
# BEAT

## Biometrics Evaluation and Testing

http://www.beat-eu.org/

Funded under the 7th FP (Seventh Framework Programme)

Theme SEC-2011.5.1-1
[Evaluation of identification technologies, including Biometrics]

# D3.5: Advanced Metrics for the Evaluation of Biometric Performance

**Due date:** 31/03/2014 **Submission date:** 24/02/2014
**Project start date:** 01/03/2012 **Duration:** 48 months
**WP Manager**: Julian Fierrez **Revision:** 0

**Author(s):** Norman Poh (UNIS), Chi Ho Chan (UNIS), Josef Kittler (UNIS),Javier Galbally (UAM), Julian Fierrez (UAM),

| **Project funded by the European Commission in the 7th Framework Programme (2008-2010)** | | |
|---|---|---|
| **Dissemination Level** | | |
| PU | Public | No |
| RE | Restricted to a group specified by the consortium (includes Commission Services) | Yes |
| CO | Confidential, only for members of the consortium (includes Commission Services) | No |


# D3.5: Advanced Metrics for the Evaluation of Biometric Performance

**Abstract:**

One of the key objectives of the BEAT project and its associated platform is the performance evaluation of biometric systems following reproducible, standard protocols. For this task, it is of the utmost importance to make availability well defined metrics that enable fair and objective comparison of results across several experimental settings or different data sets. In this regard, deliverable D3.3 of the project presented an inventory of well established metrics for the evaluation of biometric performance that have been used for quite some time within the specialized literature. Those metrics were later implemented and documented in deliverable D3.4 for their inclusion in the platform. However, there are still some evaluation scenarios for which no clear assessment methodology has yet been defined. This includes, for instance, the inference of biometric performance between datasets with clearly different samples in terms of biometric quality. The present deliverable extends the metrics already introduced in D3.3 and D3.4 with a novel biometric probabilistic test that permits the estimation of a system accuracy accross different databases.

# Contents

# 1 Introduction

Biometric authentication is a process of verifying an identity claim using a person's behavioral and physiological characteristics. There are several factors that can affect a biometric system's performance. Some of these factors are the deformable nature of biometric traits, corruption by environmental noise, variability of biometric traits over time, the state of users (especially behavioral biometrics) and occlusion by the user's accessories. As a consequence, even if two biometric samples are acquired from the same user, the system cannot produce *exactly* the same output score. Therefore, when assessing the performance, the uncertainty introduced by these numerous and often uncontrolled distortions

The goal of this report is to deliver a tool that allows a user to explore the degree of different factors on the resultant system performance. For instance, if the user knows that there are certain proportion of "good", "bad", and "ugly" samples, then the proposed tool enables the user to freely mix the prior probability so as to match a target application. This allows a certain degree of generalisation to a typical biometric performance curve, in the form of Detection Error Trade-off (DET) or equivalently Receiver's Operating Characteristics (ROC) curve. This curve is a plot of False Nonmatch Rate (FNMR) or False Rejection Rate in the Y-axis versus the False Match Rate (FMR) or False Acceptance Rate in the X-axis.

We present here a few use-case scenarios where our biometric simulation tool can be used:

- **Assessing biometric system performance operating with multiple sensors**: In border controls, it is expected that a biometric system will operate with several sensors. For instance, a biometric sensor that is installed at the port of entry may necessarily be the same as the one installed at a corresponding port of exit. In practice, it is common to have several ports of entry and ports of exit. In another scenario, older worn sensors may be replaced by newer ones but of a different type. In both examples, one has to face a practical problem whereby a biometric system has to compare two samples acquired by two different sensors. Under cross-sensor comparison, the system performance may be suboptimal, e.g., [**?**, **?**]. When a biometric system may operate in an environment wherein there is a certain mixing of proportions of same-device versus cross-device comparisons, the proposed algorithm can be used to simulate the performance. This allows one to estimate the number of false alarm cases more realistically.

- **Assessing biometric system performance under varying sample quality or operating conditions**: It is now well accepted that the biometric system performance is dependent on a sample population to some extent. For instance, it has been documented that the fingerprint recognition rates of older women and workers in certain industries are likely to be lower than the general population. When the proportion of demographics of a design data set is significantly different from that of a target operating environment, it is unlikely that the biometric performance curve,

as measured on the design data set, is representative of the target operating environment. By setting the prior of the demographics appropriate, the proposed algorithm can be used to produce a certain generalised performance that better matches the target condition.

One pre-requisite to predicting or modelling the biometric system performance under different operating conditions is the need to quantify the certainty of the predicted performance. This addresses the upper and lower bound of the performance. For this reason, we derive the confidence interval around a predicted DET curve, at the desired proportion of mixing as set by the user, using a two-step bootstrap strategy as documented in [**?**]. This approach explicitly considers the *correlation structure* or dependency of the matching scores. If $sim(\mathcal{T}, \mathcal{Q})$ is the similarity between a template and a query sample, the similarity scores $sim(\mathcal{T}, \mathcal{Q}_i)$ for all query samples belonging to the same user $i$ are correlated, whether or not the comparison is a match or a nonmatch (where the template belongs another person).

An example of bootstrapped demonstration is shown in Figure 1. For a video animation, check out `http://youtu.be/VUgJ1xh4sOU`.

# 2 Methodology

## 2.1 Overview

To realise the prediction of performance on a new operating condition, we first need to determine the number of noise factors in the matching scores for both the match and nonmatch comparisons separately. Let us consider the following scenarios:

- **Matching with multiple sensors**: In a multi-sensor environment, it is common to have the template produced by one sensor to be matched by a query sample produced by another sensor. If there are $N$ sensor, then the total number of factors is "$n$ choose 2", or $n(n-1)/2+n$ for any combination of two sensors *and* the template and query generated by the same sensor. Thus, if there three sensors enumerated by $S_1, S_2, S_3$, we have to consider up to three cross-device combinations, namely, $(S_1, S_2)$, $(S_1, S_3)$, and $(S_2, S_3)$, and three matching sensor comparisons, namely $(S_1, S_1)$, $(S_2, S_2)$, and $(S_3, S_3)$. Therefore, there are six score sets for the match (or genuine) comparisons and another six score sets for nonmatch (or impostor) comparisons.

- **Matching with varying quality**: With different sample quality, it is possible to quantise the samples into several categories. For example, the NIST Fingerprint Quality Asessment software, NFIQ, quantises a biometric sample into five levels of quality. Based on this automatically derived category, we divide the matching scores – for the match and nonmatch comparisons – into their respective quality levels.

- **Matching under different demographics**: Some biometric systems may exhibit performance bias by age or gender. This gender bias is evident in the past evaluation

