A Three-Dimensional Receiver Operator Characteristic Surface Diagnostic Metric

Donald L. Simon

NASA Glenn Research Center, Cleveland, OH, 44135, USA
Donald.L.Simon@nasa.gov

ABSTRACT

Receiver Operator Characteristic (ROC) curves are commonly applied as metrics for quantifying the performance of binary fault detection systems. An ROC curve provides a visual representation of a detection system's True Positive Rate versus False Positive Rate sensitivity as the detection threshold is varied. The area under the curve provides a measure of fault detection performance independent of the applied detection threshold. While the standard ROC curve is well suited for quantifying binary fault detection performance, it is not suitable for quantifying the classification performance of multi-fault classification problems. Furthermore, it does not provide a measure of diagnostic latency. To address these shortcomings, a novel three-dimensional receiver operator characteristic (3D ROC) surface metric has been developed. This is done by generating and applying two separate curves: the standard ROC curve reflecting fault detection performance, and a second curve reflecting fault classification performance. A third dimension, diagnostic latency, is added giving rise to three-dimensional ROC surfaces. Applying numerical integration techniques, the volumes under and between the surfaces are calculated to produce metrics of the diagnostic system's detection and classification performance. This paper will describe the 3D ROC surface metric in detail, and present an example of its application for quantifying the performance of aircraft engine gas path diagnostic methods. Metric limitations and potential enhancements are also discussed.[*]

1 INTRODUCTION

Diagnostic system designers rely on metrics to assess and compare the quality of candidate diagnostic methods. One such metric is the Receiver Operator Characteristic (ROC) curve. An ROC curve provides a technique for visualizing and evaluating the performance of binary classification systems (Fawcett, 2005; Hanley and McNeil, 1982; Metz, 1978). Historically, ROC curves have been applied in the fields of communication signal detection theory, medical diagnostics, and machine learning. Recently, they have grown in popularity as a metric for machinery diagnostics (Davison and Bird, 2008; SAE, 2008; Vachtsevanos *et al*., 2006).

As a point of introduction to ROC curves, first consider the multi-fault class diagnostic process illustrated in Figure 1. Shown is a system that can operate in either a nominal state, or in a faulty state where it has encountered one of N potential fault types. Also shown is a diagnostic method applied to produce a diagnostic inference of the current system state based on acquired sensed measurements. The diagnostic method consists of the three-step process of: 1) data conditioning—processes the acquired system sensor measurements to produce signal(s) used for fault detection and classification purposes; 2) fault detection—monitors produced signal(s) for threshold exceedance, which classifies the system as being in either a nominal or faulty state; and 3) fault classification—invoked upon fault detection to classify the system as being of one of the N possible fault states or fault types. This third step is commonly referred to as fault isolation.

A challenge in developing reliable machinery diagnostic methods is that the process is not deterministic. System measurement noise, variations in ambient conditions, operating load, deterioration, and nonlinear dynamics are all factors that contribute random variation to the process. This can lead to incorrect diagnostic inferences. For example, consider

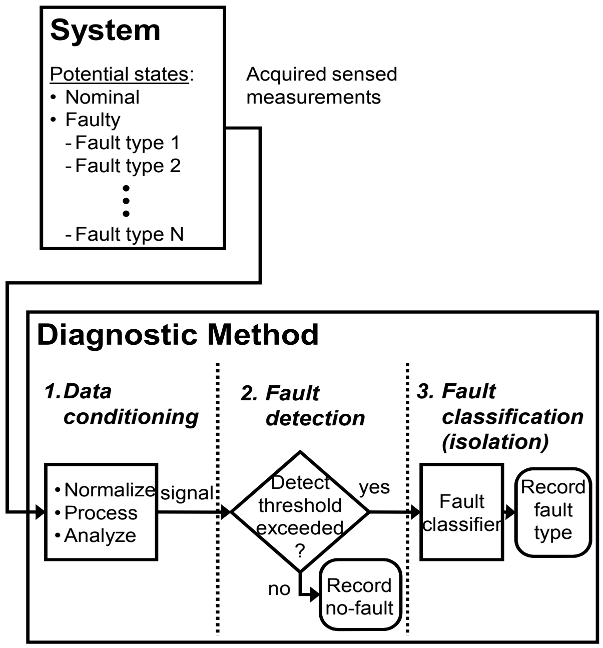

Figure 1: Multi-fault diagnostic process

the fault detection logic, shown in step 2 of Figure 1, that is tasked with performing the binary classification problem of declaring a system as either "nominal" or "faulty" based on a provided signal. The distributions in this signal under nominal and faulty operation are shown in Figure 2. Also shown is the placement of the threshold applied for detecting nominal versus faulty operation. Due to the overlap in the two distributions it will not be possible to attain 100% fault detection decision accuracy. In fact, there are four possible detection decision process outcomes including a true positive, a false positive, a false negative, or a true negative (see Figure 3). The probability of each outcome is defined as follows:

- True Positive Rate (TPR): proportion of faulty cases that trigger a threshold exceedance.
- False Positive Rate (FPR): proportion of nominal cases that trigger a threshold exceedance.
- False Negative Rate (FNR): proportion of faulty cases that do not trigger a threshold exceedance.
- True Negative Rate (TNR): proportion of nominal cases that do not trigger a threshold exceedance.

The above probabilities will depend on two factors: the separation between the nominal and faulty distributions of the detection signal, and the applied detection threshold. Reducing the detection threshold will have the desired effect of increasing true positives, but this will come at the expense of increased false positives. This can pose a dilemma in attempting to compare the merits of candidate detection strategies.

