# QUANTITATIVE EVALUATION OF NEURAL NETWORKS FOR NDE APPLICATIONS USING THE ROC CURVE

Mackay A. E. Okure and Michael A. Peshkin
Department of Mechanical Engineering
Northwestern University
Evanston, IL 60208

## ABSTRACT

The relative operating characteristic (ROC) method is applied to performance evaluation of neural networks. The study was motivated by the need to objectively evaluate neural networks for flaw waveform identification in NDE equipment, and to compare neural network performance with other methods. NDE applications are characterized by noisy real-world data, less-than-perfect detection and a serious problem of false alarm indications. The ROC method is explained by modeling neural network output as exponential probability distributions with two peaks, one near 1 (flaw) and one near 0 (no flaw). 100% POD (probability of detection) can only be achieved when the POFA (probability of false alarm) is also 100%, and if a POFA of 0% is required, the POD also falls to 0%. The ROC curve presents all intermediate performance information in an objective form and depicts the inevitable trade-off in every interpreter, human, neural, or otherwise. The ROC method is applied to the comparison of the performance of a neural network and a threshold-based scheme in classifying real-world eddy current data collected from an aircraft wheel NDE system.

KEY WORDS: performance evaluation, ROC method, probability of detection, neural networks, fundamentals of QNDE methods.

## INTRODUCTION

This paper arose from the need to quantitatively evaluate performance of a neural network example used to distinguish corrosion from noise. Although this is a simple GO/NO-GO situation, many NDE systems can be reduced to this level; for example, corrosion/no corrosion, crack/no crack, dangerous crack/ineffective crack. Even when the size of a crack is to be determined, the instrument or operator must first be able to detect it. Many NDE managers would probably prefer this GO/NO-GO decision since it can easily be verified, is likely to be more accurate, predictable and fast.

In comparing simple signal amplitude thresholding to neural networks, one quickly finds that POD measurement alone does not give enough information. For example, whereas a probability of detection (POD) of 90% for an NDE system may be good, it’s not enough on its own unless the false alarm rate is given too.

We use the ROC method adopted from signal detection theory, [1] [2] [3]. We also show how to transform NDE system output, such as signal measurement or derived computations, into appropriate conditional probability distributions, that is, probability of detection and probability of false alarm. These are the building blocks of the ROC curve. Similar analyses have recently been reported by Swets[4], Sturges [5] and Nockemann [6]. A hypothetical distribution is used to explain the derivation of the ROC curve and show how quantifying the ROC curve leads to a single measure. We present results of applying the method to neural network interpretation of eddy current NDE signal segments and comparison with a threshold-based scheme. The data used is taken from a database of signals collected from a commercial machine used to inspect aircraft wheels during periodic maintenance .

## ROC MODELING USING THEORETICAL PDFS

Explanation of the principles of POD, POFA and the ROC curve given below uses a simple mathematical model of imperfect discrimination between noise and signal-plus-noise by a neural network-based or any classifier. Specifically, noise and signal-plus-noise inputs to a neural network produce values of a decision variable, such as the neural network output, that varies from one occasion to another, with overlapping distributions of values associated with the two classes of events. These distributions, with respect to a decision variable x, may be modeled as class conditional probability density functions $p_{noise}|x$ and $p_{corrosion}|x$, for noise alone and for corrosion signal, respectively.

Let the probability density functions be modeled as

$$p_{noise}|x = Ae^{-\alpha x} \qquad (1)$$

and

$$p_{corrosion}\Big|x = Be^{-\beta(1-x)} \qquad (2)$$

where $0 \le x \le 1$.

Figure 1 shows such a distribution with exponential factors selected arbitrarily. A and B are introduced to satisfy the axioms of probability and determined following [7].

The corresponding POD and POFA are

$$POD = \int_{x_0}^{1} p_{corrosion}|xdx \qquad (3)$$

and

$$POFA = \int_{x_0}^{1} p_{noise}|xdx\,, \qquad (4)$$

where $x_0$ is the alarm threshold.