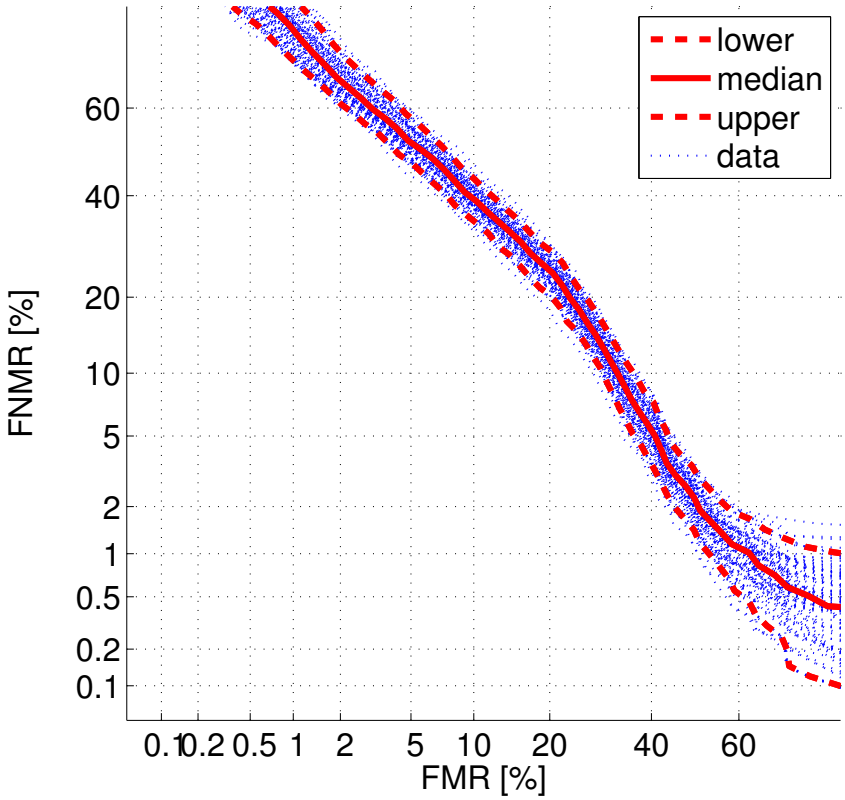



Figure 1: An exmample of boostrapped DET curve. The thin blue dashed curves are bootstrapped sample of DET curves. The thick red dashed lines indicate the lower and upper bound of the median curve (continuous red line).

of speaker and face biometric systems [**?**], for instance. Therefore, under demographic shifts, it is reasonable to expect that the system may perform differently for different demographic sectors. For this reason, it is sensible to use the demographic information as a factor.

We now describe the overall architecture of the proposed system. Let us consider the nonmatch comparison scores first. For each of the score sets representing a given factor, one can proceed to estimating its cumulative density function *cdf*, from which a number of bootstraps can be generated. The bootstrapped curves are then combined in such a way that if there are $D$ score set each containing $B$ bootstrapped *cdf* curves, one obtains $B$ combined nonmatch *cdf*s. The combination module weighs the $D$ factors using mixing coefficients set by the user. The process is then repeated for the match comparison scores, hence obtaining another $B$ combined match *cdf* curves. The two sets of curves are combined to form $B$ *bootstrapped* DET curves from which the confidence intervals of the DET curves are estimated. Figure 2 illustrates a data flow diagram of the proposed algorithm.

The basis for which the *cdf*s of different factors are combined is rooted in the Bayesian theorem, which is described in Section 2.2. The *cdf*s-combination procedure is described

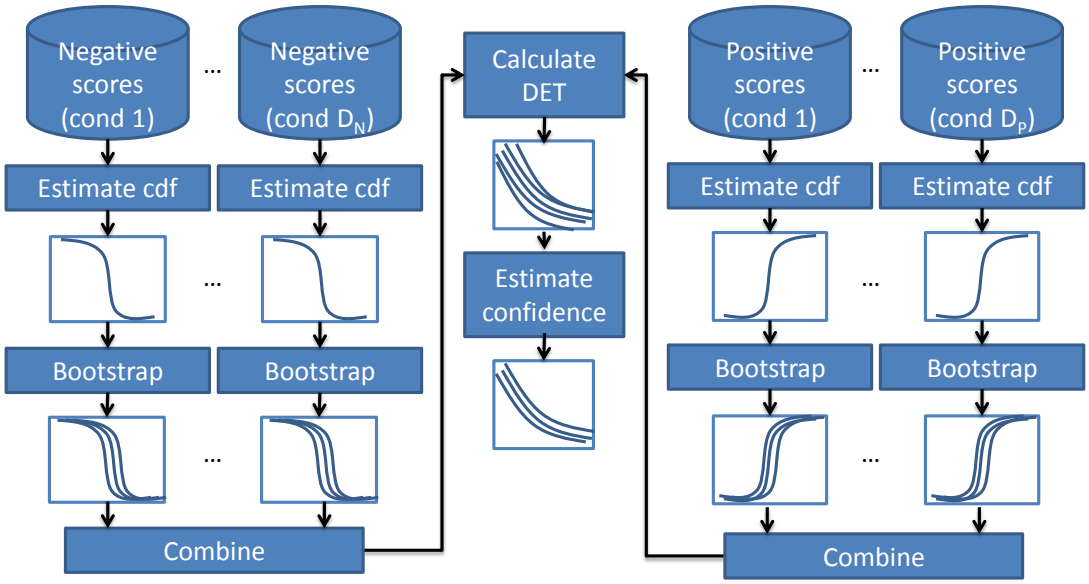



Figure 2: The architecture of the proposed system. There are $D_N$ negative score sets and $D_P$ positive score sets.

first because it is the key ingredient in making performance generalisation possible. The bootstrapping procedure that considers the correlation structure of matching scores is described in Section 2.3. Finally, we present the procedure of estimating DET confidence given a set of bootstrapped DET curves in Section 2.4. These three components, namely, combining *cdf*s, bootstrapping, and DET confidence, form the basis of the proposed performance generalisation algorithm.

## 2.2 Conditional estimate of *cdf*

We shall consider a quality factor, $Q$, to be discrete and disjoint from one another. The distribution of a factor-specific class-conditional score can then be described by $p(y|\omega_k, Q)$ for a biometric matching score $y$ and class $\omega_k$ where $\omega_1$ denotes a match (genuine) comparison and $\omega_0$ denotes a nonmatch (impostor) comparison. Therefore, if the score is similarity score, we expect that $E[y|\omega_1] > E[y|\omega_0]$, that is the mean match score is greater than the mean nonmatch score. The score distribution independent of any factor, $p(y|\omega_k)$, is a simply a mixture of the factor-specific score distributions weighted by their respective prior probability of the factor, $P(Q)$ , i.e.,

$$p(y|\omega_k) = \sum_Q p(y|\omega_k, Q)P(Q) \tag{1}$$

Very often, one has to assume the form of distribution of $p(y|\omega_k, Q)$, which can be estimated via a parametric technique such as any exponential family of one dimensional distribution or else a non-parametric technique such as the Parzen window or kernel density.

Fortunately, since our ultimate purpose is to estimate biometric performance, we do not need to estimate any density. Instead, we need only to estimate of cumulative density function (*cdf*) of the scores for a given decision threshold $\tau$. Starting from Eqn. (1), we can see that the *cdf* of *factor-independent* score, $P(y < \tau|\omega_k)$, is the summation of the mixture of the *cdf* of factor-specific score, $P(y < \tau|\omega_k, Q)$, weighted by their respective prior probability of the factor, $P(Q)$ , i.e.,

$$P(y < \tau|\omega_k) = \sum_Q P(y < \tau|\omega_k, Q)P(Q) \tag{2}$$

The False Match Rate (FMR) is defined by the *cdf* of similarity scores belonged to non-match class, $\omega_0$ greater than the decision threshold, $\tau$, whereas the the False Nonmatch Rate (FNMR) is defined by the *cdf* of similarity scores belonged to match class, $\omega_1$, smaller than the decision threshold, $\tau$, as follow:

$$\begin{aligned} \text{FMR}(\tau) &= 1 - P(y < \tau|\omega_0) & (3)\\ \text{FNMR}(\tau) &= P(y < \tau|\omega_1) & (4) \end{aligned}$$

We note that FMR is a monotonic decreasing function of the decision threshold whereas FNMR is a monotonic increasing function of the decision threshold in the similarity score space.

In order to estimate FMR and FNMR for a given test or target operation, we can plug Eqn. (2) into the above two equations. In this process, we need to further precise that $\hat{P}_{train}(y < \tau|\omega_1, Q)$ comes from the training or design data whereas $P_{test}(Q)$ depends only on the target test or operational condition. The resultant predicted FMR and FNMR are given by:

$$\begin{aligned} \hat{\text{FMR}}_{test}(\tau) &= 1 - \sum_Q \hat{P}_{train}(y < \tau|\omega_0, Q)P_{test}(Q) & (5)\\ \hat{\text{FNMR}}_{test}(\tau) &= \sum_Q \hat{P}_{train}(y < \tau|\omega_1, Q)P_{test}(Q) & (6) \end{aligned}$$

respectively, for all possible $\tau$ values. We can then plot a ROC or DET curve based on the pair (FMR, FNMR) for all possible $\tau$ values.