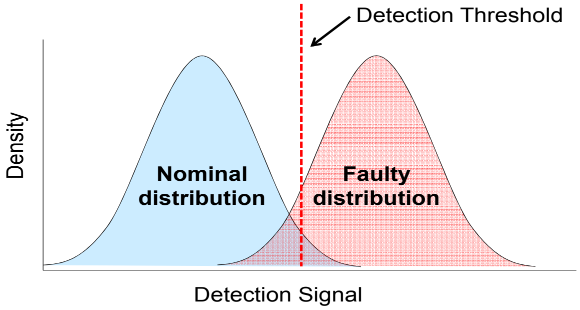

Figure 2: Distribution in fault detection signal under nominal and faulty conditions

		Actual Fault State	
		"Faulty"	"Nominal"
Classified Fault State	Positive (threshold exceeded)	True Positive	False Positive
	Negative (threshold not exceeded)	False Negative	True Negative

Figure 3: Detection decision matrix

The ROC curve is an effective tool for quantifying and comparing the TPR vs. FPR tradeoff of binary classification methods because it is independent of the applied detection threshold. An example ROC curve is shown in Figure 4. This curve is generated by plotting a detection method's TPR vs. FPR, and illustrates the interrelationship between the two parameters as the applied detection threshold is varied over the full range of possible settings. Applying a detection threshold of ∞ produces TPR and FPR values of 0. Conversely, applying a detection threshold of 0 will produce TPR and FPR values of 1.0. The area under the curve (AUC) forms a metric of detection performance, and ranges from 0.50 to 1.0. An AUC of 0.50 would be produced by applying a random guess, while an AUC of 1.0 reflects perfect detection performance. It is important to

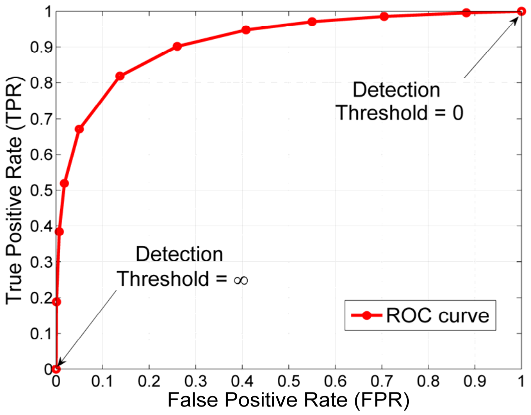

Figure 4: An example ROC curve

emphasize that an ROC curve, and its associated AUC, is not only dependent on the applied detection approach, but also the fault severity distribution. Larger faults, which are easier to detect, will generally result in higher AUC values.

While the standard two-dimensional ROC curve is a suitable metric for binary classification problems, it only partially captures the salient characteristics of the multi-fault diagnostic problem presented in Figure 1. It will provide an indication of how well fault detection is performed, but it does not provide an indication of how well fault classification is performed given a multi-fault classification problem. Furthermore, the ROC does not reflect diagnostic latency, the time required for the diagnostic system to produce a correct diagnostic inference. To address these shortcomings a new three-dimensional ROC (3D ROC) surface metric has been developed. It includes a second curve of Correct Classification Rate (CCR) versus FPR to quantify multi-fault classification performance. Additionally, a third dimension, diagnostic latency, can be added to reflect a measure of time within the diagnostic assessments.

The remaining sections of this paper are organized as follows. First, the 3D ROC surface diagnostic metric is described, and a step-by-step approach for generating and applying the metric is discussed. Next, results from the application of the 3D ROC surface metric for quantifying the diagnostic performance of several aircraft engine gas path diagnostic methods are presented. This is followed by a discussion of practical considerations for applying the new metric, including potential enhancements. Finally, conclusions are given.

2 THREE-DIMENSIONAL RECEIVER OPERATOR CHARACTERISTIC (3D ROC) SURFACE METRIC

The 3D ROC surface metric provides two enhancements to the standard ROC curve—quantification of multi-fault classification performance and quantification of diagnostic latency. These enhancements are discussed in the following sections along with a description of the steps for applying the metric.

2.1 Quantifying Correct Classification Performance

Fault classification performance is captured by generating a second curve reflecting a diagnostic method's correct classification rate (CCR) of a given fault type versus its FPR over the range of possible detection thresholds. Figure 5 shows an example of this CCR versus FPR curve, ROC$_{CCR}$, along with the original TPR versus FPR ROC curve, hereafter in this paper referred to as ROC$_{TPR}$.

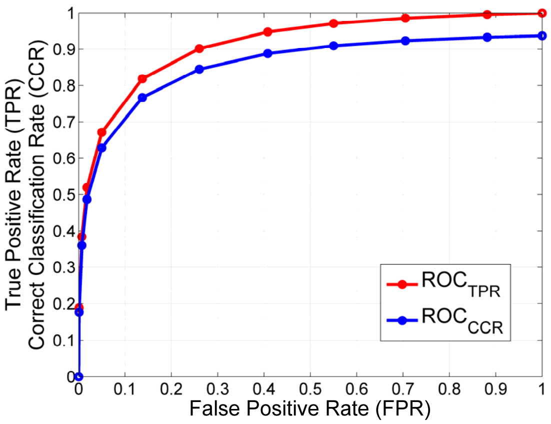

Figure 5: An example of the standard ROC curve reflecting true positives (ROC$_{TPR}$) and the ROC curve reflecting correct classifications (ROC$_{CCR}$)

Since fault detection is a prerequisite for fault classification, the ROC$_{CCR}$ curve will always reside at or below the ROC$_{TPR}$ curve. The ROC$_{CCR}$ curve, like the ROC$_{TPR}$ curve, will initiate at the origin corresponding to a detection threshold of ∞, and will monotonically increase as the detection threshold is reduced. However, unlike the ROC$_{TPR}$ curve, there is no guarantee that the ROC$_{CCR}$ curve will reach a final value of 1.0 once the detection threshold is reduced to zero. This is due to the fact that even if a fault is detected with 100% accuracy there is no guarantee that it will be classified with 100% accuracy.

