Receiver Operating Characteristic Curve Confidence Intervals and Regions

John Kerekes, *Senior Member, IEEE*

***Abstract*—Many researchers have presented results showing the empirical performance of target detection algorithms using hyperspectral or synthetic aperture radar imagery. In nearly all cases, these probabilities of detection and false alarm are presented as precise values as opposed to their true nature as estimates of random values. In this letter, we provide analytical tools and examples of computing confidence intervals and regions around these estimates commonly presented as points on receiver operating characteristic (ROC) curves. It is suggested that these tools be adopted by researchers when presenting their results to provide their audience with a quantitative metric for proper interpretation of empirically estimated ROC curves.**

***Index Terms*—Confidence intervals, confidence regions, receiver operating characteristic (ROC) curves, target detection.**

I. Introduction

THE ULTIMATE metric of detection algorithm performance is the receiver operating characteristic (ROC) curve. This curve, with its origins in communication theory, describes the tradeoff between true detections and false alarms, as a threshold is swept across a detection statistic. While theoretical performance can be described analytically, given assumptions on the target and background statistical distributions, empirical estimates are often presented to demonstrate algorithm performance when applied to real data [1]–[4].

However, it can be challenging to find data sets containing accurate truth for the targets, and many studies resort to analyzing just a single image or two as a demonstration. Additionally, the numbers of targets (or target pixels) are usually modest, and the achievable false alarm rates are so low that very few samples are often used in estimating the detection and false-alarm rates. Despite these small sample sizes, detection rates and ROC curves have been often presented as precise values with no error bars. This lack of error bars implies a high level of accuracy in the ROC curve estimates that is often not justified by the sample sizes considered.

Confidence intervals provide a well-founded method to describe the range of outcomes possible due to the underlying statistical variation of the data under evaluation. That is, they allow interpretation of a given result to be extended to the world of possible replications of the experiment, given the random nature of the data.

Confidence intervals have been studied in the context of radar target detection [5] and, quite extensively, in intelligent medical systems [6]–[8]. However, there are several motivations in preparing this letter. First, many of the previous publications (particularly in the medical field) are focused on trials with small sample sizes and presented tools that have numerical problems when dealing with very low false-alarm rates $(10^{-5}$–$10^{-6})$ achievable with remotely sensed imagery. The second motivation is that very few publications deal with the 2-D aspect of the ROC curve. That is, they present confidence intervals about either probability of detection (P_{D}) or probability of false alarm (P_{FA}) but ignore the connection between the two and the need to define a confidence region around the estimated point on the ROC curve. Lastly, by publishing this letter, we hope to call attention to the topic within the remote sensing target detection community and encourage researchers to exercise appropriate caution in drawing conclusions based on small sample sizes.

This letter presents mathematical tools for calculating the confidence interval and region about estimates of target detection and false-alarm probabilities derived from empirical analysis of remotely sensed imagery. Given these confidence intervals and regions, one can make statistically significant comparisons between empirically estimated ROC curves resulting from different algorithms or experiments. That is, if the confidence region around one ROC curve does not overlap that of a second ROC curve, we can say that there is a statistical difference between the results. If the regions overlap, there is no statistical difference. The size of the regions will vary inversely with the sample sizes used in the empirical evaluations. Section II presents the fundamental assumptions used in this letter. Section III provides background and the techniques to calculate 1-D confidence intervals. Section IV addresses the extension of these techniques to 2-D ROC-curve confidence regions. Section V provides a summary.

II. ROC Curves

A. ROC Curves

We consider the general problem of binary hypothesis testing, where a decision must be made between two mutually exclusive hypotheses: H_0 (target not present) or H_1 (target present).

The target may be an object of military interest, a diseased plant in an agricultural field, or a cancerous area in a medical image. The term "object of interest" is more appropriate and all encompassing, but for brevity, we will use the generic term "target" throughout this letter.