This gives

$$POD = \frac{1-e^{-\beta(1-x_0)}}{1-e^{-\beta}} \qquad (5)$$

and

$$POFA = \frac{e^{-\alpha x_0} - e^{-\alpha}}{1-e^{-\alpha}}\,. \qquad (6)$$

Figure 2 shows the two plots of POD and POFA versus alarm threshold.

In Figure 3, a scatter plot of (POD, POFA) pairs at corresponding $x_0$ is made on a POD versus POFA graph. This shows the spectrum of the possible operating points of the system.

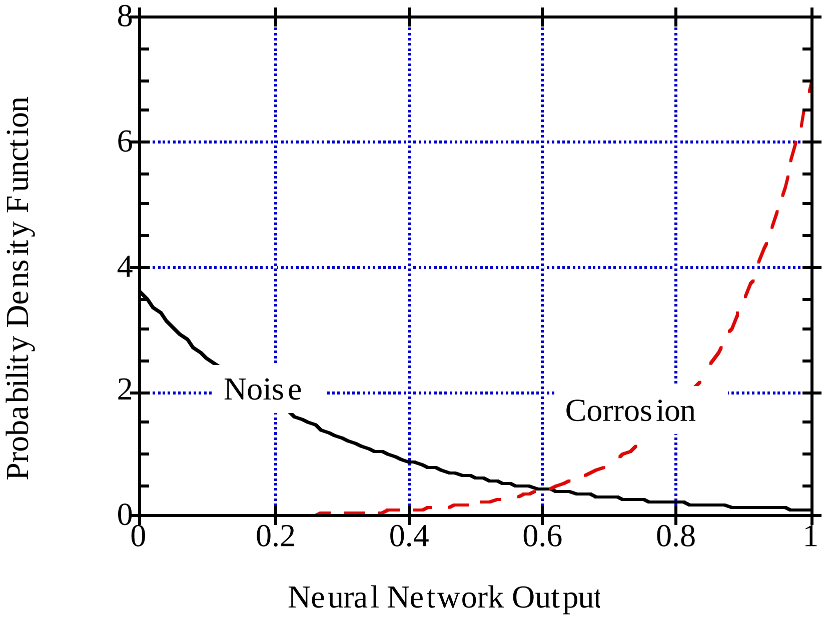

Figure 1 Hypothetical probability density distributions for noise and corrosion as a function of neural network output. Exponential factor for noise is 3.5 while that for corrosion is 6.9. Their shapes are selected to reflect the fact that the network is trained to output 0 for noise and 1 for corrosion.

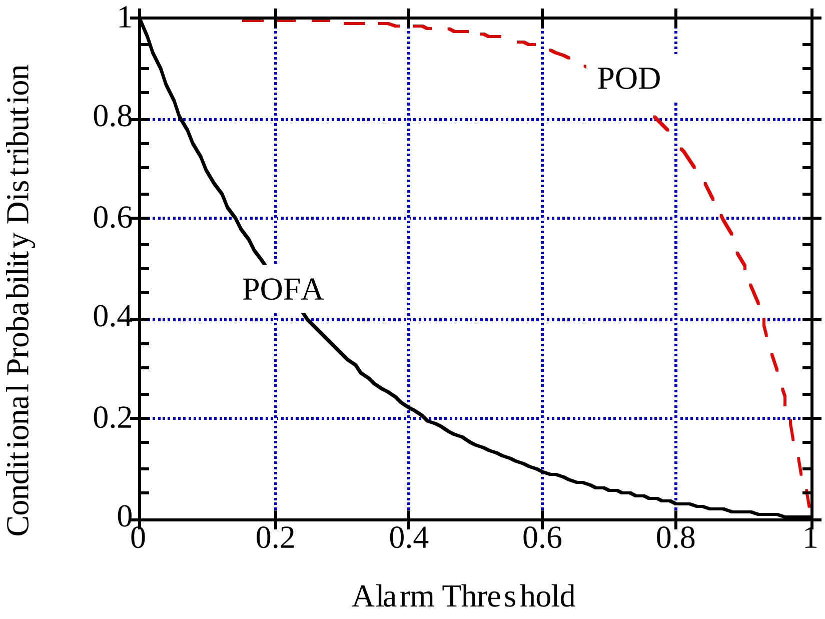

Figure 2 POD and POFA distributions as a function of alarm threshold. At an alarm threshold of 0, POD and POFA are at their maximum. The values then decrease together as the alarm threshold increases. At the lowest value of the alarm threshold both POD and POFA are 0.