The areas under the ROC$_{TPR}$ and ROC$_{CCR}$ curves, referred to as AUC$_{TPR}$ and AUC$_{CCR}$, provide metrics of the diagnostic method's true positive detection performance and correct classification performance, respectively. Once AUC$_{TPR}$ and AUC$_{CCR}$ are obtained, additional metrics reflective of the system's misclassification rate can be calculated. The area between the curves (ABC) reflects the diagnostic method's probability of misclassification given that a fault has occurred. This metric is calculated as AUC$_{TPR}$ minus AUC$_{CCR}$. A normalized measure of system misclassification performance can be produced by dividing ABC by AUC$_{TPR}$. This metric, referred to as ABC$_{NORM}$ reflects the system's probability of misclassification given that a true positive detection has occurred. ABC$_{NORM}$ is perhaps preferred over ABC because it reflects the rate of misclassification given the opportunity for a misclassification.

In general, better diagnostic performance is indicated by maximizing AUC$_{TPR}$ and AUC$_{CCR}$, while minimizing ABC$_{NORM}$. Given an N-fault classification problem, the AUC$_{CCR}$ can in general be expected to range from AUC$_{TPR}$/N (i.e., random classification) to AUC$_{TPR}$ (i.e., perfect classification). However, a lower bound of AUC$_{TPR}$/N is by no means guaranteed, especially if the classifier is designed to place more emphasis on certain fault types, or takes the rate of occurrence of fault types into consideration when making an inference.

Given a multi-fault classification problem, individual $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves can be generated for each fault type. Alternatively, single $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves reflective of the system's overall fault detection and classification performance across all fault types can be generated. However, in generating $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves reflective of all fault types, users are cautioned of the need to consider the relative frequency of occurrence of the faults. The $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves for a single fault type are invariant to the fault's frequency of occurrence relative to other fault types—they reflect the probability of true positive detection and correct classification given that a fault of the specified type is present, versus the probability of a false positive given that no fault is present. Conversely, multi-fault $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves are not invariant to changes in the relative frequency of occurrence of the different fault types. Changes in component design or in the usage profile of a machine can make certain fault types more/less likely. The more frequent occurrence of easily diagnosable faults will result in higher metric values, while the more frequent occurrence of difficult to diagnose faults will result in lower metric values. The article (Webb and Ting, 2005), and the corresponding response (Fawcett and Flach, 2005) provide an excellent discussion of causal dependence and the impact of varying class distributions, or frequency of occurrence, on ROC analysis.

2.2 Quantifying Diagnostic Latency

In developing machinery diagnostic systems, designers must deal with random variations in the measurement process, which can make discriminating between nominal and anomalous conditions challenging. It is common to apply some form of data filtering or detection threshold persistency checks to help reduce the occurrence of false alarms. While beneficial for false alarm reduction, such filtering or persistency checks can introduce undesirable delay, or latency, into the diagnostic process.

To quantify a diagnostic method's latency, a third dimension reflecting this latency, labeled as T_L, is added to the previously described $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves. This gives rise to three-dimensional ROC surfaces as shown in Figure 6. Here, true positive detection performance is reflected by the top red surface, $\mathrm{ROCSURF}_{\mathrm{TPR}}$, and correct classification performance is reflected by the blue surface below, $\mathrm{ROCSURF}_{\mathrm{CCR}}$. Unlike the TPR and FPR axes that range from 0.0 to 1.0, the T_L axis will range from 0.0 (time of first available sample after fault occurrence) to some user-specified upper bound. This upper bound can be treated as the maximum acceptable diagnostic latency for the system. In other words, diagnostic inferences made beyond this point in time are treated as having no value. For some applications the acceptable diagnostic latency may be on the order of milliseconds, whereas for others it may be on the order of months. Regardless of the specified value, the coordinates on the T_L axis are normalized by dividing them by the user-specified upper bound. This produces a T_L axis ranging from 0.0 to 1.0, and ensures that the total volume contained within the three-dimensional ROC space is 1.0. In addition to normalizing the T_L axis, it is often desirable to apply a scaling to the axis to place more emphasis on early diagnosis and less emphasis on more latent diagnosis.

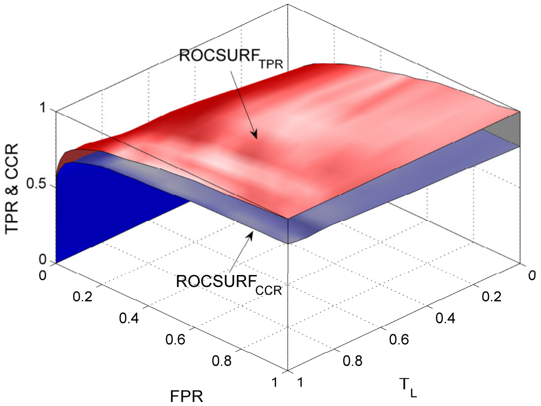

Figure 6: Example 3D ROC surfaces

Diagnostic performance metrics can be produced by calculating the volumes under, and between the two surfaces. The volume under $\mathrm{ROCSURF}_{\mathrm{TPR}}$, $\mathrm{VUS}_{\mathrm{TPR}}$, reflects true positive performance, while the volume under $\mathrm{ROCSURF}_{\mathrm{CCR}}$, $\mathrm{VUS}_{\mathrm{CCR}}$, reflects correct classification performance. The volume between the two surfaces, VBS, reflects the misclassification rate given that a fault has occurred, and is calculated as $\mathrm{VUS}_{\mathrm{TPR}}$ minus $\mathrm{VUS}_{\mathrm{CCR}}$. A normalized measure of misclassification, $\mathrm{VBS}_{\mathrm{NORM}}$, can be calculated by dividing VBS by $\mathrm{VUS}_{\mathrm{TPR}}$.

