Significance Test in Operational ROC Analysis

Jin Chu Wu*[a], Alvin F. Martin[a], Raghu N. Kacker[b] and Charles R. Hagwood[c]
[a]Information Access Division, [b]Mathematical and Computational Sciences Division, [c]Statistical Engineering Division, Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

To evaluate the performance of fingerprint-image matching algorithms on large datasets, a receiver operating characteristic (ROC) curve is applied. From the operational perspective, the true accept rate (TAR) of the genuine scores at a specified false accept rate (FAR) of the impostor scores and/or the equal error rate (EER) are often employed. Using the standard errors of these metrics computed using the nonparametric two-sample bootstrap based on our studies of bootstrap variability on large fingerprint datasets, the significance test is performed to determine whether the difference between the performance of one algorithm and a hypothesized value, or the difference between the performances of two algorithms where the correlation is taken into account is statistically significant. In the case that the alternative hypothesis is accepted, the sign of the difference is employed to determine which is better than the other. Examples are provided.

Keywords: Receiver operating characteristic (ROC) curve; Fingerprint; Biometrics; Nonparametric bootstrap; Standard error; Confidence interval; Significance test; Comparison

1. Introduction

To evaluate the performances of algorithms for fingerprint technologies on large data sets in particular, and for biometrics in general, a receiver operating characteristic (ROC) curve is used as an important tool. In analyzing fingerprint data, genuine scores are generated by comparing two different fingerprint images of the same subject, and impostor scores are created by matching two fingerprint images of two different subjects. Both scores may be referred to as similarity scores in this article. These two sets of similarity scores constitute two distributions, respectively, as schematically depicted in Figure 1 (A) for continuous similarity scores. These two distributions are interrelated with each other by the matching algorithm that generates them. All statistics of interest derived from them are influenced under the combined impact of these two samples.

The cumulative probabilities of genuine scores and impostor scores from the highest similarity score down to a specified similarity score (i.e., the threshold score) are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. As the threshold moves from the highest similarity score down to the lowest similarity score, an ROC curve is then constructed in the FAR-and-TAR coordinate system, as drawn in Figure 1 (B). Thus, biometric evaluation is a two-

* Corresponding author. Tel: + 301-975-6996; fax: + 301-975-5287. E-mail address: jinchu.wu@nist.gov.

Biometric Technology for Human Identification VII, edited by B. V. K. Vijaya Kumar, Salil Prabhakar, Arun A. Ross,
Proc. of SPIE Vol. 7667, 76670I · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.849529

distribution, one-curve, and four-domain (i.e., true accept, false accept, true reject, and false reject) issue. The accept region is where similarity scores are greater than the threshold score and the reject region is on the other side. Different scoring systems can be converted to integer scores, if they are not. Thus, the probability distribution functions of similarity scores are all discrete and an ROC curve is no longer a smooth curve [1].

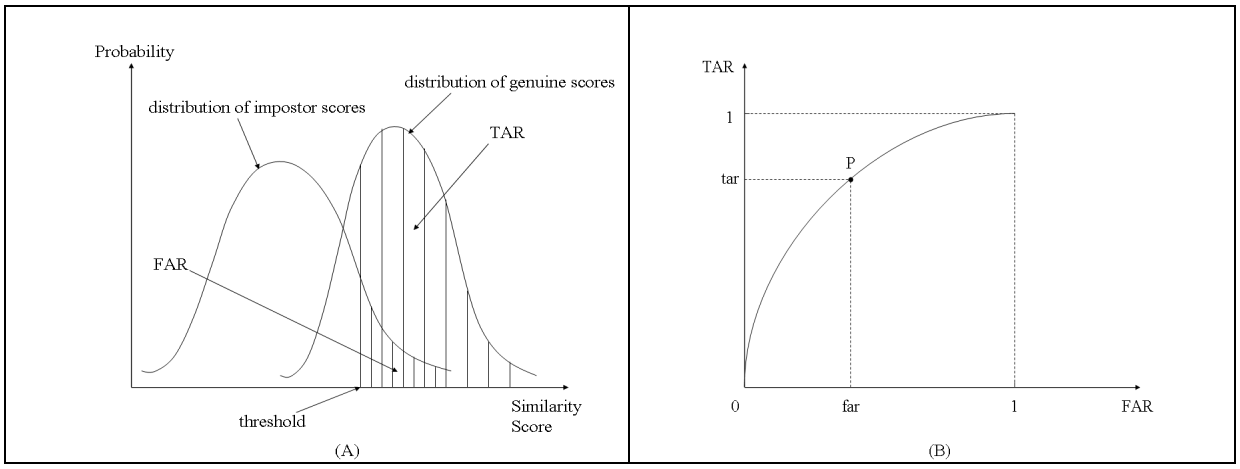


Figure 1. (A): A schematic diagram of distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. (B): A schematic drawing of an ROC curve constructed by moving a threshold from the highest similarity score down to the lowest similarity score.

As extensively explored in our previous studies [1], it was revealed that 1) usually there is no underlying parametric distribution function for genuine scores and impostor scores; 2) the distribution of genuine scores and the distribution of impostor scores are considerably different in general; and 3) the distributions vary substantially from algorithm to algorithm in ways that differentiate algorithms in terms of matching accuracy. This suggests that the nonparametric analysis be appropriate for evaluating fingerprint-image matching algorithms on large-scale data sets. Therefore, the empirical distribution is assumed for each of the observed similarity scores.

