eye movement biometric's intra-session dataset



(a) First Gait Dataset (b) Second Gait Dataset

Figure 7: Distribution of the FRR between users for the two gait datasets.

FAR is caused by very few extremely successful attackers.

Reducing security through strong features: It is interesting to note that the distribution of errors, and thereby the GC, does not simply depend on the biometric modality, but also the type of features used. When removing the pupil diameter, one of the most distinctive features of the eye

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

Table 2: Results of applying the new metrics to our datasets. As evidenced by the Gini coefficient, random errors are particularly prevalent for the touch input biometric, while eye movements are prone to systematic errors. We can also observe that not using the pupil diameter results in fewer systematic errors, as evidenced by a lower GC and lower kurtosis.

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

3.2.2 False Reject Rate

For the FAR, it is easy to agree on the fact that systematic errors are more problematic, as it leads to some attackers perpetually escaping detection. Determining the most favorable error distribution is not quite as obvious for the FRR. If most of the FRR is due to extreme outliers it might suggest that this is due to erratic user behavior, such as a bad calibration for eye tracking. In that sense, this scenario might be preferable, as this indicates a problem with a small number of users, rather than an overall problem of the system which manifests itself in all users. When the deployed system shows high error rates for some users, it might be possible to further explore the root cause of the errors (which could involve educating the user, but could also aid in improving the system itself). Reporting the fraction of users perfectly recognized by the system (given as "0's" in Table 2) would be an obvious approach to reflect this property, but Figure 7 shows why it would be quite noisy in practice. Using a combination of kurtosis and standard deviation would also suffer from the same problems as for the FAR, namely the difficulty of establishing a total order between systems.

Following the shortcomings of the other metrics, the Gini Coefficient can again be used to quantify where exactly a biometric recognition system lies between the extremes of purely systematic and purely random errors. Our data shows that the touch input biometric has the most even distribution of false rejects, exhibiting a GC of 0.55. The eye movement biometric generally shows the highest GC, with little change due to feature sets, time distance or tasks used. This might be explained by the fact that the biometric strongly relies on controlled user behavior, specifically requiring a good calibration and as few head movements as possible. If some users are better at achieving this optimal behavior it would explain this rather extreme concentration of errors. In addition, this type of behavior would likely be regardless of the feature set used or increased time distance between sessions.

3.3 Lessons Learned

The previous subsections have shown that error distributions vary wildly across different datasets. This observation is valid for both the FRR and the FAR, leading to different consequences. Out of the set of the metrics we analysed to augment the FAR/FRR the Gini Coefficient is the most promising due to its compactness and ability to provide an absolute ordering of systems. For the FAR, systems with a lower GC are desirable as this indicates false accepts that are spread relatively evenly across attackers, rather than enabling few attackers to perpetually escape detection. Our data shows that adding distinctive features, such as the pupil diameter for eye movement biometric decreases the EER, but at the same time increases the GC. This suggests