As with the previously described $\mathrm{ROC}_{\mathrm{TPR}}$ and $\mathrm{ROC}_{\mathrm{CCR}}$ curves, $\mathrm{ROCSURF}_{\mathrm{TPR}}$ and $\mathrm{ROCSURF}_{\mathrm{CCR}}$ surfaces can be generated for individual fault classes, or combined to produce 3D ROC surfaces reflective of average fault detection and classification performance. Obviously, the same considerations regarding a fault's relative frequency of occurrence hold for the three-dimensional surfaces as were previously discussed for the two-dimensional ROC curve metric.

As a point of emphasis it is noted to readers that the inclusion of the third dimension is only warranted for diagnostic methods where diagnostic latency is an important design consideration. If diagnostic latency is not of concern, the two-dimensional ROC_{TPR} and ROC_{CCR} curves are sufficient.

2.3 Step-by-Step Approach for Generating 3D ROC Surface Diagnostic Metrics

The following is a more detailed description of the steps applied for generating the 3D ROC surface diagnostic metrics:

1. Obtain system data: Gain access to a suitable database of system data under both nominal and all faulty operating scenarios. This should contain stochastic variations such as sensor measurement noise and variations in operating conditions, performance deterioration, and load conditions. The faulty scenario data should be selected to represent the actual fault severity level distribution so that the ROC surface is not incorrectly biased towards a particular fault size. This data should be of suitable quantity to generate ROC surfaces of the desired fidelity. Richer data sets will allow higher fidelity surfaces to be generated, while sparser data sets will generate coarser surfaces with less precise metric results.

2. Specify FPR and latency axis coordinates: The user specifies *m* coordinates along the FPR axis ranging from 0.0 to 1.0, and *n* coordinates along the T_L axis spanning the defined latency range. This forms an $m \times n$ grid covering the 3D ROC surface. Uniform spacing of the coordinates is not required, and finer coordinate spacing often proves to be beneficial in regions where the detection and classification surfaces tend to undergo the most rapid rate of change, e.g., low FPR and latency levels. The only limitations are that attainable FPR axis coordinate spacing is dependent on the number of nominal (no fault) scenarios available, and latency axis coordinate spacing must be at, or a multiple of, the diagnostic inference update rate of the system.

3. Determine the detection thresholds corresponding to the specified FPR and latency axis coordinates: The nominal system data collected in Step 1 can next be processed (using just the detection logic portion of the diagnostic method) to determine the detection threshold required to produce the specified FPR and latency at each of the $m \times n$ grid points defined in step 2. If the detection logic is a function of previous detection assessment(s), as is often the case in diagnostic methods, different thresholds will be required to maintain a given FPR as the T_L axis is traversed.

4. Evaluate fault detection and fault classification performance: The next step is to evaluate the diagnostic method's fault detection and fault classification performance at each of the $m \times n$ grid points defined in step 2 when applying the corresponding thresholds determined in step 3. This is accomplished by applying the diagnostic method to the faulty operational data obtained in Step 1. Average TPR and CCR values are determined at each grid point. The TPR and CCR information, along with FPR and T_L coordinates, form three-dimensional $\mathrm{ROCSURF}_{TPR}$ and $\mathrm{ROCSURF}_{CCR}$ surfaces similar to the example previously shown in Figure 6.

5. Calculate the volumes under and between the 3D ROC surfaces: Once the $\mathrm{ROCSURF}_{TPR}$ and $\mathrm{ROCSURF}_{CCR}$ surfaces have been generated, the corresponding volumes under each surface, VUS_{TPR} and VUS_{CCR}, can be calculated. This can be accomplished by partitioning the VUS_{TPR} and VUS_{CCR} volumes into polyhedrons as illustrated in Figure 7 and Figure 8, respectively, and then applying a Riemann sum numerical integration technique to calculate and sum the individual polyhedron volumes to produce VUS_{TPR} and VUS_{CCR}. In the example given in Figures 7 and 8 a logarithmic scaling has been applied to the T_L axis. This places an emphasis on the importance of early diagnosis, with a decaying emphasis over time, and results in the observed non-uniform spacing of grid points along the latency axis.