Four outcomes of the binary hypothesis test are possible.

TT True target labeled as a target (correct detection).

TB True target labeled as background (missed detection).

Manuscript received March 31, 2007; revised September 22, 2007.

The author is with the Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623 USA (e-mail: kerekes@cis.rit.edu).

Digital Object Identifier 10.1109/LGRS.2008.915928

BT True background labeled as target (false alarm).
BB True background labeled as background (correct nondetection).

The ROC curve describes detection performance with P_{D} plotted as a function of P_{FA}. When applied to a known empirical situation (with ground truth), these probabilities are estimated by the maximum-likelihood estimates shown in

$$\hat{P}_{\mathrm{D}} = \frac{\text{Number of observed true detections } (N_{\mathrm{TT}})}{\text{Number of possible targets } (N_{\mathrm{TT}} + N_{\mathrm{TB}})} \tag{1}$$

$$\hat{P}_{\mathrm{FA}} = \frac{\text{Number of observed false detections } (N_{\mathrm{BT}})}{\text{Number of background pixels } (N_{\mathrm{BT}} + N_{\mathrm{BB}})}. \tag{2}$$

It is common for target detection algorithms to produce a scalar test statistic h for each observation. An observation x_j is labeled (correctly or incorrectly) as a target when its associated test statistic h_j exceeds a threshold. The ROC curve is obtained by varying the threshold across the range of the test statistic with points on the curve obtained by counting the number of true and false detections and computing (1) and (2) at each threshold. For cases where overlap occurs in the test statistic for target and background samples, increasing the number of true detections will necessarily increase the number of false alarms. The ROC curve characterizes this tradeoff.

In synthetic aperture radar or hyperspectral image target detection studies, detection decisions can be made on a per-target or per-pixel basis. For cases where the target is subpixel or there is a desire to have fine granularity in the results and larger sample sizes, the per-pixel method is used, and we will consider that case in our examples, although the methodology applies with either approach.

B. Assumptions and Approximations

The fundamental assumption we make is that the detections are statistically independent events with probability P stationary over the set of trials (pixels). In this case, the number of detections N_{D} (true detections or false alarms[1]) out of N pixels is well known to be binomially distributed with

$$\Pr\{N_{\mathrm{D}}\} = \frac{N!}{N_{\mathrm{D}}!(N - N_{\mathrm{D}})!} P^{N_{\mathrm{D}}}(1-P)^{N-N_{\mathrm{D}}}. \tag{3}$$

As will be discussed in Section III, the 1-D confidence intervals for the binomial distribution of (3) can be calculated directly using the F-distribution, even for large sample sizes. Our method presented in Section IV for 2-D confidence regions requires the computation of the explicit probability density functions. In cases with N large, the numerical computation of the binomial probability density function becomes difficult and unnecessary, as approximations become sufficiently accurate. In particular, two other well-known distributions can be used to approximate the binomial under the following conditions [9].

[1]To be clear, when using (3) to calculate the probability of N_{D} target detections, $N = N_{\mathrm{TT}} + N_{\mathrm{TB}}$ is the number of true target pixels and $P = P_{\mathrm{D}}$ is the probability of detecting a target, but when calculating the probability of N_{D} false alarms, $N = N_{\mathrm{BT}} + N_{\mathrm{BB}}$ is the number of true background pixels and $P = P_{\mathrm{FA}}$ is the probability of a false alarm.