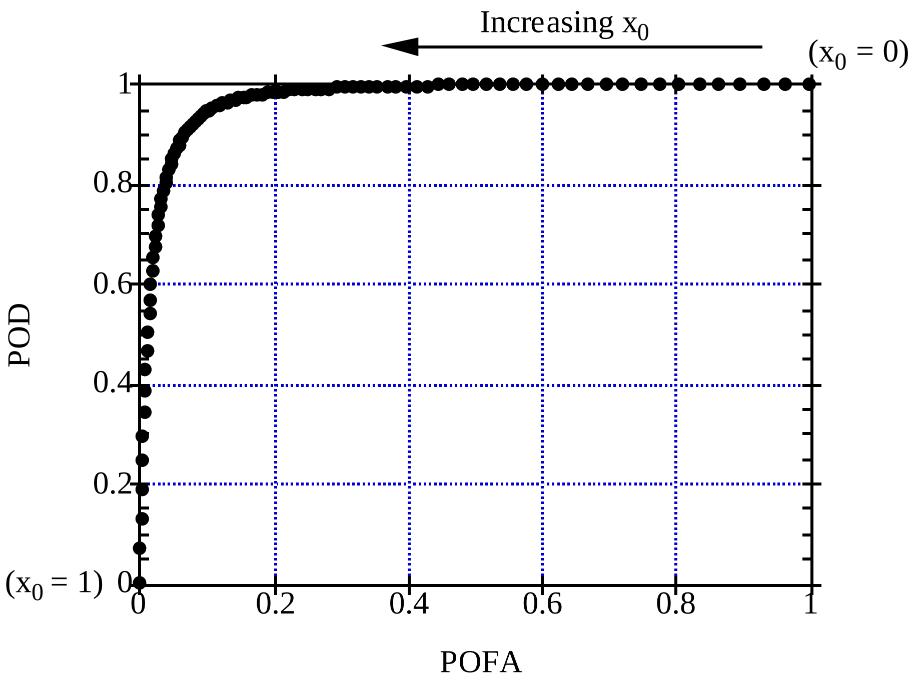

Figure 3 ROC curve for Hypothetical Distribution. The ROC curve is a plot of (POD, POFA) pairs on a POD versus POFA graph parameterized by alarm threshold, $X_{0,}$ the decision variable.

It is evident from Figure 3 that the alarm threshold, $X_{0,}$ is an important parameter. It is observed that at low values of $X_0$, both a high POD and a high POFA are obtained whereas at high values, both the POD and the POFA are low. POD and POFA thus increase and decrease together, and by varying the alarm threshold, different levels of POD and POFA are obtained.

## CONVERSION OF THE ROC CURVE TO A SINGLE MEASURE

A single-measure may be useful for expressing the performance of a classifier following the derivation of the ROC curve. The purpose of such a measure is to allow the location of different ROC curves on a common spectrum and facilitate comparison among systems.

The two extremes of such a measure should correspond to the worst and the best or ideal classifier. The worst classifier, on one hand, may be defined as that which has no discrimination between positives and negatives. A positive will have equal chance of being interpreted as a positive or a negative , and vice-versa. This means true-positive and false-positive frequencies are equal, that is, POD = POFA. This is a straight-line between points (0.0, 0.0) and (1.0, 1.0). The best classifier, on the other hand, represents perfect interpretation which follows a POD = 1.0 for all values of POFA. This corresponds to an ideal performance of POD = 1.0 and POFA = 0.0, that is, the top-left corner.