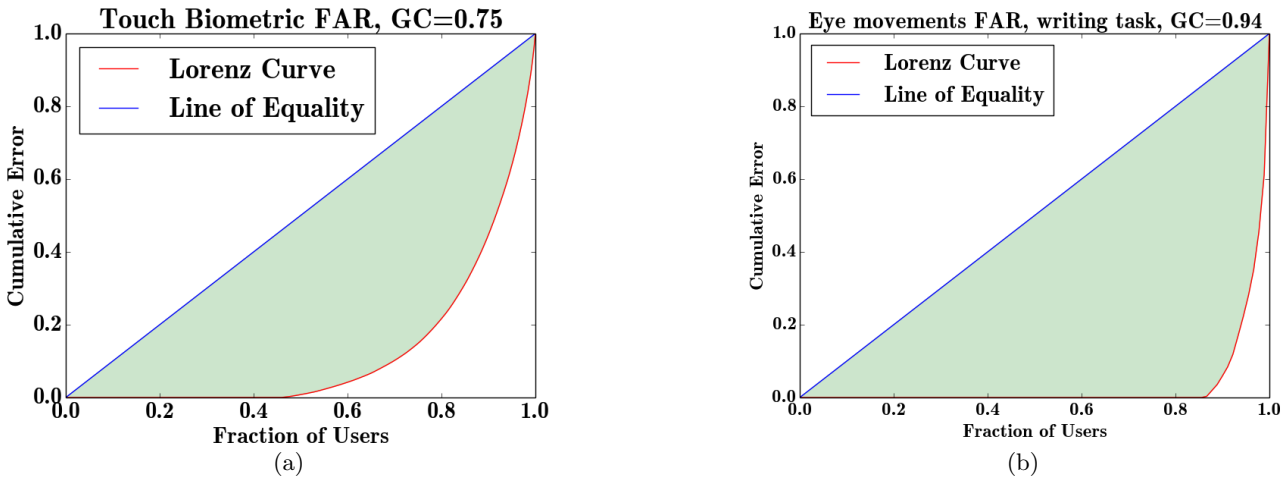


Figure 8: The different Gini coefficients draw attention to different error distributions. The touch biometric has a comparatively low GC of 0.75, which indicates largely random errors, while the eye movement biometric's higher GC of 0.94 suggests systematic errors which will lead to attackers consistently fooling detection.

that features that change little over the system's operation might be suitable to tell users apart in general, but confuses similar users more consistently, thereby leading to the aforementioned systematic errors. This insight is crucial during feature selection, at which point some distinctive features should even be dropped completely to avoid this scenario. As such, it is important to remember that not every change to a system that lowers the average error is actually beneficial to its security. For the FRR a high GC indicates erratic user behavior for a small number of users, an insight that can help improve either the system design or aid in avoiding this behavior during system operation. Overall, we recommend to closely monitor changes to GC when experimenting with different feature sets to evaluate whether any of them consistently lead to systematic errors. When publishing results, the GC should always be reported together with the mean EER/FAR/FRR in order to allow readers to take error distributions into account during their evaluation.

4. INFLUENCING ERROR RATES THROUGH TRAINING DATA SELECTION

In Section 2 we have shown that the majority of papers either randomly sample training data from the entire available dataset or merge data from all users (including the attacker) to form the negative class. It is well-known in related fields that error rates are systematically under-estimated when the temporal order of samples is not preserved when selecting training data. The precise impact is well-researched in the context of malware analysis, in which case past malware can be classified more accurately when signatures of future malware is included in the training data [3, 4]. Nevertheless, the precise impact has, to the best of our knowledge, not been quantified for biometric-based continuous authentication. Knowing the precise influence of these methodologies is important in order to assess whether a lower EER is due to a better system or excessive optimization through non-functional design decisions.

4.1 Quantifying the Impact on the EER

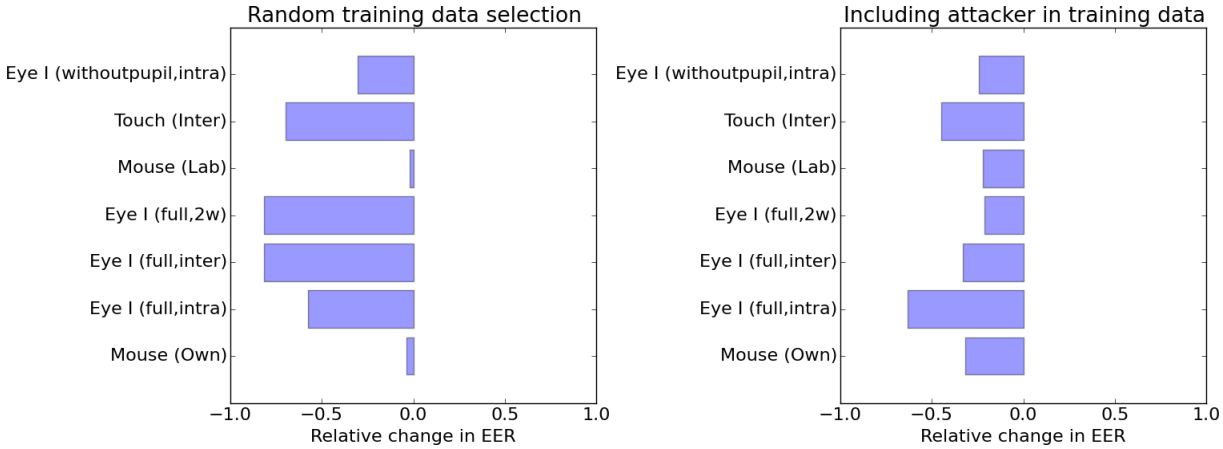
The two non-functional parameters most likely to impact error rates are the attacker modelling process and the division of training data. There are a number of valid choices for both, raising the question whether there is a seemingly "best" choice that leads to a minimization of (reported) error rates. In order to answer this question, we compute the EER for a number of datasets under different assumptions. We consider all combinations of the below parameters:

Number of aggregated samples: We statically choose a value of 100 for eye movement datasets, and 15 for the others (to reflect the lower sampling rate). We then aggregate samples based on a simple majority voting. Aggregating samples is common practice and a technique used in the original evaluation of all datasets we consider.

