Robust Performance Metrics for Authentication Systems

Shridatt Sugrim, Can Liu, Meghan McLean, Janne Lindqvist
Rutgers University

***Abstract*—Research has produced many types of authentication systems that use machine learning. However, there is no consistent approach for reporting performance metrics and the reported metrics are inadequate. In this work, we show that several of the common metrics used for reporting performance, such as maximum accuracy (ACC), equal error rate (EER) and area under the ROC curve (AUROC), are inherently flawed. These common metrics hide the details of the inherent trade-offs a system must make when implemented. Our findings show that current metrics give no insight into how system performance degrades outside the ideal conditions in which they were designed. We argue that adequate performance reporting must be provided to enable meaningful evaluation and that current, commonly used approaches fail in this regard. We present the unnormalized frequency count of scores (FCS) to demonstrate the mathematical underpinnings that lead to these failures and show how they can be avoided. The FCS can be used to augment the performance reporting to enable comparison across systems in a visual way. When reported with the Receiver Operating Characteristics curve (ROC), these two metrics provide a solution to the limitations of currently reported metrics. Finally, we show how to use the FCS and ROC metrics to evaluate and compare different authentication systems.**

I. Introduction

Many authentication systems utilizing machine learning have been proposed (see Table II). However, there is no clear agreement in the community about how these systems should be evaluated or which performance metrics should be reported. Specifically, publications often report misleading single-number summaries which include true positive rate (TPR), false positive rate (FPR), equal error rate (EER), area under ROC curve (AUROC), and maximum accuracy (ACC). Figure 1 enumerates the reporting rates of common metrics for thirty-five recent publications.

Improving the metrics and reporting methods can resolve two primary obstacles to the evaluation of authentication systems. These obstacles are (1) skew in the distributions used to train and evaluate the systems, and (2) misleading comparisons that arise from the reported metrics. For example, skew within the population of study participants can artificially inflate the maximum accuracy. Additionally, misleading comparisons can result from commonly reported metrics. For example, it is inappropriate to conclude that one system performs better than another by comparing an EER of 0.05 to 0.10. Similarly, an accuracy of 80% versus 90% does not allow clear inferences about the system performance.

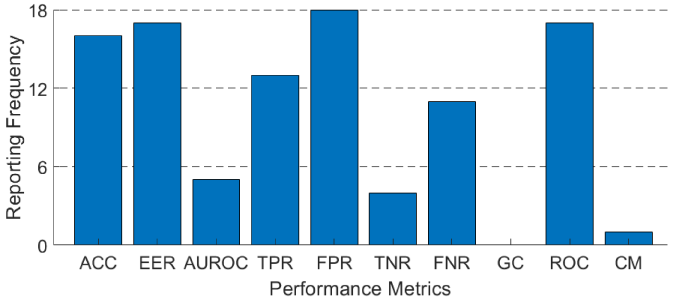

Fig. 1: The frequency of reported metrics from Table I for thirty-five recent publications surveyed from venues listed in Table II. Classical biometric and detection summaries such as EER and TPR are often reported because they are widely used in the literature. The FPR is the most reported because it is the common element for two frequently reported metric pairs, the (TPR,FPR) and (FPR,FRR).

We show the following three primary flaws with existing metrics: 1) It is incomplete to report performance using solely single-number metrics, like the ACC, FPR, TPR, FAR, and FRR. Single-number summaries hide the details of how and what errors occurred. For example, because if a system was trained on mostly unauthorized users' data, it will learn to recognize unauthorized users very well and may not recognize authorized users. 2) Reporting performance results without the parameters of the model hinders the implementation of the system. The system cannot be faithfully replicated when only the performance is reported. 3) Performance comparisons cannot be made when using single-number summaries derived from the ROC. One cannot conclude that one system will perform better than another in a target application by direct comparison of the EER and other ROC-derived summaries.

In this paper, we uniquely propose and demonstrate how the ROC, combined with the unnormalized Frequency Count Scores (FCS) (shown in Figure 2), aids in the ability to understand the trade-offs for authentication performance and adequately evaluate the proposed approach. The major contributions of the paper are as follows:

- We show how commonly used metrics in authentication systems, including TPR, FPR, EER, AUROC, ACC, and GC are inherently flawed metrics for understanding authentication system performance.
- We survey eleven top publication venues where au-

Network and Distributed Systems Security (NDSS) Symposium 2019
24-27 February 2019, San Diego, CA, USA
ISBN 1-891562-55-X
https://dx.doi.org/10.14722/ndss.2019.23351
www.ndss-symposium.org

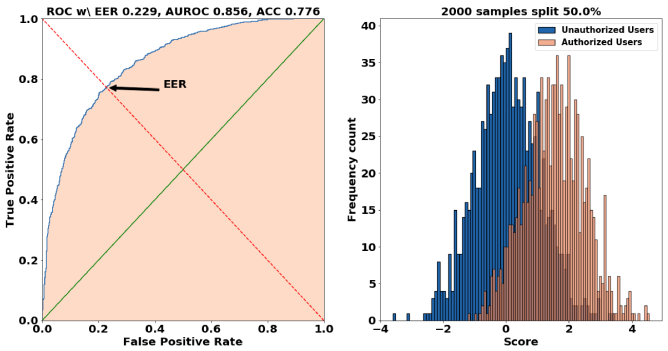

Fig. 2: The left figure is the ROC curve of an authentication system. The right figure is its corresponding FCS that displays score distributions for measurements of authorized and unauthorized users. The ROC curve is computed by varying the value of the threshold required to grant access and computing the true positive rate (TPR) and the false positive rate (FPR). The FCS is a histogram of measurement scores separated by ground truth. In this FCS figure, the blue histogram represents unauthorized users' scores, determined by the ground truth of the measurement. The red histogram in the FCS figure represents authorized users' scores.

thentication systems based on machine learning are proposed. We identified thirty-five recent (2016-2018) proposed authentication systems and identified their reporting themes and common flaws.

- We show how any single-number summary provides incomplete information for the evaluation of authentication systems.
- We propose unnormalized Frequency Count Scores (FCS) as an augmentation to current authentication metrics that enable visual comparison and identification of some errors.
- We show how using the FCS with the ROC can further solve the limitations associated with current authentication system metrics.
- We demonstrate flawed comparisons with existing datasets by reimplementing the proposed systems.

II. Review of Recent Authentication Systems

To determine the current state of performance metric reporting, we surveyed recent research published in top venues. The selection criteria for including papers in the review was the following: (1) The article was published in a top venue for systems security, mobile computing, human-computer interaction, or pattern recognition for authentication. These venues included NDSS, CCS, CHI, IMWUT/UbiComp, INFOCOM, MobiCom, MobiSys, SOUPS, SS&P (Oakland), USENIX Security and Pattern Recognition journal. Machine learning venues (e.g. NIPS) were not included due to their primary focus on algorithms and lack of attention to system applications. (2) In order to evaluate current practices, the paper had to be published within the last 2 years (2016 to 2018). (3) The paper had to propose an authentication scheme. Specifically, the paper had to use machine learning to label users as authorized and unauthorized (as opposed to identifying users from a group). Although we identified many related papers (n = 58), only 35 proposed an authentication scheme and were included in the review. We note that we did not find any publications matching our criteria from USENIX Security.

	Definition	Description
TP	True Positive	Authorized legitimate users count.
FP	False Positive	Authorized illegitimate users count.
TN	True Negative	Denied illegitimate users count.
FN	False Negative	Denied legitimate users count.
CM	Confusion Matrix	Table of contingency counts.
ACC	Maximum Accuracy	Probability of a correct declaration.
TPR	True Positive Rate	How often a legitimate users is authorized.
TNR	True Negative Rate	How often an illegitimate user is denied.
FPR (FAR)	False Positive Rate (False Accept Rate)	How often an illegitimate user is authorized.
FNR (FRR)	False Negative Rate (False Reject Rate)	How often a legitimate user is denied.
ROC	Receiver operator characteristic curve	Curve of (TPR, FPR) by varying threshold
ERR	Equal Error Rate	The point that TPR equals 1-FPR
AUROC	Area under the ROC curve	Probability of scores of random legitimate users are higher than illegitimate user.
GC	Gini-coefficient	Calculated from the AUROC
FCS	Unnormalized frequency count of scores	Histogram of scores separated by ground truth

TABLE I: Performance metric name abbreviations

Venue	References
CCS	[34], [44], [55]
CHI	[12], [33], [41], [42]
IMWUT/UbiComp	[11], [20], [25], [35], [48]
INFOCOM	[8], [36], [45], [51], [53]
MobiCom	[15], [32]
MobiSys	[7], [31]
NDSS	[5], [17], [49], [50]
Pattern Recognition	[2], [9], [18], [24], [38], [40], [54]
SS&P (Oakland)	[46]
SOUPS	[10], [30]

TABLE II: Publications surveyed grouped by venues

In order to find these articles, one researcher used the Google Scholar search engine to limit results to these venues and included the following search terms: authentication, behavioral, biometric, machine learning, password, recognition, and access. A second researcher separately reviewed the venue proceedings using the search terms in order to generate a complete list of related work.