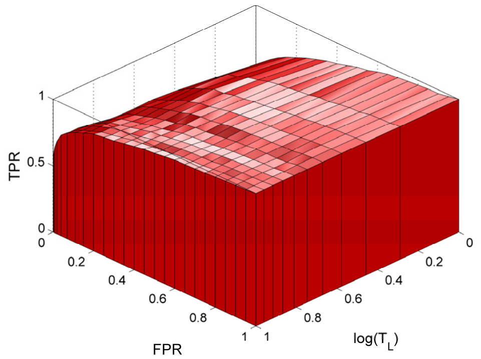

Figure 7: An example of VUS_{TPR} partitioning

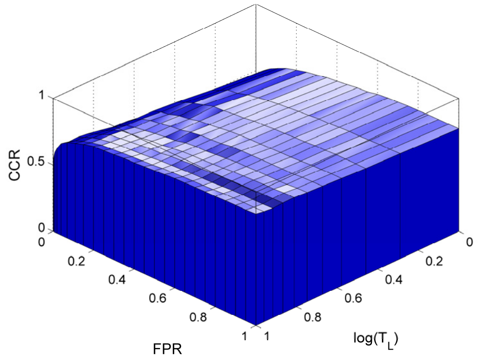

Figure 8: An example of VUS_{CCR} partitioning

Once the VUS_{TPR} and VUS_{CCR} volumes are obtained the previously mentioned VBS and VBS_{NORM} metrics reflective of misclassification performance can be calculated. VBS, which provides an indication of the diagnostic method's misclassification rate given that a fault has occurred, is calculated as

$$\mathrm{VBS} = \mathrm{VUS}_{\mathrm{TPR}} - \mathrm{VUS}_{\mathrm{CCR}} \tag{1}$$

VBS_{NORM}, which provides a measure of the misclassification rate given that a fault has been positively detected, is calculated as

$$\mathrm{VBS}_{\mathrm{NORM}} = \frac{\left(\mathrm{VUS}_{\mathrm{TPR}} - \mathrm{VUS}_{\mathrm{CCR}}\right)}{\mathrm{VUS}_{\mathrm{TPR}}} \tag{2}$$

3 EXAMPLE: APPLICATION OF THE 3D ROC SURFACE METRICS FOR QUANTIFYING AIRCRAFT ENGINE DIAGNOSTIC PERFORMANCE

An example application of the 3D ROC surface diagnostic metrics is given in the form of a simulated aircraft turbofan engine gas path diagnostic problem. The following subsections describe the diagnostic problem, the diagnostic methods applied to the problem, and each method's corresponding 3D ROC surface metrics results.

3.1 Description of the Aircraft Engine Gas Path Diagnostic Problem

Aircraft operators rely on gas path diagnostic methods to assist them in managing the health of their gas turbine engine assets. It is conducted by monitoring sensed measurements collected from the engine flow path, and utilizing this information to detect and classify engine faults that can impact engine flow path performance (Li, 2002; Volponi and Wood 2005; Von Karman Institute, 2003). Gas path diagnostics presents a classic multi-fault detection and classification problem, and thus is ideal for illustrating the application of the 3D ROC surface metrics.

In this study, a simulated gas path diagnostic problem is constructed using the NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS), a generic turbofan engine simulation (Frederick, DeCastro, and Litt, 2007). The problem is setup to emulate the collection of time-averaged engine sensor measurement snapshots from each engine, each flight, while the aircraft is operating at a cruise operating point. Excluding sensors used for parameter correction and power setting reference purposes, the snapshot measurement vector consists of the four sensors shown in Table 1. The diagnostic objective is to accurately detect and classify the occurrence of any engine gas path faults with minimal diagnostic latency. In this study it is assumed that an engine is either operating nominally (fault-free), or has experienced one of three possible turbomachinery module faults. The three faults considered are fan, high pressure compressor (HPC), and low pressure turbine (LPT) faults. These faults are simulated in C-MAPSS by adjusting the efficiency and flow capacity health parameters of the respective faulty module. Fan and HPC faults are simulated by simultaneously reducing their efficiency and flow capacity health parameters, while LPT faults are simulated by reducing LPT efficiency and increasing LPT flow capacity. Each fault, along with their uniformly distributed fault magnitudes, or health parameter adjustments, are shown in Table 2. The location of the faulty modules and sensor locations within the C-MAPSS engine are shown in Figure 9.

Table 1: Example gas path diagnostic problem—engine sensor measurements

Sensor	Description
Nc	Core Speed
Ps30	HPC exit static pressure
T48	Inter-turbine total temperature
Wf	Fuel flow

Table 2: Example gas path diagnostic problem—fault types and magnitudes

Fault type	Description	Magnitude (uniformly distributed)
Fan	Fan fault	0.5% to 3.0%
HPC	High pressure compressor fault	0.5% to 3.0%
LPT	Low pressure turbine fault	0.5% to 3.0%

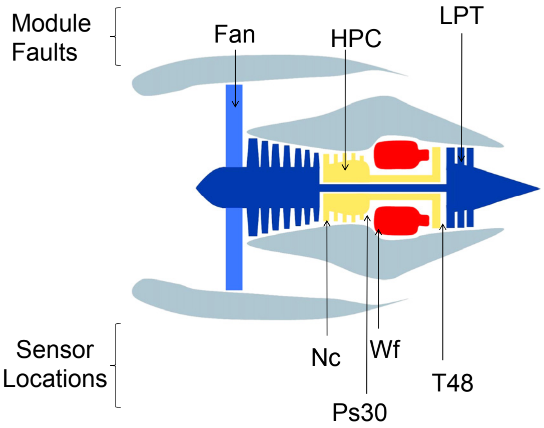

Figure 9: C-MAPSS module faults and sensor locations

Additional relevant characteristics of the fault scenarios considered are as follows:

- Only single fault scenarios are included. Multiple faults occurring in combination is not considered.
- All fault types have equal probability of occurrence.
- All faults are simulated as abrupt step changes that occur between collected snapshot measurements. Engine fault transient dynamics are not captured.
- Once a fault occurs, it will persist at a constant magnitude. Intermittent faults or faults that initiate and then undergo a change in magnitude are not considered.
- The defined problem includes random variations in sensor noise, and the flight-to-flight cruise operating point.

