LLNL-TR-642693

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance

J. V. Candy, E. F. Breitfeller

August 22, 2013

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance Analysis

James V. Candy and Eric F. Breitfeller
Lawrence Livermore National Laboratory
Livermore, CA 94551

1 INTRODUCTION

Detection theory is a fundamental tool in decision analysis [1], [2]. However, many decision functions both formal (e.g. likelihood ratio) and informal (e.g. maximum of function) evolve in a wide variety of applications. The fundamental detection problem of interest is shown in Fig. 1. That is, given a set of *data* (e.g. noisy measurements) obtained from an instrument and a *decision function*, “decide” whether or not the signal is present or not.

Detection (binary), same fiber or not, and classification methods form the basis of detection processors including modern machine learning algorithms. The major question that arises when investigating these methods [3]-[7] is their underlying performance on problems of interest. There exists a variety of metrics that can be applied to evaluate algorithm performance ranging from confusion matrices to sophisticated statistical hypothesis tests [7], but perhaps the most basic and most robust method is the calculation of the receiver operating characteristic (*ROC*) curve. The *ROC* curve is simply a graph of detection (P_{DET}) versus false alarm (P_{FA}) probabilities parameterized by threshold, τ. This particular metric has evolved from the analysis of radar systems during World War II [6], as a critical tool for diagnostic testing in medicine, to pattern recognition in forensics and a wide variety of other applications [8]. There are many individual metrics that can be extracted from a *ROC* curve including sensitivity, specificity, cost/benefit analysis along with a set of specific features like area-under-the-curve (*AUC*) and minimum probability of error (*MinE*) [1]. All of these problems have one requirement in common---they must be analyzed in some uniform manner so that their detection performance can be evaluated. This requirement leads directly to the *ROC* curve, since it provides all of the fundamental information from which most other metrics are derived (e.g. area-under-ROC curve, AUC).

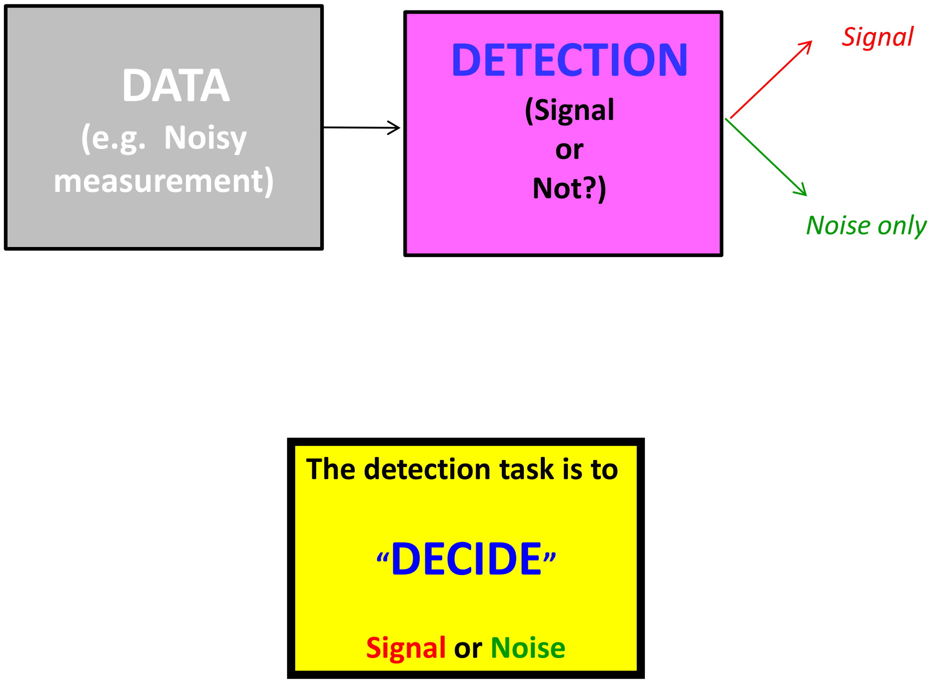


Figure 1: Basic object detection problem: Data, detection with thresholding and final decision.

A typical *ROC* curve is shown in Fig. 2 where we observe that the "*ROC* space" is defined by the 1×1 square region in the (P_{FA}, P_{DET})-plane. The graph is monotonically increasing from ($0,0$)-to-($1,1$). *Detection performance* can range from complete alternative (negative) hypothesis at ($0,0$) to complete hypothesis (positive) detection at ($1,1$) with "perfect detection" at (P_{FA}, P_{DET})=($0,1$) and "random detection" (e.g. coin toss) along the cross-diagonal or 45^0-line from ($0,0$)-to-($1,1$), that is, $P_{DET} = P_{FA}$. Curves lying above this line are considered a representation of "good" detection performance (i.e., better than a coin toss), while those lying below the line are considered "bad" detection performance (i.e., worse than a coin toss). Each (operating) point along the *ROC* curve is parameterized by a threshold value, $\tau_n, n = 1, \cdots, N$ defining a particular expected *operating point* of the *detector* under analysis, that is, ($P_{FA}(n), P_{DET}(n)$) at τ_n. *ROC* curves are a function of detection and false alarm probability density (mass) functions determined by sweeping the threshold through the decision *PDF*s and calculating the underlying area overlaps [3]. We shall discuss this in more detail in Sec. 2. *ROC* curves are generated from “known” explicit decision functions based on the particular problem (under investigation) statistics. For example, Gaussian decision functions lead to explicit error function calculations that can be calculated analytically [10]. Unfortunately most decision function probability distributions (*not* necessarily data distributions) are unknown or too complex to evaluate directly; therefore, we must resort to numerical techniques (e.g. series approximations) to calculate the *ROCs* or resort to “brute” force techniques, if possible, by generating large ensemble realizations and applying counting methods [5]. Even though it appears to be a straightforward calculation, *ROC* curves still possess a bit of “mysticism” because we rarely find simple interrelations between P_{DET}, P_{FA}, and τ_n such as a random detector, $P_{DET}(\tau_n) = P_{FA}(\tau_n)\ \forall\ n = 1, \cdots, N$. For instance, the well-known Neyman-Pearson detection [10] algorithm fixes P_{FA} and then maximizes P_{DET} for a specified threshold (τ). Typically, we know the decision function employed can perform the required integration (analytically or numerically) to determine the detection and false alarm probabilities enabling us to generate various points along the curve at each threshold, tracing out the curve. So we see that the *ROC* curve, though simple in concept, can present a challenge to calculate depending on the underlying problem scenario and availability of good measurements and known or unknown decision distributions.

In this report, we discuss the basic (well-known) theory required to comprehend (intuitively and mathematically) the receiver operating characteristic curve and its inherent features that enable us to ascertain the performance of *detection* algorithms. In Sec. 2, we develop the theory, while in Sec. 3, we develop a variety of metrics that can be extracted from the *ROC* to increase our understanding of the embedded information and help present a picture. We summarize the results in Section 4.