The main assumption deployed in Eqn. (5) and Eqn. (6) is that the *cdf* of a given factor remains the same in the design and operational (test) conditions. This is to say $\hat{P}_{train}(y < \tau|\omega_0, Q)$ and $\hat{P}_{test}(y < \tau|\omega_0, Q)$ are the same. This condition is satisfied if and only if *all other factors* remain the same between the design (train) and operational (test) conditions. What is not required to remain the same is the prior probability of the factors, which can vary across the data sets. In addition, the subjects in the design and operational conditions may also be mutural exclusive. Because of the difference in subjects, the only way to validate the assumption is by carrying out experiments. This will be further discussed in Section 3.

## 2.3 Subset Bootstrapping of *cdf*

The conventional confidence interval estimation assumes that all samples are independent and identically distributed. This assumption is violated in any experimental outcome of a biometric experiment. This is because the comparison scores originating from the same template are dependent on each other. For this reason, if there are $U$ users, one should sample the user identity set with replacement. The scores that are associated with the bootstrapped user set will then constitute a bootstrapped sample of the scores which are then constitute the *cdf*. A set of these *cdf*'s form the basis of the estimating FMR and FNMR.

Each bootstrapped score set consists of all the scores belonging to the bootstrapped user identity set. Suppose there are $U$ enrolled users in the set, $u \in \mathcal{U} = \{1, \ldots, U\}$. The query scores of these users are, $\mathcal{Y}_u$ for $u \in \mathcal{U}$. At this point, it is useful to distinguish the match and nonmatch scores; they are denoted as $\mathcal{Y}_u|\omega_k$ for $k \in \{0, 1\}$, corresponding to nonmatch and match comparisons, respectively.

The bootstrapping procedure is shown Algorithm 1. The function "bootstrap" takes a set of identity and returns another set of identities with possible repetitions.

**Algorithm 1** Subset bootstrapping of *cdf*

**INPUT**:

- $\Upsilon$, the range of decision thresholds at a regular interval
- $N$, the number of bootstraps
- $\mathcal{U}$, a set of user identities

**OUTPUT**: $(\mathrm{FMR}_i(\tau), \mathrm{FNMR}_i(\tau))$ for $i = 1, \ldots, N$ bootstraps and $\tau \in \Upsilon$
**for** $i = 1$ **to** $N$ **do**
  $\mathcal{U}' = \mathrm{bootstrap}(\mathcal{U})$
  $\mathrm{FMR}_i(\tau) = 1 - \mathrm{cdf}(\{\mathcal{Y}_u|\omega_0 u \in \mathcal{U}'\})$
  $\mathrm{FNMR}_i(\tau) = \mathrm{cdf}(\{\mathcal{Y}_u|\omega_1 u \in \mathcal{U}'\})$
**end for**

## 2.4 Deriving confidence of a set of Bootstrapped DET Curves

The objective of this section is to characterise the confidence interval of $\hat{\mathrm{FMR}}_{test}(\tau)$ and $\hat{\mathrm{FNMR}}_{test}(\tau)$ respectively. Since the target chart we will visualise is a DET curve, following Martin *et al*'s work we will work on the inverse *cdf* of the Gaussian distribution. If $\Psi(\cdot)$ is the *cdf* of a Gaussian distribution, and $\Psi^{-1}(\cdot)$ its inverse, a DET curve is plotted in the coordinate system of

$$\mathbf{v} \equiv [v_{FMR}, v_{FNMR}] = [\Psi^{-1}(\hat{\mathrm{FMR}}(\tau)), \Psi^{-1}(\hat{\mathrm{FNMR}}(\tau))].$$

There are three ways to define the confidence intervals of a DET curve, as discussed in Poh and Bengio [**?**]. For example, one can fix the FMR and then define the confidence intervals of the corresponding FNMR. This is called vertical averaging. One can also average the FMR and FNMR for a given threshold. This strategy is called threshold averaging. A third method is called "simultaneous joint confidence regions" which does not fix any threshold nor any axes on the DET plan but instead estimates a confidence region based on a set of paired (FMR,FNMR) data points directly. Two variants were reported in [**?**], i.e., fixed-width band [**?**] and working-hotelling band [**?**]. The fixed-width band method, in our context, obtains a confidence region that is defined by two parallel DETs[1] with a fixed width distance such that the original observed DET is fully contained inside the region. The working-hotelling band fits the best regression line in the DET plan. Therefore, it assumes that the class-conditional scores follow a Gaussian distribution.

We follow the third approach as documented in Poh and Bengio [**?**]. We first work in the polar coordinate of $\mathbf{v}$, which can be expressed in $(\theta, r)$ where

$$\theta = \tan^{-1}\left(\frac{v_{FNMR}(\tau) - v_{FNMR}(-\infty)}{v_{FMR}(\tau) - v_{FMR}(-\infty)}\right),$$

and

$$r = \sqrt{(v_{FNMR}(\tau) - v_{FNMR}(-\infty))^2 + (v_{FMR}(\tau) - v_{FMR}(-\infty))^2},$$

for $\theta \in [0, \pi/2]$, $r \in [-\infty, \infty]$ and $(v_{FMR}(-\infty), v_{FNMR}(-\infty))$ is the origin. Since $\Psi^{-1}(-\infty) = -\infty$, in practice, we replace the origin with the point $(\Psi^{-1}(1/N), \Psi^{-1}(1/N))$ where $N$ is the total number of nonmatch (impostor) comparisons rounded to the nearest and the larger power of 10. For example, if the number of impostor attempts is 3,800, then 10,000 can be used.

In order to inverse the process from the polar coordinate to the Cartesian coordinate $\mathbf{v}$, we can apply the following equations,

$$v_{FMR}(\tau) = r \times cos(\theta) + v_{FNMR}(-\infty) \quad (7)$$

$$v_{FNMR}(\tau) = r \times sin(\theta) + v_{FMR}(-\infty) \quad (8)$$

The FMR($\tau$) and FNMR($\tau$) is then obtained by applying the *cdf* of the Gaussian distribution, $\Psi(\cdot)$ , on $\mathbf{v}$, i.e., FMR $= \Psi(v_{FMR})$ and FNMR $= \Psi(v_{FNMR})$, respectively.

To obtain $\alpha \times 100\%$ confidence given the set of bootstrapped DET curves in polar coordinates, we estimate the upper and lower bounds:

$$\frac{1-\alpha}{2} \leq \Psi_\theta(r) \leq \frac{1+\alpha}{2},$$

where $\Psi_\theta(r)$ is the empirical *cdf* of the radius $r$ observed from the $U \times S$ bootstrapped curves for a given $\theta$ since each bootstrapped curve cuts through $\theta$ exactly once. The lower and upper $r$ will be given by $r_{lower} = \Psi_\theta^{-1}(\frac{1-\alpha}{2})$ and $r_{upper} = \Psi_\theta^{-1}(\frac{1+\alpha}{2})$, respectively. Note

[1]The original method applies to the ROC plan.

that the inverse of $\Psi_\theta$, i.e., $\Psi_\theta^{-1}$, requires linear interpolation[2]. The corresponding lower (more optimistic) DET curve is given by $(r_{lower}\cos(\theta), r_{lower}\sin(\theta))$ across all $\theta \in [0, \pi/2]$. The upper (less optimistic) DET curve is defined similarly. By convention, the significance threshold $\alpha$ is set to 0.05 so that one obtains a 95% level of confidence. Note that DET angle was reported in [**?**] to combine several DET curves into a single one. Although DET angle seems to be an uncommon choice, three $\theta$ values are extremely commonly used: $\{0, \frac{\pi}{4}, \frac{\pi}{2}\}$. They correspond respectively to the estimate of confidence interval of FMR at FNMR=0, EER and that of FNMR at FMR=0. Therefore, the procedure described here can be seen as a generalization to this practice.