In the thirty-five publications that were surveyed, we discovered no uniform approach for reporting metrics. However, there were several recurring themes. Figure 1 shows that the most common metric reported is the FPR since the FPR is the common element in two different but related metric pairs: the (TPR,FPR) and (FPR,FRR) pairs. These pairs are often reported when one value is held constant and the other is minimized (e.g. fixing the FPR and adjusting the system parameters until the TPR is optimized). There is often no justification for the value that is chosen to be held constant. Another frequently reported metric is the EER. It is often reported for comparison with existing systems in the literature. Unfortunately, without a uniform approach, we cannot make comparative quantitative conclusions about the performance across all the proposed systems.

Sixteen (less than half) of the publications reported the ROC. Eleven of the publications that reported the ROC had

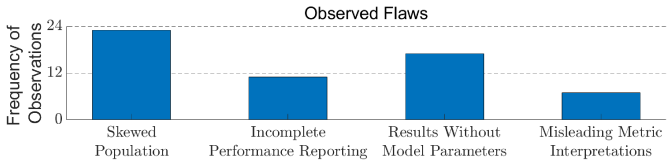

Fig. 3: The flaws noted in Section II-A occur several times in many of the publications. We note that the common practice of recruiting N participants and then electing one as the authorized user often leads to skewed measurement populations. Twenty-three of the thirty-five papers surveyed had skewed measurement populations.

measurement populations that were skewed (here measurement population is defined as the set of measurements from users). In some of these cases we concluded that the effects of the skew did not impact the validity of the claims because the performance claims were based on the ROC.

A. Common Flaws

We observed three flaws that were common to many publications. The three flaws are as follows:

1) Flaw 1: Incomplete performance reporting. Reporting based solely on single-summary metrics is incomplete. For example, the maximum accuracy (ACC) metric does not identify the type of user (authorized or unauthorized) an error was made on. For example, an accuracy of 90% does not mean that the system makes errors 10% of the time. A system may have been tested on data it would almost always get correct. Specifically, if a system was trained on mostly unauthorized users' data, it will learn to recognize unauthorized users very well. If this system is only tested on unauthorized users, the ACC metric will be very high. However, because this system was built on only unauthorized users' data, it may not recognize authorized users very well. The system will not be trained to identify authorized users' data because it has no model for authorized users. The model is thus incomplete when trained on data that is either mostly authorized or unauthorized users.

Comparisons made solely based on TPR, FPR, or FRR are also incomplete. It is not possible to tell if the optimized value is the result of better discrimination between users or simply an adjustment of system parameters for the purpose of inflating a metric. Since the (TPR,FPR) and (FPR,FRR) are the result of a specific compromise between the two kinds of errors. One can trade one error type for another by adjusting the parameters (e.g. threshold), without improving overall performance.

It is imperative that authors incorporate this knowledge into the interpretation of their metrics. If the authors do not report the frequency of authorized and unauthorized users' data, then the ACC metric provides little information about the system. Of the thirty-five publications surveyed, eleven directly exhibited this flaw. Several others made performance claims based only on one of these metrics but also reported other metrics for comparison.

2) Flaw 2: Results without model parameters. Performance results reported without the parameters of the model hinders the implementation of the system. The confusion matrix (CM), and all metrics derived from it (e.g. the ACC or TPR), depend on the threshold used to obtain the results. For metrics derived from the confusion matrix, the system cannot be implemented when only the performance is reported. Researchers tend to determine their own thresholds in isolation when they design a system (they cannot know what an implementer would need). If the threshold is unknown, the conditions under which the original research was completed cannot be replicated. We recommend that the ROC and FCS are reported to surpass this limitation. The ROC and FCS show how the system responds to changes in the threshold and covers a range of possible thresholds enabling implementers to choose the threshold that is right for their application.

Every machine learning algorithm has parameters that, when adjusted, significantly change their behavior (e.g. slack factor, size and number of hidden layers, K). Since machine learning usage is often off-the-shelf, it is easy to overlook the necessary parameters required to use the system even though these parameters control the behavior of one of the most critical parts of the system. In our survey, seventeen of the thirty-five publications left out the parameters of the machine learning algorithms used. Without these parameters, the task of replicating the system becomes much more difficult.

In one case, a publication studied the effects of environment on the system performance by comparing the accuracy of the system in several environments. Based on the described methods, the test data had a positive bias as the authors elected a very small subset of the participants to be attackers. The publication did not disclose whether the threshold remained the same in all environments tested. The authors concluded that since the accuracies were less than 5% from each other, that the environmental effects were negligible. However, because the population was skewed, that less than 5% difference could also be accounted for by shifting the threshold in each of the environments. Since the data was mostly comprised of positive samples from authorized users, moving the threshold up or down could yield more or less correct decisions purely due to sampling effects.

3) Flaw 3: Misleading interpretations. Performance comparisons cannot be made when using single-number summaries derived from the ROC. Direct comparison of the EER (and other ROC-derived summary metrics) does not show whether one system performs better than another. Two systems with similar values for these metrics can have very different ROCs. Because the ROCs are different, the performance will be different when implemented. When a system is implemented, it must be implemented with a target application in mind (e.g. banking or loyalty discounts). These different target applications come with requirements for the amount of false positives (wrong people allowed in) or the amount of false negatives (correct people kept out) that the application can tolerate. In our survey, seven of the thirty-five papers drew direct comparisons about system performance by comparing one of these summary values. Twenty-two of the thirty-five papers reported one of the metrics with the expectation that similar systems could be compared using these summaries.

Systems with different ROCs behave differently when implemented. These differences can lead to unexpected behavior because the error rates of the implemented system may differ from the error rates the implementer expected.

As a toy example, consider a case in which an implementer desired to improve the performance of an application that used Touchalytics [16] and required the false positive rate to be 0.1. In this scenario, the implementer may consider switching to SVC2004 [28]. The implementer may look at the EER of SVC2004 which is 0.185 and compare this to the Touchalytics EER of 0.198. From this comparison, an implementer would conclude that the system's performance will improve. He or she would be unpleasantly surprised when their true positive rate dropped from 0.8 to 0.6 at the false positive rate their application required. Therefore, comparing systems on these metrics can lead to detrimental, real-world security consequences.

In several of the surveyed publications the EER and AUROC were used in two different ways. In the first use case, authors made direct claims about the relative behavior of two or more systems either by comparing the proposed system to an existing system which uses the same types of measurements or by adjusting the parameters for their own measurements, to determine the impact on the metric. In some cases, authors concluded that a change in parameter had no impact on performance because the metric was unchanged. However, because they did not include the ROC, we do not know what effect the changes had on the ends of the curve. In contrast, some authors reported the ROC for multiple parameterizations of their system demonstrating that their systems behavior was predictable over a wide range of parameterizations, even though they did not make this claim.

In the second use case, the EER or AUROC is reported with the expectation that these metrics will be used to compare the proposed system to other competing systems. This use case only allows for naïve comparisons such as those detailed in Section II-B. Since the second use case requires knowing what the authors intended the reported metric to be used for, we did not count these cases as evidence of flaw 3.

B. A naïve comparison

Table III enumerates the top five systems based on the reported performance metrics. We consider the three most common metrics: the EER, ACC, and FPR. For any single metric, this comparison fails to produce a meaningful result for several reasons.

- Not all publications report the same metric. Comparison across different metrics does not have a meaningful interpretation.
- Individually using any of the three above-mentioned metrics can lead to flawed conclusions because no individual metric captures the complete performance.
- Just because a system optimizes a metric, does not mean that it can be utilized in the target application (e.g. an approach implemented for keyboards may not work on touchscreens directly).

It is clear that such a naïve comparison cannot lead to an informed comparison of the proposed systems. It is even difficult to identify if a system is suitable because some of the metrics fail to provide information that is relevant to the context of the target application (e.g. an FPR may not be achieved at a target TPR).

EER %	ACC %	FPR %
0.00 [9]	99.30 [55]	0.00 [11]
0.34 [24]	98.61 [32]	0.01 [9]
0.59 [2]	98.47 [51]	0.10 [35]
0.95 [40]	98.00 [53]	0.10 [5]
1.26 [55]	97.00 [42]	0.10 [40]

TABLE III: The top five authentication systems according to a naïve comparison of their best reported values for EER, ACC, and FPR metrics. These metrics are reported the most often but rarely yield a meaningful comparison. There is no clear winner as each of the top five performers in each category varies significantly.

Of importance, system evaluation requires the ability to evaluate the potential security trades-offs of a system. Instead of answering the question, "Has the system produced a single metric that has surpassed a seemingly adequate threshold?" implementers need to answer the question, "Can this system be tuned to meet the needs of my application such that reported metrics show possible security trade-offs?" Many of the current metrics that are reported fail to answer the latter question.

These flaws occur in many of the publications surveyed. Figure 3 enumerates the observation frequency of each of the described flaws. We also note that almost two-thirds of publications have skewed measurement populations. It is common practice to recruit N participants, take M number of measurements from each of them, and then elect one participant as the authorized user. When this is done, the measurement populations are skewed because there are $N-1 \times M$ measurements from unauthorized users, and only M measurements from authorized users. In some cases, we were unable to assess whether the measurement population was skewed because the publication did not report the sources of measurements. This was the case in *23 of the 35 publications* we reviewed. However a few reported balanced accuracy in an attempt to compensate for the skew. As we will see, had the FCS been reported, we would have been able to visually assess if there was measurement population skew. One of the publications actually reported a normalized FCS but did not use it for analysis.