An ROC curve can be measured by computing the area under ROC curve (AURC) [1, and references therein]. If using the trapezoidal rule, this area is equivalent to the Mann-Whitney statistic formed by genuine and impostor scores. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of AURC. Because the Mann-Whitney statistic is asymptotically normally distributed regardless of the distributions of genuine and impostor scores thanks to the Central Limit Theorem, the Z statistic formulated in terms of areas under two ROC curves along with their variances and the correlation coefficient can be used to test the significance of the difference of these two ROC curves.

However, from the operational perspective, the measures TAR, FAR, and threshold are used. It is illustrated in Figure 1 that these three variables are related to each other, and any one of these three variables can determine the other two variables [2]. First, in practice it is never required that TAR be specified in the first place. Second, different algorithms may invoke different threshold scores to generate TAR and FAR. It is hard to compare algorithms using these two metrics TAR (the larger the better) and FAR (the smaller the better) simultaneously. So, TAR and FAR for a given threshold

are not good metrics for comparing two algorithms [3]. As a result, for comparison purpose, the metric TAR for a specified FAR is the one that is investigated.

The two rates 1 – TAR and FAR, which are analogous to the probabilities of type I error and type II error, respectively, are traded off with each other. When these two rates are equal, such a rate is defined as the equal error rate (EER). Generally speaking, the smaller the EER is, the more apart the two distributions of genuine scores and impostor scores are, thus the higher the ROC curve is and the more accurate the matching algorithm is [1, 4]. Hence, the EER can be used as a metric to evaluate and compare the performances of matching algorithms.

Because of the discreteness of two probability distribution functions as stated above, as opposed to continuous distribution, some concepts and definitions need to be established and modified accordingly [2, 5]. For instance, first, the ties of genuine scores and/or impostor scores at a threshold can often occur on large fingerprint data sets and thus must be taken into account while computing the estimated TAR at a specified FAR. Second, while computing the cumulative discrete probability at a score, the probability at this score must be taken into account [6]. Third, it seems that generally speaking there does not exist such a similarity score (range) at which the probability of type I error is exactly equal to the probability of type II error.

The sampling variability can result in uncertainties of measures in ROC analysis. As a result, while comparing the performances of two matching algorithms, the measurement uncertainties must be taken into account. Then, the question arises: how to calculate those uncertainties? In our previous studies [2, 5], the uncertainties of the measures in ROC analysis in terms of standard errors (SE) and 95% confidence intervals (CI) were computed using the nonparametric two-sample bootstrap based on our extensive investigation of bootstrap variability on large fingerprint data sets [2, 5, 7, 8]. The two samples are referred to a set of genuine scores and a set of impostor scores.

As is well-known, the bootstrap method assumes that an independent and identically distributed (i.i.d.) random sample of size n is drawn from a population with its own probability distribution. Our large government data bases used for developing similarity scores were randomly collected from the real practice rather than using multiple acquisitions and thus had no dependencies. The SEs of AURC on our data bases computed using the nonparametric two-sample bootstrap with the assumption of i.i.d. matched very well the analytical results using the Mann-Whitney statistic (this work is underway). Moreover, an example was made, in which the similarity scores were created using the random generator of normal distribution "rnorm" in R [9]. Certainly, there is no dependency among these scores at all. The result shown in the example behaved in the exactly same way as the results derived from our data bases. As a result, in our work, the random sample is assumed to be i.i.d..

Under the assumption of i.i.d., the objects of nonparametric two-sample bootstrap are individuals in the sample. As pointed out in Ref. [5], if the data base had dependencies due to multiple biometric acquisitions, then the assumption of i.i.d. could not be made. Hence, the sample may need to be regrouped into subsets according to dependencies, and the objects of nonparametric two-sample bootstrap are the subsets of the sample in order to preserve the dependencies [8, 10, 11]. However,

everything else in the bootstrap method remains intact. Certainly, how to regroup the sample into subsets will have impact on the bootstrap results.

The number of two-sample bootstrap replications in the fingerprint applications was determined to be 2 000 based on our bootstrap variability studies [5]. In our applications, the total number of genuine scores is a little over 60 000 and the total number of impostor scores is a little over 120 000 [12]. With this amount of similarity scores, the FAR was set to be 0.001 while dealing with TAR [4].

Regarding the issues of comparisons, here are two categories. The first category is the one-algorithm significance test, which is to determine whether the difference between the performance of one fingerprint-image matching algorithm and a hypothesized value is real or by chance. The second category is the two-algorithm significance test, which is to investigate whether the difference between the performances of two algorithms is statistically significant. As stated above, in this respect the metric TAR at a given FAR and/or the metric EER are typically employed.

In some applications, it is of interest to determine if the matching accuracies derived from two different samples of data are statistically different. Indeed, this case does belong to the second category, in which the performances of two different algorithms on the same dataset are replaced by the performances of a single algorithm on two different datasets.

Such comparison issues can be dealt with intuitively to some extent using 95 % CIs. But it is hard to reach any conclusion while the 95 % CIs overlap for two-algorithm significance test. Nonetheless, such an approach cannot provide any quantitative information, such as how much the *p*-value is, i.e., what the statistical significance of the difference is. Thus, the issue of determining whether the difference is real or by chance must be dealt with using the statistical hypothesis testing.