## 2.5 Overall Algorithm

Having discussed the three core components of the generalised DET prediction algorithm, this section puts the algorithms together more formally. To do so, we will use the following procedures in the form of "procedure : input $\rightarrow$ output":

- `Bootstrap` : $\{y \in \mathcal{Y}, u \in \mathcal{U}\} \rightarrow \{y\}$
  `Bootstrap` takes a set of comparison scores as well as its corresponding template indexes. The number of elements in scores and the tempalte indexes have to the same.

- `Estimate_cdf` : $\{y\} \rightarrow P(y < \tau)$
  `Estimate_cdf` takes a set of scores and produce an empirical estimate of the *cdf.*

- `DET2radius` : FMR, FNMR $\rightarrow (\theta, r)$
  `DET2radius` takes FMR and FNMR as input and produces the corresponding points in polar coordinates.

- `Percentile` : $\{r\} \rightarrow (r_{lower}, r_{median}, r_{upper})$
  `Percentile` takes a set of real-numbered data as input and produces the desired confidence intervals.

- `Radius2DET` : $(\theta, r) \rightarrow v_{FMR}, v_{FNMR}$
  `Radius2DET` convert the $v_{FMR}$,$v_{FNMR}$ from polar coordinate to cartesian coordinate.

All the four functions are used in Algorithm 2. This algorithm takes 7 arguments, namely, the number of bootstraps, $B$; $|Q_1|$ match score sets, their corresponding template identity sets, and their desired prior probabilities on the target operation; as well as $|Q_0|$ nonmatch score sets, their corresponding template identity sets, and their desired prior probabilities on the target operation. The procedure returns the confidence intervals of the predicted DET curve corresponding to the target operation.

[2]In our implementation, we verified that by projecting a DET curve into polar coordinates and then reversing the process, one obtains *exactly* the same DET curve. Therefore, there is no loss of generality by working in polar coordinates as long as the *same* origin (according to footnote 4) is used.

**Algorithm 2** Generalised DET curve with confidence intervals

**INPUT**:

- $\{y|\omega_0, Q_0\}$, $\forall_{Q_0}$ [The nonmatch score sets]
- $\{u|\omega_0, Q_0\}$, $\forall_{Q_0}$ [The set of template indexes of $\{y|\omega_0, Q_0\}$]
- $\{y|\omega_1, Q_1\}$, $\forall_{Q_1}$ [The match score sets]
- $\{u|\omega_1, Q_1\}$, $\forall_{Q_0}$ [The set of template indexes of $\{y|\omega_1, Q_1\}$]
- $\text{Prior}_0$ [The prior probabilities of the nonmatch comparison on the target operation]
- $\text{Prior}_1$ [The prior probabilities of the match comparison on the target operation]
- $B$ [The number of bootstraps]

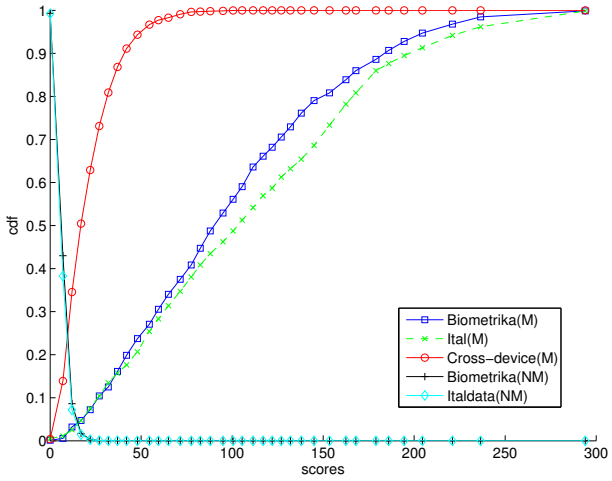
**OUTPUT**: $(r_{lower}, r_{median}, r_{upper})$ [The confidence intervals]
**for** $b = 1$ **to** $B$ **do**
  **for** $k \in \{0, 1\}$ **do**
    **for all** $Q_k$ **do**
      $P_k(:, Q_k) = \text{Estimate\_cdf}(\ \text{Bootstrap}(\{y|\omega_k, Q_k\}, \{u|\omega_k, Q_k\})\ )$
    **end for**
    $P_k^{com} = P_k \cdot \text{Prior}_k$
  **end for**
  $r_b^\theta|\forall_{\theta\in\Theta} = \text{DET2radius}(P_0^{com}, P_1^{com})$
**end for**
$(r_{lower}^\theta, r_{median}^\theta, r_{upper}^\theta) = \text{Percentile}(\{r_b^\theta|b = 1, \ldots, B\}), \forall_{\theta\in\Theta}$
$\text{FNMR}_k(\tau) = (r_k^\theta \times cos(\theta) + v_{FNMR}(-\infty)\ |k \in \{lower, median, upper\}$
$\text{FMR}_k(\tau) = (r_k^\theta \times sin(\theta) + v_{FMR}(-\infty)\ |k \in \{lower, median, upper\}$

## 2.6 Demonstrations

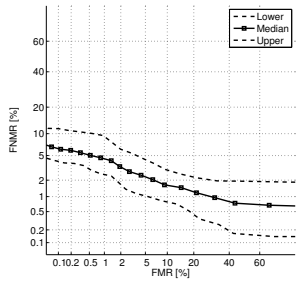
Before presenting the results, let us give a demonstration of the derived DET confidence intervals. We shall present two case studies, namely, assessing biometric performance with two sensors, and asessing biometric performance under varying levels of quality.

In the first case, we will have three sets of match comparison. In this data set, we have two sensors, namely Biometrika and Italdata sensors. We, therefore, have match comparisons due to the comparison of template and query samples of one sensor, another sensor, as well as their cross comparison wherein the template has been captured using one sensor; and the query, with another sensor. As for the nonmatch comparisons, we observed that the *cdf* does not change with the sensor type. For this reason, we have considered only the nonmatch comparison due to each of the two sensors. The five *cdf*s are shown in Figure 3(a).

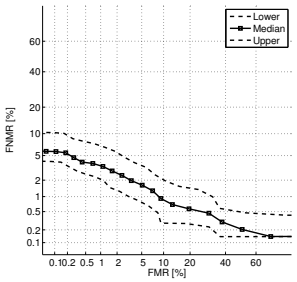
We then identify four scenarios, and in each scenario, we weigh the priors of the five *cdf*s



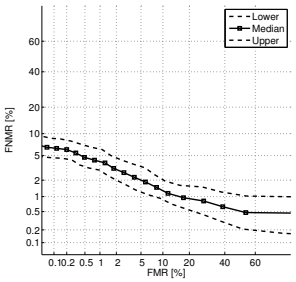

(a) FMRs and FNMRs



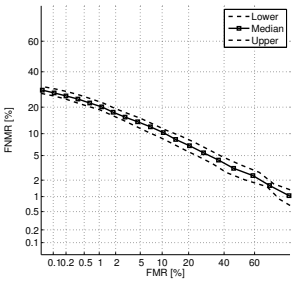

(b) Biometrika only sensor




(c) Italdata only sensor




(d) Both sensors



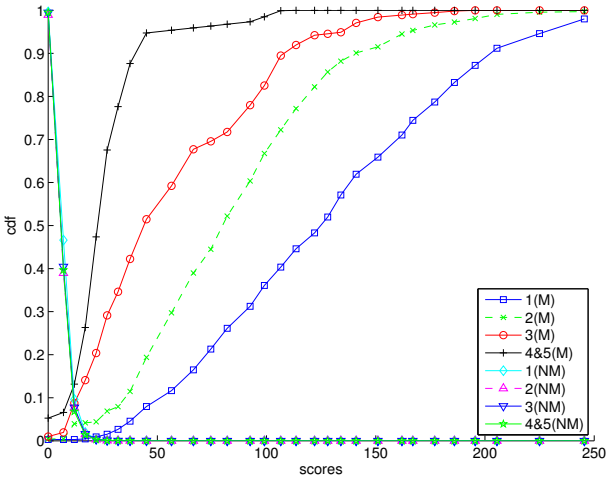

(e) Cross-device comparison

Figure 3: The *cdf* of match and nonmatch comparison for the cross-sensor setting. The DET curves of four use-cases: (b) Biometrika only sensor, (c) Italdata only sensor, (d) both sensors, (e) cross-device comparison.