III. Related Work

We discovered only one publication in the systems security community that has studied how performance metrics are reported [13]. They studied flaws that occur in reporting for continuous authentication systems. They note that the EER is among the most popular metrics and observe the misleading characteristics of only reporting the EER and false negative rate (FNR). They also note that data sets are rarely made available, which creates a barrier to follow-up analyses. They additionally advocate for the use of the Gini Coefficient (GC) which is functionally related to the AUROC. We show that the GC and AUROC are also flawed metrics that hinder the comparison and implementation of a system and we instead advocate for the combination of FCS and the ROC.

Bonneau et al. [3] compared different types of authentication systems to text passwords with qualitative metrics in usability, deployability, security. They provide a wide range of real-world constraints but they did not provide quantitative

approaches to evaluate the metrics. In contrast, we focus on quantitative metrics in this paper.

The efficacy of the EER in communicating performance has been questioned in other fields [37]: the EER has the significant disadvantage of representing only a single point in the performance continuum and that this misrepresents the capabilities of a system. The paper [37]: argues for the ROC as the main performance metric but does not consider how measurements are separated, nor the utility of looking at score range overlap (we will address how these factors limit the utility of the ROC). Others [39], [14] have argued for using the ROC curve over the accuracy as a performance metric. Papers from several fields, including clinical medicine [56], chemistry [4] and psychiatry [47], have been arguing for the use of the ROC. Although many disciplines call for the usage of the ROC, the interpretation and consequences of a classification error are distinct to each discipline. In our work, we focus on classification error in the context of authentication and show how the ROC alone is an inadequate metric.

Prior research has used normalized histograms to estimate score distributions [26]. This approach is fundamentally different from what we propose. We propose that an unnormalized metric - the Frequency Counts of Scores (FCS) can be used to diagnose security flaws for authentication systems. This approach is not widely known or applied.

Some of the flaws we discuss may be known in the machine learning community. For example, previous research in machine learning has discussed population skew [14]. However, our work clearly shows that our suggestions are unknown in this context and novel to the authentication systems community. Thus, it is imperative that the flaws and our proposed recommendations are discussed.

In summary, using EER as a performance metric has been questioned in continuous authentication systems and other fields. However, there is no work that propose a convincing alternative metrics to EER. We are the first to propose the FCS in addition to the ROC to augment the comparability of authentication systems. Although prior texts discussed the normalized histogram for estimating score distributions, we are the first to use the unnormalized FCS for diagnosing security flaws of authentication systems. The FCS addresses the deficiencies in the availability of data for analysis by enabling analyses to be done on the distribution of scores. This is true even in cases where the data may be sensitive and cannot be made available to the public. The FCS can be used to directly identify thresholds that fit the application criteria. Analysis of the scores can give insight into the modifications a score function might need to achieve better separation of users. With FCS, we can identify two types of flaws in the surveyed publications in top venues: incomplete performance reporting and misleading metric interpretation.

IV. Machine Learning in Authentication Systems

Authentication systems that utilize machine learning can use a variety of methods to distinguish users (e.g. fingerprints, visited locations, and keystroke dynamics). Regardless of the authentication method, the machine learning methods used to classify the measurements are the same. Figure 4 shows how machine learning in most authentication systems includes three major operations: preprocessing, scoring and thresholding.

In the preprocessing operation, the measurements are filtered, re-centered and scaled [43]. This operation may discard measurements that fail to meet any admissibility criteria the authentication system may have (e.g. measurements that are too short). Scoring applies a mathematical function ($f : M \rightarrow \mathbb{R}$) to the measurements to generate a numerical summary of the measurements (by numerical summary, we mean a number that is used to describe a characteristic of a dataset). Scores of measurements from authorized users by convention score higher (to the right of) than those of unauthorized users [23].

The scoring operation is the most critical part of the authentication process. Scoring measurements well enables unambiguous classifications by separating measurements. The better scoring is at separating measurements, the fewer errors will be made. The scores between different authentication systems are rarely comparable and bare no direct relevance to each other, even if the systems measure the same thing.

Thresholding uses the score (a numerical summary) as evidence for a decision. The choice threshold establishes the minimum required score to be deemed authorized. For any user's measurements, if the score is below the threshold, the user will be denied access. Similarly, if the user's score is above the threshold, the user will be granted access. The further away the score is from the threshold, the more confident we can be in our classification. Thus, user measurements that score significantly higher than the threshold are considered strong evidence for a decision to grant access. User measurements that score significantly lower than the threshold yield a confident decision to deny access. In this sense, the choice of threshold dictates the strength of the decision.

A. How authentication systems research is consumed

While there is no formula for implementing a proposed system from its publication description, there is a common theme that many publications follow. Many publications start with a description of what is measured and why it is important to be measured. A case then is made for why the measured quantities will produce good performance or have some additional benefit (e.g. easy to remember, resistant to some types of attacks, and require fewer resources). A classification algorithm is typically chosen based on criteria such as ease of implementation or good performance with available data for a chosen metric. Finally, a user study is performed to validate the design choices made, demonstrate the claims of utility or defensibility, and potentially compare to existing systems.

An implementer of these systems will have to determine what was measured from the description, and then collect those measurements. The implementer will then need to use the classification and compare the performance metrics achieved by their implementation against those reported in the publications. The implementer will have to pick a system based on a comparison of the reported performance values, the ability to recreate the measurement apparatus (e.g. collecting heart rhythms or breath sounds), and applicability of the system's benefits to their specific case (e.g. the need for resistance to shoulder surfing). As we will see, comparing performance values between publications is often challenging

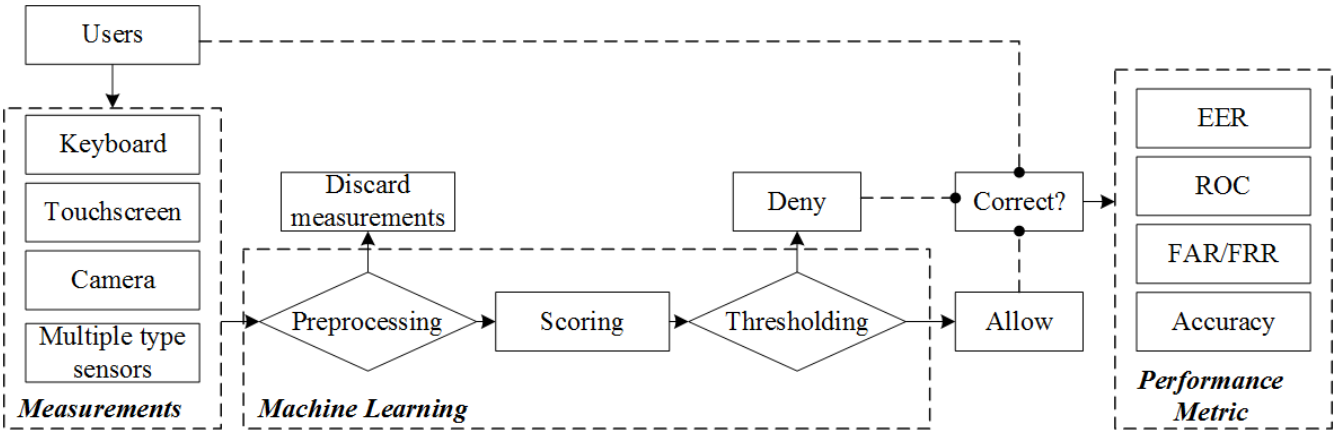

Fig. 4: The use of machine learning as a classifier is common to many authentication systems. The preprocessing phase prepares measurements by filtering, re-centering and scaling. The scoring phase applies a mathematical function to the measurements to map them to a number. The thresholding phase compares the number to a fixed threshold to make a decision. The full authentication system feeds measurements taken from the user to the machine learning classifier and then evaluates the performance based on the decisions that come out of it.

for a variety of reasons, complicating the process of choosing a system. Implementing the classification algorithm can also prove daunting because the descriptions are often inadequate (e.g. often lacking critical parameter values).

B. How classifiers work

The performance of a classifier is influenced by two major factors. The first is how well the scoring function separates the measurements from different users and the second is how well the threshold is chosen.

The scoring operation plays the most important role in the authentication system performance. The score function's ability to separate measurements via their score values reflects the underlying capability of the measurements collected to separate users. The score function can be seen as extracting information, in the form of score separation, from the measurements. If all the measurements from authorized users are distinct from the measurements of unauthorized users, an optimal score function was achieved. If the distinction between authorized and unauthorized users is inadequate, the score function will be inadequate.

The thresholding operation comes immediately after the scoring operation in importance. The selection of a threshold represents the choice of a compromise between error types that a classification system can make [27]. It cannot eliminate error; it can only trade one type of error for another (e.g. decrease the error of authorizing illegitimate users by increasing the error of denying legitimate users). If the scoring function provides good separation, there will be many choices of threshold that yield a good compromise between the error types. Several of the metrics (e.g. EER and ACC) fix a specific threshold and derive a performance metric value from this fixed point. The threshold is often chosen to optimize this metric and it is only this optimized value that is reported.

C. Why authentication systems make mistakes

From a security standpoint, what distinguishes one system from another is not the measurements they collect, but how well they tell authorized users apart from unauthorized users. For example, the problem of granting access to the wrong person has severe consequences if the target application is banking. On the other hand, the problem of denying access to the correct person is not a significant infraction if the target application is loyalty discounts.