The relationship between the two types of 95 % CIs for the statistics TAR at a given FAR and EER was examined in all cases encountered in Ref. [2, 5]. One type of 95 % CI was computed using the definition of quantile; another type of 95 % CI was calculated if the distribution of 2000 bootstrap replications of the statistic was assumed to be normal. It was found that these two types of 95 % CIs were matched up to the third to fourth decimal place. The higher the accuracy of algorithm is, the more decimal places are matched. Moreover, the Shapiro-Wilk normality test [9] was conducted on the 2000 bootstrap replications of the statistics of interest, and it was observed that the majority of *p*-values were greater than 5 %, especially for relatively high-accuracy algorithms.

All these suggest that the statistics of interest in our applications be assumed to be normally distributed regardless of the distributions of genuine and impostor scores. Under the normality assumption, the Z-test can be used to perform the significance test, as it was done for AURC [1, and references therein]. In ROC analysis, we do not know beforehand the correlated pairs of metrics, such as TAR for a given FAR, or EER, on which the hypothesis testing is conducted. Thus, the paired t-test cannot serve our purpose.

The statistics of interest of two matching algorithms may or may not be correlated, depending on how the sets of similarity scores are generated. In our applications, the sets of similarity scores were generated in a way that might cause the correlation between the two statistics of interest. Thus, an

algorithm is provided in this article to find the correlated pairs of metrics from the correlated similarity scores. Thereafter, the correlation coefficient of metrics can be computed explicitly, in order to show the relationship between the magnitudes of correlation coefficients and the accuracies of matching algorithms, and to show what the impact would be if the positive correlation coefficient were neglected. The way of computing correlation coefficient in this paper is completely different from the way in Ref. [1, and references therein], which is based on a table provided by other researchers.

Bootstrap methods have been applied widely for error estimation, and so is the use of ROC curve. Numerous references can be found [11, 13-18, and references therein]. However, employing the methods of nonparametric two-sample bootstrap in ROC analysis and conducting Z-test on ROC curve can only be found in medical applications [13-17], as pointed out in our previous work [5].

In medical applications, data samples are small. In our applications, such as biometrics and the evaluation of speaker recognition, etc., the sizes of data sets are much larger. For instance, in the fingerprint applications, hundreds of thousands of similarity scores are used. Moreover, in comparison with other applications of bootstrap methods, our statistics of interest are probabilities, such as TAR, FAR, EER, etc, rather than a sample mean [2, 5, 8] and our data samples of similarity scores have no parametric model to fit [1, 8]. Therefore, the bootstrap variability was re-studied to determine the appropriate number of bootstrap replications in our applications, in order to reduce the bootstrap variance and ensure the accuracy of the computation [5].

Further, in medical applications, the metric that is used most is AURC. From the operational perspective in our applications, the measures and accuracies of the statistics of interest, such as TAR, FAR, EER, etc. in all three scenarios were computed using the nonparametric two-sample bootstrap [2]. The significance Z-test was conducted on TAR and EER. To perform Z-test, an algorithm for computing the correlation coefficient in our application is also provided. Our methods can also be applied to dealing with AURC as well as a cost function consisting of probabilities of type I error and type II error in the evaluation of speaker recognition (this work is underway).

The general formulas of hypothesis testing for one fingerprint-image matching algorithm and two algorithms are presented in Section 2. An algorithm for computing the correlation coefficient in our applications is provided in Section 3. The results of examples involving six fingerprint-image matching algorithms[1] are shown in Section 4. Finally the conclusion and discussion is found in Section 5.

2. Significance test

As pointed out in Section 1, the hypothesis testing is performed in two categories: one-algorithm hypothesis testing and two-algorithm hypothesis testing; the statistics of interest from the operational perspective are in two scenarios: the metric TAR at a given FAR and the metric EER.

[1] Specific hardware and software products identified in this report were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

There is no reason to believe *a priori* that the performance of one algorithm is likely to be better than a hypothesized value or the performance of the other algorithm. Further, the two-tailed test is generally more conservative than the one-tailed test in the sense that the former is more difficult to reject the null hypothesis for a given significance level [19]. Thus, the two-tailed test is used in this article. In the case that the alternative hypothesis is accepted, the sign of the difference between the estimate and a hypothesized value or the two estimates is employed to determine which is better than the other.

2.1 One-algorithm hypothesis testing

Let STI denote the probability measure, such as TAR and EER, for an algorithm and μ_o denote the hypothesized value. Then, the null and alternative hypotheses are

$$\begin{aligned} H_o &: \mathrm{STI} = \mu_o \\ H_a &: \mathrm{STI} \neq \mu_o \end{aligned} \tag{1}$$

Based on the normality assumption, the Z statistic is

$$Z = \frac{\hat{\mathrm{STI}} - \mu_o}{\mathrm{SE}(\hat{\mathrm{STI}})} \tag{2}$$

where $\hat{\mathrm{STI}}$ is the estimator of the statistic of interest and $\mathrm{SE}(\hat{\mathrm{STI}})$ stands for its SE. The Z statistic is subject to the standard normal distribution with zero expectation and a variance of one. The standard error can be computed using the nonparametric two-sample bootstrap [2, 5].

While evaluating the performance of an algorithm with respect to an accuracy criterion value, besides *p*-value, other factors also need to be taken into account, such as the characteristic of the statistic of interest (the larger the better or the smaller the better) and the sign of the difference between the estimator and the accuracy criterion value. For instance, if the statistic of interest is TAR (the larger the better) and its estimator is less than μ_o, then less-than-5 % *p*-value indicates that this algorithm fails the test.