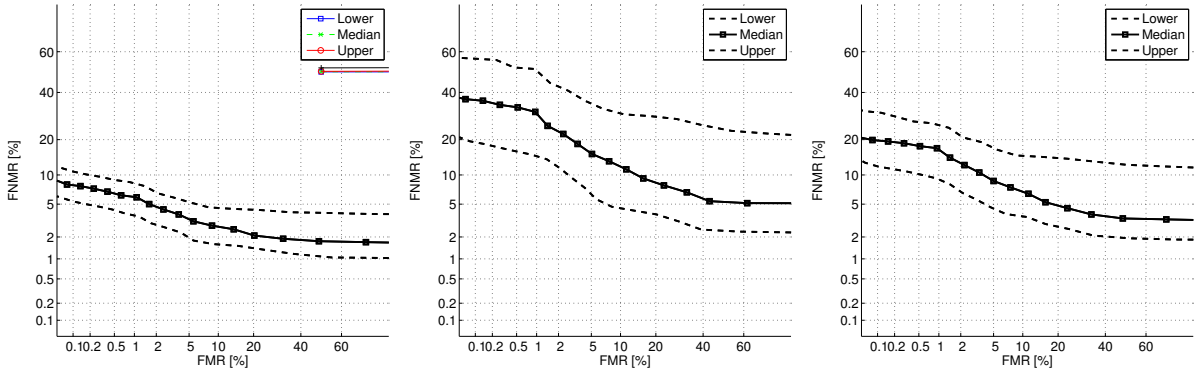
differently. The priors are listed in Table 1(b). The DET curves for these four scenarios are shown in Figures 3(b)–(e), respectively. As can be observed, the last scenario which involves cross-device comparison has the significantly worst performance than the first three scenarios which do not involve two types of single-device comparison.

We repeated the same process of varying quality levels. In this case, we have used the NIST fingerprint matcheer (bozorth3) as well as its fingerprint quality, namely, NFIQ, which gives five levels of quality. Consequently, we divided each of the match and nonmatch comparison scores into four sets, thereby, binning the quality levels 1, 2 and 3 as three sets and the fourth set contains the combined quality levels 4 and 5. This is because the number of samples in this last set is often very small. The *cdf*s of the these eight score sets are shown in Figure 4(a). As can be observed, the *cdf*s of the nonmatch comparisons do not change the quality measures, whereas the match comparisons vary significantly across the quality levels.

We then identify three scenarios with different quality tendency, namely high quality




(a) FMRs and FNMRs




(b) high quality tendency (c) low quality tendency (d) equal-prior quality

Figure 4: The *cdf* of match and nonmatch comparison for the quality-dependent setting. The DET curves of three use-cases: (b) high quality tendency, (c) low quality tendency, and (d) equal-prior quality.

tendency, low quality tendency, and balanced quality. In the first case, quality level one (the besst quality) has a stronger prior whereas the combined quality levels 4 and 5 (the worst quality) have a much smaller prior. In the second case, the priors are switched, causing the low quality samples to dominate. We, therefore, expect the DET curve of the second case to be significantly worse than the first case. In the third case, all quality levels are given the same prior, which is 1/4 in our case, since we have four sets of scores for each of the match and nonmatch comparisons. The results are shown in Figures 4(b)–(d), respectively.

# 3 Experiments

This section addresses the issue of how well one can predict an unseen DET curve using the confidence intervals derived from the proposed procedure. For this purpose, there must be a training and a test set. One requirement is that the enrolled subjects in the two sets must be mutually exclusive. This reflects a real application whereby the population in the design set is usually different from those in the operational setting.

Before we provide any details about the experiments, we shall summarize a number of properties that we know about the confidence intervals of a DET curve. These not only allow us to rule out a number of hypotheses, but also enable us to design experiments testing the unknown.

## 3.1 What we know, and we do not know about DET Confidence Intervals

Poh and Bengio [**?**] designed a number of experiments in order to study the properties of the confidence intervals around a DET curve. Their findings can be summarized as follow:

- Between the sample variability and user-variability, the latter has a larger effect. 1000 match comparison scores constituted by 100 subjects each contributing 10 match scores are more representative than 1000 match comparison scores constituted by 10 subjects each contributing 100 match scores. The DET confidence intervals estimated by the former will be more representative

- The DET confidence intervals estimated from a larger subject population will have smaller confidence intervals. Poh and Bengio showed that by increasing the size of the subject from 10, 20 and 40 to 80, the relative estimate of the corresponding DET confidence intervals are visibly reduced. They also objectively measured the reduction in terms of entropy, demonstrating that with an increasing size of subjects, the entropy decreases, hence, showing a sharper distribution of the DET curve or narrower confidence intervals.

- Using the DET curve of a completely different population of subjects for testing, the confidence intervals of a DET curve has a coverage of between 67% and 83%, depending on the choice of database and biometric systems. Coverage is a measure of proportion of the predicted DET curve that falls in the visible region of the confidence interval intervals.

In their conclusion section, the authors highlighted the challenge of generalizing a DET curve from one context to another. Poh and Bengio's two-step bootstrap procedure cannot generalize the DET curve as soon as the context of application changes because the method has no means of detecting factors or covariates that can influence the performance of a biometric system.

The above conclusion directly points to the need of evaluating, how well one can predict the unseen DET curve, when the factors are explicitly identified. In light of this, it is imperative to evaluate the coverage of the proposed DET confidence intervals under different factors. For this purpose, we will set up 7 use-case scenarios from five data sets.

## 3.2 Database, Protocols, Evaluation Criterion

We have chosen to use the LivDet 2011 data set [**?**] because this database has a number of unique features. First, the fingerprint images have been acquired using two different sensors. This allows us to perform cross-sensor comparison wherein the system is required to compare two samples obtained using two different sensors. The database also contains spoof fingerprint impressions made by five different fabrication materials. This allows us to evaluate the performance predicted using both zero-effort and nonzero-effort attacks at different proportions. Third, the database contains live fingerprint of different quality, thus, enabling us to study the effect of fingerprint quality on the system performance under zero-effort as well as nonzero effort attack. Finally, we can also evaluate the quality of prediction of liveness detection algorithms under the spoof attacks carried out by different fabrication materials.

The LivDET 2011 database contains 8000 samples. The most important key statistics relevant for our experiments are:

- 144 unique fingers containing both live and spoof samples
- 128 unique fingers containing only live samples
- 4000 fingerprints acquired using the Biometrika sensor, and another 4000 acquired using the Italdata sensor.
- 800 fake fingerprint samples for each of the five fabrication materials

## 3.3 Experimental Protocol

In order to estimate the quality of prediction, we kept the 144 fingers which have both live and spoof samples as enrolment identities (constituting the gallery set). In this way, we will have both live and spoof comparison scores which can be further divided into match and nonmatch comparisons. Let $u$ denote an identity drawn from the set of identities, $\mathcal{U}$. We divide $\mathcal{U}$ into two smaller sets $\mathcal{U}_{train}$ and $\mathcal{U}_{test}$ of equal size but containing identities that are mutually exclusive. Let $\mathcal{Y}_{train}$ denote the comparison scores generated from $\mathcal{U}_{train}$; and similarly, $\mathcal{Y}_{test}$ denotes the comparison scores generated from $\mathcal{U}_{test}$.