3.2 Description of the Evaluated Gas Path Diagnostic Methods

Each of the evaluated gas path diagnostic methods is comprised of the three-step process of data conditioning, fault detection, and fault classification (see Figure 1). These steps are implemented as follows:

Data conditioning: The incoming snapshot measurement vector is normalized to account for variations in operating condition, and referenced against a model reflective of nominal C-MAPSS engine performance to produce a vector of snapshot measurement residuals, y. Next, an exponential moving average (EMA) is applied to each of the four measurements contained in y. This places more emphasis on the most recent data, while older data is exponentially forgotten over time. The EMA of the residual in sensor a at time sample k is given as

$$\bar{y}_a(k) = \alpha \cdot \bar{y}_a(k-1) + (1-\alpha) \cdot y_a(k) \tag{3}$$

where the weighting between the EMA on the previous time step, $\bar{y}_a(k-1)$, and the current residual sample, $y_a(k)$, is established by the constant α (where $0 \leq \alpha < 1$). The EMA α values applied by the diagnostic methods evaluated in this study are:

- EMA α = 0.000 (no averaging)
- EMA α = 0.667
- EMA α = 0.905

Fault detection: The signal monitored for fault detection purposes is the Mahalanobis distance (Hall and Llinas, 2001) of the vector, $\bar{y}(k)$, relative to the origin. This distance is calculated as

$$D_M(k) = \sqrt{\bar{y}(k)^T R^{-1} \bar{y}(k)} \tag{4}$$

where R is the measurement noise covariance matrix. If $D_M(k)$ exceeds the applied detection threshold, a fault is declared. All of the diagnostic methods evaluated in this study apply the Mahalanobis distance detection approach.

Fault classification: Two different fault classification approaches were evaluated: a weighted least squares (WLS) approach and a probabilistic neural network (PNN) approach. The WLS classification approach is based on an analytically derived linear approximation of the influence of each fault type on the observed measurement vector (Gelb, 1974). Upon fault detection, the WLS classifier evaluates how well each candidate fault type matches the observed $\bar{y}(k)$ measurements in a weighted least squares sense. Weighting is applied to factor in covariance amongst the sensed measurements. The fault type that produces the closest match to $\bar{y}(k)$ is then classified as the fault type. The PNN is a radial basis neural network classifier, designed using the newpnn function of the Matlab (The MathWorks, Inc.) neural network toolbox (Demuth and Beale, 2001). This classifier is empirically trained using C-MAPSS produced fault data sets, which are different than the data sets used for testing. The PNN is designed to produce a classification of the most likely fault class, provided the $\bar{y}(k)$ vector as an input.

The logic associated with each diagnostic method is designed such that once a fault is diagnosed, be it either correctly or incorrectly, that same fault diagnosis will persist into the future. In other words, a diagnostic method is not permitted to change its diagnostic assessment from one sample to the next once a positive diagnosis has occurred. All combinations of the previously described data conditioning approaches (α = 0.000, α = 0.667, α = 0.905) and fault classification approaches (LS, PNN) were evaluated resulting in a total of six different diagnostic methods.

3.3 3D ROC Surface Diagnostic Metric Results

3D ROC surface metric results were generated for each of the six evaluated diagnostic methods following the steps previously listed in Section 2.3. The test data generated by C-MAPSS to conduct this evaluation consisted of:

- 5,000 nominal engines, each 20 snapshot measurement samples in duration.
- 600 faulty engines (200 engines of each fault type), each 20 snapshot measurement samples in duration. Each of the faulty engines experienced a fault appearing on sample 10. During evaluation, the first 9 samples were used to establish residual moving averages, and samples 10 through 20 were used for evaluating fault diagnostic performance.

The specified FPR and T_L (latency) axes coordinates consisted of:

- 26 FPR coordinates ranging from 0.0 to 1.0
- 11 T_L coordinates ranging from 0 to 10. This corresponded to samples 10 through 20 of the faulty engine data sets. The T_L axis coordinates were scaled (applying a base 2 logarithmic scaling), and normalized to range from 0 to 1.

For each diagnostic method the 3D ROC surface metrics of VUS_{TPR}, VUS_{CCR}, and VBS_{NORM} were generated applying a Riemann sum numerical integration technique. This was done considering the detection and classification of each fault type individually, as well as overall performance across all fault types. The results are shown in Table 3 for the six diagnostic methods considered.

These results reveal several findings. Based on the VUS_{TPR} results it can be seen that fan faults are the most difficult to detect, followed by HPC faults and then LPT faults. The results also indicate that applying an EMA α of 0.667 provides superior detection results compared to EMA α values of 0.000 or 0.905. This holds for all three fault types (fan, HPC, and LPT). The relative performance of the two fault classification approaches is mixed. The VUS_{CCR} and VBS_{NORM} results indicate that the PNN classifier provides superior classification of fan faults, LPT faults, and all faults collectively. However, the WLS classifier provides better classification of HPC faults. The best overall performance is obtained when applying an EMA α of 0.667 coupled with the PNN classification approach (shown in bold font in Table 3).