Measurements collected from users are often random with an unknown distribution. When these random measurements are fed into a scoring function, the resulting scores will also be randomly distributed. These variable scores will be compared against a fixed threshold and a decision will be made to grant or deny access based on this comparison. If the randomness of the measurements causes the score to incorrectly cross the threshold, an error is made.

The scores of measurements from authorized users score higher than measurements from unauthorized users; therefore, if a score falls above the threshold, it is assumed to have come from an authorized user. If a score is below the threshold, it is assumed to have come from an unauthorized user. In the ideal case, all authorized users' measurements will score much higher than those of unauthorized users. This, however, is rarely the case. Often the scores from both types of users overlap (see Figure 2) because of the randomness in the measurements. The greater the overlap between scores, the more likely it is that the system will make a mistake and thus make more errors [19].

V. Common Performance Metrics Used in Authentication

For every decision a classifier must make, there are four possible contingencies: (1) authorize a legitimate user (true positive or TP), (2) authorize an illegitimate user (false positive or FP), (3) deny an illegitimate user (true negative or TN), and

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

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

$$ACC = \frac{TP+TN}{TP+TN+FN+FP}$$

TABLE IV: For a single value of the threshold, the confusion matrix (CM) arranges the counts of all possible contingencies in a grid.

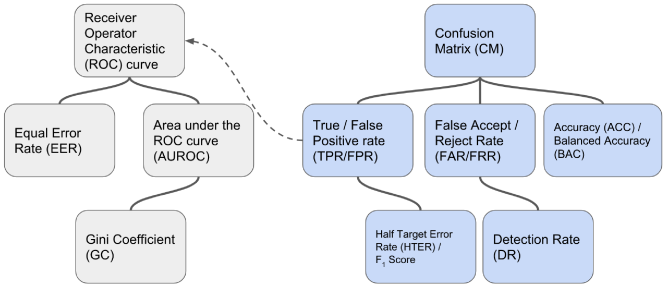

Fig. 5: Many of the commonly reported metrics are derived from the CM or the ROC. The ROC represents multiple CMs under varying thresholds. The connection between the ROC and the CM is realized through the (TPR, FPR) pairs. Each point on the ROC is one (TPR, FPR) pair for a fixed value of the threshold.

(4) deny a legitimate user (false negative or FN). The decision counts (TP, FP, TN, FN) are the fundamental components of all performance metrics. To compute these counts, the authentication system is used on a set of measurements where the ground truth is known. Once the scoring and thresholding is complete, the authentication system will produce a set of decisions based on those measurements. The counts are then computed by comparing the decisions with the ground truth.

There are two families of metrics which differ in the metric from which they are derived (shown in Figure 5). The families are: (1) confusion matrix (CM) derived metrics which depend on the threshold and (2) ROC curve derived metrics which may not depend on the threshold. The CM is a count of all possible contingencies arranged in a grid. This contingency table is computed for a fixed value of the threshold and thus depends on it. All related metrics inherit this dependence. Many of the other performance metrics are ratios of the counts enumerated in the confusion matrix (e.g. ACC).

The ROC represents many confusion matrices under varying values of the threshold and thus does not depend on it. Metrics derived from the ROC may be specific points on the ROC, such as the EER, or functions of the ROC (e.g. AUROC). Since the EER is a specific point on the ROC, it corresponds to a specific value of the threshold. Figure 5 shows the relationships between the CM related metrics and ROC curve related metrics.

A. Confusion Matrix (CM) related metrics

Table IV shows a confusion matrix and the related metrics derived from it. The true positive rate (TPR) and false positive rate (FPR) are two key metrics that are computed from a confusion matrix to evaluate authentication system performance. TPR is interpreted as the probability that an authorized user will successfully authenticate and FPR is the probability that an unauthorized user will successfully authenticate. FPR is sometimes called the false accept rate (FAR). Other ratios often reported include the false negative rate (FNR) which is alternatively called the false reject rate (FRR), and the true negative rate (TNR).

The maximum accuracy (ACC) is another key metric for authentication system performance. It is interpreted as the relative frequency of a correct classification of a measurement source, regardless of its origin. *Since the accuracy is a function of the threshold, often the value of accuracy that is reported is the maximum across all thresholds.* The maximum accuracy represents the best performance the classifier can offer, however, solely reporting accuracy can be misleading. Because only a single threshold is represented in this performance metric, consumers of the research cannot know how the accuracy will change if the threshold changes. This may lead to the conclusion that a system is unfit for an application because the accuracy achieved is below an error requirement even though a judicious choice of threshold would satisfy an FPR requirement (at the cost of some TPR).

1) Other accuracy metrics: The maximum accuracy is not the only accuracy metric reported. There are several other metrics that are functions of the values across the columns of the CM, such as the balanced accuracy (BAC), F_1 score and half total error rate (HTER). Some of these metrics, such as BAC and HTER, attempt to weight the ratios to adjust for skews within the measurement populations. We note that these metrics are also functions of the confusion matrix, and thus still dependent on the value of the threshold. These metrics are reported less frequently than others considered in this paper, and they share many issues with the metrics that we consider.

B. ROC curve related metrics

Despite the overwhelming reliance on single threshold metrics generated by a confusion matrix, they have limited utility. Single threshold metrics present an incomplete picture of the system's performance. All of these metrics give no indication of how changes to the threshold affect the behavior of the metric. If the thresholds that were used to derive the metric are not reported, it is not possible to repeat the experiment to determine if the achieved metric values can be obtained in subsequent trials. To implement a system, some insight into the relationship between the performance metrics and the threshold is needed.

The ROC is computed by varying the authentication threshold from the maximum to the minimum possible values of the score and calculating the TPR and FPR for each threshold value. As the threshold lowers, scores that were not initially high enough to grant access will eventually rise above the threshold. Consequently, the probability that an unauthorized user is erroneously granted access because the random variation caused their measurements to score above the threshold. However, if the random variation caused an

authorized user's measurements to score lower than is typical, the user would still be granted access because the threshold is lower. Eventually, as the threshold is lowered, both the number of TPs and the number of FPs will increase. Each value of threshold represents a specific trade-off between the TPR and FPR. Taken as a pair, the (FPR, TPR) is a parametric function of the threshold. This parametric curve drawn as the threshold varies is the ROC curve (see Figure 2). Also shown is the line of indifference, $y = x$ (green line of Figure 2). If the ROC is close to this line, the system performance is comparable to blind guessing.

There are three common single-number performance metrics for summarizing the ROC curve: the EER, the AUROC (or AUC in some texts), and the GC.

1) Equal Error Rate (EER): As Figure 2 shows, it is the point on the ROC where the $FPR = 1 - TPR$. It is easily identified as the intersection of the line $y = 1 - x$ (red dashed line of Figure 2) and the ROC curve. It represents the probability of making an incorrect positive or negative decision in equal probability. Since the ROC is a parametric curve, there is a specific value of the threshold that corresponds to the EER.

2) Area Under the ROC Curve (AUROC): The AUROC is defined as the area below the ROC curve and is depicted as the shaded region in Figure 2. It reflects the probability that a random unauthorized user's measurement is scored lower than a random authorized user's measurement [23]. It can be interpreted as a measure of how well a classifier can separate measurements of an authorized user from their unauthorized counterparts. In Figure 2, the AUROC is computed for the given ROC.

3) Gini Coefficient (GC): The Gini Coefficient (GC) is functionally related to the AUROC as follows [22]: $\text{GC} = 2 \times \text{AUROC} - 1$. It also tries to quantify how much separation there will be in the measurements.

VI. Proposed Metric: The Frequency Count of Scores (FCS)

We propose the Frequency Count of Scores (FCS) as an additional performance metric to be reported with the ROC curve. Figure 2 shows examples of the FCS coupled with the ROC. The FCS provides the ability to visually diagnose and explain the achieved performance reported in the ROC because the ROC can be constructed from the FCS. By examining the distribution skew and overlap of the score frequencies, we can determine if the proposed systems exhibit any biases towards a positive or negative decision. We can also justify the reported performance observed in the ROC by examining how well the score distributions are separated. Good score separation will yield good system performance which will be reflected in an ROC with a low EER. Sensitivity to changes in the threshold can be assessed by looking at how the scores are spread relative to each class. The two metrics complement each other.

The FCS is a fundamental metric that is different from the ROC and the confusion matrix. It is considered fundamental in this context because it is not derived from the CM or the ROC. It can be used to diagnose model problems, compare systems, and validate implementations. The FCS is constructed by identifying the maximum and minimum scores across all measurements and then choosing a common bin width over this range. Scores are separated by the ground truth and then plotted as separate histograms which are binned using the common bin width. The bin width is a free parameter that can be chosen to reflect the amount of data available and the observed score variability.

The FCS should *not* be normalized to make it look like a distribution. The unnormalized version makes the population skews, score distribution imbalances and score overlap regions visually apparent. The FCS is a useful addition to the reported metrics because it allows a research consumer to visually perform additional analyses which would not be possible with the ROC or CM metrics alone.

How the scores are distributed plays a central role in the performance of a system. A large majority of the decision errors are made because the random variation in measurements causes the scores to erroneously cross a chosen threshold. Because the measurements are not deterministic, the scores are variable even if they are a deterministic function of the measurements. How well the score function separates measurements in the presence of this variability dictates the range of possible error trade-offs between TPR and FPR for a system. If the separation of scores is large, then it is possible to achieve high TPR while keeping the FPR low. Since each choice of threshold represents a compromise between TPR and FPR, larger score separation implies better choices of threshold.