The data division procedure is repeated 100 times in order to obtain 100 pairs of $\mathcal{U}_{train}$ and $\mathcal{U}_{test}$; and their corresponding comparison scores, $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$. From the training score set, $\mathcal{Y}_{train}$, we will obtain the confidence intervals, the accuracy of which is then assessed using the DET curve derived from $\mathcal{Y}_{test}$.

Table 1: The data set and scenarios used to benchmark the quality of prediction in terms of coverage.

(a) Five score data sets used to measure coverage

| **Scenario** | | **Class** | **Description** |
|---|---|---|---|
| Cross-sensor | $\omega_1$ | Genuine comparison with Biometrika sensor, Genuine comparison with Italdata sensor, Cross-sensor genuine comparison | Cross-sensor comparison involving two sensors |
| | $\omega_0$ | Nonmatch zero-effort comparison involving either of the two sensors (Biometrika and Italdata) | |
| Varying quality | $\omega_1$ | Genuine comparison conditional on NFIQ quality levels 1-3 and combined levels of 4 and 5 | Match and nonmatch comparisons conditional upon different sample quality levels |
| | $\omega_0$ | Nonmatch zero-effort attack comparison conditional on NFIQ quality levels 1-3 and combined levels of 4 and 5 | |

(b) 7 scenarios for assessing coverage

| **Scenario** | **Scenario** | **Prior ratio,** $P(Q_1\|\omega_1)$ | **Prior ratio,** $P(Q_0\|\omega_0)$ |
|---|---|---|---|
| Cross-sensor | Biometrika only sensor | [1 0 0] | [1 0] |
| | Italdata only sensor | [0 1 0] | [0 1] |
| | Both sensor | [1 1 0] | [1 1] |
| | Cross sensor operation | [1 1 1] | [1 1 1] |
| Varying quality | High quality tendency | [8 4 2 1] | [8 4 2 1] |
| | Low quality tendency | [1 2 4 8] | [1 2 4 8] |
| | Equal prior quality | [1 1 1 1] | [1 1 1 1] |

## 3.4 4 Data sets and 7 Use-case scenarios

In order to test the generalization ability of the DET confidence intervals, we have prepared five data sets in order to examine 7 use-cases of predicting DET confidence intervals. The four data sets are described below; and are summarised in Table 1.

1. **Cross-sensor matching**: Insofar as the data set that permits us, we can consider the cross matching due to two sensors. This gives us three possibilities, namely, matching involving the same sensor, such as Biometrika-vs-Biometrika (Biometrika template vs Biometrika query), Italdata-vs-Italdata, and Biometrika-vs-Italdata. We shall use a *symmetric comparison score* such as Biometrika-vs-Italdata and Italdata-vs-Biometrika give exactly the same result. Therefore, the use-case scenarios considered are:

   (a) Biometrika only comparisons,

   (b) Italdata only comparisons,

   (c) Both sensors – comparisons involving both sensors excluding cross-sensor comparisons.

   (d) cross-sensor operation – all possible comparisons including the cross-sensor ones.

   Refer to Table 1s (a) and (b) for the prior probabilities considered involving the above four scenarios.

2. **Comparisons under varying levels of quality**: In the case of discrete quality measures, it is possible to separate the biometric samples into various categories. The NFIQ that is applied to fingerprint enables us to divide the fingerprint samples into five levels of quality. However, the proportion of levels 4 and 5 are so small that we have combined both of them. This gives us the following four categories: 1, 2, 3, and 4&5. The match and nonmatch score sets are thus divided into these four subsets. We then considered three scenarios among the many possibilities:

   (a) High quality tendency: This is the default case where in the majority of the samples are of high quality (level 1). The next level of quality (level 2) has half the samples, and so on. This gives us the ratio of $P_k$ of $[8, 4, 2, 1]$ for both match and nonmatch comparisons.

   (b) Low quality tendency: By the same token of argument, we also consider a realistic worst case scenario of $[1, 2, 4, 8]$.

   (c) Balanced: Finally, we also considered a balanced scenario of $[1, 1, 1, 1]$.

## 3.5 Evaluation Criteria

### 3.5.1 Coverage

Coverage is defined by the proportion of the DET curve contained within the confidence intervals of the DET curve derived from the training set. Let $r_{test}(\theta)$ be the DET curve of the test set represented in the polar coordinate $(r, \theta)$, and $r^{upper}_{train}(\theta)$ and $r^{lower}_{train}(\theta)$ be the upper and lower confidence bound. Coverage is defined as the proportion of the test DET curve that falls inside the *computable angles* of the confidence bound, $\Theta \equiv \{\theta | r^{upper}_{train}(\theta) < \infty \wedge -\infty < r^{lower}_{train}(\theta), \theta \in [0^o, 90^o]\}$:

$$\text{coverage} = \frac{|\{\theta | r^{lower}_{train}(\theta) < r_{test}(\theta) < r^{upper}_{train}(\theta), \forall_{\theta \in \Theta}\}|}{|\{\theta | \forall_{\theta \in \Theta}\}|}$$

Figure 5 illustrates an example of a test DET curve where a portion of the curve falls inside the DET confidence intervals and another portion falls outside the intervals. The portion that falls inside the intervals is the coverage. Although the computation of coverage is carried out in the polar coordinates, we have back-projected them to the (FMR,FNMR) coordinates for visualisation here. Although there is a one-to-one correspondence between the two representations, the polar coordinates can sometimes capture points on the

### 3.5.2 Test of difference in distribution between the training and test scores

A key assumption that the proposed algorithm makes is that the factor-specific *cdf* of the training and the test sets are the same, that is, $P_{train}(y < \tau | \omega_k, Q)$ and $P_{test}(y < \tau | \omega_k, Q)$ comparable. Equivalently, we want to test the two samples that have been used to derive the above probabilities, namely, $\{y | Q_k, \text{training}\}$ and $\{y | Q_k, \text{test}\}$, are the same. There are a number of tests that can be used, such as measuring relative entropy, test of means