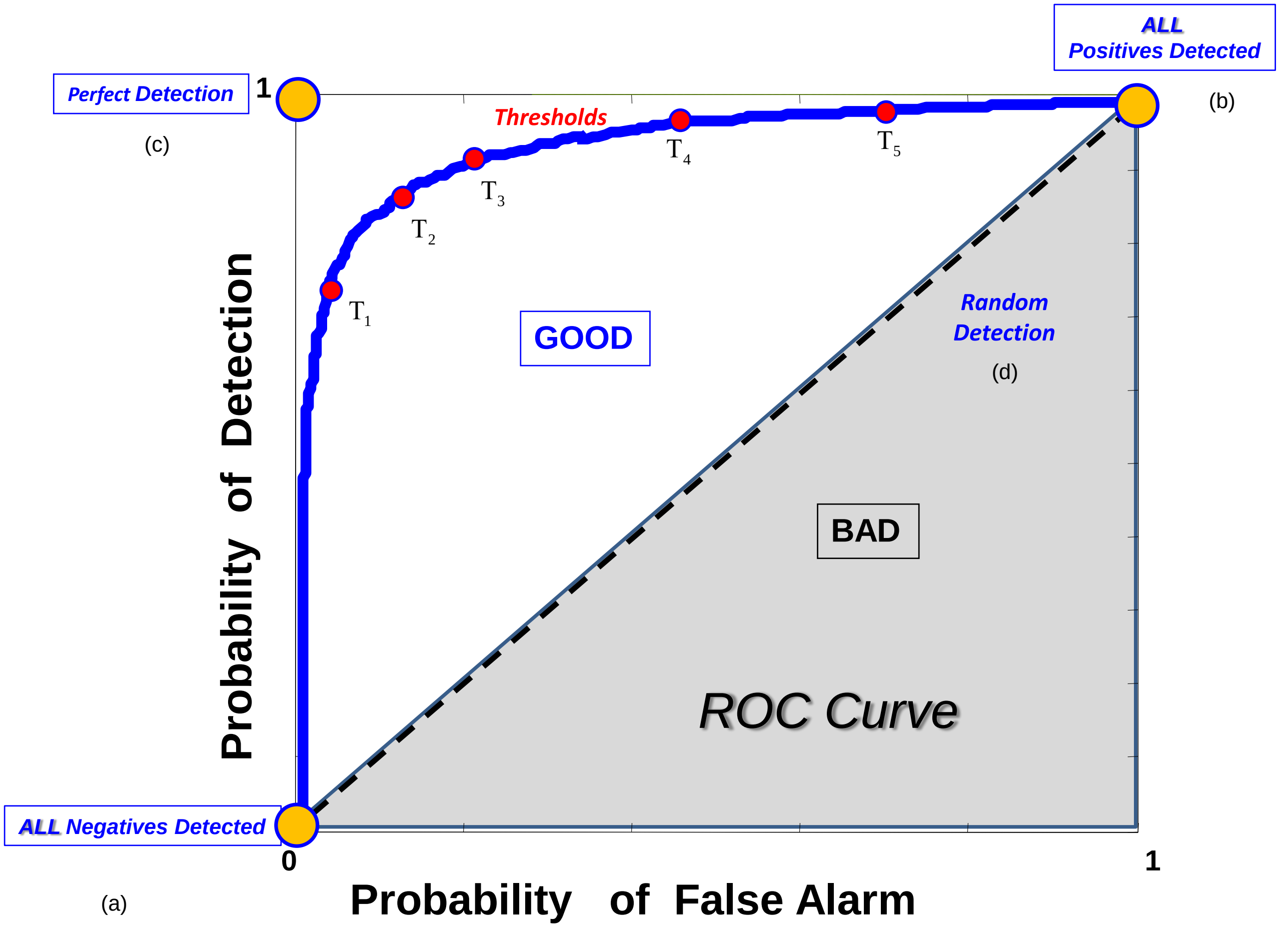


Figure 2: Receiver operating characteristic (*ROC*) curve: Detection (P_{DET}) versus false alarm probability (P_{FA}) for selected thresholds (τ) indicating various performance metrics including: (a) All negative (P_{FA}, P_{DET})=$(0,0)$ detection. (b) All positive (P_{FA}, P_{DET})=$(1,1)$ detection. (c) Perfect (P_{FA}, P_{DET})=$(0,1)$ detection. (d) Random (coin toss) detection.

2 THEORY: DECISION FUNCTIONS AND *ROC* CURVES

2.1 Mathematical Theory

Underlying the mathematical description of the *ROC* curve is the basic detection (decision) problem (see Fig. 1). Here we are given a number of choices to make our decision. These choices are in the form of hypotheses for our detection problem that are used to enable us to decide whether the measurement evolved from a signal or disturbance and noise. If our measurements, $\boldsymbol{\mu}(k)$, came from a signal ($\boldsymbol{\mu}_1$) or a non-signal or disturbance ($\boldsymbol{\mu}_0$), then the hypothesis test can be defined by

$$\mathcal{H}_0 : \boldsymbol{\mu}(k) = \boldsymbol{\mu}_0(k) + \mathbf{v}(k) \qquad [NON - SIGNAL/\text{DISTURBANCE}]$$

$$\mathcal{H}_1 : \boldsymbol{\mu}(k) = \boldsymbol{\mu}_1(k) + \mathbf{v}(k) \qquad [\text{SIGNAL}]$$

where the subscript notation refers to the non-signal/disturbance (“0”) or alternate signal hypothesis (“1”) of the test corresponding to the two respective hypotheses ($\mathcal{H}_0, \mathcal{H}_1$)and corresponding decision regions, $\mathbf{D}_0$, $\mathbf{D}_1$ that partition the observation or measurement space [1], [8], [9].

The typical approach to solving this decision problem is to define a decision function, $\mathcal{D}(\boldsymbol{\mu}(k))$ and its underlying distributions (where possible) under each hypothesis, that is, the conditional probability (decision function given the known hypothesis), is specified by

$$Pr[\mathcal{D}(\boldsymbol{\mu}(k)) \,|\, \mathcal{H}_\ell];\, \ell = 0,1 \qquad (1)$$

Based on this information, we are able to calculate the corresponding *probability of detection* as

$$P_{DET}(\tau) := \int_{\tau:\mathbf{D}_1}^{\infty} Pr[\mathcal{D}(\boldsymbol{\mu}(k)) \,|\, \mathcal{H}_1]\, d\mathcal{D} \qquad (2)$$

with $\mathbf{D}_1$ a partition (signals) of the measurement (decision) space. The *probability of false alarm* is defined by

$$P_{FA}(\tau) := \int_{\tau:\mathbf{D}_1}^{\infty} Pr[\mathcal{D}(\boldsymbol{\mu}(k)) \mid \mathcal{H}_0]\, d\mathcal{D} \qquad (3)$$

Note that this integral is performed over the region $\mathbf{D}_1$ leading to false alarms [8] as shown in Fig. 3.

Also related to these two probabilities is the *probability of a miss* given by

$$P_{MISS} := \int_{-\infty}^{\tau:\mathbf{D}_0} Pr[\mathcal{D}(\boldsymbol{\mu}(k)) \mid \mathcal{H}_1]\, d\mathcal{D} = 1 - P_{DET} \qquad (4)$$

Summarizing there are actually four probabilities associated with our decision problem:

- P_{DET} ------a detection declared when item is actually a signal (right)
- P_{FA} --------a false alarm when detection is declared and item is actually a non-signal (wrong)
- P_{MISS} ------a non-detection declared when item is actually a signal (miss)
- P_{REJECT} ----a non-detection declared when item is actually a non-signal (right)

There are errors associated with each "wrong" decision. Suppose the decision space is partitioned into two regions (binary: signal/non-signal problem) with region $\mathbf{D}_0$ corresponding to the non-signals ($\mathcal{H}_0$) and region $\mathbf{D}_1$ corresponding to the signals ($\mathcal{H}_1$). There are two possible mechanisms in which an error can occur: (1) either a decision, $\mathcal{D}(\mu(k))$ falls into region $\mathbf{D}_1$ and the item is a non-signal or (2) the item is a signal and falls into $\mathbf{D}_0$. Since these events are mutually exclusive, then we can define the *total probability of error* as

$$P_{\varepsilon} := Pr[\mathcal{D} \in \mathbf{D}_1, \mathbf{D}_0] + Pr[\mathcal{D} \in \mathbf{D}_0, \mathbf{D}_1] \qquad (5)$$

applying Bayes' rule ($Pr[A,B] = Pr[A \mid B] \times Pr[B]$) to the joint distribution gives

$$P_{\varepsilon} = Pr[\mathcal{D} \mid \mathcal{H}_1] \times Pr[\mathbf{D}_0] + Pr[\mathcal{D} \mid \mathcal{H}_0] \times Pr[\mathbf{D}_1] \qquad (6)$$

where $Pr[\mathbf{D}_0]$, $Pr[\mathbf{D}_1]$ are the prior probabilities (prevalence, [17]) associated with each hypothesis. The total error is sometimes called the *Bayes' error*, since the Bayes' detector is developed by minimizing this error to obtain the appropriate decision function.

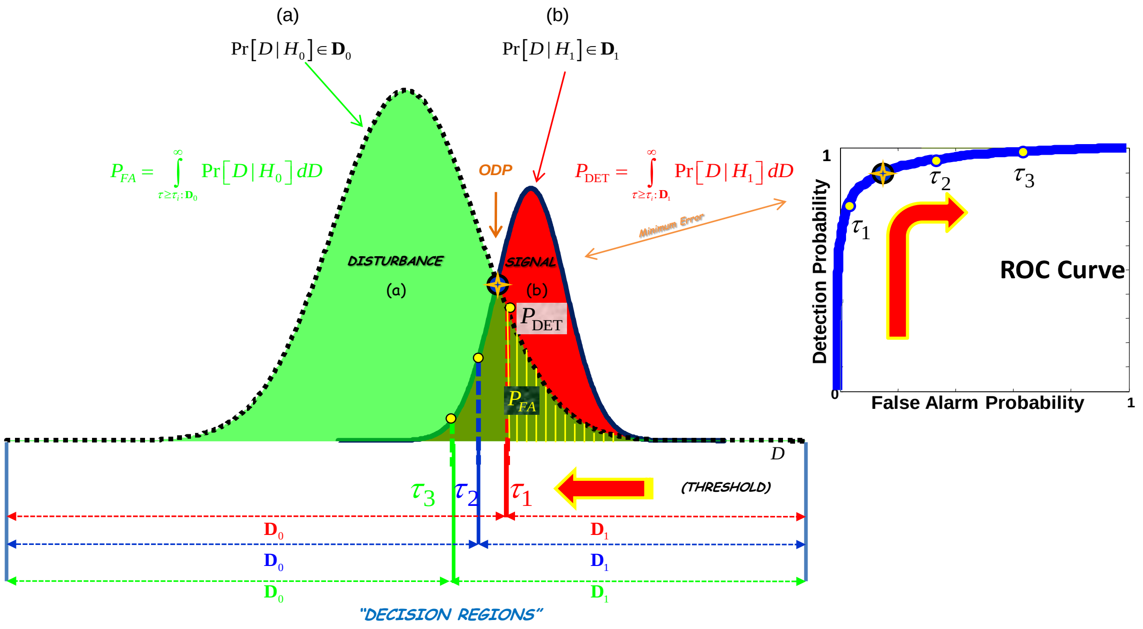


Figure 3: Decision function probability distributions for selected thresholds (τ_n) and defined decision regions: (a) Disturbance probability distribution: $Pr[\mathcal{D}(\boldsymbol{\mu}(k)) | \mathcal{H}_0]$. (b) Signal probability distribution: $Pr[\mathcal{D}(\boldsymbol{\mu}(k)) | \mathcal{H}_1]$. Note that the star is location of minimum error probability (maximum detection/minimum false-alarm and optimal decision point (*ODP*)).