Figure 4 illustrates examples of ROC curves for an ideal, good, poor and worst classifier.

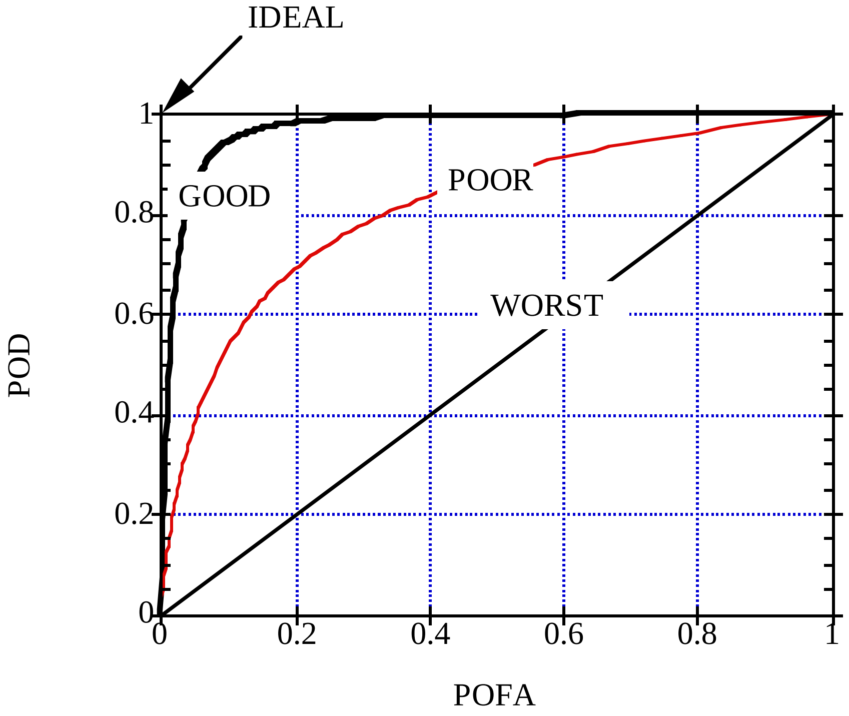

Figure 4: Illustrating an ideal, good, poor and worst classifier. The worst classier has no discriminating ability between a flaw and noise. Any input has equal probability of being called a flaw or noise. Improving performance shifts the ROC curve towards (0, 1), the ideal point.

### Area Under the ROC Curve

This is the area of the entire graph that lies beneath the curve, and is designated $A(P)$. It is bounded by the axes POD = 0.0 and POFA = 1.0, and the (POD, POFA) pairs that are generated during testing of the classifier and is often computed by the trapezoidal rule.

$A(P)$ values vary from 0.5 to 1.0 where a value of $A(P) = 0.5$ corresponds to the case of no discrimination while $A(P) = 1.0$ represents a perfect classifier.

The advantage of $A(P)$ is that it is objective, that is, it does not depend on the relative importance attached to POD and POFA values. One disadvantage of $A(P)$ is that it underestimates the area beneath a complete ROC, especially when the points are not well spread across the ROC space. Further more, it depends a lot on the uninteresting part of the ROC curve, that is, where both POD and POFA tend to 1.0.

### Distance from (0.5, 0.5) to the Point of Intersection of the ROC Curve with the Minor Diagonal

This is designated $d^c$ and measures the distance from the center (0.5, 0.5) to the point of intersection of the ROC curve and the minor diagonal, POD + POFA = 1. This may also be interpreted as that point where probability of detecting positives (POD) equals the probability of correctly classifying negatives (1 - POFA), the complement of probability of false alarm. This is illustrated in Figure 5.