1) The FCS can be used to gain performance insights beyond what the existing metrics show: Many of the existing metrics can actually be derived from the FCS. For example, the TPR can be computed as the relative frequency of positive scores that lie beyond the threshold.The proportion of the score range from authorized and unauthorized users that overlaps is important because many of the performance metrics, such as the EER and ACC, attempt to summarize system performance by quantifying how often a measurement from either user will get a score in this overlapping range. The ACC and EER both depend on the width of the overlapping region as well as the relative frequency of the scores that fall within this overlap. Neither metric considers the portion of the scores that lie outside the overlapping score region for either score distribution (authorized and unauthorized). It is this lack of consideration for these other aspects of the scoring that cause these metrics to be incomplete. By reporting the FCS, difficult concepts can be easily visualized. Consider the AUROC: its definition is very technical, and is thus difficult to interpret. However, if we look at two different FCS and note the score overlaps are smaller in one vs. the other, we have captured the essence of what the AUROC is trying to measure.

Some insights into system performance that only the FCS can provide are gained by considering scores that lie outside the overlapping region. Performance metrics, such as the TPR, are directly impacted by the portion of these types of scores. For example, authorized users' scores that lie outside the overlap can only contribute to the TPR. If this portion is not empty, then the TPR may never practically reach zero (e.g. there is a threshold for (A) of Figure 7 that achieves non-zero TPR at zero FPR because scores above this threshold could not have come from unauthorized users). Although this can be visually confirmed on the ROC, it would be difficult to identify why it happens from the ROC.

The differing slopes of the ROCs in Figure 7 do not indicate how sensitive the classifier is to changes in the threshold. A research consumer cannot ascertain how far along the ROC a change in the threshold will move them purely by looking at the ROC. However, this information can be gathered by looking at the spread of the score distributions in the FCS. If the scores are spread wide relative to the width of the overlapping region, then the classifier is not particularly sensitive to the threshold. If the width of the overlapping region is small compared to the score distributions, small changes in the threshold will cause significant movement along the ROC.

If scores from authorized users that are above the overlap occur with higher relative frequency than scores within the overlap, then the authentication system will produce more positive declarations. A similar result holds for the TNR and unauthorized users' scores which are below the overlap. When examining the FCS of a proposed system, if only one user has scores that lie outside the overlap (e.g. FCS (F) in Figure 8), the system may be biased towards decisions in favor of that type of user (e.g denying access since most of the scores range comes from unauthorized users). Without the FCS, it is difficult to determine if a proposed system has this kind of flaw, even if the ROC is reported.

VII. Flaws with Existing Metrics

In this section, we discuss in detail the implications of the observed flaws we summarized in Section II. We note the cases where the FCS aids in diagnosing whether a flaw is present or explains why the flaws occur.

A. *Incomplete performance reporting*

Skews within the measurement population can artificially inflate some CM derived metrics. Measurements are often split into training and testing data. Training data is used to build a model and testing data is used to compute performance metrics. If the measurement population is skewed, both data sets will exhibit this skew. If this approach is coupled with a report that uses only a single performance metric, misconceptions arise. For example, in one of the papers we reviewed, the authors only report the FNR. Unfortunately, their measurement population was skewed. From their reporting, we cannot know whether the low FNR is due to their system's ability to discern users or due to a skew in the measurement population.

If we only have a single metric available, such as the ACC for two systems we are trying to compare, the system with the better value can be deemed superior. On the surface this seems like a perfectly fine criteria. For example, given that the interpretation of the ACC as an approximation of the probability of a correct classification, a higher ACC would seem to indicate superior performance. Unfortunately, relying on the ACC as the sole criteria can be very misleading. It is possible that the ACC value was inflated by skew in the measurement population.

When a classifier has poor ACC, the scores from authorized and unauthorized measurements will have significant overlap (as seen in the left side of Figure 6). These overlapping scores are ambiguous and thus difficult to classify. If we skew the measurement population to have mostly unauthorized users, the scores from the unauthorized users overshadow the

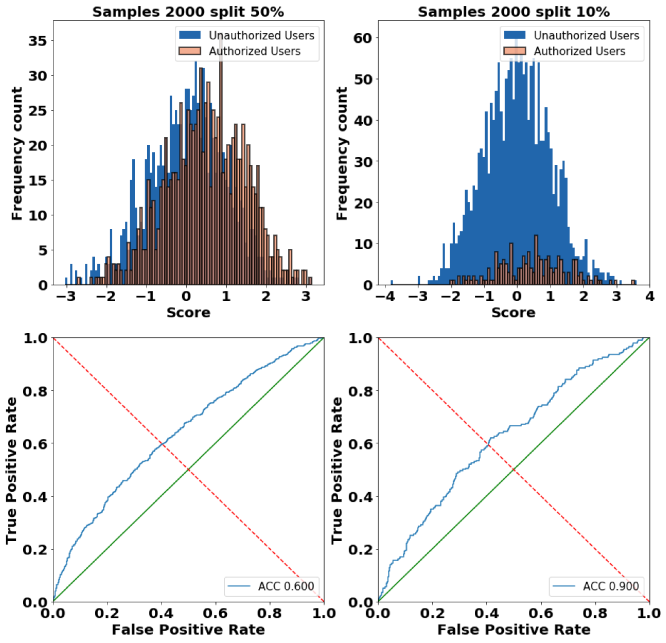

Fig. 6: Measurement population skew can cause low accuracy classifiers to have artificially high accuracy values. On the left side, the set of 2000 measurements is evenly split between authorized and unauthorized. We drew the FCS and computed maximum accuracy $ACC \approx 60\%$ and the ROC under this measurement split. On the right side, the set of 2000 measurements is skewed to include 10% authorized and 90% unauthorized measurements. Because of this skew, the FCS shows that the positive scores are effectively buried in the negative scores. The maximum accuracy achieved is $ACC \approx 90\%$ which is reached by choosing a threshold that results in mostly negative declarations. This reported accuracy is misleading because the scoring function was the same.

scores from the authorized users' measurements (right side of Figure 6). In this instance, possible values of the threshold that cause the classifier to make mostly negative decisions (denials of access) will be favored because most classifiers are optimized by minimizing the error over the data on which they are trained [1]. Since the test data is skewed in the same way, a classifier that returns mostly negative decisions will be correct most of the time. This skew in the test data will make the classifier appear to be more accurate because it is being tested on data it would always get correct.

While the detrimental effects of skew are evident for the ACC, any metric that depends on both N and P at the same time (cross column in the confusion matrix of Table IV) will be affected by population skew [14]. For reference, $N = TN + FN$ and $P = TP + FP$. Figure 6 also demonstrates how the ROC is mostly unaffected by population skew. Population skew can mask the poor score separation by providing performance numbers that are artificially high. However, these flaws are easily identified by the frequency count of scores. The unnormalized counts in FCS show that the total volume of scores from unauthorized users vastly outnumber those of authorized users. This visualization helps designers easily examine the skewed measurement population.

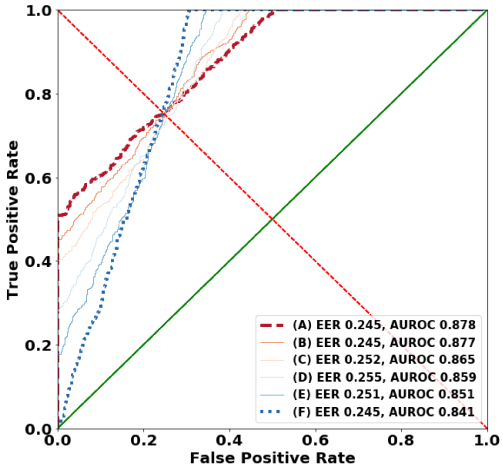

Fig. 7: The EER does not represent a single ROC curve. Instead, it represents a family of ROC curves. While each member of this family have similar EER, their performance varies significantly across the range of possible thresholds. Unexpected sensitivity to changes in the threshold can lead to surprises in system behavior when the thresholds used in the implementation deviate from the published values. In this case, ROC (F) is mostly inferior to ROC (A) because (A) achieves a higher TPR at 0.0 FPR. However, according to the EER, they are essentially the same. If an implementer has a specific TPR or FPR target, the EER may be of little value to them as they cannot determine how the TPR/FPR may vary between the EER and their target. It should be noted that if the target application can tolerate an FPR $>$ EER then (F) is the superior choice, however this tolerance cannot be known to the researcher. There is no skew in these examples.

B. Results without model parameters

In the ideal case, the code used to derive the results of a study would be published along with the proposed system. This may not be feasible in all situations. However, most of the systems that use machine learning do not implement the algorithms from scratch. Instead, they often apply to existing implementations in libraries such ask Weka [21] or libSVM [6]. The novelty of these proposed systems often lies in the complete end-to-end performance, not the algorithm used to make decisions. Implementations of the proposed systems can be simplified by having the model parameters (e.g. SVM slack factor, number of hidden layers in the NN, and learning rate) that were used to derive the reported results. These parameters control the behavior of the algorithm and reflect a value judgment made by the researcher based on their understanding of the how the algorithm interacts with the measurements.