2.2 *ROC* Curve Generation

Next let us investigate just how a *ROC* curve is generated, first, intuitively (pictorially) and then numerically. Consider the two (binary problem) decision distribution functions shown in Fig. 3----the disturbance (green) and the signal (red). From the figure we observe that their means (distribution peaks) differ providing an offset and they overlap one another. It is this overlapping that dictates the shape of the *ROC* and the performance of the detector. If we allow the decision variable (threshold) to assume various values and sweep it from $+\infty$ to $-\infty$, a *ROC* curve is traced out. That is, for each threshold value, τ_n, a vertical line drawn from $\mathcal{D}=\tau_n$ intersects either curve associated with the underlying hypothesis, $Pr[\mathcal{D}\,|\,\mathcal{H}_\ell];\ \ell=0,1$ defining the lower integration limit for both P_{DET} and P_{FA} (see Eqs. 2, 3). Performing the integration at τ_n calculates the areas under both decision distributions (see Fig. 3) providing a corresponding operating point $(P_{FA}(\tau_n), P_{DET}(\tau_n))$ on the *ROC* curve for the selected threshold. Sweeping $\tau_n;\ n=1,2,\cdots$ the lower integration limits (thresholds) are varied and the corresponding areas defined by these boundary limits trace out the entire *ROC* curve. From Fig. 3 we also see that selecting the thresholds defines the new boundaries for the decision regions $\mathbf{D}_0$ and $\mathbf{D}_1$ as well as the corresponding areas of integration for all of the operating points. There exists a particular point of *minimum error* corresponding to the intersection of the conditional distributions (orange star in Fig. 3) providing the maximum detection and minimum false-alarm probabilities for an optimum (Bayes) detector design [3]. Note also that for a given threshold both P_{MISS} and P_{REJECT} can be calculated by changing the integration limits to $-\infty$-to-τ, if desired. So we see that knowledge of the decision function *PDF*s can provide us with some insight about functions generating the *ROC* curve and its construction.

Numerical calculation of the *ROC* curve can evolve from a variety of methods. Perhaps the simplest is through integration, either analytically (where possible) or numerically, with the particular availability of the detection and false-alarm probabilities and their known forms (analytic function or numerical values). Recall that it is necessary for both probabilities to evolve from “known” functions or data.

One particular technique that can be applied to calculate the *ROC* directly from known data sequences generated either through simulations or controlled experiments, is the so-called *brute force method*. Once the *decision sequences* are calculated, then this method can be applied using simple counting methods. For this method, known data are generated and input to the decision function which is varied according to selected threshold values similar to sweeping the thresholds illustrated in Fig. 3, that is, for the decision problem, we must *decide*

$$\mathcal{D}(\boldsymbol{\mu}(k)) \underset{NON-SIGNAL\,/\,DISTURBANCE}{\overset{SIGNAL}{\begin{matrix} > \\ < \end{matrix}}} \tau_n$$

Once this decision function is obtained, then for any threshold value, we can calculate the probability of detection and the probability of false alarm. These are obtained empirically by estimating the probability of detection as the ratio of the number of correct signal decisions declared for that threshold with the signal present over the total number of signal samples. The false alarm probability is the ratio of the number of signal decisions declared for that threshold without the signal present over the total number of non-signal or disturbance samples. These estimated probabilities provide a single point on the *ROC* curve at the selected threshold. Varying the threshold and performing the same calculation for both detection and false alarm probabilities generates the entire *ROC* curve. That is, we calculate the detection and false alarm probabilities based on the decisions made for each realization such that

$$\mathrm{P}_{DET}(n) = \frac{\text{No. Detections (Signal Present)}}{\text{TOTAL No. Signal Realizations (K)}};\; P_{\mathrm{FA}}(n) = \frac{\text{No. Detections (Signal NOT Present)}}{\text{TOTAL No. Non-Signal Realizations (K)}}$$

$$\text{with realizations } \left\{\mathrm{D}(\mu(\mathrm{k}))\right\};\, k = 1, \cdots, K \text{ and } @\, \tau_{\mathrm{n}} \text{ for } n = 1, \cdots, N$$

enabling the generation of a single point $\left(\mathrm{P}_{\mathrm{F}A}(\tau_n), \mathrm{P}_{\mathrm{D}ET}(\tau_n)\right)$ on the *ROC* curve for the specified threshold value (τ_n).

Summarizing, we can now see that the selection of the decision threshold clearly specifies an operating point of the detector by its location on the *ROC* curve. It is the overlap of the decision function distributions (conditional) that determine the corresponding detection and false-alarm probabilities. For instance, If there is *no* overlap then perfect deteciton can be achieved ((0,1) on *ROC*). Therefore, by sweeping through the thresholds, loci of operating points are defined tracing out the entire *ROC* curve.

The steps required to generate the *ROC* curve from “known” data are:

ROC GENERATION: Brute Force Method

- *Generate* two ensembles of system realizations (or disturbance or non-signal) through simulations or controlled experiments;
- For each realization, *calculate* the corresponding decision function and compare its value to the threshold;
- *Estimate* the detection and false alarm probabilities at the specified threshold (above);
- Continue to choose new thresholds and compare the decision function while *accumulating* the *counts* of detection and false alarm probabilities generating the *ROC* curve.

So we see that the dynamic *ROC* curve can easily be used to evaluate individual *detector* performance as well as compare techniques.

2.3 Average *ROC* (*AROC*)

One of the questions with a *ROC* curve is whether or not it is actually the "true" curve. Perhaps a better way to pose the problem is to ask what is the best estimate of a *ROC* curve from synthesized or actual measurement data? One way to approach this problem is to ask for the estimate, but also include a measure of precision (standard deviation) along with it.

One technique is to create an ensemble of *ROC*s which can be accomplished in a variety of ways such as through (1) simulation; (2) controlled experimental data; or (3) sectioning of actual data. When data is sparse and a simulation is not possible then "bootstrap" methods can be applied to generate the known (signal or non-signal) data to generate the *ROC* curves [19].

A simple method that can easily be applied would be to generate an *ensemble* of realizations (when possible) and create an ensemble average along with its corresponding confidence bounds to assess the quality of the *ROC* curve estimates as long as the thresholds are the same for each member [7]. For a Gaussian problem we show an ensemble of 100-curves with the mean estimate (blue) $\overline{\mathcal{R}}$ along with the $\pm 2\overline{\sigma}$ bounds (red) in Fig. 4. It must be noted that simple averaging or so-called *vertical* averaging [14] assumes that *all* of the threshold values for both detection and false-alarm probabilities are identical (aligned) for each member. If the thresholds are not available and we only have values of detection and false-alarm probabilities, then the value of the P_{FA} is fixed and the values of the P_{DET} for each ensemble value are interpolated and then averaged.