For large sample sizes $(N > 100)$ and moderate or high probabilities $(P > 0.1)$, the binomial distribution can be accurately approximated by a Gaussian distribution with mean NP and variance $NP(1-P)$

$$\Pr\{N_{\mathrm{D}}\} = \frac{1}{\sqrt{2\pi}\sqrt{NP(1-P)}} \exp\left[\frac{(N_{\mathrm{D}} - NP)^2}{2\left[NP(1-P)\right]}\right]. \tag{4}$$

For large sample sizes $(N > 100)$ but low probabilities $(P < 0.1)$, we can approximate the binomial with the Poisson distribution. In the limit of $N \to \infty$ and $P \to 0$, but with $NP = \lambda$, a finite nonzero value, the binomial converges to the Poisson

$$\Pr\{N_{\mathrm{D}}\} = \frac{\lambda^{N_{\mathrm{D}}}}{N_{\mathrm{D}}!} e^{-\lambda}. \tag{5}$$

III. Confidence Intervals on $P_{\mathrm{D}}/P_{\mathrm{FA}}$ Estimates

A. Confidence Interval Theory

The theory of confidence intervals is well understood and has seen wide application. We illustrate the concept with the simple case of estimating the mean value $\overline{x}$ from a sample of N Gaussian distributed random variables with true mean μ_x and known true variance σ_x^2. The $100(1-\alpha)\%$ confidence interval with significance level α can be described by

$$\Pr\left\{z_{1-\alpha/2} < \frac{(\bar{x} - \mu_x)\sqrt{N}}{\sigma_x} \le z_{\alpha/2}\right\} = 1 - \alpha \tag{6}$$

where z_β is the cutoff for the standardized normal variable with area under the Gaussian curve from z_β to ∞ equal to β. Here, α is split evenly between the upper and lower limits.

We can rearrange (6) to provide the confidence interval derived for the sample set as

$$\left[\overline{x} - \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}} \le \mu_x < \overline{x} + \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}}\right]. \tag{7}$$

We interpret (7) as "the true mean μ_x that lies within the noted interval with a confidence of $100(1-\alpha)\%$." Note that, for the case of unknown mean and unknown variance, we use the estimated sample variance with the Student t distribution instead of the standard normal [10].

The above is generalized by defining $\hat{\theta}$ as the estimate of the desired parameter with a confidence interval defined by lower limit θ_{L} and upper limit θ_{U}. This interval is shown in Fig. 1, where the upper limit is found by assuring the lower tail of a distribution centered on θ_{U}, which has area $\alpha/2$, with a complementary argument used to find the lower limit.

B. One-Dimensional Confidence-Interval Calculations

Following the theory outlined earlier, the confidence interval limits for parameter θ (either P_{D} or P_{FA}) can be determined by

$$\int_{\hat{\theta}}^{1} p\left(\theta|\theta_{\mathrm{L}}\right) d\theta = \int_{0}^{\hat{\theta}} p\left(\theta|\theta_{\mathrm{U}}\right) d\theta = \alpha/2 \tag{8}$$

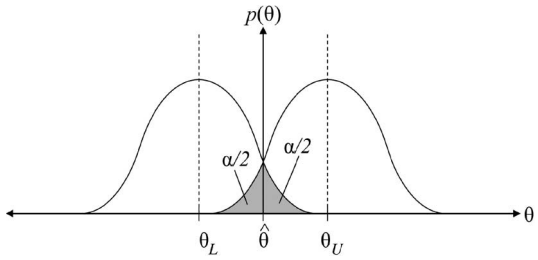

Fig. 1. Definition of lower limit $\theta_{\rm L}$ and upper limit $\theta_{\rm U}$ for θ using observed estimate $\hat{\theta}$, assumed distribution, and evenly split confidence level of $1-\alpha$.

which states that the integrated area under the probability density function $p(\theta)$ from the observed value $\hat{\theta}$ to the extreme is equal to $\alpha/2$. The exact upper and lower limits have been derived for the three distributions discussed and are calculated as follows, where N is the number of samples [9].