Dataset Division: We consider ordered and random division. For a single session, an ordered split uses the first half for training and the second half for testing. If two sessions are available, only the first is used for training. For the random split, we randomly select half the data for training. The process maintains the relative proportions of the classes to ensure roughly equal amounts of training data for each user. We then repeat the sampling and classification process 20 times to measure the effects of this selection.

Attacker Modelling: Anomaly detection requires a specialized classifier (such as a oneclass SVM), which makes it difficult to isolate the effects of this parameter alone. As such, we consider the "all users" and "except attacker" approaches. For the latter, we perform classifier training separately for each user and each attacker, while excluding the attacker from the training set. The negative class is instead created by the combination of all other users. The "all users" approach instead trains a single model per user, which includes positive data (from the legitimate user) and a single negative class (all other users). In both cases, we balance the positive and negative class as to not bias the classifier.

The results of our analysis are shown in Figure 9. Selecting training data randomly provides the biggest improvement,



(a) Data Division (b) Attacker Model

Figure 9: EERs decrease up to 80% when randomly selecting training data. Including the actual attacker in the negative class provides a reduction of up to 63%. The impact of random training data selection is particularly strong for datasets collected over longer time spans.

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

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

These results show that simply looking at the EER of a proposed system is insufficient, as it is skewed greatly by non-functional parameters that would not affect the performance of the system in a production environment. For example, if the exact same dataset (i.e., identical features and classifiers) were evaluated with random and ordered training data selection, one might favor one over the other (even though their practical performance would be identical). This is particularly alarming as our analysis (see Section 2) shows that out of 25 papers, 13 use at least one of the methodologies that we have shown to lead to systematic underestimation of error rates. In addition, a further 6 do not report how the error rates were obtained, which not only decreases confidence in the results but also impedes reproducing them and comparing them to related work. In order to inspire the highest confidence in their results researchers should exclude attackers from the negative class in their training data and choose the first part of their entire dataset for training, rather than sampling it randomly. In order to allow an easier comparison with some earlier work it would also be advisable to report error rates for different methodologies (such as random sampling) as well.

5. CONCLUSION

In this paper we have provided a systematic analysis of the methodology used to evaluate behavioral biometrics for continuous authentication. Our analysis shows that most papers present the mean of standard metrics, specifically the Equal Error Rate (EER) and False Accept Rate (FAR), but don't give any insights of their precise distributions. We argue that some errors, specifically systematic false negatives, are particularly severe in the context of continuous authentication. The analysis of 16 real-world datasets shows that some biometrics, such as touchscreen inputs, exhibit mostly random errors, leading to the eventual detection of attackers due to the process of continuous authentication. Others, such as gait patterns, tend to produce more systematic errors, thus allowing some attackers to consistently avoid detection. In order to allow the comparison of different systems with regard to this property without requiring manual inspection, we discuss a number of candidate metrics. As a result of this discussion we propose the use of the Gini Coefficient (GC) to capture different distributions of both the FAR and FRR. The application of the GC to our datasets reveals that the addition or removal of certain features can greatly impact the biometric's error distribution. Specifically, using the pupil diameter for classification reduces the system's average EER, but also greatly contributes to systematic errors, thereby suggesting it might even reduce overall security. Based on these insights, the GC can not only be used to compare the security of different systems, but can also guide researchers during evaluation of different classifiers, bioemtrics and featuresets. We therefore recommend that authors report the GC as well as established metrics in order to provide information about error distributions as well.

We also quantified the impact of a number of different machine learning methodologies on a system's error rates.

We identify two main factors, the selection of training data (specifically, random versus ordered split) and the inclusion of imposter data in the negative class. While these effects are somewhat well-known in other fields, their precise impact has not been quantified in the context of continuous authentication. Our analysis shows that random sampling of training data can reduce the EER by up to 80%, while inclusion of imposter data provides a reduction of up to 63%. These results highlight a particular problem, as 13 of the 25 papers we analyzed used a methodology that we have shown to lead to systematic underestimation of error rates and a further 6 did not report which methodology was used at all.

Our results highlight that it is inadequate to compare biometric systems simply by their EERs. Instead, it is crucial to take into account both the distribution of errors, as well as the design decisions that were made when simulating system operation on a static dataset.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M50659X/1].