Table 3: 3D ROC surface metric results

EMA α	**Fault type**	**VUS_{TPR}**	**VUS_{CCR}**		**VBS_{NORM}**	
			WLS	**PNN**	**WLS**	**PNN**
α = 0.000	*FAN*	*0.787*	*0.562*	*0.686*	*0.286*	*0.128*
	HPC	*0.944*	*0.879*	*0.855*	*0.069*	*0.094*
	LPT	*0.946*	*0.789*	*0.909*	*0.166*	*0.039*
	ALL	0.892	0.743	0.817	0.167	0.084
α = 0.667	*FAN*	*0.820*	*0.652*	*0.760*	*0.205*	*0.074*
	HPC	*0.943*	*0.901*	*0.892*	*0.044*	*0.054*
	LPT	*0.955*	*0.851*	*0.914*	*0.109*	*0.044*
	ALL	**0.906**	0.801	**0.855**	0.115	**0.056**
α = 0.905	*FAN*	*0.785*	*0.597*	*0.694*	*0.241*	*0.117*
	HPC	*0.894*	*0.831*	*0.814*	*0.071*	*0.089*
	LPT	*0.928*	*0.779*	*0.855*	*0.160*	*0.079*
	ALL	0.869	0.736	0.788	0.154	0.094

4 DISCUSSION

The diagnostic latency dimension of the 3D ROC surface metric provides a means of measuring diagnostic performance over a range of latencies suitable for the given application. In contrast, one must assume a fixed diagnostic latency (i.e., a fixed number of samples available for producing diagnostic inferences) when applying the standard 2D ROC curve. Consequently, the 2D ROC curve does not reflect the percentage of correct inferences that could have been made based on fewer/more samples. Furthermore, it does not emphasize the importance of early diagnosis. To illustrate this refer to Figure 10 showing the variation in AUC_{TPR} and AUC_{CCR} versus latency for the three EMA α's coupled with PNN classification considered in the previous example. Here, the latency axis reflects the applied base 2 logarithmic scaling. These plots show the change in the area under the two-dimensional ROC_{TPR} and ROC_{CCR} curves as additional samples are considered. Based on one sample it would be concluded that applying an α of 0.000 (no EMA) is superior given the three α options. If 11 samples are considered it would be concluded that applying an α of 0.905 is superior. Over the range of samples shown, an α of 0.667 becomes the superior choice. This illustrates the benefit of including the third dimension within the 3D ROC surface metric. Instead of simply capturing diagnostic performance at a fixed latency as the 2D ROC curve does, the 3D ROC surface provides an assessment of diagnostic performance across a range of potential latencies. That being said, it is acknowledged that the evaluated latency range and scaling is somewhat subjective as it is user specified. Integrating over a shorter or longer length of the latency axis, or applying a different scaling, will produce different results. Therefore, users are reminded of the importance of specifying a latency range and scaling that is appropriate and relevant for their given application.

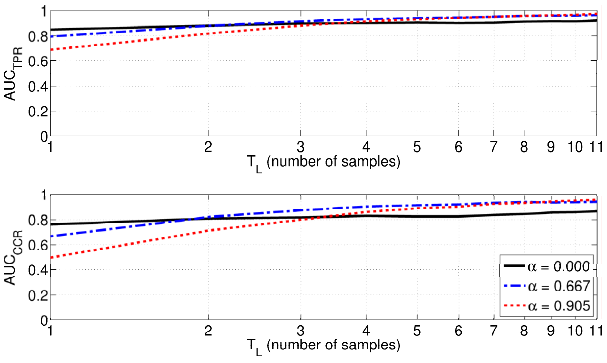

Figure 10: Variation in AUC$_{TPR}$ and AUC$_{CCR}$ versus latency

As an additional point of emphasis, readers are cautioned that the volume under the 3D ROC surface metric, similar to the area under the 2D ROC curve metric, provides a general measure of diagnostic system performance. It does so by condensing performance over the entire surface into single metrics of detection and classification performance. While in practice these metrics perform very well in evaluating candidate diagnostic methods, it is possible for a diagnostic method receiving superior metric results to perform worse than a lower-scoring diagnostic method at specific detection threshold settings. This is a particularly relevant concern in machinery diagnostic applications that place high emphasis on maintaining low false positive rates. Therefore, it is often advisable to conduct additional analysis and comparison of candidate diagnostic methods applying detection thresholds that produce a false positive rate deemed appropriate for the given application.

Several potential enhancements to the presented 3D ROC surface metrics are possible. The presented approach has assumed that once a fault is detected, it must be classified to a specific fault type. This is not always a requirement in real-world diagnostic applications. For example, there could be cases where a fault is detected and then classified to be one of several faults within an ambiguity group—a group of faults known to produce similar sensed measurement signatures. There may also be cases where the diagnostic system simply declares an anomaly—an indication that the system is exhibiting abnormal behavior indicative of a fault, although no classification of the specific fault type is given. Such scenarios could be captured by adding an additional 3D ROC surface reflective of ambiguity group, and/or anomaly classification. Such a surface would reside between the TPR surface and the CCR surface. Collectively the three surfaces would show the demarcation between detections, classification to an ambiguity group or anomaly level, and classification to the individual fault level.

The presented approach has also taken the rather simplistic view of assuming that individual fault types can only occur in isolation. Two or more faults occurring in combination has not been discussed. However, the presented metric could be readily extended to encompass such scenarios. It would however require expanding the number of fault classifications to include all possible combinations of the N different fault types, which adds complexity.

Finally, readers are reminded that the presented metrics only provides a measure of diagnostic performance. As such the ROC surface metrics should be coupled with other metrics that provide measures of cost and complexity to thoroughly consider all aspects of the diagnostic method decision.