Binomial :

$$\theta_{\rm L}=\frac{N\hat{\theta}}{N\hat{\theta}+(N-N\hat{\theta}+1)f^2_{\alpha/2}} \qquad \begin{matrix} d_1=2(N-N\hat{\theta}+1) \\ d_2=2N\hat{\theta} \end{matrix} \tag{9}$$

$$\theta_{\rm U}=\frac{(N\hat{\theta}+1)f^2_{\alpha/2}}{N-N\hat{\theta}+(N\hat{\theta}+1)f^2_{\alpha/2}} \qquad \begin{matrix} d_1=2(N\hat{\theta}+1) \\ d_2=2(N-N\hat{\theta}) \end{matrix} \tag{10}$$

where $f^2_{\alpha/2}$ is the cutoff value of an F-distributed random variable with d_1 and d_2 degrees of freedom such that the probability that the F-distributed random variable is greater than the cutoff value is $\alpha/2$. These equations represent solutions to the Clopper–Pearson equations for the binomial confidence interval [11].

Poisson :

$$\theta_{\rm L}=\frac{1}{2N}\chi^2_{1-\alpha/2} \qquad d=2N\hat{\theta} \tag{11}$$

$$\theta_{\rm U}=\frac{1}{2N}\chi^2_{\alpha/2} \qquad d=2(N\hat{\theta}+1) \tag{12}$$

where $\chi^2_{\alpha/2}$ is the cutoff value of a Chi-squared-distributed random variable with d degrees of freedom such that the probability that the Chi-squared-distributed random variable is greater than the cutoff value is $\alpha/2$.

Gaussian :

$$\theta_{\rm L}=\hat{\theta}-\frac{\hat{\sigma}t_{\alpha/2}}{\sqrt{N}} \qquad d=N-1 \tag{13}$$

$$\theta_{\rm U}=\hat{\theta}+\frac{\hat{\sigma}t_{\alpha/2}}{\sqrt{N}} \qquad d=N-1 \tag{14}$$

where $t_{\alpha/2}$ is the cutoff value of a Student t distributed random variable with d degrees of freedom such that the probability that the Student t random variable is greater than the cutoff value is $\alpha/2$. In addition, $\hat{\sigma}=\sqrt{\hat{\theta}(1-\hat{\theta})}$ is the standard deviation for the estimated parameter.

C. Examples

Table I shows several examples of exact 1-D 95% confidence intervals ($\alpha=0.05$) for a range of values for $\hat{P}$ (detection or false-alarm rate), the sample size N, and the distribution assumed based on the discussion in Section II-B and N. This table uses (9)–(14), where $\hat{\theta}=\hat{P}=N_{\rm D}/N$, the probability of a detection or false-alarm event. Note that, even in the case of $N_{\rm D}=0$, an interval can be defined, but the cutoff value must use α instead of $\alpha/2$, since the tail area is one-sided.

TABLE I
EXAMPLE OF 95% CONFIDENCE INTERVALS FOR PROBABILITY ESTIMATE

N_D	N	Method	$\hat{P}$	P_L	P_U
5	10	Binomial	0.500	0.187	0.813
8	10	Binomial	0.800	0.444	0.975
50	100	Binomial	0.500	0.398	0.602
500	1000	Gaussian	0.500	0.469	0.531
800	1000	Gaussian	0.800	0.775	0.825
0	10 000	Poisson	0.0	0.0	$2.996\text{x}10^{-4}$
1	10 000	Poisson	$1.000\text{x}10^{-4}$	$2.532\text{x}10^{-6}$	$5.572\text{x}10^{-4}$
10	100 000	Poisson	$1.000\text{x}10^{-4}$	$4.795\text{x}10^{-5}$	$1.839\text{x}10^{-4}$