6. REFERENCES

[1] A. A. E. Ahmed and I. Traore. A new biometric technology based on mouse dynamics. *Dependable and Secure Computing, IEEE Transactions on*, 4(3):165–179, 2007.

[2] H. J. Ailisto, M. Lindholm, J. Mantyjarvi, E. Vildjiounaite, and S.-M. Makela. Identifying people from gait pattern with accelerometers. In *Defense and Security*, pages 7–14. International Society for Optics and Photonics, 2005.

[3] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon. Are your training datasets yet relevant? In *International Symposium on Engineering Secure Software and Systems*, pages 51–67. Springer, 2015.

[4] K. Allix, T. F. D. A. Bissyande, J. Klein, and Y. Le Traon. Machine learning-based malware detection for android applications: History matters! Technical report, University of Luxembourg, SnT, 2014.

[5] S. Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC)*, 3(3):186–205, 2000.

[6] P. Bours and S. Mondal. Performance evaluation of continuous authentication systems. *IET Biometrics*, 4(4):220–226, 2015.

[7] A. Brajdic and R. Harle. Walk detection and step counting on unconstrained smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and ubiquitous computing*, pages 225–234. ACM, 2013.

[8] Ş. Budulan, E. Burceanu, T. Rebedea, and C. Chiru. Continuous user authentication using machine learning on touch dynamics. In *International Conference on Neural Information Processing*, pages 591–598. Springer, 2015.

[9] 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*, pages 1393–1402. ACM, 2015.

[10] Z. Cai, C. Shen, M. Wang, Y. Song, and J. Wang. Mobile authentication through touch-behavior features. In *Biometric Recognition*, pages 386–393. Springer, 2013.

[11] M. O. Derawi, C. Nickel, P. Bours, and C. Busch. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, pages 306–311. IEEE, 2010.

[12] B. Draffin, J. Zhu, and J. Zhang. Keysens: Passive user authentication through micro-behavior modeling of soft keyboard interaction. In *International Conference on Mobile Computing, Applications, and Services*, pages 184–201. Springer, 2013.

[13] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *Proceedings of the Network and Distributed Systems Security Symposium (NDSS)*, 2015.

[14] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. Looks like eve: Exposing insider threats using eye movement biometrics. *ACM Transactions on Privacy and Security*, 19(1):1, 2016.

[15] T. Feng, J. Yang, Z. Yan, E. M. Tapia, and W. Shi. Tips: Context-aware implicit user identification using touch screen in uncontrolled environments. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, page 9. ACM, 2014.

[16] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *Information Forensics and Security, IEEE Transactions on*, 8(1):136–148, 2013.

[17] D. Gafurov, K. Helkala, and T. Søndrol. Biometric gait authentication using accelerometer sensor. *Journal of computers*, 1(7):51–59, 2006.

[18] H. Gascon, S. Uellenbeck, C. Wolf, and K. Rieck. Continuous authentication on mobile devices by analysis of typing motion behavior. In *Sicherheit*, pages 1–12. Citeseer, 2014.

[19] C. Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1, 1912.

[20] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon. Self-calibrating view-invariant gait biometrics. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(4):997–1008, 2010.

[21] Z. Jorgensen and T. Yu. On mouse dynamics as a behavioral biometric for authentication. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 476–482. ACM, 2011.

[22] T. Kinnunen, F. Sedlak, and R. Bednarik. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 187–190. ACM, 2010.

[23] J. Mäntyjärvi, M. Lindholm, E. Vildjiounaite, S.-M. Mäkelä, and H. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *Acoustics, Speech, and Signal Processing, 2005.*

Proceedings.(ICASSP'05). IEEE International Conference on, volume 2, pages ii–973. IEEE, 2005.

[24] S. Mondal and P. Bours. Continuous authentication using mouse dynamics. In *Biometrics Special Interest Group (BIOSIG), 2013 International Conference of the*, pages 1–12. IEEE, 2013.

[25] M. Pusara and C. E. Brodley. User re-authentication via mouse movements. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 1–8. ACM, 2004.

[26] K. B. Rasmussen, M. Roeschlin, I. Martinovic, and G. Tsudik. Authentication using pulse-response biometrics. In *NDSS*, 2014.

[27] L. Rong, D. Zhiguo, Z. Jianzhong, and L. Ming. Identification of individual walking patterns using gait acceleration. In *2007 1st International Conference on Bioinformatics and Biomedical Engineering*, pages 543–546. IEEE, 2007.