2.2 Two-algorithm hypothesis testing

Let STI_1 and STI_2 denote the probability measures, such as TAR and EER, for Algorithms 1 and 2, respectively. Then, the null and alternative hypotheses are

$$\begin{aligned} H_o &: \mathrm{STI}_1 = \mathrm{STI}_2 \\ H_a &: \mathrm{STI}_1 \neq \mathrm{STI}_2 \end{aligned} \tag{3}$$

Based on the normality assumption, the Z statistic is expressed as

$$Z = \frac{\hat{\mathrm{STI}}_1 - \hat{\mathrm{STI}}_2}{\sqrt{\mathrm{SE}^2(\hat{\mathrm{STI}}_1) + \mathrm{SE}^2(\hat{\mathrm{STI}}_2) - 2\,r\,\mathrm{SE}(\hat{\mathrm{STI}}_1)\,\mathrm{SE}(\hat{\mathrm{STI}}_2)}} \tag{4}$$

where $\hat{\mathrm{STI}}_1$ and $\hat{\mathrm{STI}}_2$ are two estimators of the statistics of interest, $\mathrm{SE}(\hat{\mathrm{STI}}_1)$ and $\mathrm{SE}(\hat{\mathrm{STI}}_2)$ stand for their SEs, respectively, and r is the correlation coefficient between STI_1 and STI_2. The Z statistic is subject to the standard normal distribution with zero expectation and a variance of one. The

standard errors can be computed using the nonparametric two-sample bootstrap [2, 5]. If the two statistics of interest are positively correlated and the correlation coefficient r is not taken into account, it can leave the denominator of Eq. (4) larger and the Z score smaller; thereby reduce the chance of detecting a difference between the performances of two algorithms.

3. An algorithm for computing the correlation coefficient

The two statistics of interest of any two algorithms may or may not be correlated, depending on how the sets of similarity scores are generated. In our tests, different fingerprint-image matching algorithms generated different sets of similarity scores, respectively, using the same set of fingerprint images. Any two scores with the same ordinal number of entry in the two sets of similarity scores were generated using the same two images, and thus co-varied. All algorithms have the same tendency to assign a higher (or lower) similarity score to the match where two fingerprint images are more (or less) similar. Such a characteristic may cause positive correlation between two sets of similarity scores of two algorithms. Subsequently, it may eventually result in the positive correlation between the statistics of interest of two algorithms. On the other hand, this correlation may be reduced due to the large magnitude of the size of datasets.

The genuine score sets of matching Algorithms A and B are denoted as

$$\mathbf{G^i} = \{ m^i_j \mid i \in \{ A, B \} \text{ and } j = 1, \ldots, N_G \} , \quad (5)$$

and the impostor score sets of Algorithms A and B are expressed as

$$\mathbf{I^i} = \{ n^i_j \mid i \in \{ A, B \} \text{ and } j = 1, \ldots, N_I \} , \quad (6)$$

where N_G and N_I are the total numbers of genuine scores and impostor scores, respectively. It is assumed that Algorithms A and B generate the same amount of genuine scores as well as impostor scores. As stated above, the two j-th genuine scores m^i_j where $i \in \{ A, B \}$ are generated using the same two images but employing different algorithms and thus co-vary. So do the two j-th impostor scores n^i_j where $i \in \{ A, B \}$.

An algorithm for computing the correlation coefficient of the statistic of interest STI, i.e., either TAR or EER, of Algorithms A and B in our applications is as follows.

Algorithm

1: **for** i = 1 **to** M **do**
2: Synchronized_WR_Random_Sampling (N_G, $\mathbf{G^A}$, $\mathbf{\Theta^A_i}$, $\mathbf{G^B}$, $\mathbf{\Theta^B_i}$)
3: Synchronized_WR_Random_Sampling (N_I, $\mathbf{I^A}$, $\mathbf{\Xi^A_i}$, $\mathbf{I^B}$, $\mathbf{\Xi^B_i}$)
4: the new genuine score set $\mathbf{\Theta^A_i}$ and the new impostor score set $\mathbf{\Xi^A_i}$ => statistic $\hat{STI}^{A_i}$
5: the new genuine score set $\mathbf{\Theta^B_i}$ and the new impostor score set $\mathbf{\Xi^B_i}$ => statistic $\hat{STI}^{B_i}$
6: **end for**
7: $\{ \hat{STI}^{A_i} \mid i = 1,\ldots,M \}$ and $\{ \hat{STI}^{B_i} \mid i = 1,\ldots,M \}$ => the correlation coefficient r^{AB}_{STI}
8: **end**
1.1: **function** Synchronized_WR_Random_Sampling (N, $\mathbf{S^A}$, $\mathbf{\Gamma^A}$, $\mathbf{S^B}$, $\mathbf{\Gamma^B}$)
1.2: **for** j = 1 **to** N **do**
1.3: randomly select WR an index $k \in \{ 1, \ldots, N \}$
1.4: $\gamma^A_j = s^A_k$

1.5: $\gamma^{B}_{j} = s^{B}_{k}$
1.6: **end for**
1.7: **end function**

where s^{A}_{k}, γ^{A}_{j}, s^{B}_{k}, and γ^{B}_{j} are members of the score sets $\mathbf{S^A}$, $\mathbf{\Gamma^A}$, $\mathbf{S^B}$, and $\mathbf{\Gamma^B}$, respectively. Based on our previous bootstrap variability studies [5], the number of iterations M is set to be 2000.