IV. CONFIDENCE REGIONS ON ROC CURVES

A. Confidence Region Theory

The direct extension of the discussion in Section III-A to the 2-D case of ROC curves is not straightforward. The explicit solution would require finding the set of points around an estimated point on the ROC curve such that the volume under the joint probability density function (centered at a confidence region boundary point) integrated from the estimated point to the appropriate extremes in all four cardinal directions is $\alpha/4$. To explain this further, let $p(P_{\rm FA},P_{\rm D})$ be the joint probability density function for our probabilities of false alarm and detection. First, consider the area on the ROC graph below and to the left of an empirically estimated point $(\hat{P}_{\rm FA},\hat{P}_{\rm D})$. Following (8), the boundary of the confidence region in that area would be defined by all points $(P_{\rm FA\_L},P_{\rm D\_L})$ that satisfy the following:

$$\int_{\hat{P}_{\rm FA}}^{1}\int_{\hat{P}_{\rm D}}^{1} p\left(P_{\rm FA},P_{\rm D}|P_{\rm FA\_L},P_{\rm D\_L}\right)dP_{\rm FA}dP_{\rm D}=\alpha/4. \tag{15}$$

The boundary points in the other three quadrants surrounding the estimated point would be determined in a similar manner.

While this approach could lead to accurate identification of the confidence region boundaries, its implementation is excessively complex and computationally burdensome. Essentially, one would have to perform the earlier integration for each possible point on the ROC curve plane, keeping track of only those that satisfy the equality to define the boundary.

B. Intuitive Approximation to Confidence Regions

An alternative to the excessively complex method outlined in Section IV-A is to consider an intuitive definition of a confidence region. That is, given the joint probability density function at an estimated point $(\hat{P}_{\rm FA},\hat{P}_{\rm D})$ on the ROC curve, define the confidence region around that point as the boundary that encompasses $100(1-\alpha)\%$ of the volume under the function. This is equivalent to the common approximate view of 1-D confidence intervals which is that the lower and upper limits are defined such that $100(1-\alpha)\%$ of the area under the probability

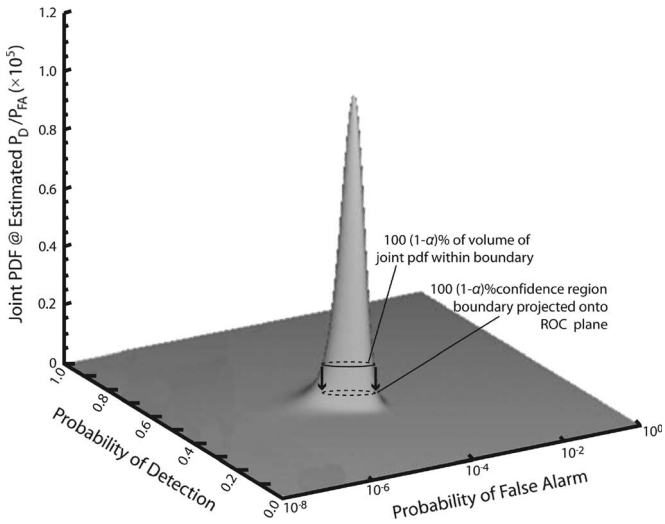

Fig. 2. View of joint probability density function at a given point on the ROC curve. Also shown is the boundary for the $100(1-\alpha)\%$ volume and its projection down to the plane of the ROC curve.

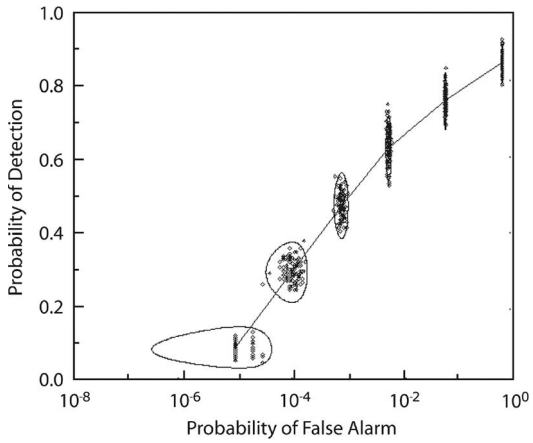

Fig. 3. 95% confidence regions drawn around each empirical point $(+)$ on an example estimated ROC curve using $N_{\mathrm{T}} = 200$ and $N_{\mathrm{B}} = 100\,000$. Also plotted are point estimates simulated from 100 trials of binomially distributed pseudorandom variables.

density function lies between the limits. An example is shown in Fig. 2.