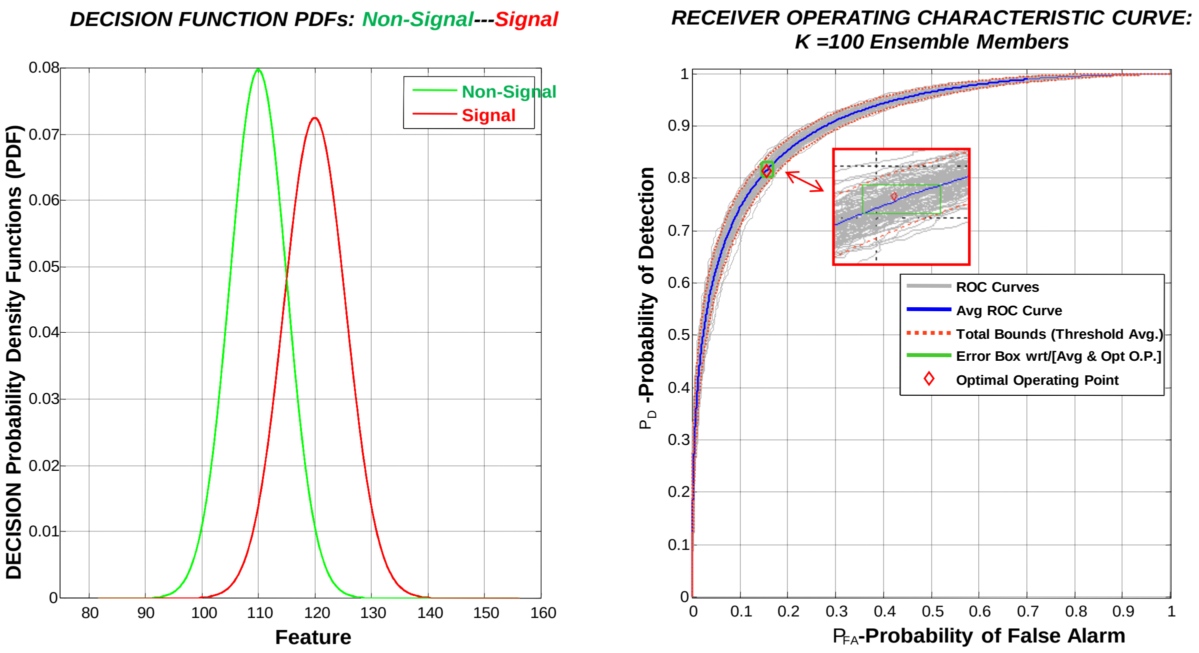
Another method of averaging is called *threshold averaging* where the detection, false-alarm and threshold values are available for "each" realization of the *ROC* curve. Here the thresholds are chosen and the corresponding detection probabilities selected (or interpolated) to provide a threshold and false-alarm probability with the resulting *L*-ensemble of *ROC* curves averaged. One way that is employed is to: (1) find the global minimum/maximum thresholds for each ensemble member, and (2) generate a set of thresholds based on these values using them to generate all *ROC* curves in the ensemble. This way the *ROCs* are forced to "threshold align" and then averaging across the ensemble is accomplished. This is the method of choice, when possible, that is, if we generate an ensemble of *ROC* curves, $\mathcal{R}_\ell(P_{DET}(\tau_n), P_{FA}(\tau_n));\ \ell = 1, \cdots, L, \forall n,$ then the ensemble *average ROC*, $\overline{\mathcal{R}}$ is given by

$$\overline{\mathcal{R}}(P_{DET}, P_{FA}) = \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{R}_\ell(P_{DET}(\tau_n), P_{FA}(\tau_n));\ \forall n \tag{7}$$

with the ensemble standard deviation given by

$$\overline{\sigma}(P_{DET}, P_{FA}) = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}(\mathcal{R}_\ell(P_{DET}(\tau_n), P_{FA}(\tau_n)) - \overline{\mathcal{R}}(P_{DET}, P_{FA}))^2} \tag{8}$$

As an example using a set of Gaussian distributions, we perform simple vertical averaging on an ensemble of 100-members (gray lines). The average (blue line) *ROC s*hown in Fig. 4 along with the corresponding confidence limits or bounds (red dots). So, we see that we have an estimated *ROC* with its associated precision metric indicating the uncertainty in the estimate. Of course if desired, the uncertainty can be decreased by incorporating more and more realizations (ensemble members) of the *ROC* ensuring tighter and tighter bounds and therefore a more precise estimate of the *ROC* curve. Next let us consider a variety of metrics that can be derived and extracted from the *ROC* curve. Also shown in the figure (red zoom box) is the "optimal decision (threshold) point" (*ODP*) and its corresponding 2σ-by-1σ uncertainty box (green). This *ODP* value is the *best* operating point for the detector. We shall discuss in more detail this in the next section.



(a)

(b)

Figure 4: Threshold averaging of *ROC* curves: (a) Example decision *PDF*s. (b) *ROC* ensemble (100-members) results with average in blue and $\pm 2\sigma$ bounds (red dots). The *ODP* (red diamond) and 2σ-by-1σ uncertainty box (green) is also shown (red zoom box).

3 *ROC* PROPERTIES

In this section we discuss a variety of properties that can be extracted from a *ROC*. Knowledge of the *ROC* specifies detection performance and enables us to compare various *detection* techniques. The shape and location of points along the curve can be used for performance analysis and specifications.

The basic properties of a *ROC* curve are well-known [1], [3], [8], and can be summarized succinctly by:

- The *ROC* curve is monotonically increasing;
- The slope of the *ROC* curve is identical to its threshold value at that point: $\tau_n = dP_{DET}(\tau_n) / dP_{FA}(\tau_n)$;
- Decision functions have *ROC*s that typically lie above the random detection curve (coin toss);
- A sufficient statistic (contains all of the statistical information required) is desirable for decisions.

Besides these basic properties, there are certain metrics that can be determined from the *ROC* curve that are used to give a "feel" towards a single quantitative number that can be used to assess the *detection* performance such as:

3.1 Distance, d'

Consider a family of *ROC* curves generated from a set of individual non-signal items, assumed Gaussian, with different location parameters (means and variances) relative to a signal item (red) as shown in Fig. 5. We know if the distance, say d', between the non-signal and signal (red) distribution is large [8], then "perfect detection" can be achieved as shown in the figure by the dark green *PDF* and its corresponding *ROC* curve [2], [3]. If the distribution is Gaussian, then the distance is simply the Euclidean distance of the ratio of individual means (M_0) to standard deviations ($\sqrt{V_0}$) relative to the signal distribution parameters ($M_1, \sqrt{V_1}$), that is,

$$d' = \left| \frac{M_1}{\sqrt{V_1}} - \frac{M_0}{\sqrt{V_0}} \right| \qquad (9)$$

As the non-signal mean of each individual (non-signal) item increases (in the figure), the corresponding non-signal distribution moves closer and closer to the signal distribution (red) increasing the overlap and decreasing desirable detector performance. This is illustrated in the figure by the non-signal decision function *PDF*s migrating closer to the signal (red) decision function *PDF* and the corresponding *ROC* curves approaching the random detector (coin toss) performance (cross-diagonal or 45^0-line) signifying complete overlap of the distributions. From the distance metric, we gain insight into the *separability* of the signal and non-signal *PDF*s and we can observe performance degradation as they overlap more and more or equivalently as the distance gets smaller and smaller. A large distance metric means a large separation (smaller overlap) and higher expected detection performance. From Fig. 5, we see an example of this for the respective distances: $d^{'} = 4.6, 3.2, 2.4, 1.6, 0.8, 0.01, 0.0$ with the *ROC* curve moving closer to the random detector (coin toss) performance as d' decreases.

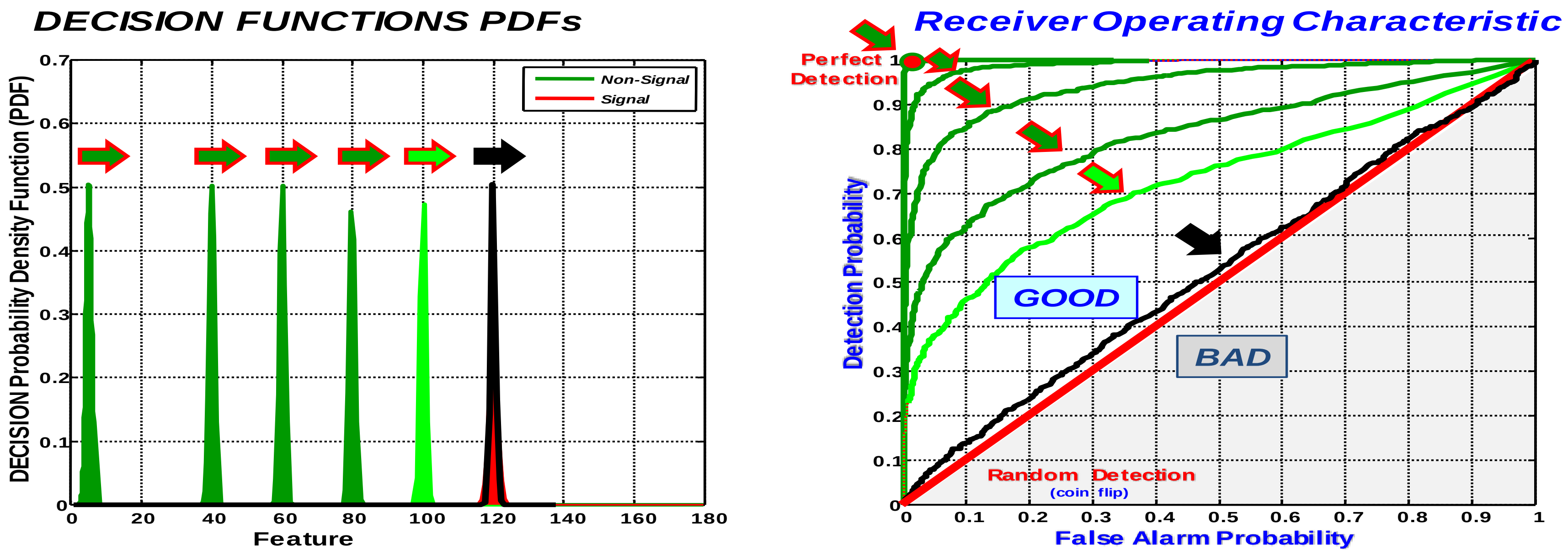


Figure 5: Family of *ROC* curves (color coded): As the decision function *PDF*s move closer to the signal (red) *PDF*, the *ROC*s move closer to the random detctor performance. All parameters of the *ROC* are listed below in (1)-(6).