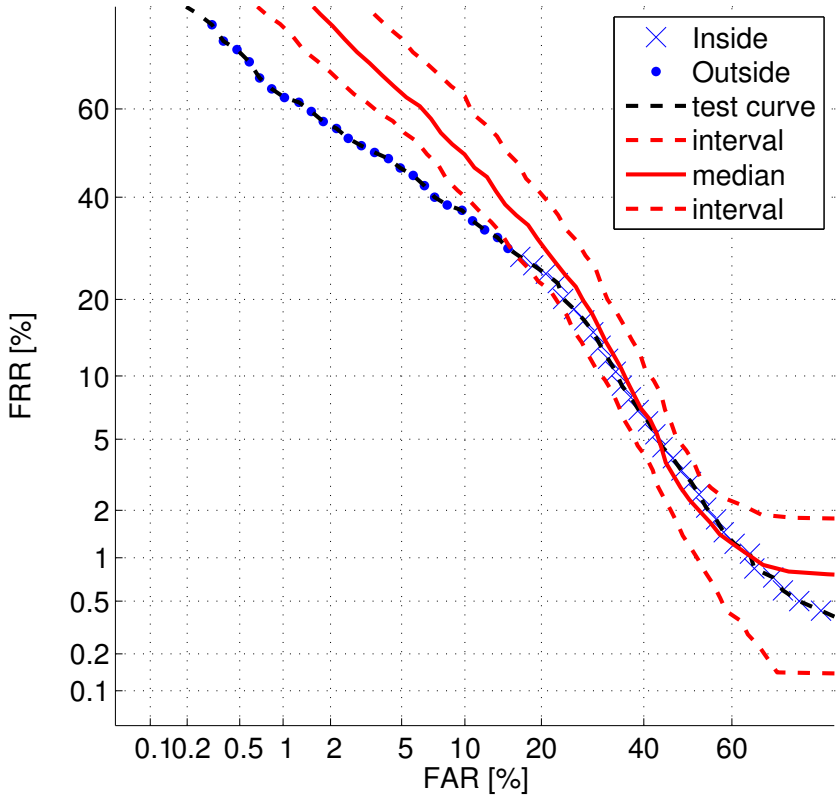


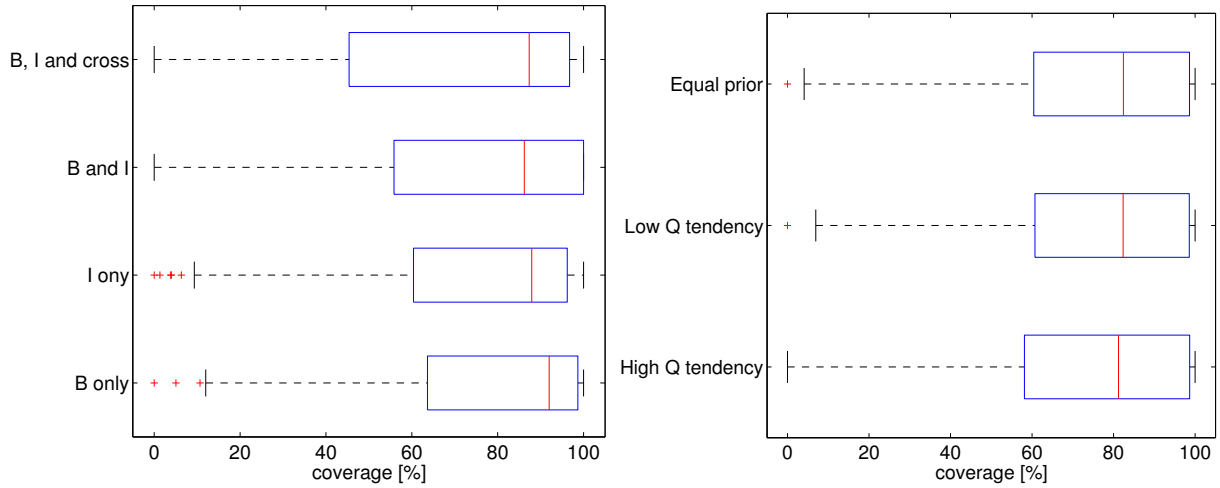

Figure 5: An example demonstrating how coverage is calculated. The portion of DET curve falling inside the DET 95%confidence intervals is used to calculcate the coverage of the test curve.

difference (student's t-test), or two-sample Komolgorove-Smirnov test (KS-test). While the first method requires an estimate of density, the second method is useful only for data where their means are meaningful, hence, implicitly assuming a single-mode distribution. The third method is non-parametric and does not require any estimate of the density. KS-test simply takes two *cdf*s and find the maximum vertical distance between them, that is,

$$\text{KS-stat} = \arg\max_x |F_1(x) - F_2(x)|$$

In our context, we can directly use $P(y < \tau | Q_k, set\}$ where *set* can either be the training or the test set, each replacing $F_1$ and $F_2$ above.

The null hypothesis of the KS-test is that the two score sets being compared are drawn from the same distribution whereas the alternative hypothesis is that they are drawn from two different distributions. The KS-stat is compared to a critical value that is dependent on the number of samples, as well as the significant level, $\alpha$, which is set to 0.05. Alternatively, one can calculate the P-value of the KS-stat and check if this value is exceeds the $\alpha$ level or not. A P-value that is smaller than the significant $\alpha$ level suggests that one can reject the null hypothesis. In summary, a high KS-stat is likely to lead to small P-value, which




(a) Cross sensor (b) Quality

Figure 6: Boxplots of coverage across two tasks and 7 use-case scenarios. Each boxplot contains 100 bootstrap of experiments for one of the 7 use-case scenarios. Coverage is measured on the unseen test curve with a different population of subjects than the training set from which the confidence intervals have been derived.

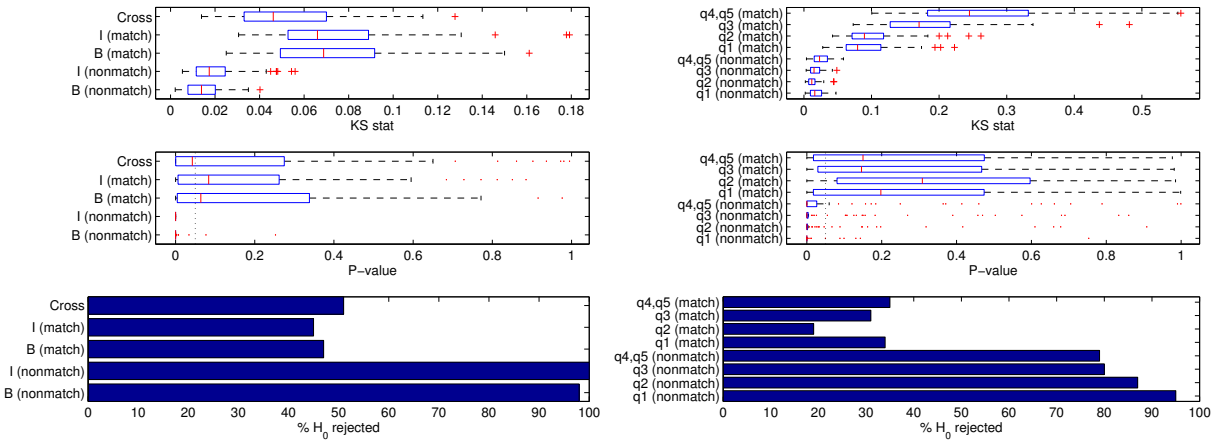
in turns leading to the rejection of the null hypothesis, suggesting that the sets of data are different.

## 3.6 Results

We shall present three sets of experiments. The first one aims to study the generalisation ability of the DET confidence intervals under population mismatched. The second set of experiments assesses the stability of factor-specific score distribution across two different population of subjects. This represents a secondary but important analysis because it provides an explanation to why perfect predictability cannot be achieved by explicitly measuring the discrepancy between training and test factor-specific score distributions. The third set aims to study if there is a particular DET angle that is harder to predict than others, that is, if FNMR is harder to predict than FMR, or vice-versa.

### 3.6.1 Coverage

The coverage for all the 7 use-case scenarios are shown in Figure 6. Each boxplot contains 100 bootstrapped samples. The expected coverage is more than 75% for the first four score data sets and more than 60% for the liveness detection task. The range of coverage values obtained here is consistent with Poh and Bengio's study [**?**]. If we were to measure the coverage using the same training set, we would have obtained 100%. The discrepancy between the training and the test sets is possibly due to the mismatch in subjects between the training and the test sets. This is further verified in the next set of experiments.




(a) Cross sensor comparison

(b) Comparison involving varying quality

Figure 7: Two-sample Komolgorov-Smirnov test of whether or not the training and test score distributions are the same.

The following observations can be made:

- The DET curves with cross-device comparison are somewhat harder to predict than those with a single-device only curves.
- The DET curves with high or low quality tendency curves can be predicted at the same or comparable level of accuracy

### 3.6.2 Secondary analysis

The KS-stat and P-value for the 7 use-case scenarios are shown in Figures 7(a) and (b). In (c), the proportion of the null hypothesis being rejected, out of the 100 bootstrapped samples, is also plotted. As it turns out, most of the tests rejected the null hypothesis.