C. Practical Implementation of ROC Confidence Regions

While the earlier discussion illustrates the theory behind ROC confidence regions, we now present a practical way to produce a plot of the regions given an empirically estimated ROC curve. The basic idea is to find the boundary of the region for each estimated point on the ROC curve and then, if there are sufficient points on the curve and the regions overlap, merge them together to produce a single region around the entire interpolated curve.

To define the confidence region boundary for a given point on the empirically estimated ROC curve, we first recall our assumption that detection and false-alarm events are statistically independent. We then select the appropriate statistical model for P_{D} and P_{FA} based upon the discussion in Section II-B and form the joint density function as the product of the 1-D density functions. A discrete sampling grid across the ROC curve plane is formed (at a sufficient sampling density), and a numerical computation of the joint density function centered on the estimated ROC point is made similar to the one plotted as a shaded surface in Fig. 2. Then, to find the $1-\alpha$ boundary, we follow the approach presented in [8], and we reform this 2-D joint density function as a 1-D vector and sort the values from high to low, keeping track of their locations in the ROC curve plane. We then integrate this sorted probability density function until we find the index i_α for which the integrated probability corresponds to $1-\alpha$. At this point, all indexes less than i_α in the sorted density function correspond to points inside the $1-\alpha$ boundary.

While there are several ways to then determine and plot the actual points in the confidence region boundary, we present the following simple and straightforward method which is easily implemented in high-level programming tools. We create a binary image in the ROC curve plane by placing a one in locations corresponding to all indexes in the sorted probability density function less than i_α and a zero elsewhere. This produces a filled-in area corresponding to the confidence region for that estimated point on the ROC curve. A contour plot of the edge around this region can then be made and overlaid on the ROC curve to show the confidence region boundary at that point.

This process is repeated for all estimated points on the ROC curve. If there are sufficient points on the ROC curve and the regions overlap, the individual regions can be merged by taking the union of the binary images and plotting the contour of the filled-in areas as the confidence region for the interpolated ROC curve.

D. Examples

We demonstrate and verify the method by first presenting confidence regions for a ROC curve with six estimated points. We consider a case with 100 000 background pixels and 200 target pixels. We define six points with (20, 60, 95, 125, 150, 170) as the number of target pixels detected and (1, 10, 75, 500, 5000, 50 000) as the number of background pixels detected as false alarms for corresponding thresholds. Fig. 3 shows the resulting ROC curve with 95% confidence regions drawn around each point on the estimated ROC curve.

We also have shown in Fig. 3 point estimates from 100 trials at each of the six locations on the ROC curve generated using simulated binomially distributed pseudorandom variables generated using the estimated P_{D} and P_{FA} at each point. We can see that these simulated ROC points generally lie within the 95% confidence regions, with just a few falling outside as one would expect. Note that the left side of the confidence region near the bottom of the plot does not have any simulated points. This is because the lowest possible simulated empirical false-alarm rate is 10^{-5} (one false alarm in the 100 000 pixels). However, the theory developed shows the theoretical confidence interval for the true P_{FA} given the observed estimates and finite sample size.

Two observations can be made from this example. One is that, given the limited number of target pixels $(N_{\mathrm{T}} = 200)$,