(1) $ROC_1 : d_1^{'} = 4.6, AUC_1 = 1, MinE_1 = 2.1\%$;
(2) $ROC_2 : d_2^{'} = 3.2, AUC_2 = 0.99,, MinE_2 = 8.4\%$;
(3) $ROC_3 : d_3^{'} = 2.4, AUC_3 = 0.87, MinE_3 = 14.4\%$;
(4) $ROC_4 : d_4^{'} = 1.6, AUC_4 = 0.74, MinE_4 = 21.5\%$;
(5) $ROC_5 : d_5^{'} = 0.8, AUC_5 = 0.53, MinE_5 = 27.4\%$;
(6) $ROC_6 : d_6^{'} = 0.5, AUC_6 = 0.05, MinE_6 = 29.7\%$.

3.2 Area Under the Curve (*AUC*)

Another method to compare the performance of *detection* techniques is to calculate the area under the *ROC* curve [14], [15], [17]. In fact, it has been shown that the *AUC* is statistically consistent and more discriminating than accuracy (see Sec. 3.4) [18]. The *AUC* is a portion of the area of the *ROC*space (1×1 or unit square), its value is always between 0 and 1; however, since the 45^0-line from $(0,0)$-to-$(1,1)$ represents the random guess, no pragmatic *detector* would have a *AUC* less than 0.5; therefore, *AUC* ≥ 0.5. The *AUC* has an important statistical property, that is, the *AUC* of a *detector* is equivalent to the probability that the classifier will *rank* a randomly selected signal ($\mathcal{H}_1$) measurement "higher" than a randomly selected non-signal ($\mathcal{H}_0$) measurement [15] implying that it is very sensitive to detecting signals. In practice, the *AUC* performs very well and is frequently employed as a general metric of detection performance. It is calculated numerically by simple trapezoidal integration as

$$AUC = \sum_n P_{DET}(\tau_n)\Delta P_{FA}(\tau_n) + \frac{1}{2}\Delta P_{DET}(\tau_n) \times \Delta P_{FA}(\tau_n) \qquad (10)$$

$$\text{where } \Delta P_{DET}(\tau_n) = -(P_{DET}(\tau_n) - P_{DET}(\tau_{n-1}))$$
$$\Delta P_{FA}(\tau_n) = (P_{FA}(\tau_n) - P_{FA}(\tau_{n-1}))$$

Referring to Fig. 5 the area under each *ROC* curve is given by: $AUC = 1.00, 0.99, 0.96, 0.87, 0.74, 0.53, 0.50$ clearly demonstrating the robustness of this metric and its ability to predict overall performance. Thus, the larger the *AUC* or equivalently the closer its value is to *unity* (perfect detection), the better the expected detection performance.

3.3 Minimum Probability of Error (*MinE*)

The minimum (attainable) probability of error or equivalently Bayesian error corresponds to the intersection point of the decision probabilities as shown as the “star” in Fig. 3. It evolves directly from the Gaussian distribution assumption and requires the solution of an optimization problem. We would like to investigate the decision errors more closely, that is, the probability of error in making the decision---right or wrong. Assuming that we have a binary decision problem with the usual two hypotheses: $\mathcal{H}_0$ and $\mathcal{H}_1$, as before and the data set, $\mathcal{U} := \{\mu_1, \cdots, \mu_K\}$ of attenuation coefficients; therefore, the *total probability of error* is (as before in Sec. 2)

$$\Pr(\varepsilon \mid \mathcal{U}) = \Pr(\mathcal{U}, \mathcal{H}_0 \mid \mathcal{H}_1) + \Pr(\mathcal{U}, \mathcal{H}_1 \mid \mathcal{H}_0) \tag{11}$$

or applying Bayes' rule, we have

$$\Pr(\varepsilon \mid \mathcal{U}) = \int_{\mathcal{H}_1} \Pr(\mathcal{U} \mid \mathcal{H}_0)\Pr(\mathcal{H}_0) d\mathcal{U} + \int_{\mathcal{H}_0} \Pr(\mathcal{U} \mid \mathcal{H}_1)\Pr(\mathcal{H}_1) d\mathcal{U} \tag{12}$$

Thus, the probability of error is based on making the *wrong* decision, that is,

$$\Pr(\text{Error} \mid \text{Data}) := \Pr(\varepsilon \mid \mathcal{U}) = \begin{cases} \Pr(\mathcal{H}_1 \mid \mathcal{U}) & \text{if we decide on } \mathcal{H}_0 \qquad [MISS] \\ & \text{or} \\ \Pr(\mathcal{H}_0 \mid \mathcal{U}) & \text{if we decide on } \mathcal{H}_1 \quad [FALSE\ ALARM] \end{cases} \tag{13}$$

Bayesian decisions are based on the principle of selecting the hypothesis corresponding to the largest posterior probability such that

$$\begin{array}{lll} \text{If} & Pr(\mathcal{H}_0 \mid \mathcal{U}) > \Pr(\mathcal{H}_1 \mid \mathcal{U}) & \text{decide on } \mathcal{H}_0 \\ & \text{Otherwise} & \text{decide on } \mathcal{H}_1 \end{array} \tag{14}$$

Under this decision function (Bayesian) the error probability becomes

$$\Pr(\varepsilon \mid \mathcal{U}) = min\left[\Pr(\mathcal{H}_0 \mid \mathcal{U}), \Pr(\mathcal{H}_1 \mid \mathcal{U})\right] \le \Pr(\mathcal{H}_0 \mid \mathcal{U})^{\beta} \times \Pr(\mathcal{H}_1 \mid \mathcal{U})^{1-\beta} \;\; for \;\; 0 \le \beta \le 1 \tag{15}$$

A general error integral can be upper bounded [3] using this inequality to give

$$\Pr\left(\varepsilon \mid \mathcal{U}\right) \leq P_{\mathcal{U}}^{\beta}(\mathcal{H}_0 \mid \mathcal{U}) \times P_{\mathcal{U}}^{1-\beta}(\mathcal{H}_1 \mid \mathcal{U}) \times \int \mathrm{p}_{\mathcal{U}}^{\beta}(\mathcal{U} \mid \mathcal{H}_0) \times \mathrm{p}_{\mathcal{U}}^{1-\beta}(\mathcal{U} \mid \mathcal{H}_1)\, d\mathcal{U};\ \ \textit{for}\ \ 0 \leq \beta \leq 1 \tag{16}$$

where $\mathrm{p}_{\mathcal{U}}(\bullet)$ is a probability density function and $\mathrm{P}_{\mathcal{U}}(\bullet)$ the corresponding distribution. If the conditional probabilities are Gaussian, then the integral in this expression can be simplified to

$$\int \mathrm{p}_{\mathcal{U}}^{\beta}(\mathcal{U} \mid \mathcal{H}_0) \times \mathrm{p}_{\mathcal{U}}^{1-\beta}(\mathcal{U} \mid \mathcal{H}_1)\ \ d\mathcal{U} = e^{-\kappa(\beta)} \tag{17}$$

where $\kappa(\beta)$ is a function of the means and variances of the distributions (see [4] or [5] for details) leading to

$$\Pr\left(\varepsilon \mid \mathcal{U}\right) \leq P_{\mathcal{U}}^{\beta}(\mathcal{H}_0 \mid \mathcal{U}) \times P_{\mathcal{U}}^{1-\beta}(\mathcal{H}_1 \mid \mathcal{U}) \times e^{-\kappa(\beta)};\ \ \textit{for}\ \ 0 \leq \beta \leq 1 \tag{18}$$

and this is called the *Chernoff upper bound* on the error probability, $\Pr(\varepsilon)$ [4]. The bound is calculated analytically or numerically by finding the value of β that minimizes $e^{-\kappa(\beta)}$ with the error calculated by substituting the $\beta_{\min}$ from the optimizer into the integrals of Eq. 17.