By providing the parameters of the algorithms used, authors enable replication of the research, benefiting the community in two ways. First, the data analysis can be replicated exactly to determine if other factors contributed to the reported results. If the data is also available, follow-up analyses can be more easily performed. Updated versions of the libraries that may have fixes for vulnerabilities or performance enhancements can be validated against existing results. Second, any potential implementers of the system only need to replicate the measurement collection and data processing portions of the proposed systems. The properly parameterized software library can essentially be treated as a black box.

C. Misleading performance metric interpretations

A key issue with reporting only a single performance metric as a summary of the system is that the value of the metric does not uniquely identify the classifier from which it came. This information is lost, and with it all knowledge of how the system performs when the parameters are adjusted. In Figure 7, we can directly observe this issue in the EER and AUROC performance metrics. The graph shows several ROC curves from different score functions that all have very similar EERs and AUROCs. Each of the ROC's linear portions have different slopes. These differing slopes reflect different sensitivities to the threshold, due to the difference in how the scores are distributed. A change in the threshold has much more impact on the ROC (F) than on the ROC (A) curve, thus an implementation can fail in unexpected ways because the implementers were unaware of this difference.

If we are only given the EER to evaluate a system and we have a specific target for our TPR or FPR, we are unable to determine from the EER if our target will be met. This information is not knowable because many ROCs (and thus many classifiers) have the same EER. Because we do not know which classifiers were used, there are many possibilities for how the performance can vary with the threshold. As we note in the Introduction, many applications have specific requirements for the TPR or FPR, which are attained by controlling the threshold.

By definition, all ROCs must connect the point $(0, 0)$ to the point $(1, 1)$ (i.e. setting the threshold of a classifier higher than the maximum achieved score will result in 0 TPs and 0 FPs, whereas setting it below the minimum will have the opposite effect). The EER fixes a third point that the curve must pass through. However, as depicted in Figure 7, these three points do not uniquely determine the curve. There are many ROCs that correspond to a small range of EERs due to what the EER is measuring.

The EER is the point on the ROC in which the probability of an incorrect denial of access is equal to the probability of an incorrect granting of access. Both of these probabilities are proportional to the width of the region of overlap in the scores. In Figure 8, the corresponding FCS for each ROC in Figure 7 is depicted. Each score distribution has the same width of overlap region, however, the overlapping region moves to the right as the figures are read left to right, top to bottom. As the overlapping region moves right, it consumes more of the score range for the authorized users' measurements.

As the authorized users' score range shrinks, the unauthorized users' score range grows to maintain the width of the overlap. Because all of the overlapping regions are the same width in all cases, a threshold can be identified for each score set that strikes the same balance between the two error types (FP and FN). This threshold will be in a different place for each of the different score distributions. However, each classifier can be tuned to achieve the same EER by picking the correct threshold, even though their individual tolerances to

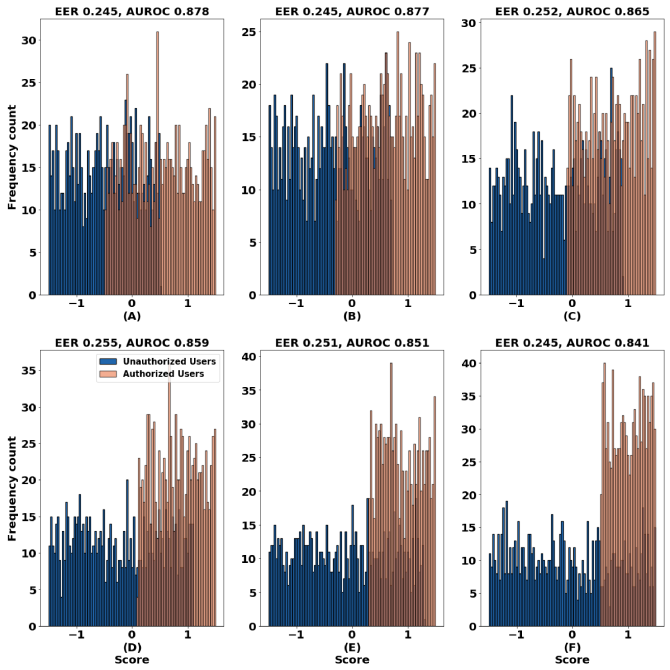

Fig. 8: These FCS graphs show where the shapes in Figure 7 come from. They were constructed so that as we move from left to right and top to bottom, the region of scores that an authorized source has to definitively identify as authorized is shrinking, however the width of the overlap stays the same. This explains why (A) has non-zero TPR at FPR(0.0) but (F) does not. As the authorized score region is consumed by the overlap in scores, there are fewer distinct scores from the authorized user and thus fewer ways to get a purely true declaration. This difference explains why the range of possible trade-offs is worse for (F) than (A), as reflected in the slope of the ROC. The ROCs are linear because these score distributions are uniform (purely for example purposes). If the distribution shape changes but the overlap region remains the same, the EER behavior will be unchanged, however the ROC will be more curved. There is no skew in these examples, only the score distributions change.

threshold shifts vary greatly. Since there are fewer authorized users' scores, the sensitivity of the classifier to changes in the threshold goes up because each change has a greater impact on the classifications that are made causing the distinct differences in slope in Figure 7.

A key shortcoming of the EER is that it focuses only on the overlap between the score ranges. It fails to consider the proportion of the score range that lies outside the overlapping region for either measurement source or any asymmetries in the distribution. Only scores that are within the overlap contribute to classification errors because they can be confused with scores from the alternate class. The proportion of scores outside the overlap is as important as the width of the overlap itself because it governs the probability that an easily confused score will be observed. The EER fails to account for asymmetries in the score distribution. The graph of FCS depicted in Figure 8 makes the proportion and the asymmetry visually apparent.

There is no measurement population skew in any of the graphs in Figure 8; instead, the authorized range is shrinking. The unnormalized frequency count shows that the probability mass across authorized measurement scores is being redistributed over a smaller range. Thus, the score distributions are becoming more asymmetric. The classifier is becoming more biased because the probability of observing a score that could only have come from an authorized user's measurements is getting smaller. Similar to the skew accuracy problem of Section VII-A, observing the weight and range of the scores from both measurement sources allows for the identification of problems, with the authentication system making the overlapping region width and distribution asymmetry apparent.

1) AUROC and GC are also flawed: The AUROC is interpreted as the probability that scores of differing measurement sources separate well [23]. This probability is proportional to the width of the overlap in score range. As the width of the overlap gets smaller, the probability increases. Thus bigger values of the AUROC are desirable. If the AUROC $\rightarrow 1$, then all authorized users' measurements will score higher than unauthorized users' measurements. The probability of separation is maximum, thus there may be a threshold that achieves perfect classification. Since the Gini Coefficient (GC) is functionally related to the AUROC, it is also functionally tied to the width of the overlap in score range.

Unfortunately, the AUROC and GC also exhibit interpretation flaws because they are summary metrics. They also mask the complexity of the classifier performance. In Figure 7, the AUROC was computed for each of the curves (the GC can be computed from the formula). As expected, the range of the AUROC does not vary significantly even though the resulting ROCs are very different. The AUROC is within the range $(0.84-0.88)$ across the different classifiers. The AUROC does not vary due to the width of the overlapping region which is held constant, as seen in Figure 8. Like the EER, the AUROC focuses heavily on the overlap of scores.

VIII. Recommendations for Reporting: Solutions to the Pitfalls of Current Reported Metrics

In the ideal case, authors would make all source material available, including data and code. This approach would yield the best results for system evaluation because evaluators and implementers could verify their implementations against the reference provided by the researcher.

Although there is no one-size-fits-all strategy for analysis, we propose guidelines that can be followed to simplify the task of evaluating the proposed system. The following three suggestions may aid consumers of research, including an implementer who needs to choose the best authentication system for their target application.

First suggestion: Report as many metrics as possible including both the ROC and the FCS. These two graphs enable comparisons across many parameterizations and serve as a visual check for biases. The FCS enables both researcher and reader to diagnose issues with population skew and score distribution via immediate visual analysis. It can also serve as a diagnostic tool for implementations to verify that the scores produced by the implemented system are within the range the researcher originally reported. Other metrics such as the EER,

AUC and ACC should also be reported for comparison with existing literature. These metrics can be added to abstracts and introductions for glanceability but are not a substitute for a complete analysis that includes an ROC and FCS.

Second suggestion: If the FCS cannot be reported, report the ROC curve to enable an implementer to decide if the system has a threshold that meets the error performance requirements of their target application. Implementers can find specific points on the ROC curve that satisfy their requirements and be assured that if the implementation is faithful to the proposed system, the can find a value of the threshold that yields the chosen error rates.

Third suggestion: If the ROC cannot be reported (e.g. for space constraints), report multiple summary metrics that are not functionally dependent. Since each of the summary metrics, EER, ACC, and AUROC only represent a single aspect of the system's performance, the reader can obtain a more thorough evaluation of the performance if all three are reported. Reporting all three gives readers the ability to compare the proposed systems from the existing body of literature that often only report one or two of these metrics.

A. Case Study

To demonstrate how to use the ROC and FCS to compare systems, we evaluated the authentication performance of three existing datasets via the ROC and FCS. We will first describe the datasets and classifiers that were built. Each publication provides a dataset and a system model to test their dataset. When the classifier is used on the dataset, an FCS and ROC will be computed. Because each publication's dataset and classifier has its own population distribution and score function, we expect the FCS and ROC from each publication's proposed system to be very different. We will show how to use the ROC and FCS to decide between these systems.