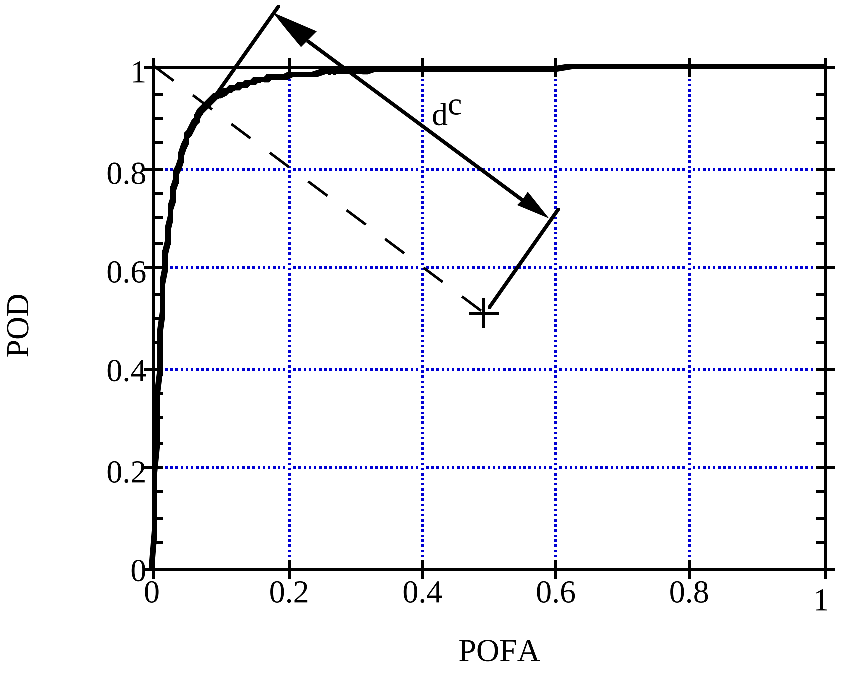

Figure 5 Using Distance from (0.5,0.5) along the Minor Diagonal to the ROC Curve to Measure Efficacy

Values of $d^c$ vary from 0.0 to 0.707 ($1/\sqrt{2}$) with $d^c$ = 0.0 corresponding to the case of no discrimination and $d^c$ = 0.707 a perfect classifier.

Although commonly used, $d^c$ does not strictly meet the objectivity criteria because it assumes a value system that attaches equal utility to correct classification of positives and negatives. It is thus inappropriate for applications in which a higher cost may be attached to misclassification of negatives, especially when the negatives-to-positives ratio is high. By using a different slope, however, it is possible to take into account different criteria.

## Transformation of Neural Network Output To ROC

The derivation of the ROC curve for a given classifier depends on the nature of the classifier, the test data and the test method. For a neural network, defining the output as 1.0 for the positive class and 0.0 for the negative class simplifies the problem. It also allows the use of the model described earlier.

Points on the ROC curve are obtained by counting the true-positives and false-positives at different levels of the alarm threshold. The number of true positives is the number of positives with output values exceeding the alarm threshold. Similarly, the number of false-positives equals the number of negatives whose output exceeds the alarm threshold. By dividing the number of true-positives and false-positives by the corresponding number of positives and negatives, respectively, POD and POFA values are obtained. A plot of such corresponding values of POD and POFA provides the ROC curve of the neural network.

It is instructive to note that an alarm threshold of 0.0 leads to all positives classified as positives (POD = 1.0) and all negatives misclassified as positives (POFA = 1.0). Conversely, an alarm threshold at the maximum value of 1.0 leads to all positives misclassified as negatives (POD = 0.0) and all negatives correctly classified as negatives (POFA = 0.0). Different values of alarm threshold, from 0.0 to 1.0, give different POD and POFA values in a monotonic way, thus generating the ROC curve for the neural network.

Different neural networks will in turn generate different ROC curves which can easily be compared on the same platform.

## APPLICATION TO AIRCRAFT WHEEL NDE

ROC evaluation has been applied to two methods for interpretation of data collected from an NDE machine for inspecting aircraft wheels. The first uses a threshold applied on the signal amplitude, as is done now in practice. The second uses a neural network to analyze the data after preprocessing and then applying an alarm threshold on the neural network output.