If the decision functions are Gaussian $(\mathcal{N}(M_0,V_0),\mathcal{N}(M_1,V_1))$, then the bound expression becomes [4]

$$\kappa(\beta) = \frac{\beta(1-\beta)}{2}\left(M_1 - M_0\right)^{'}\left[\beta V_0 - (1-\beta)V_1\right]^{-1}\left(M_1 - M_0\right) + 1\,/\,2\,\ln\frac{\left|\beta V_0 - (1-\beta)V_1\right|}{\left|V_0\right|^{\beta}\left|V_1\right|^{1-\beta}} \tag{19}$$

For the Gaussian example of Fig. 5, we have the following sequence of bounds in percentage of error:

$$\mathit{MinE}\ (\%) = 2.1\%, 8.4\%, 14.4\%, 21.5\%, 27.4\%, 29.7\%$$

which occurs because of the overlapping in the tails of the Gaussian distributions.

3.4 Confusion Matrix

Another set of metrics that can be applied to the decision problem has been developed by diagnosticians (medicine, finance, etc.) and is based on the so-called confusion matrix which is shown in Fig. 6 [16]. For a binary (signal/non-signal) decision problem, it is a 2×2 matrix with critical *ROCs*statistics extracted at a particular operating point on the *ROC* curve. On the diagonals of the matrix we have the detection and rejection probabilities at a selected threshold, τ_n with (slight notational change: $\tau_n \to n$) $P_{DET}(n), P_{REJECT}(n)$ and on the off-diagonals, we have the corresponding false-alarm and miss probabilities $P_{FA}(n), P_{MISS}(n)$.

A common jargon in diagnostics maps these probabilities into:

$$\begin{aligned}
&P_{DET}(n) \Rightarrow P_{TP} \quad &&--- \textit{True Positive Rate}\\
&P_{REJECT}(n) \Rightarrow P_{TN} \quad &&--- \textit{True Negative Rate}\\
&P_{FA}(n) \Rightarrow P_{FP} \quad &&--- \textit{False Positive Rate}\\
&P_{MISS}(n) \Rightarrow P_{FN} \quad &&--- \textit{False Negative Rate}
\end{aligned}$$

From these distribution values extracted from the *ROC* curve, we can calculate other meaningful statistics.

We define a positive instance or event as a “signal” and a negative event as a “non-signal.” The true positives (*TP*) are the number of correct detections declared and the corresponding true positive rate (*tpr*) is defined by:

$$\mathrm{P}_{\mathrm{DET}}(n) = tpr = \frac{TP}{P} = \frac{\text{No. Correct Threat Detections}}{\text{TOTAL No. Threat Realizations}} \tag{20}$$

corresponding to the detection probability at τ_n .

Likewise, the false positives (*FP*) are the number of detections declared when the signal is *not* present and the corresponding false positive rate (*FPR*) is

$$\mathrm{P}_{FA}(n) = fpr = \frac{FP}{N} = \frac{\text{No. Incorrect Threat Detections}}{\text{TOTAL No. Non-Threat Realizations}} \tag{21}$$

corresponding to the false alarm probability at τ_n .

The true negatives (*TN*) are the number of true non-signals declared and the true negative rate (*tnr*) is given by

$$\mathrm{P}_{REJECT}(n) = tnr = \frac{TN}{N} = \frac{\text{No. Correct Non-Threat Detections}}{\text{TOTAL No. Non-Threat Realizations}} \tag{22}$$

corresponding to the rejection probability at τ_n .

Finally, the false negatives (*FN*) are the number of non-detections declared, then the non-signal is *not* present and the false negative rate (*fnr*) is defined to be

$$\mathrm{P}_{MISS}(n) = fnr = \frac{FN}{P} = \frac{\text{No. Incorrect Non-Threat Detections}}{\text{TOTAL No. Threat Realizations}} \tag{23}$$

corresponding to the miss detection probability at τ_n .

Examining the confusion matrix further, we have that the column summations (P_{TOTAL}, N_{TOTAL}) are the frequencies of occurrence of the truth (actual) items (signal/non-signal) and the row summations ($P^{'}_{TOTAL}$, $N^{'}_{TOTAL}$) are the frequencies of the choices (estimates) of the item. The total sample size, N_{ALL} is the overall sum of all the occurrences.

From these values extracted at a given operating point on the *ROC* curve, we can calculate a variety of useful metrics such as:

1. *ACCURACY* (*ACC*): is defined as the total number of correct decisions (signals & non-signals) divided by the total possible correct, that is,

$$ACC = \frac{\text{Total No. Correct Decisions}}{\text{TOTAL No. Possible Correct}} = \frac{TP+TN}{P+N} \tag{24}$$

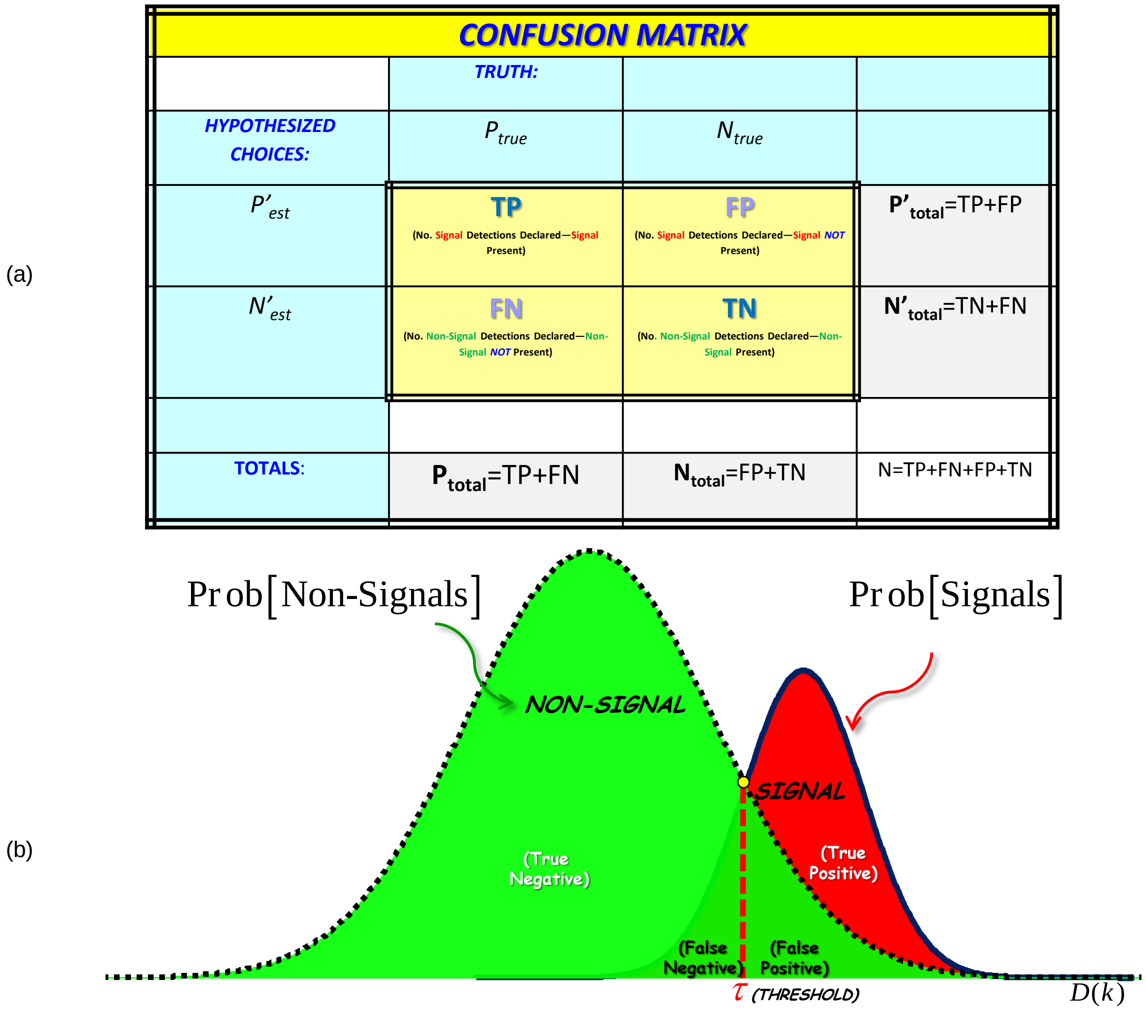
2. *PRECISION*: is usually defined as the “positive (signal)” predictive value (*PPV*) or the “negative (non-signal)” predictive value (*NPV*), since they offer an indication of how well the detector can predict signals or non-signals and is given by:

$$PPV = \frac{\text{No. Threat Decisions}}{\text{Total Threat Decisions}} = \frac{TP}{TP+FP}$$

$$NPV = \frac{\text{No. Non-Threat Decisions}}{\text{Total Non-Threat Decisions}} = \frac{TN}{TN+FN}$$