From Step 1 to 6, this algorithm runs M iterations. As indicated in Steps 2 and 3, in the i-th iteration, the synchronized WR (with replacement) random sampling is carried out on $\mathbf{G^A}$ and $\mathbf{G^B}$ ($\mathbf{I^A}$ and $\mathbf{I^B}$) to generate two new genuine (impostor) score sets $\mathbf{\Theta^A_i}$ and $\mathbf{\Theta^B_i}$ ($\mathbf{\Xi^A_i}$ and $\mathbf{\Xi^B_i}$), respectively.

From Step 1.1 to 1.7, the function, Synchronized_WR_Random_Sampling, runs N iterations, where N is the total number of genuine/impostor scores. As indicated in Step 1.3, in the j-th iteration, an index k is randomly drawn WR from the integer set { 1, …, N }. Then as indicated in Steps 1.4 and 1.5, the k-th score of the input score set $\mathbf{S^A}$ is assigned to the j-th score of the new score set $\mathbf{\Gamma^A}$, and the k-th score (i.e., synchronized) of the input score set $\mathbf{S^B}$ is also assigned to the j-th score of the new score set $\mathbf{\Gamma^B}$. With such synchronized random sampling, the co-varying similarity scores (i.e., with the same ordinal number of data entry) between Algorithms A and B are selected simultaneously, and the correlation in the similarity scores between two algorithms is preserved if there is any.

In Steps 4 and 5, after sampling, the i-th estimated statistic $\hat{STI}^{A}_{i}$ ($\hat{STI}^{B}_{i}$) of Algorithm A (B) is computed from the new score sets $\mathbf{\Theta^A_i}$ and $\mathbf{\Xi^A_i}$ ($\mathbf{\Theta^B_i}$ and $\mathbf{\Xi^B_i}$). Finally after M iterations in Step 7, the correlation coefficient r^{AB}_{STI} of the statistic of interest STI of Algorithms A and B can be calculated from the two sets of estimated statistics of interest.

This algorithm involves a synchronized random sampling. Thus, it is a stochastic process. In practice, if the *p*-value is not considerably different from the critical values, such as 5 %, 1 %, etc., then in order to reduce the computational fluctuation this algorithm needs to run multiple times. In this article, the average out of 10 runs was taken to be the resultant correlation coefficient for significance test.

4. Results

Relatively high-accuracy fingerprint-image matching Algorithms 1 through 3 and relatively low-accuracy Algorithms 4 through 6 were taken to be examples.[2] These algorithms used different types of scoring systems. Algorithms 1 through 3 were taken as examples for both one-algorithm hypothesis testing and two-algorithm hypothesis testing while the statistic of interest was assumed to be TAR at a specified FAR. Algorithms 4 through 6 were used only for two-algorithm significance test while the statistic of interest was set to be EER. The method applied to TAR can be applied to EER, and vice versa. The only difference is that for TAR the larger the better, but for EER the smaller the better.

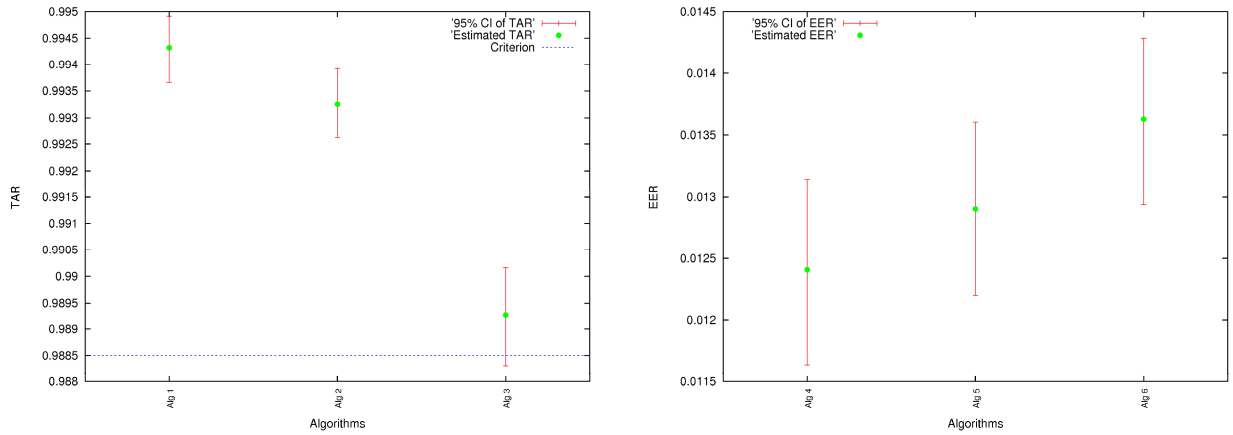
[2] The algorithms are proprietary. Hence, they cannot be disclosed.

Algorithm	TÂR (f)	SÊ	95 % Confidence interval
1	0.994322	0.000324	(0.993662, 0.994918)
2	0.993255	0.000325	(0.992622, 0.993922)
3	0.989263	0.000470	(0.988307, 0.990159)

Table 1. The estimates of TARs, SEs, and 95 % CIs for relatively high-accuracy Algorithms 1 through 3, while FAR was specified at 0.001.

Algorithm	EÊR	SÊ	95 % Confidence interval
4	0.012409	0.000378	(0.011638, 0.013148)
5	0.012903	0.000360	(0.012205, 0.013609)
6	0.013634	0.000338	(0.012940, 0.014287)

Table 2. The estimates of EERs, SEs, and 95 % CIs for relatively low-accuracy Algorithms 4 through 6.