The ROC curves are generated using corrosion signal segments, as positives, and noise signal segments, as negatives. A total of 1010 signal segments were used for the test; 263 were corrosion signal segments and 747 were noise signal segments. The data is part of a database of NDE signal segments from a real-world inspection environment [8].

Figure 6 shows a comparison of the thresholder and neural network 56-23-1. The neural network performs much better than a thresholder; especially at low POFA levels.

Using a single measure, the ROC curves can be compared as shown in Table 1.

## CONCLUSION

The wide range of methods for NDE, from the mundane to the esoteric, and the process itself do not lend themselves to easy performance measurement. The major problem seems to be documentation of the signal indications and the decision taken. This paper has taken advantage of a database of such indications to present a procedure for quantitatively measuring such performance. Various advantages are foreseen such as objective equipment comparison, monitoring and standardization.

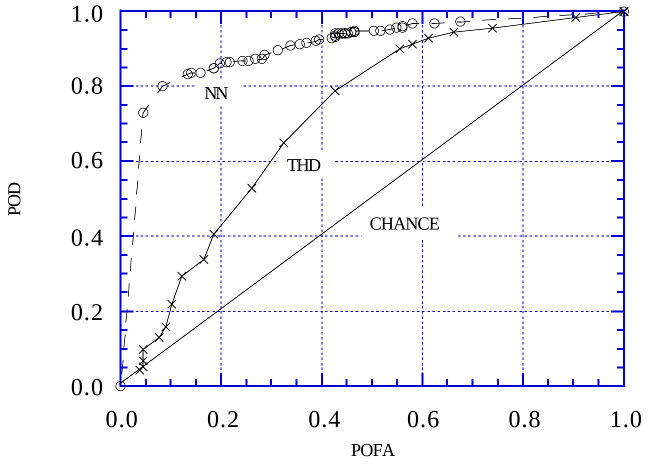

Figure 6 ROC of neural network (NN) 56-23-1, thresholder (THD), and chance signature classifiers. The neural network achieves a significant improvement on performance. At low POFA for example the THD loses classification ability while the NN still performs well.

Table 1 Comparison of various ROC Curves

| ROC Curve | $A(P)$ | $d^c$ |
|---|---|---|
| CHANCE | 0.5 | 0 |
| THD | 0.717 | 0.24 |
| NN | 0.904 | 0.483 |
| HYPOTHETICAL (with exponents $\alpha = 3.5$ and $\beta = 6.9$) | 0.971 | 0.595 |
| IDEAL | 1.00 | 0.707 |

ACKNOWLEDGMENTS

This work is supported by the FAA Center for Aviation Systems Reliability.

REFERENCES

1. Van Trees, H., *Detection, Estimation, and Modulation Theory* Part I ( J. Wiley and Sons, New York, 1971).
2. Swets, J. A., and R. M., Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, (Academic Press, New York, 1982).
3. Egan, J. P., *Signal Detection Theory and ROC Analysis,* (Academic Press, New York, 1975).
4. Swets, J. A., *Materials Evaluation* 41 (May 1983), p. 1294-1293.
5. Sturges, D. J., in Paper Summaries of the *ASNT Spring Conference and Third Annual Symposium*, March 21-25, 1994 New Orleans, LA, (The American Society for Nondestructive Testing, Columbus 1994), p. 229-231.
6. Nockemann, C., G. R. Tillack and H. Wessel, in Paper Summaries of the *ASNT Spring Conference and Third Annual Symposium*, March 21-25, 1994 New Orleans, LA, (The American Society for Nondestructive Testing, Columbus 1994), p. 52-56.
7. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, (McGraw-Hill, New York, 1984). .
8. Okure, M. A. E. and M. A. Peshkin., in Paper Summaries of the *ASNT Spring Conference and Third Annual Symposium*, March 21-25, 1994 New Orleans, LA, (The American Society for Nondestructive Testing, Columbus 1994), p. 232-234.