2. *SPECIFICITY*: $SPEC = P_{REJECT} = tnr = 1 - P_{FA} = 1 - fpr$
3. *SENSITIVITY*: $SENS = P_{DET} = tpr$

It should be noted that *ACC* can be a misleading metric that should be used with caution [11]. This completes the section on metrics derived from *ROC* curves


(a)

CONFUSION MATRIX

	TRUTH:		
HYPOTHESIZED CHOICES:	P_{true}	N_{true}	
P'_{est}	**TP** (No. Signal Detections Declared—Signal Present)	**FP** (No. Signal Detections Declared—Signal *NOT* Present)	$\mathbf{P'}_{\mathbf{total}}$=TP+FP
N'_{est}	**FN** (No. Non-Signal Detections Declared—Non-Signal *NOT* Present)	**TN** (No. Non-Signal Detections Declared—Non-Signal Present)	$\mathbf{N'}_{\mathbf{total}}$=TN+FN
TOTALS:	$\mathbf{P}_{\mathbf{total}}$=TP+FN	$\mathbf{N}_{\mathbf{total}}$=FP+TN	N=TP+FN+FP+TN

(b)

Figure 6: Confusion matrix for binary decision problem: (a) Matrix entries: Diagonals are detection (TP) and rejection (TN) probabilities; off-diagonals are false alarm (FP) and miss (FN) probabilities, totals are colums and row sums, rows headers are "truth" and columns are "estimates" (guesses or predictions). (b) Decision probabilities distinguishing regions mapped to matrix at a given threshold location.

3.5 Optimum Decision (Threshold) Point (*ODP*)

After obtaining the *ROC* curve from a particular *detector* or equivalently a detection algorithm, it is natural to ask the question: What is the “best” threshold τ to set and its corresponding operating point (P_{FA}, P_{DET})? That is, what is the best tradeoff between cost (P_{FA}) and benefit (P_{DET}) for the particular detection scheme? In order to answer this question, we can cast the problem into one of Bayes' risk by first defining the various costs of making a decision, defining the criterion in terms of these costs and determining the threshold value (operating point) that minimizes this risk. Thus, in this section we develop the relations to calculate the optimum decision (threshold) point (*ODP*) required to choose the best operating point from the estimated or average *ROC* curve. The Bayes' risk criterion for this *detector* has the following costs (weights) associated with decision making:

$$\begin{aligned}
&C_{00}: \quad cost\ of\ accepting\ \mathcal{H}_0,\ when\ \mathcal{H}_0\ is\ true \quad [REJECTION] \\
&C_{01}: \quad cost\ of\ accepting\ \mathcal{H}_0,\ when\ \mathcal{H}_1\ is\ true \quad [MISS] \\
&C_{10}: \quad cost\ of\ accepting\ \mathcal{H}_1,\ when\ \mathcal{H}_0\ is\ true \quad [FALSE\ ALARM] \\
&C_{11}: \quad cost\ of\ accepting\ \mathcal{H}_1,\ when\ \mathcal{H}_1\ is\ true \quad [DETECTION]
\end{aligned}$$

Also required are the prior probabilities associated with the underlying hypotheses $Pr[\mathcal{H}_1]$ (signal) and $Pr[\mathcal{H}_0]$ (non-signal). With this information, we define the *Bayes' risk criterion* [9], [18] as:

$$\begin{aligned}
\mathcal{B} := {} & C_{00} Pr[accept\ \mathcal{H}_0, \mathcal{H}_0\ true] + C_{01} Pr[accept\ \mathcal{H}_0, \mathcal{H}_1\ true] + \\
& C_{10} Pr[accept\ \mathcal{H}_1, \mathcal{H}_0\ true] + C_{11} Pr[accept\ \mathcal{H}_1, \mathcal{H}_1\ true]
\end{aligned} \tag{25}$$

Applying Bayes' rule ($Pr[A,B] = Pr[A \mid B] \times Pr[B]$) to this expression, we obtain

$$\begin{aligned}
\mathcal{B} := C_{00} Pr[\mathcal{H}_0] \times Pr[accept\ \mathcal{H}_0 \mid \mathcal{H}_0\ true] + C_{01} Pr[\mathcal{H}_0] \times Pr[accept\ \mathcal{H}_0 \mid \mathcal{H}_1\ true] + \\
C_{10} Pr[\mathcal{H}_1] \times Pr[accept\ \mathcal{H}_1 \mid \mathcal{H}_0\ true] + C_{11} Pr[\mathcal{H}_1] \times Pr[accept\ \mathcal{H}_1 \mid \mathcal{H}_1\ true]
\end{aligned}$$

Recognizing the last term in each summand as known probabilities, we have

$$\begin{aligned}
\mathcal{B} := {} & C_{00} Pr[\mathcal{H}_0] \times P_{REJECT} + C_{01} Pr[\mathcal{H}_0] \times P_{MISS} + \\
& C_{10} Pr[\mathcal{H}_1] \times P_{FA} + C_{11} Pr[\mathcal{H}_1] \times P_{DET}
\end{aligned} \tag{26}$$

Substituting for P_{REJECT} and P_{MISS} in terms of false alarm and detection probabilities and gathering like-terms, we obtain the risk as:

$$\mathcal{B} = C_{00} Pr[\mathcal{H}_0] + C_{01} Pr[\mathcal{H}_1] + (C_{10} - C_{00}) Pr[\mathcal{H}_0] \times P_{FA} + (C_{11} - C_{01}) Pr[\mathcal{H}_1] \times P_{DET} \tag{27}$$

This expression is minimized by differentiating the risk criterion with respect to the false alarm probability, setting the result to zero and solving for the threshold or equivalently the differential or slope of the *ROC* curve to give

$$\frac{d\mathcal{B}}{dP_{FA}} = (C_{10} - C_{00}) Pr[\mathcal{H}_0] + (C_{11} - C_{01}) Pr[\mathcal{H}_1] \times \frac{dP_{DET}(\tau)}{dP_{FA}(\tau)} = 0 \tag{28}$$

where solving for the slope of the *ROC* with $\tau \to \tau^*$

$$\tau^* := \frac{dP_{DET}(\tau)}{dP_{FA}(\tau)}\Big|_{\tau=\tau^*} = \frac{(C_{10} - C_{00}) Pr[\mathcal{H}_0]}{(C_{01} - C_{11}) Pr[\mathcal{H}_1]} \tag{29}$$

gives the slope leading to the desired *ODP*. We illustrate this calculation in the following example.

3.6 EXAMPLE: Gaussian Decision Function Performance Analysis

In this section we return to the example illustrated in Fig. 4 where we show two Gaussian decision functions: $\mathcal{N}(110,5)$ and $\mathcal{N}(120,5.5)$ where the notation, $\mathcal{N}(M,V)$, is a normal distribution with mean *M* and variance *V*. An ensemble of 100-members was generated for the non-signal or disturbance (green) and signal (red) distributions, respectively. Using the "brute force" method described in Sec. 2.2, each member *ROC* curve was estimated and threshold averaged. The results are shown in Fig. 7. First we note the (average) *AROC* (blue) obtained using the threshold averaging method discussed in Sec. 2.3 along with its corresponding $\pm 2\sigma$ confidence limits (bounds) in red. We also observe the optimal decision threshold point (red diamond) and its associated uncertainty (2σ-by-1σ) box (green) indicating the maximum and minimum uncertainty for both P_{DET} (vertical sides) and P_{FA} (horizontal sides). In Fig. 7, we see some of the other metrics (e.g. *AUC*, *ODP*, etc.) calculated as well as the inset of the *ODP*. These metrics are summarized in Table 1.