(A) (B)

Figure 2. (A): The estimates of TARs and the corresponding 95 % CIs for relatively high-accuracy Algorithms 1 through 3, while FAR was specified at 0.001, along with the hypothesized value μ_o to be set at 0.988500. (B): The estimates of EERs and the corresponding 95 % CIs for relatively low-accuracy Algorithms 4 through 6.

The estimated TÂR (f) at a given FAR and the estimated EÊR along with their uncertainties in terms of SE and 95 % CI can be computed using the nonparametric two-sample bootstrap with 2000 bootstrap replications [2, 5]. They are shown in Table 1 and Table 2, and drawn in Figure 2 (A) and (B), respectively. In Figure 2 (A), a hypothesized value μ_o is also depicted.

4.1 One-algorithm hypothesis testing

For one-algorithm hypothesis testing, the estimate of the statistic of interest of an algorithm is compared against a hypothesized value, i.e., the accuracy criterion value, to see whether the difference is real or by chance. Suppose that the metric TAR at a given FAR is considered in the testing and the hypothesized value μ_o is set to be 0.988500.

In Figure 2 (A), the two 95 % CIs of Algorithms 1 and 2 are way above 0.988500. Thus, the performances of Algorithms 1 and 2 measured by the metric TAR are better than the accuracy criterion value 0.988500. This observation is supported by applying Eq. (2). Using the estimated TARs and SEs for these two algorithms from Table 1, it was found that the two two-tailed *p*-values were all equal to 0.0000 as presented in Table 3. This indicates that the alternative hypothesis H_a : $\hat{\mathrm{STI}} \neq \mu_o$ is very strongly accepted. Further, with the positive sign of the difference between the estimated TAR and the hypothesized value 0.988500, it is concluded that the $\hat{\mathrm{TAR}}$ (f) of Algorithm 1 and Algorithm 2 are all very significantly greater than the accuracy criterion value 0.988500. In other words, Algorithms 1 and 2 pass the test.

In Figure 2 (A), the horizontal line of the hypothesized value μ_o = 0.988500 intersects the 95 % CI of Algorithm 3. After using Eq. (2) by substituting the estimates of TAR and SE for Algorithm 3 from Table 1, it was found that the two-tailed *p*-value was 0.1049 as presented in Table 3, which is greater than 5 %. This suggests that the null hypothesis H_o : $\hat{\mathrm{STI}} = \mu_o$ be accepted. That is to say, the difference between the estimator $\hat{\mathrm{TAR}}$ (f) = 0.989263 of Algorithm 3 and the hypothesized value 0.988500 is not real but by chance at the significance level 10 %. Therefore, Algorithm 3 fails the test, if the performance is required to be better than the accuracy criterion value μ_o.

Algorithms	*p*-value
1	0.0000
2	0.0000
3	0.1049

Table 3. The two-tailed *p*-values of the statistic of interest TAR with respect to the hypothesized value μ_o = 0.988500 for relatively high-accuracy Algorithms 1 through 3.

4.2 Two-algorithm hypothesis testing

The hypothesis testing for two algorithms is not as straightforward as the one for a single algorithm. It cannot be judged merely using the confidence interval approach. In order to determine whether the difference between the performances of two fingerprint-image matching algorithms is statistically significant, the two-algorithm hypothesis testing must be carried out.

4.2.1 TAR for a given FAR

As shown in Figure 2 (A), the 95 % CI of Algorithm 1 slightly overlaps the one of Algorithm 2. But both of them are above the 95 % CI of Algorithm 3. What is the statistical significance of the differences among the performances of these three algorithms?

The correlation coefficient of the statistic of interest TAR between two matching algorithms can be computed using the algorithm presented in Section 3. For relatively high-accuracy Algorithms 1 to 3, the average correlation coefficients out of ten runs are listed in Table 4. The positive correlation coefficients for TARs are near 0.5. This indicates that all high-accuracy fingerprint-image matching algorithms have the same tendency to assign higher (lower) similarity scores to the matching results of more (less) similar images.

Algorithms	1	2	3
1	1.000000	0.496089	0.454423
2		1.000000	0.493979
3			1.000000

Table 4. The average correlation coefficients of the statistic of interest TAR out of ten runs among relatively high-accuracy Algorithms 1 through 3.

Algorithms	1	2	3
1	1.0000	0.0011	0.0000
2		1.0000	0.0000
3			1.0000

Table 5. The two-tailed *p*-values of two statistics of interest TARs for relatively high-accuracy Algorithms 1 through 3, where the correlation coefficient was taken into account.

For relatively low-accuracy Algorithms 4 to 6, the average correlation coefficients of the statistic of interest TAR out of ten runs are 0.223933, 0.240295, and 0.266922, respectively. They are not as high as those for the high-accuracy algorithms. It is expected that the tendency of assigning higher (lower) similarity scores to the matching results of more (less) similar images for relatively low-accuracy algorithms is not as strong as the tendency for high-accuracy algorithms. Thus, these results provide evidence that the synchronized algorithm for computing the correlation coefficient is quite reasonable.

After applying Eq. (4) using the estimates of TARs and SEs from Table 1 and the correlation coefficients from Table 4, the two-tailed *p*-values of two statistics of interest TARs for Algorithms 1 through 3 can be computed and are shown in Table 5. The two-tailed *p*-value between Algorithms 1 and 2 is 0.0011, and other two are 0.0000.