However, what is not expected is that the null hypotheses for the nonmatch comparisons are also rejected. This is because the KS-stat for the nonmatch comparisons is very small, indicating that the pair of data sets should come from the same distribution. Consequently, we would have expected that its P-value to be relative large. See Figures 7(a) for Biometrika and Italdata nonmatch comparisons; and (b) for quality-dependent q1–q5 nonmatch comparisons. However, in each case, their P-value turns out to distribute around zero. After a careful investigation, we found that this is because there are a lot more nonmatch samples, in the order of 400 thousand samples. As a result, the large KS-stat is offset by the large samples, leading to very small P-values for this class. This phenomenon is further discussed in the appendix (Section A).

### 3.6.3 Angle-dependent Coverage

As a final analysis, we look at the DET angle-dependent coverage. This looks at the 100 bootstraps of 7 use-case scenarios. This enables to study the behaviour of 700 observations

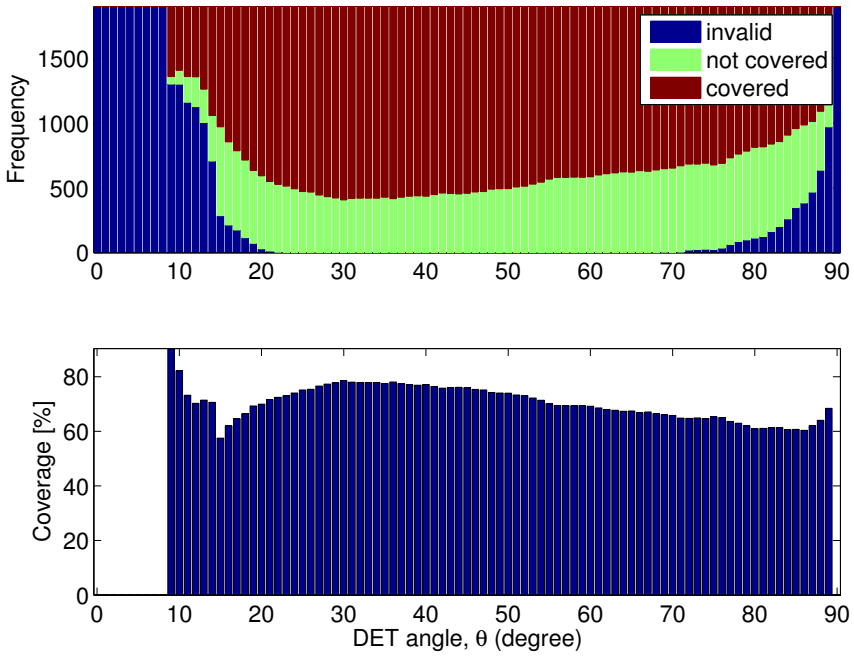



Figure 8: (a) The number of invalid or unobserved values, uncovered, and covered cases out of the 7 use-case scenarios each having 100 bootstrap experiments. (b)The coverage of 700 use-case experiments for each DET angle. The minimum angle-dependent coverage is 57.5% and the maximum is 90.3%.

of DET coverage for each of the angles, $\theta \in \{0^0, \ldots, 90^0\}$. Each DET angle is divided into three parts: invalid or unobserved angle, covered, and uncovered angle. The invalid angles are those for which no value can be calculated simply because no data is available. The covered angles are those whose test DET curve is covered by the confidence intervals derived from the training data; whereas the uncovered angles are those falling outside the DET confidence intervals. Figure 8(a) shows the frequency of the three types of DET angles derived from the 700 bootstrap experiments.

We note that the low DET angles which correspond to the low FNMR has no value. This is because the precision FNMR, is often significantly lower than that of the FMR. The low precision is due to the disproportionately smaller number match samples than the nonmatch samples.

All the valid DET angles have a coverage between 57.5% and the maximum is 90.3%, as shown in Figure 8. By using only the DET-angles that are valid, the 2.5-th, 50-th and the 97.5-th percentiles of the coverage across the valid DET-angle dependent curves are 60.5%, 70.1%, and 80.3%, respectively.

Our result here shows that there is little bias as to whether or not a particular angle of a DET curve is harder to predict. Put differently, every DET angle has equal chance of being correctly predicted, and the probability of this is around 60% to 80% under mismatched population; and they are statistically significantly better than random, which is 50%.

# 4 Conclusions

In this deliverable, we proposed a novel metric for biometric performance evaluation that complements those already presented in D3.3 and implemented in D3.4. This new tool permits the computation of a DET curve with confidence intervals, aiming to predict the most likely performance of a biometric system on an unseen target population of subjects. This is achieved by explicitly modelling the *cdf*s of the identified factors. We have demonstrated the feasibility of this approach on 4 different tasks across 7 use-case scenarios. These tasks assess the following scenarios: the impact of the performance of the system under multiple sensors and the impact of quality on the system performance. For each use-case scenario of a given task, we sampled two mutually exclusive sets of subjects, simulating a design and a target test environment with different subjects. The design data set is used to derive a DET confidence intervals whereas the test data set is used to assess to what extent the DET confidence intervals can predict the unseen DET curve. The prediction quality is measured in terms of coverage. Across 700 experiments, we found that the coverage at the curve level, as well as at the DET angle level is between 60 and 90%.

# A Supplements

In the two-sample KS test as implemented in Matlab, it takes two samples of data. Let $n_1$ and $n_2$ be the number of samples; and their KS-statistics to be represented by KS-stat. We shall first calculate $n$ and $\lambda$ from which the P-value can be calculated:

$$n = \frac{n_1 \times n_2}{n_1 + n_2}$$

$$\lambda = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}\ \text{KS-stat}}$$

It follows that the P-value is calculated by:

$$\text{P-Value} = 2\sum_{j=1}^{101}((-1)^{j-1}\exp(-2\lambda^2 j^2)$$

Figure 9 shows how a function of P-value when $n$ is allowed to vary from 50, to 200, and 1000. As can be observed, under very large samples, any large KS-statistsic will tend to have very small P-value.

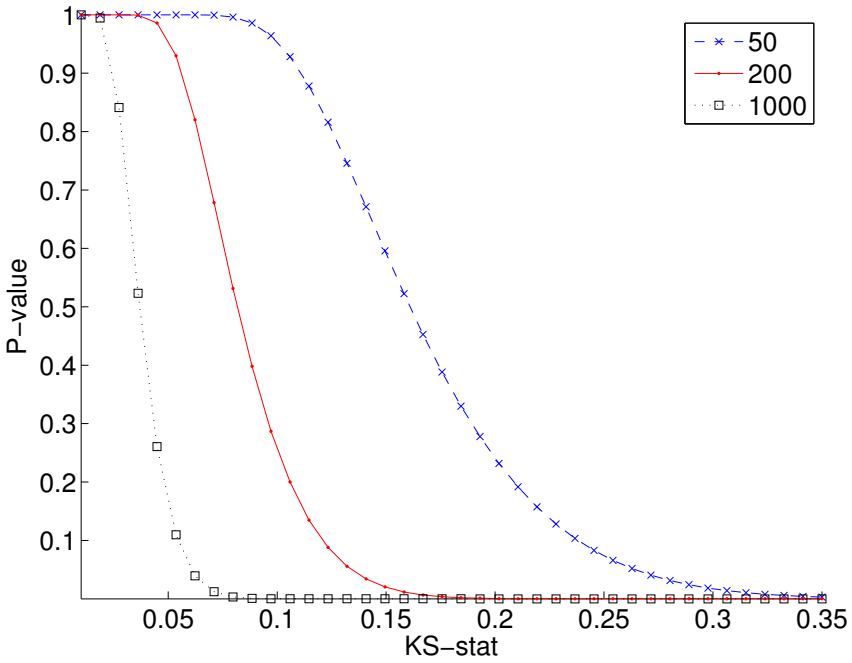



Figure 9: The relationship between P-value and the KS-statistic for various sample size