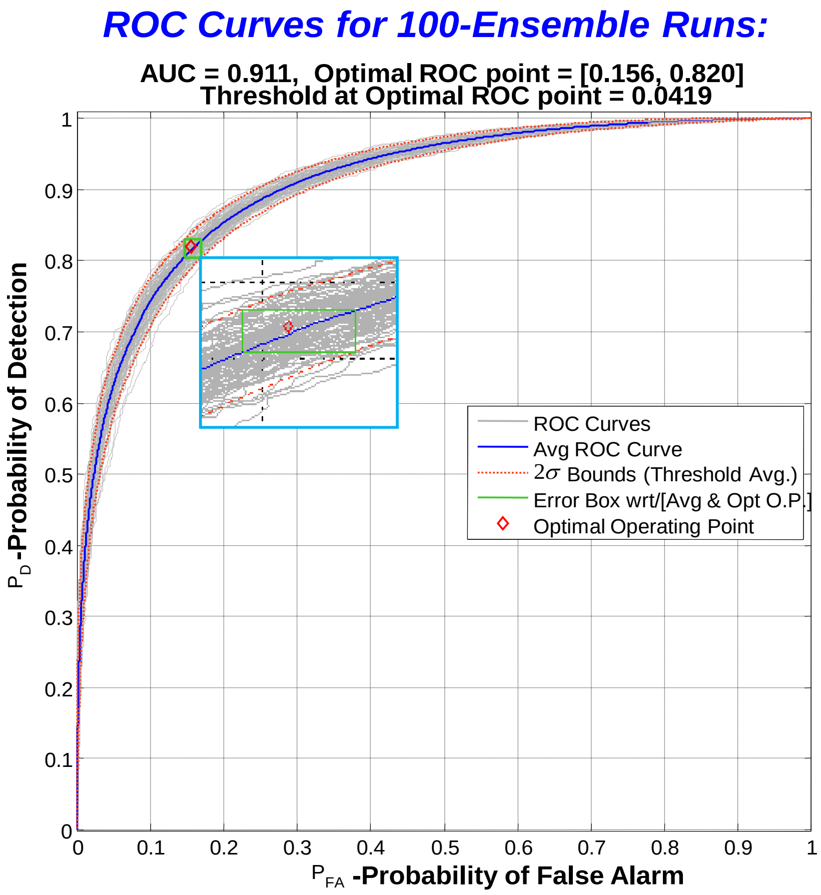
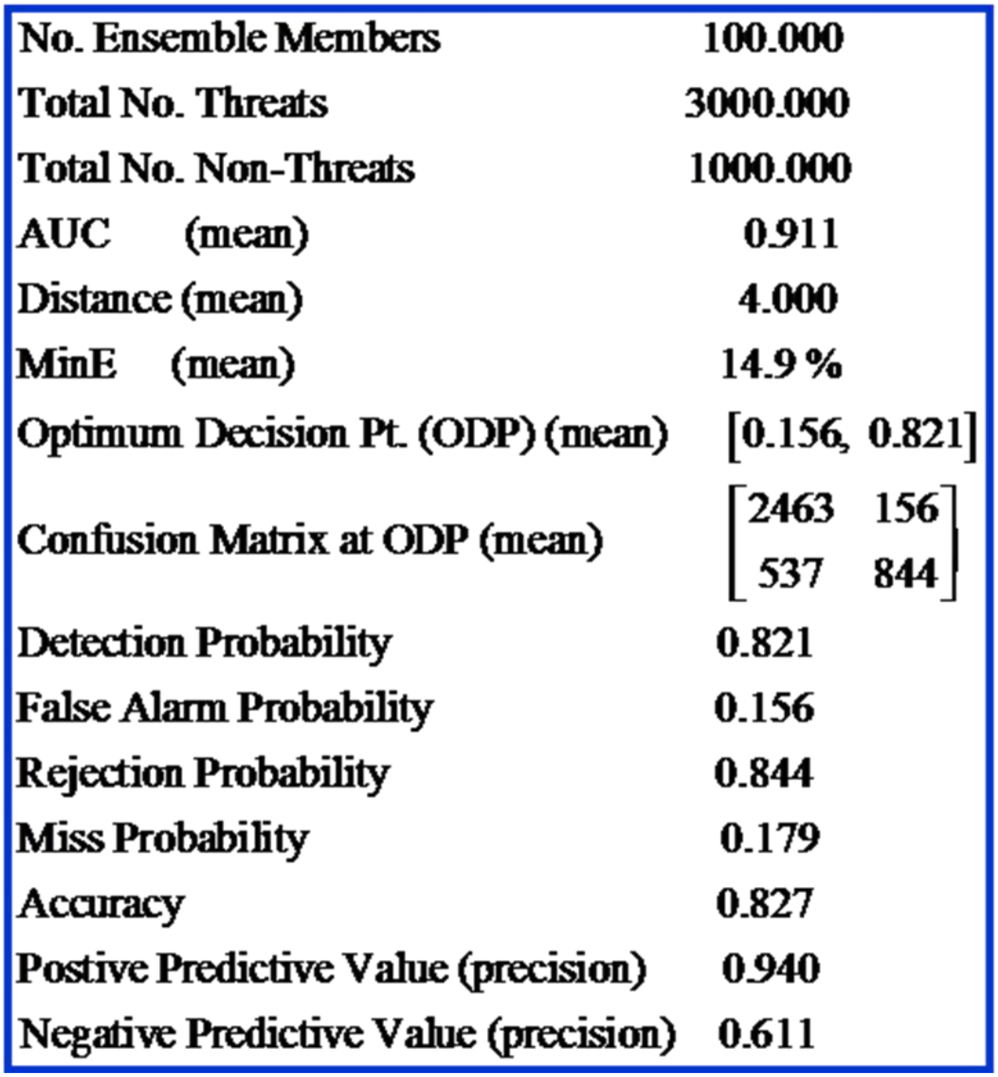


Figure 7: Final *ROC* curve for Gaussian decision function example including metrics along with zoom box for optimal decision function operating point, average (threshold averaging) and bounds.

TABLE 1: *Gaussian Decision Function Example Results*



No. Ensemble Members	100.000
Total No. Threats	3000.000
Total No. Non-Threats	1000.000
AUC (mean)	0.911
Distance (mean)	4.000
MinE (mean)	14.9 %
Optimum Decision Pt. (ODP) (mean)	[0.156, 0.821]
Confusion Matrix at ODP (mean)	$\begin{bmatrix} 2463 & 156 \\ 537 & 844 \end{bmatrix}$
Detection Probability	0.821
False Alarm Probability	0.156
Rejection Probability	0.844
Miss Probability	0.179
Accuracy	0.827
Postive Predictive Value (precision)	0.940
Negative Predictive Value (precision)	0.611

Table 1: Results for Gaussian decision function example: various metrics obtained by averaging over the ensemble are shown for detector performance analysis.

5 SUMMARY

In this report, we have developed the concept of performance metrics for automated detection systems. We have shown both pictorially and mathematically (briefly) how the *ROC* curve evolves from the detection and false-alarm probabilities.

We have also shown how a variety of performance metrics can be extracted from the *ROC* curve computation and how some enable a "single number" point (*AUC*, *ODP*, *ACC*, etc.) that can be used for performance ranking the performance of *detection* systems.

References

[1] R. O. Duda, P.E. Hart and D. G. Stork *Pattern Classification*, 2^{nd} Ed., Hoboken, New Jersey: John Wiley & Sons, 2001.

[2] K. Fukunaga, *Statistical Pattern Recognition*, 2^{nd} Ed., New York, New York: Academic Press, 1990.

[3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, New York, New York: Academic Press, 1998.

[4] I. T. Nabney, *NETLAB Algorithms for Pattern Recognition*, New York, New York: Springer, 2003.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, New York: Springer, 2006.

[6] W. W. Peterson, T. G. Birdsall and W . C. Fox. "The theory of signal detectability," Trans, IRE, Prog. Group Inform. Theor., PGIT-4, pp. 171-212, 1957.

[7] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed., New York, New York: McGraw-Hill, 2002.

[8] H. Van Trees, *Detection, Estimation and Modulation Theory*, pt. 1, New York, New York: John Wiley, 1968.

[9] A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*, New York, New York: McGraw-Hill, 1971.

[10] C. W. Therrien, *Decision, Estimation, and Classification: An Introduction to Pattern Recognition and Related Topics*, New York, New York: John Wiley, 1989.

[11] C. E. Metz, "Basic principles of *ROC* analysis," *Seminars Nuclr. Med.*, Vol. VIII, No. 4, 1978.

[12] T. Fawcett, *ROC Graphs: Notes and Practical Considerations for Researchers*, Technical Report, Palo Alto, CA: HP Laboratories, 2004.

[13] T. Fawcett, “An introduction to *ROC* analysis,” *Pattern Recogn. Letters*, Vol. 27, pp. 861-874, 2006.

[14] J. A. Swets,”Indices of discrimination or diagnostic accuracy: their ROCs and implied models,” *Psycho. Bulletin*, Vol. 99, No. 1, pp. 100-117, 1986.

[15] A. P. Bradley, “AUC: a statistically consistent and more discriminating measure than accuracy,” *Pattern Recogn.*, Vol. 30, No. 7. pp.1145-1159, 1997.

[16] C. X. Ling, J. Huang and H. Zhang, “The use of the area under the *ROC* curve in the evaluation of machine learning algorithms,” *Proc. 16th Canadian Confr. Art. Intell.*, Springer, 2003.

[17] K. Horsch, M. L. Giger and C. E. Metz, “Prevalence scaling: application to an intelligent workstation for the diagnosis of breast cancer,” *Acad. Radiol.*, Vol. 15, pp. 1446-1457, 2008.

[18] R. J. Irwin and T, C, Irwin, “A principled approach to setting optimal diagnostic thresholds: Where receiver operating characteristic and indifference curves meet,” *Europ. J. of Inter. Med.*, EJIM-02024. PP. 1-5, 2011.

[19] A. M. Zoubir and D. R. Iskander, *Bootstrap Techniques for Signal Processing*, Cambridge UK: Cambridge University Press, 2004.