These two-tailed *p*-values are all much less than 5 %. It suggests that the alternative hypothesis H_a : $\hat{\mathrm{STI}}_1 \neq \hat{\mathrm{STI}}_2$ be strongly accepted. In other words, the differences of performances among Algorithms 1 through 3 are very significant, even though the 95 % CI of Algorithm 1 does slightly overlap the one of Algorithm 2. It follows from the sign of the difference between the two corresponding estimated TARs that the performance of Algorithm 1 is better than the performance of Algorithm 2; and the performances of both of them are better than the performance of Algorithm 3.

4.2.2 EER

As shown in Figure 2 (B), the three 95 % CIs of EERs of Algorithms 4 through 6 mutually overlap. In such a circumstance, how can the statistical significance of the differences among the performances of these three algorithms be determined?

For relatively low-accuracy Algorithms 4 through 6, the average correlation coefficients of the statistic of interest EER out of ten runs are presented in Table 6. For high-accuracy Algorithms 1 to

3, the corresponding average correlation coefficients are 0.513037, 0.529609, and 0.567842, respectively. They are all larger than those for relatively low-accuracy Algorithms 4 through 6. This is expected as discussed in Subsection 4.2.1, and also supports the synchronized algorithm for computing the correlation coefficient.

Algorithms	4	5	6
4	1.000000	0.360888	0.398198
5		1.000000	0.453439
6			1.000000

Table 6. The average correlation coefficients of the statistic of interest EER out of ten runs among relatively low-accuracy Algorithms 4 through 6.

Using Eq. (4) by substituting the estimates of EERs and SEs from Table 2 and the correlation coefficients from Table 6, the two-tailed *p*-values of two statistics of interest EERs for Algorithms 4 through 6 can be calculated and are presented in Table 7.

Algorithms	4	5	6
4	1.0000	0.2370	0.0019
5		1.0000	0.0457
6			1.0000

Table 7. The two-tailed *p*-values of two statistics of interest EERs for relatively low-accuracy Algorithms 4 through 6, where the correlation coefficient was taken into account.

The two-tailed *p*-value between Algorithms 4 and 5 is 0.2370, which is much greater than 5 %. It suggests that the null hypothesis $H_o : \hat{STI}_1 = \hat{STI}_2$ be accepted. In other words, the difference between the performances of Algorithms 4 and 5 is by chance, i.e., not statistically significant, even though the estimated $\hat{EER}$ 0.012409 of Algorithm 4 is lower than the estimated $\hat{EER}$ 0.012903 of Algorithm 5. To some extent, this conclusion is supported by the fact that the 95 % CI of Algorithm 4 heavily overlaps the one of Algorithm 5, as illustrated in Figure 2 (B).

The two-tailed *p*-value between Algorithms 5 and 6 is 0.0457. Without considering the correlation coefficient, it increases to 0.1392. As pointed out in Subsection 2.2, neglecting the correlation coefficient can reduce the chance of detecting a difference between the performances of two algorithms. Since 0.0457 is slightly less than 5 %, the alternative hypothesis $H_a : \hat{STI}_1 \neq \hat{STI}_2$ is accepted with reasonably strong evidence, despite that the 95 % CI of Algorithm 5 quite overlaps the 95 % CI of Algorithm 6. Further due to the sign of the difference between the two estimated EERs, the performance of Algorithm 5 is reasonably better than the performance of Algorithm 6.

The two-tailed *p*-value between Algorithms 4 and 6 is 0.0019, which is less than 5 %. It suggests that the alternative hypothesis $H_a : \hat{STI}_1 \neq \hat{STI}_2$ be strongly accepted, although the 95 % CIs of these two algorithms slightly overlap. Moreover, because of the sign of the difference between the

two estimated EERs, the performance of Algorithm 4 is considerably better than the performance of Algorithm 6.

In addition, the *p*-value 0.0019 between Algorithms 4 and 6 is much smaller than the p-value 0.0457 between Algorithms 5 and 6. It indicates that the difference between the performances of Algorithms 4 and 6 is more statistically significant than the difference between the performances of Algorithms 5 and 6. To some extent, this conclusion can be supported by the relationship among the 95 % CIs of Algorithms 4 to 6 as illustrated in Figure 2 (B).

5. Conclusion and discussion

In operational ROC analysis of fingerprint-image matching algorithms, it is very important to determine whether the difference between the performance of one algorithm and an accuracy criterion value, or the difference between the performances of two algorithms where the correlation is taken into account is statistically significant. In this regard, no study was found to date. For such comparison issues, the two statistics of interest, TAR at a specified FAR and EER, are typically employed.

These two statistics of interest can be assumed to be normally distributed regardless of the distributions of genuine scores and impostor scores. This assumption is supported by the matches in various cases between two types of 95 % CIs. One is computed using the definition of quantile, and the other is calculated if the distribution of 2000 bootstrap replications of the statistic of interest is assumed to be normal. It is also partly supported by the Shapiro-Wilk normality test.

Under the normality assumption, the Z-test can be applied. The Z statistic is formulated using the estimated TAR at a specified FAR or EER of one algorithm or two algorithms along with their variances and correlation coefficient, and it is subject to the standard normal distribution with zero expectation and a variance of one. All the standard errors can be computed using the nonparametric two-sample bootstrap with 2000 bootstrap replications based on our variability study of bootstraps.

In this article, an algorithm is provided to calculate the correlation coefficient between two statistics of interest of two fingerprint-image matching algorithms, under the assumption that for these two algorithms any two scores with the same ordinal number of entry in the two sets of similarity scores were generated using the same two images. Otherwise the user needs to provide the correlation coefficient, if they are correlated. Further, in our case, if the orders in the two score sets were changed manually, i.e. the similarity scores with the same ordinal number did not co-vary anymore, the correlation coefficients computed using the algorithm in Section 6.3 were close to zero. This also supports the synchronized algorithm for computing the correlation coefficient.