1) Datasets and classifiers used to create ROCs and FCSs: The SVC2004 dataset [52] is a public signature dataset with 40 types of signatures and 20 genuine samples for each signature. We implemented the linear classifier of Principal Component Analysis, proposed by Kholmatov [28]. The Touchalytics [16] is a public touching behavioral biometric dataset with 41 participants' continuous touching behavior data. We built the authentication system with k-nearest-neighbors as described in their paper. We selected k=100 and each user contributed 150 periods of touching behavior as templates. The dynamic keystroke dataset [29] is a set of keystroke features that was collected while users input a password. Typing behavior was observed for 51 users, and each user contributed 400 typing samples. The proposed system was built with a one-vs.-all-classifier for each user. In our study, we randomly chose a user and built the authentication system with Manhattan (scaled) similarity that was described in the paper.

2) Analysis of these three systems with the FCS and ROC: In Figure 9, we display the ROC curves for all three systems. We assumed that the implementer has a fixed requirement on the FPR of 0.1. To choose a system that meets our requirements, we drew a solid black vertical line at our FPR limit. Thus, we can visually identify the system that has the highest TPR for our FPR limit. In this case, Keystroke is the clear winner, even though it does not have the lowest EER. Thus, potential implementers would not able to assess a proposed system when given only the EER.

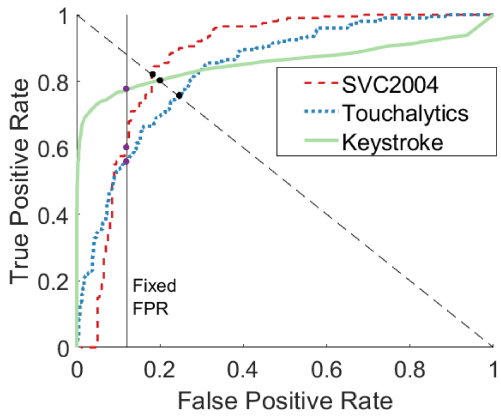

Fig. 9: Comparing systems with a specific FPR target in mind is done by finding the highest TPR for that FPR. To find the highest, TPR draw a vertical line at that FPR and then identify the ROC that crosses the line at the highest point. A similar procedure works for specific TPR targets with horizontal lines.

We also see the slope of the ROC near the fixed FPR target. Observe how quickly the TPR degrades if we need to make the FPR tolerance lower. For example, the segment of the ROC from SVC2004 classifiers around the FPR target is very steep, indicating that a small change in threshold will lead to a significant change of the system's TPR and FPR rates. In contrast, the ROC of the Keystroke dataset is very stable because the TPR changes slightly with the change of FPR. If we have an upper bound on the FPR and want a system that gracefully degrades when the FPR target is lowered, the Keystroke classifier is the clear winner. It should be noted that if we could tolerate an FPR of 0.3 or higher, the Keystroke system would be inferior to both SVC2004 and Touchalytics when considering both the the TPR value for a fixed FPR and the slope around a fixed FPR.

The FCS can be used to show the asymmetry of the score distributions in detail. Figure 10 displays the ROC curves of the three systems along with their corresponding FCS. For example, the ROCs of SVC2004 and Touchalytics are similar, but their FCS shows that the SVC2004 classifier has an advantage because the authorized and unauthorized scores are more separable than the Touchalytics classifier. Additionally, the EERs of SVC2004 and Keystroke classifier are similar, but their FCS shows that the SVC2004 classifier is superior because the unauthorized and authorized users' scores overlap significantly in the Keystroke dataset.

From the FCS of Figure 10, we can observe the asymmetry in the scores. The unauthorized users' scores in the Touchalytics classifiers almost cover the entire score range, indicating that the classifiers can never be certain about granting access. Every authorized user's score could have come from an unauthorized user, thus this system may be biased to deny access more often. The Keystroke classifiers are biased in the other direction: all unauthorized users' scores could have come from an authorized user. The SVC2004 classifiers have some score range which does not overlap, and thus can

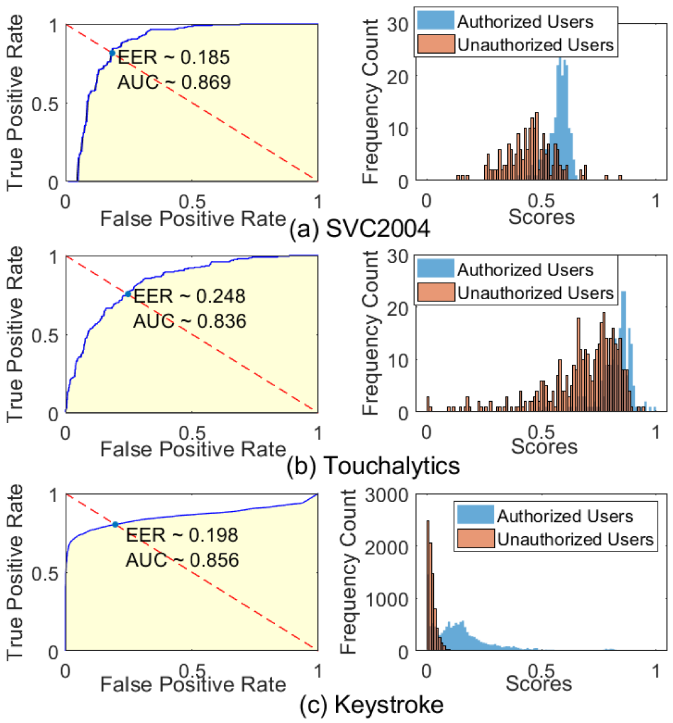

Fig. 10: The FCS can be used to decide between different systems that have similar EERs and ROCs.

make some decisions with certainty. If our application needs to be balanced, SVC2004 is our best choice. If the application needs to be biased towards denials, then we should choose Touchalytics. If we want more positive decisions, Keystroke has a higher probability of delivering them.

We have showed how relying on the EER can produce erroneous conclusions about authentication systems. We have provided evidence for the limitations of reporting the EER in three systems: Keystroke, SVC2004, and Touchalytics. We have also showed how the ROC and FCS should be implemented as a solution to the limitations of single number summaries. In Figure 10, we show how the FCS compliments ROC to improve reporting authentication system metrics.

IX. Discussion and Conclusions

We have proposed robust metrics for evaluating machine learning-based authentication systems: ROC curves and their corresponding FCS. We argue for the ROC as a method of reporting classification performance because it is able to provide an overview of the authentication performance across all thresholds. However, the ROC misses some scoring details, such as the difference in the width of score ranges and the asymmetry of score distributions. This scoring detail can indicate whether a classification bias is present in the scoring function and how sensitive the error rates are to changes in the threshold. Therefore, we introduce the FCS as an augmentation to the ROC curve. We believe reporting the ROC and FCS together provides a robust metric for evaluating the performance of authentication systems.

The commonly used authentication performance metrics, such as EER, AUROC, GC and ACC, are inherently flawed. EER only focuses on the overlap between the score ranges and does not consider the proportion of the score range that lies outside of the overlapping region. Since the scores inside the overlapping region are the reason errors are made in the authentication system, there always needs to be balance between the two types of errors.

The two types of error rates of the system depend heavily on the thresholds in the overlapping region of scores. If the proportion of scores inside the overlap is large, one will likely encounter a score that is difficult to classify. We show in Figure 8 that with a similar EER, system (A) is much better than system (F) because system (A) is able to completely separate some of the measurements from different user types. Therefore, EER, AUROC, and ACC hide important information that could be used for comparison between authentication systems. Even the ROC by itself provides a limited analysis. While the differences between (A) and (F) visually manifest in the ROC as a higher TPR at 0.0 FPR and a different slope, it is not visually obvious why this happens or how the TPR changes as the threshold changes. Reporting practices that focus on a single metric limit the ability to compare systems by ignoring these factors.

We introduce the FCS to augment the ROC in order to evaluate and compare the performance of authentication systems. The FCS is fundamentally different from the ROC curve and the CM because it is not derived from either and thus brings additional information into the analysis. We can use FCS to detect the measurement population skew, asymmetries in the scoring distribution and assess sensitivity to threshold changes. Since the scores in FCS are not normalized, the population skews are visually apparent. Usage of the FCS is not limited to authentication systems. The ability to identify distribution imbalance and threshold sensitivity is relevant to any applications that use machine learning to decide where their measurements come from.

We have illuminated the problems with current reporting practices in authentication system research. Reporting these single-number summaries alone is a barrier to comparison between systems and can misrepresent a system's potential. For example, some metrics do not show the performance trade-offs or whether performance degrades outside the conditions for which the system was designed. We proposed a solution to the limitations of current metrics: reporting a full set of metrics that includes the FCS and the ROC. We argue that performance reporting should be as comprehensive as possible and that the the FCS and ROC can help in this regard by provides additional information to evaluate authentication systems. We believe it is crucial for our community to adopt more transparent reporting of metrics and performance.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number 1750987. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional material available at http://scienceofsecurity.science.

REFERENCES

[1] Y. S. Abu-Mostafa, *Learning From Data*, 1st ed. Pasadena, CA: AMLbook.com, 2012, ch. Training versus Testing, pp. 39–69.