5 CONCLUSION

A three-dimensional Receiver Operator Characteristic (3D ROC) surface metric designed for visualizing and quantifying the performance of multi-fault class diagnostic methods has been presented. It enhances the standard True Positive Rate (TPR) versus False Positive Rate (FPR) ROC curve by adding a second curve reflecting Correct Classification Rate (CCR) versus FPR. A third dimension, diagnostic latency, is added to construct two 3D ROC surfaces. The volumes under and between the two surfaces give rise to a unified set of metrics indicative of a diagnostic method's fault detection, classification, and misclassification performance. These metrics are independent of the applied fault detection threshold, and inherently reflect diagnostic latency. The metrics can be used to assess a method's ability to diagnose a single fault type, or used to assess average diagnostic capability over all fault types. However, in the latter case the metrics are susceptible to changes in fault type distributions. Results from the application of the metric to aircraft engine diagnostic methods have shown that it is an effective tool for evaluating diagnostic performance in multi-fault detection and classification problems.

ACKNOWLEDGMENT

This research was conducted under the NASA Aviation Safety Program, Integrated Vehicle Health Management Project.

NOMENCLATURE

3D ROC	three-dimensional ROC
ABC	area between curves
$\mathrm{ABC_{NORM}}$	normalized area between curves
AUC	area under curve
$\mathrm{AUC_{CCR}}$	area under $\mathrm{ROC_{CCR}}$ curve
$\mathrm{AUC_{TPR}}$	area under $\mathrm{ROC_{TPR}}$ curve
CCR	correct classification rate
C-MAPSS	Commercial Modular Aero-Propulsion System Simulation
$\mathrm{D_M}$	Mahalanobis distance
EMA	exponential moving average
FNR	false negative rate
FPR	false positive rate
HPC	high pressure compressor
HPT	high pressure turbine
i	index of FPR axis coordinates
j	index of $\mathrm{T_L}$ axis coordinates
k	time sample index
LPC	low pressure compressor
m	number of FPR axis coordinates
N	number of fault classes
n	number of $\mathrm{T_L}$ axis coordinates
PNN	probabilistic neural network
R	measurement covariance matrix
ROC	receiver operator characteristic
$\mathrm{ROC_{CCR}}$	CCR ROC curve
$\mathrm{ROCSURF_{CCR}}$	CCR ROC surface
$\mathrm{ROCSURF_{TPR}}$	TPR ROC surface
$\mathrm{ROC_{TPR}}$	TPR ROC curve
$\mathrm{T_L}$	diagnostic latency
TNR	true negative rate
TPR	true positive rate
VBS	volume between surfaces
$\mathrm{VBS_{NORM}}$	normalized volume between surfaces
$\mathrm{VUS_{CCR}}$	volume under $\mathrm{ROCSURF_{CCR}}$
$\mathrm{VUS_{TPR}}$	volume under $\mathrm{ROCSURF_{TPR}}$
WLS	weighted least squares
y	measurement residual vector
$\bar{y}$	average measurement residual vector
α	exponential moving average weighting

REFERENCES

(Davison and Bird, 2008) C. R. Davison and J. W. Bird. Review of Metrics for Gas Path Diagnostic Health Management Techniques for Gas Turbine Engines, *Proceedings of the ASME Turbo Expo 2008*, GT2008-50849, 2008.

(Demuth and Beale, 2001) H. Demuth and M. Beale. *Neural Network Toolbox for Use with Matlab*, 7th printing. Natick, Massachusetts: The MathWorks, Inc., 2001.

(Fawcett, 2006) T. Fawcett. An Introduction to ROC Analysis, *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

(Fawcett and Flach, 2005) T. Fawcett and P. A. Flach. A Response to Webb and Ting's On the Application ROC Analysis to Predict Classification Performance Under Varying Class Distributions, *Machine Learning*, vol. 58, pp. 33-38, 2006.

(Frederick, DeCastro, and Litt, 2007) D. K. Frederick, J. A. DeCastro, and J. S. Litt. User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS), *NASA Technical Memorandum* TM-2007-215026, 2007.

(Gelb, 1974) A. Gelb, *Applied Optimal Estimation*. Cambridge, Massachusetts: The MIT Press, 1974.

(Hall and Llinas, 2001) D.L. Hall and J. Llinas. *Handbook of Multisensor Data Fusion*, Boca Raton, Florida: CRC Press LLC, 2001.

(Hanley and McNeil, 1982) J. A. Hanley and B. J. McNeil. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, *Radiology*, vol. 143, pp. 29-36, 1982.

(Li, 2002) Y. G. Li. Performance-Analysis-Based Gas Turbine Diagnostics: A Review, *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, vol. 216, No. 5, pp. 363-377, 2002.

(Metz, 1978) C. E. Metz, Basic Principals of ROC Analysis, *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283-298, 1978.

(SAE, 2008) *Health and Usage Monitoring Metrics—Monitoring the Monitor*, SAE ARP 5783, 2008.

(Vachtsevanos *et al.,* 2006) G. Vachtsevanos, F. L. Lewis, M. Roemer, A. Hess, and B. Wu. *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*, 1st ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2006.

(Volponi and Wood 2005) A. Volponi and B. Wood. Engine Health Management for Aircraft Propulsion Systems, *The Forum on Integrated System Health Engineering and Management (ISHEM) in Aerospace*, November 7-10, Napa, CA, 2005.

(Von Kalman Institute, 2003) Gas Turbine Condition Monitoring and Fault Diagnostics, *Von Karman Institute for Fluid Dynamics Lecture Series* 2003-01, 2003.

(Webb and Ting, 2005) G. I. Webb and K. M. Ting. On the Application ROC Analysis to Predict Classification Performance Under Varying Class Distributions, *Machine Learning*, vol. 58, pp. 25-32, 2005.