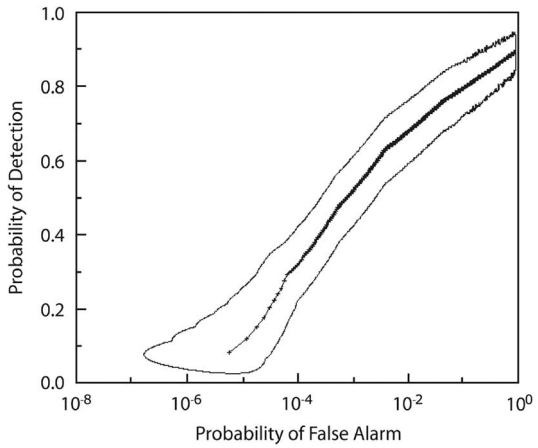

Fig. 4. 95% confidence regions drawn around each empirical point $(+)$ for a continuously varying threshold using interpolated detection and false-alarm numbers from the example shown in Fig. 3 with $N_{\mathrm{T}} = 200$ and $N_{\mathrm{B}} = 100\,000$.

the confidence regions span a large range of $P_{\mathrm{D}}(\sim 0.2)$. The other is that, while the confidence regions are quite wide (orders of magnitude) at low P_{FA} values $(\sim 10^{-5})$, they become very narrow (a few percent) at high P_{FA} values $(\sim 10^{-2})$.

Fig. 4 shows the confidence region around an estimated ROC curve developed using 1400 points (spaced nonlinearly) interpolated from the six points used in the previous example. In this result, the confidence regions drawn around each point overlap. Since their union was used in the contour-drawing program, the confidence regions are merged and appear continuous.

Note that the exact 1-D confidence intervals for P_{D} or P_{FA} can be calculated at any point on the ROC curve using the equations and methods outlined in Section III-B.

V. Summary and Discussion

Methods have been presented to compute and plot confidence intervals and regions around points on empirically estimated ROC curves. A new method has been presented to obtain confidence regions on ROC curves using a tractable approximation to true confidence regions. This new method was shown to be consistent with numerical simulations of random trials for estimated points on the ROC curve.

The methods and results demonstrated in this letter should provide researchers wishing to compare the performance of target detection systems the tools by which statistically significant conclusions can be made regarding performance estimates. These tools also provide a mechanism for assessing the validity of simulation and modeling tools by allowing predicted and empirical performance to be compared through confidence regions rather than point estimates. Future work will include the application of these tools to the validation of our spectral imaging model [12] by exploring whether its predicted ROC curves lie within confidence regions developed from the method presented here and corresponding empirical data.

References

[1] E. Ashton, "Detection of subpixel anomalies in multispectral infrared imagery using an adaptive Bayesian classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 2, pp. 506–517, Mar. 1998.

[2] C.-I Chang and H. Ren, "An experiment-based quantitative and comparative analysis of target detection and image classification algorithms for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 1044–1063, Mar. 2000.

[3] R. Mayer, F. Bucholtz, E. Allman, D. L. von Berg, and M. Kruer, "A metric of background candidate assessment for spectral target signature transforms," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 113–117, Apr. 2005.

[4] T. Wang, J. Keller, P. Gader, and O. Sjahputera, "Frequency subband processing and feature analysis of forward-looking ground-penetrating radar signals for land-mine detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 718–729, Mar. 2007.

[5] J. Echard, "Estimation of radar detection and false alarm probabilities," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 27, no. 2, pp. 255–260, Mar. 1991.

[6] C. Metz, "ROC methodology in radiological imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[7] S. Mackassy, F. Provost, and S. Rosset, "ROC confidence bands: An empirical evaluation," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 537–544.

[8] J. Tilbury, P. Van Eetvelt, J. Garibaldi, J. Curnow, and E. Ifeachor, "Receiver operating characteristic analysis for intelligent medical systems—A new approach for finding confidence intervals," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 952–963, Jul. 2000.

[9] A. Hald, *Statistical Theory With Engineering Applications*. New York: Wiley, 1952.

[10] J. Bendat and A. Piersol, *Random Data: Analysis and Measurement Procedures*. New York: Wiley-Interscience, 1971.

[11] C. Clopper and E. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, Dec. 1934.

[12] J. Kerekes and J. Baum, "Spectral imaging system analytical model for subpixel object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1088–1101, May 2002.