[2] I. Bhardwaj, N. D. Londhe, and S. K. Kopparapu, "A spoof resistant multibiometric system based on the physiological and behavioral characteristics of fingerprint," *Pattern Recognition*, vol. 62, pp. 214 – 224, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316302576

[3] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *2012 IEEE Symposium on Security and Privacy*, May 2012, pp. 553–567.

[4] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24 – 38, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743905000766

[5] C. Castelluccia, M. Dürmuth, M. Golla, and F. Deniz, "Towards implicit visual memory-based authentication," in *Network and Distributed System Security Symposium (NDSS 2017)*, 2017. [Online]. Available: http://dx.doi.org/10.14722/ndss.2017.23292

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 278–291.

[8] Y. Chen, J. Sun, X. Jin, T. Li, R. Zhang, and Y. Zhang, "Your face your heart: Secure mobile face authentication with photoplethysmograms," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.

[9] F. Cheng, S.-L. Wang, and A. W.-C. Liew, "Visual speaker authentication with random prompt texts by a dual-task cnn framework," *Pattern Recognition*, vol. 83, pp. 340 – 352, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318302152

[10] H. Crawford and E. Ahmadzadeh, "Authentication on the go: Assessing the effect of movement on mobile device keystroke dynamics," in *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, 2017, pp. 163–173. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/crawford

[11] M. T. Curran, N. Merrill, J. Chuang, and S. Gandhi, "One-step, three-factor authentication in a single earpiece," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ser. UbiComp '17. New York, NY, USA: ACM, 2017, pp. 21–24. [Online]. Available: http://doi.acm.org/10.1145/3123024.3123087

[12] S. Das, G. Laput, C. Harrison, and J. I. Hong, "Thumprint: Socially-inclusive local group authentication through shared secret knocks," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 3764–3774. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025991

[13] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic, "Evaluating behavioral biometrics for continuous authentication: Challenges and metrics," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17. New York, NY, USA: ACM, 2017, pp. 386–399. [Online]. Available: http://doi.acm.org/10.1145/3052973.3053032

[14] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[15] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '17. New York, NY, USA: ACM, 2017, pp. 343–355. [Online]. Available: http://doi.acm.org/10.1145/3117811.3117823

[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," *Trans. Info. For. Sec.*, vol. 8, no. 1, pp. 136–148, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1109/TIFS.2012.2225048

[17] D. Freeman, S. Jain, M. Dürmuth, B. Biggio, and G. Giacinto, "Who are you? a statistical approach to measuring user authenticity," in *Network and Distributed Systems Security (NDSS) Symposium 2016*, 2016, pp. 1–15. [Online]. Available: http://dx.doi.org/10.14722/ndss.2016.23240

[18] M. Gadaleta and M. Rossi, "Idnet: Smartphone-based gait recognition with convolutional neural networks," *Pattern Recognition*, vol. 74, pp. 25 – 37, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317303485

[19] R. G. Gallager, *Stochastic Processes, Theory for Applications*, 1st ed. Cambridge, UK: Cambridge University Press, 2013, ch. 8, pp. 1–1000.

[20] H. Gomi, S. Yamaguchi, K. Tsubouchi, and N. Sasaya, "Towards authentication using multi-modal online activities," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ser. UbiComp '17. New York, NY, USA: ACM, 2017, pp. 37–40. [Online]. Available: http://doi.acm.org/10.1145/3123024.3123097

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.

[23] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[24] J. Ho and D.-K. Kang, "Mini-batch bagging and attribute ranking for accurate user authentication in keystroke dynamics," *Pattern Recognition*, vol. 70, pp. 139 – 151, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031730184X

[25] C. Huang, H. Chen, L. Yang, and Q. Zhang, "Breathlive: Liveness detection for heart sound authentication with deep breathing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 12:1–12:25, Mar. 2018. [Online]. Available: http://doi.acm.org/10.1145/3191744

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[27] S. M. Kay, *Fundamentals of statistical signal processing: Detection theory, vol. 2*. Prentice Hall Upper Saddle River, NJ, USA:, 1998, ch. 3, pp. 61–65.

[28] A. Kholmatov and B. Yanikoglu, "Identity authentication using improved online signature verification method," *Pattern Recogn. Lett.*, vol. 26, no. 15, pp. 2400–2408, Nov. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2005.04.017

[29] K. S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *2009 IEEE/IFIP International Conference on Dependable Systems Networks*, June 2009, pp. 125–134.

[30] K. Krombholz, T. Hupperich, and T. Holz, "Use the force: Evaluating force-sensitive authentication for mobile devices," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. Denver, CO: USENIX Association, 2016, pp. 207–219. [Online]. Available: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/krombholz

[31] F. Lin, K. W. Cho, C. Song, W. Xu, and Z. Jin, "Brain password: A secure and truly cancelable brain biometrics for smart headwear," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '18. New York, NY, USA: ACM, 2018, pp. 296–309. [Online]. Available: http://doi.acm.org/10.1145/3210240.3210344

[32] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '17. New York, NY, USA: ACM, 2017, pp. 315–328. [Online]. Available: http://doi.acm.org/10.1145/3117811.3117839

[33] C. Liu, G. D. Clark, and J. Lindqvist, "Where usability and security go hand-in-hand: Robust gesture-based authentication for mobile systems," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 374–386. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025879

[34] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. ACM, 2017, pp. 73–87.

[35] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 19:1–19:23, Mar. 2018. [Online]. Available: http://doi.acm.org/10.1145/3191751

[36] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018 - The 37th Annual IEEE International Conference on Computer Communications*, April 2018.

[37] J. Oglesby, "What's in a number? moving beyond the equal error rate," *Speech Communication*, vol. 17, no. 1, pp. 193–208, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939500017I

[38] M. Okawa, "Synergy of foregroundbackground images for feature extraction: Offline signature verification using fisher vector with fused kaze features," *Pattern Recognition*, vol. 79, pp. 480 – 489, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318300803

[39] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 445–453. [Online]. Available: http://dl.acm.org/citation.cfm?id=645527.657469

[40] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, "Improved ear verification after surgery - an approach based on collaborative representation of locally competitive features," *Pattern Recognition*, vol. 83, pp. 416 – 429, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318302188

[41] M. Sato, R. S. Puri, A. Olwal, Y. Ushigome, L. Franciszkiewicz, D. Chandra, I. Poupyrev, and R. Raskar, "Zensei: Embedded, multi-electrode bioimpedance sensing for implicit, ubiquitous user recognition," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 3972–3985. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025536

[42] S. Schneegass, Y. Oualil, and A. Bulling, "Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: ACM, 2016, pp. 1379–1384. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858152

[43] M. E. Schuckers, *Computational Methods in Biometric Authentication*. Springer London, 2010. [Online]. Available: https://doi.org/10.1007%2F978-1-84996-202-5

[44] I. Sluganovic, M. Roeschlin, K. B. Rasmussen, and I. Martinovic, "Using reflexive eye movements for fast challenge-response authentication," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1056–1067. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978311

[45] C. Song, A. Wang, K. Ren, and W. Xu, "Eyeveri: A secure and usable approach for smartphone user authentication," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[46] Y. Song, Z. Cai, and Z.-L. Zhang, "Multi-touch authentication using hand geometry and behavioral information," in *Security and Privacy (SS&P), 2017 IEEE Symposium on*. IEEE, 2017, pp. 357–372.

[47] D. L. Streiner and J. Cairney, "What's under the roc? an introduction to receiver operating characteristics curves," *The Canadian Journal of Psychiatry*, vol. 52, no. 2, pp. 121–128, 2007, pMID: 17375868. [Online]. Available: http://dx.doi.org/10.1177/070674370705200210

[48] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 36:1–36:18, Mar. 2018. [Online]. Available: http://doi.acm.org/10.1145/3191768

[49] E. Uzun, S. P. H. Chung, I. Essa, and W. Lee, "rtcaptcha: A real-time captcha based liveness detection system," in *Network and Distributed Systems Security (NDSS) Symposium 2018*, 2018. [Online]. Available: http://dx.doi.org/10.14722/ndss.2018.23253

[50] W. Xu, G. Lan, Q. Lin, S. Khalifa, N. Bergmann, M. Hassan, and W. Hu, "Keh-gait: Towards a mobile healthcare user authentication system by kinetic energy harvesting," in *Network and Distributed Systems Security (NDSS) Symposium 2017*, 2017. [Online]. Available: http://dx.doi.org/10.14722/ndss.2017.23023

[51] Y. Yang and J. Sun, "Energy-efficient w-layer for behavior-based implicit authentication on mobile devices," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.

[52] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "Svc2004: First international signature verification competition," in *Biometric Authentication*, D. Zhang and A. K. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 16–22.

[53] S. Yi, Z. Qin, E. Novak, Y. Yin, and Q. Li, "Glassgesture: Exploring head gesture interface of smart glasses," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[54] L. Zhang, L. Li, A. Yang, Y. Shen, and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," *Pattern Recognition*, vol. 69, pp. 199 – 212, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317301681

[55] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: ACM, 2017, pp. 57–71. [Online]. Available: http://doi.acm.org/10.1145/3133956.3133962

[56] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine." *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993. [Online]. Available: http://clinchem.aaccjnls.org/content/39/4/561