[28] A. Roy, T. Halevi, and N. Memon. An hmm-based behavior modeling approach for continuous mobile authentication. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3789–3793. IEEE, 2014.

[29] P. Saravanan, S. Clarke, D. H. P. Chau, and H. Zha. Latentgesture: active user authentication through background touch analysis. In *Proceedings of the Second International Symposium of Chinese CHI*, pages 110–113. ACM, 2014.

[30] D. A. Schulz. Mouse curve biometrics. In *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, pages 1–6. IEEE, 2006.

[31] C. Shen, Y. Zhang, Z. Cai, T. Yu, and X. Guan. Touch-interaction behavior for continuous user authentication on smartphones. In *2015 International Conference on Biometrics (ICB)*, pages 157–162. IEEE, 2015.

[32] M. Soriano, A. Araullo, and C. Saloma. Curve spreads-a biometric from front-view gait video. *Pattern Recognition Letters*, 25(14):1595–1602, 2004.

[33] E. Vildjiounaite, S.-M. Mäkelä, M. Lindholm, R. Riihimäki, V. Kyllönen, J. Mäntyjärvi, and H. Ailisto. Unobtrusive multimodal biometrics for ensuring privacy and information security with personal devices. In *International Conference on Pervasive Computing*, pages 187–201. Springer, 2006.

[34] A. Weiss, A. Ramapanicker, P. Shah, S. Noble, and L. Immohr. Mouse movements biometric identification: A feasibility study. *Proc. Student/Faculty Research Day CSIS, Pace University, White Plains, NY*, 2007.

[35] H. Xu, Y. Zhou, and M. R. Lyu. Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 187–198, 2014.

[36] X. Zhao, T. Feng, and W. Shi. Continuous mobile authentication using a novel graphic touch gesture feature. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–6. IEEE, 2013.

[37] N. Zheng, A. Paloski, and H. Wang. An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 139–150. ACM, 2011.

APPENDIX

A. DATASETS

In this section we describe the datasets used to carry out to evaluate our new metrics (see Section 3). We use 13 datasets obtained from the authors of previously published work and augment them with three datasets gathered specifically for this study.

A.1 Gait Biometric

There are a number of approaches to capture human bait patterns, they are typically based on video recordings [32, 20] or accelerometer data [23, 17, 2]. As accelerometer data can be readily captured with smartphones (and then be used to protect the device after a theft), we focus on this approach. We adapt the classification process of [17] to support continuous authentication.

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

Using this data we obtained an average EER of 8.44%.

A.2 Second Gait Dataset

The second gait dataset was obtained from the authors of [7]. The set contains data from 27 participants that walked along a footpath at three different paces. While the data was collected for the purpose of evaluating step-counting algorithms, the data format makes it suitable for authentication as well. The data was collected through the accelerometer of a smartphone held in various positions (in a front or back trouser pocket, in a backpack/handbag, or in a hand with or without simultaneous typing). Not all sensor positions are available for each subject. In order to remove potential distinguishing information resulting purely from the sensor position, we only use the subset of traces in which the device was held by the subject without simultaneous typing, limiting the number of subjects to 24. The data was collected at a rate of 100Hz, with an average of 4400 samples (or 44 seconds) per subject. For each subject we extract the portion of the trace during which the subject was walking, using the timestamps provided as part of the dataset. As the first half of the data is used for training it contains mostly slow movements, unlike the testing timeframe during most of which the subjects were moving at a quicker pace.

The system shows an EER of 28.4%. This relatively high value (especially compared to the dataset collected by us)