This algorithm is a stochastic process, since it involves a synchronized sampling. In practice, if the *p*-value is not considerably different from the critical values, such as 5 %, 1 %, etc., then in order to reduce the computational fluctuation this algorithm needs to run for several times (ten in our case). The average correlation coefficient out of these correlation coefficients is taken to be the resultant correlation coefficient for significance test.

In Ref. [20], the false non-match rate (FNMR) for a given FAR was employed as a metric to evaluate the fingerprint technologies. FNMR, analogous to the probability of type I error, is equal to 1 – TAR. It is trivial to prove that the standard errors of TAR and FNMR for an algorithm are equal, the correlation coefficients of TAR and FNMR given two algorithms are also the same, and so are the Z scores and the *p*-values of TAR and FNMR for two algorithms. The only difference is that the upper (or the lower) bound of 95 % CI of FNMR is one minus the lower (or the upper) bound of 95 % CI of TAR [2, 5]. As a result, every method for TAR stated in this article can be applied to FNMR, and every result obtained for TAR holds true for FNMR.

While conducting comparisons, in some cases the 95% CIs can be applied to some extent. Nonetheless, the issue of determining quantitatively whether the difference is real or by chance must be dealt with using the significance test. As presented in Subsection 4.2.1, although the 95 % CIs of Algorithms 1 and 2 did slightly overlap, the hypothesis testing showed that the difference of performances between these two algorithms was very statistically significant. And also as discussed in Subsection 4.2.2, all three 95 % CIs were mutually overlapped to a certain degree, but the hypothesis testing showed that the statistical significances of the differences in performances among the three algorithms were quite different accordingly in terms of *p*-values.

Conventionally, if the two-tailed *p*-value is greater than or equal to 5 %, the null hypothesis is accepted; if it is less than 5 %, the alternative hypothesis is accepted. In the literature [8], it suggested: If *p*-value is less than 0.10, borderline evidence is against H_o; if *p*-value is less than 0.05, reasonably strong evidence is against H_o; if *p*-value is less than 0.025, strong evidence is against H_o; if *p*-value is less than 0.01, very strong evidence is against H_o.

References

1. Wu, J.C. and Wilson, C. L., "Nonparametric analysis of fingerprint data on large data sets", Pattern Recognition 40(9), 2574-2584 (2007).
2. Wu, J.C., "Operational measures and accuracies of ROC Curve on large fingerprint data Sets", NISTIR 7495, National Institute of Standards and Technology, (May 2008).
3. Wu, J.C., Martin, A. F. and Kacker, R. N., "Hypothesis test of fingerprint-image matching algorithms in operational ROC analysis", NISTIR 7586, National Institute of Standards and Technology, (June 2009).
4. Wu, J.C. and Garris, M. D., "Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion", in Biometric Technology for Human Identification IV, Proceedings of SPIE Vol. 6539, 65390N (2007).
5. Wu, J.C., "Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap", NISTIR 7449, National Institute of Standards and Technology, (September 2007).
6. Ostle, B. and Malone, L. C., Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, (1988).
7. Efron, B., "Bootstrap methods: Another look at the Jackknife", Ann. Statistics 7, 1-26, (1979).
8. Efron, B. and Tibshirani, R. J., An Introduction to the Bootstrap, Chapman & Hall, New York, (1993).

9. R: A Language and Environment for Statistical Computing, The R Development Core Team, Version 2.8.0, (2008), at http://www.r-project.org/.
10. Liu, R.Y. and Singh, K., "Moving blocks jackknife and bootstrap capture weak dependence", Exploring the limits of bootstrap, ed. by LePage and Billard. John Wiley, New York, (1992).
11. Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K. and Senior, A. W., Guide to Biometrics, Springer, New York, 269-292 (2003).
12. Wu, J.C. and Wilson, C. L., "An empirical study of sample size in ROC-curve analysis of fingerprint data", in Biometric Technology for Human Identification III, Proceedings of SPIE Vol. 6202, 620207 (2006).
13. Hanley, J. A. and McNeil, B. J., "A method of comparing the areas under receiver operating characteristic curves derived from the same cases", Radiology 148, 839-843 (1983).
14. DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", Biometrics 44, 837-845 (1988).
15. Mossman, D., "Resampling techniques in the analysis of non-binormal ROC data", Medical Decision Making 15(4), 358-366 (1995).
16. Platt, R. W., Hanley, J. A. and Yang, H., "Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test", Statistics in Medicine 19(3), 313-322 (2000).
17. Jensen, K., Muller, H.-H. and Schafer, H., "Regional confidence bands for ROC curves", Statistics in Medicine 19(4), 493-509 (2000).
18. Dass, S. C., Zhu, Y.F. and Jain, A. K., "Validating a biometric authentication system: sample size requirements", IEEE Trans. Pattern Analysis and Machine Intelligence 28(12), 1902-1913 (2006).
19. Box, G. E. P., Hunter, J. S. and Hunter, W. G., Statistics for experimenters: design, innovation, and discovery, second ed., John Wiley & Sons, Inc., New York, (2005).
20. Cappelli, R., Maio, D., Maltoni, D., Wayman, J. L. and Jain, A. K., "Performance evaluation of fingerprint verification systems", IEEE Trans. Pattern Analysis and Machine Intelligence 28(1), 3-18 (2006).