			Hyperparameters					Available Online	
Ref	**Biometric**	**Classifier**	**Values**	**Method**	**att-model**	**training data**	**sample agg**	**Data**	**Code**
[16]	Touch	SVM,knn	✓	CV	all users	ordered	weighted	✓	✓[1]
[15]		knn	✓	✗	✗	ordered	✗	✗[2]	✗
[35]		SVM	✗	✗	subset	CV-10	majority	✓[3]	✗
[8]		DT	✓	GS+CV	✗	CV-3	✗	✗[2]	✗
[18]		SVM	✗	✗	✗	✗	✗	✗	✗
[36]		sim-score	N/A	N/A	AD	random	N/A	✗	✗
[12]		NN	✓	✗	AD	ordered	N/A	✗	✗
[9]		knn	(✓)	✗	no-attacker [4]	✗	✗	✗	✗
[10]		NN,SVM	✓	✗	all users	random [5]	✗	✗	✗
[29]		SVM,RF	✗	✗	AD	✗	✗	✗	✗
[28]		HMM	✓	CV-5	all users	ordered	mean	✗[2]	✗
[31]		SVM	✓	✓	AD	ordered	N/A	✗	✗
[13]	Gaze	SVM,knn	✓	GS+CV	all users	CV-5	majority	✗	✗
[22]		UBM	✓	✓	all users	✗	N/A	✗	✗
[14]		SVM,knn	✓	GS+CV	AD, all users	CV-5	majority	✗	✗
[26]	Pulse Response	SVM,knn	✓	✓	all users	CV-5	N/A	✗	✗
[11]	Gait	sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[33]		sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[2]		sim-score	N/A	N/A	AD	ordered	N/A	✗	✗
[27]		sim-score	N/A	N/A	AD	random	N/A	✗	✗
[25]	Mouse	DT	N/A	N/A	all users	ordered	weighted	✗	✗
[1]		NN	✓	✓	no-attacker	random	N/A	✗	✗
[30]		sim-score	N/A	N/A	AD	✗	N/A	✗	✗
[37]		SVM	✗	✓	no-attacker	ordered	mean	✗	✗
[24]		SVM	✗	✗	all users	random	N/A	✗	✗

✓Reported (✓) Partially reported ✗Not reported
Unless indicated otherwise, only the mean of each metric is reported
[1] Only feature extraction
[2] Uses data from [16]
[3] Dead URL 10/07/2016
[4] Training data consists of all other users, excluding the attacker
[5] Sampling repeated 10 times

Table 3: Simulation Design Choices in Related Work

might also be a consequence of a mismatch between training and test data (which were gathered at different walking speeds).

A.3 Mouse Movement Biometric

In addition to the gait data, we conduct an experiment to collect volunteers' mouse movements. Our experimental design is conceptually close to that in [34]. During the experiment, each participant was shown 25 rectangles arranged in a 5x5 grid, one of which was red. The user is then asked to click on the red rectangle. This task is repeated 200 times, with the red rectangle appearing in a new, random location for each iteration. The random seed to generate the sequence was kept identical for all users in order to limit the effects of the rectangle's position on our features. The size of the window displaying the rectangles was fixed in order to avoid any distinctiveness created solely by different screen resolutions. In order to control for artificial bias created by different input devices [21] we collect two datasets. The first set was obtained by sending our software to subjects, to be run on their own home or work machine. For the second set we invited a (different) set of volunteers to take part in the experiment on our lab machine. If any features are more distinctive in the first set this would imply that their distinctiveness is at least partially due to the properties of different devices, rather than differences in user behavior.

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

A.4 Eye Movement Biometric

The eye movement biometric, as proposed in [13], is based on involuntary fixational eye movements. The distinctiveness of eye movements is not limited to a certain task and features can be computed regardless of screen content. As such, the biometric can be used in a continuous authentication scenario without limiting the user. The set of 20 features used in the

paper reflect the properties of microsaccades (high velocity and acceleration), the steadiness of the gaze and both static pupil diameter as well as the pupil diameter's changes over a short time. The pupil diameter generally outperforms the remaining features in terms of distinctiveness.

A.5 First Eye Movement Dataset

The first dataset was obtained from the authors of [13]. In order to test the features' time stability, three identical sessions are performed, with a time distance of one hour and two weeks, respectively. In line with the presentation in the paper we form three datasets from the sessions: The intra-session set contains data only from the first session and involved 30 subjects. The inter-session set combines the second and third session (i.e., with the two parts being one hour apart) and the first and second session form the 2-weeks dataset.

In order to reflect different threat models the authors propose the use of different featuresets, specifically describing a set that excludes features based on the pupil diameter. Using this reduced feature set increases the EER from 6.9% to 19.83% as some identifying information is lost. The combination of three sessions and two featuresets results in six distinct datasets.

A.6 Second Eye Movement Dataset

The second dataset was provided by the authors of [14] and extends the previous study with several real-world tasks. These tasks include reading, writing, web browsing and watching two different videos. We consider each of these tasks separately (by sampling training and testing data from the same task) and jointly (by merging all tasks before sampling training and testing data).

A.7 Touch Input Biometric

The touch input dataset is based on the data shared in [16]. The biometric's features describe the properties of swiping motions on touchscreens, including their position, curvature and pressure. Data was collected over two weeks, resulting in an intra-session, inter-session and 1-week dataset. The error rates range from 0% for intra-session authentication to 4% when authentication is performed a week after enrolment. As we are interested in determining the distribution and causes of errors we do not use the intra-session dataset for our comparison.