Case Studies: Use of Big Data for Condition Monitoring

Raj Bharadwaj [1], Dinkar Mylaraswamy [1], Andrew Vechart [1], Marshall Smith [2],
Peter Figliozzi [3], Prof. Gautum Biswas [4], Daniel Mack [4]

[1] *Honeywell Aerospace Advanced Technology, 1985 Douglas Drive N., Golden Valley, MN 55442, USA*
[2] *Clockwork Solutions Inc., 805 Las Cimas Pkwy #100, Austin, TX 78746, USA*
[3] *Blitztrade, Austin, TX, USA*
[4] *Vanderbilt University, EECS Dept. Box 1824 Sta B, Nashville, TN 37325, USA*

Abstract

This paper describes two case-studies that exercised a big data framework, data mining, machine learning (ML) algorithms, and predictive analytics to improve the accuracy of aircraft engine health monitoring and HUMS-based condition indicator's (CI) ability to predict gearbox removals. In the first example machine learning algorithms operating on multiple data sources produce useful insights to increase our ability to predict engine fuel controller failures prior to the in-flight auto-shutdown. In the second example, we demonstrate the use of aggregated HUMS CI to predict the intermediate gearbox removal event. Using statistical hypothesis testing, we found three identifiable distributions of aggregate (Super) CI values: normal/good condition, anomalous with more than 100 flight hours remaining, and anomalous with less than 100 flight hours remaining. From these results, we conclude that the value of the condition indicator cannot provide high-resolution condition information; for example, the amount of remaining useful life down to flight-hour accuracy. However, it does provide potentially valuable low-resolution information, such as when an intermediate gearbox has less than 100 hours of remaining useful life.

Keywords: Predictive Data Analytics, Machine Learning, Remaining Useful Life

Introduction

The Vehicle Health Management (VHM) group within Honeywell Aerospace Advanced Technology continues to explore and evaluate data-driven methods and technologies that support aircraft health management and maintenance. In particular, the work strives to enhance existing and create new diagnostic and prognostic capabilities to improve aviation safety and enable condition-based maintenance. This paper includes a discussion of two case studies related to this work. The first case comprises support of development of a Vehicle Integrated Prognostic Reasoner (VIPR) which is a model-based vehicle level reasoning system (VLRS). Through a combination of unsupervised and supervised machine learning, system reference models were updated in a systematic way to improve the performance of the reasoning system. The second case included analysis of HUMS data from the Honeywell Intelligent Machinery Diagnostic System (iMDS) to explore how data mining might support improved diagnostic performance based on CIs. Techniques were used to aggregate CI information into a more effective diagnostic metric to recognize component issues. In addition, the prognostic capabilities of this enhanced metric were evaluated. The paper will conclude with a brief discussion tying these two case studies to the general big data framework.

Case I: Improving Vehicle Level Reasoning System Performance

Vehicle Integrated Prognostic Reasoner (VIPR) is an on-board reasoning system designed by Honeywell that provides fault detection and isolation capabilities at the aircraft level and estimates remaining useful life of components and subsystems of the aircraft [1]. VIPR is composed of four main functional modules as illustrated in Fig. 1.

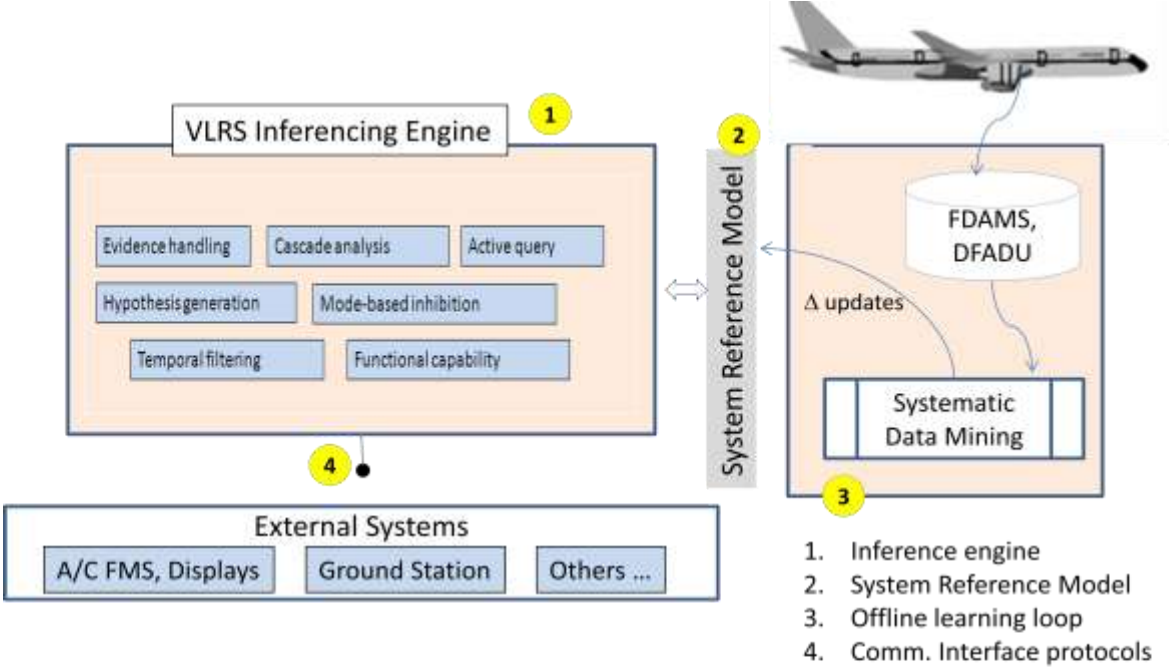

Fig. 1: Functional modules within VIPR

Briefly, the inference engine of VIPR uses health evidence generated by components and subsystems to produce a current diagnostic state or predict future fault evolution. The system reference model contains the relationships between evidence generated by components or subsystems and failure modes that can be mapped to specific corrective actions. The offline learning loop, the main subject of this section, is used to generate an incremental update to the system reference model to improve the performance of VIPR. Finally, the figure also shows the interfaces to the external systems, both for input of health information as well as output to the flight crew, ground maintainer, and/or a flight management system.

A main goal of the work was to explore ways to systematically update the system reference model based on results of the offline learning loop/data mining. The system reference model is generated based on expert knowledge and experience and includes a wealth of valuable system information. An unconstrained approach to the offline learning loop could easily result in a completely different construction of the system reference model from that generated by the experts. The newly generated reference model may be obscure in the sense that relationships generated are not obviously understood by the expert. In the end the expert will have to verify that the results of the update are sensible and reasonable. Therefore, the approach here was to explore ways to create an incremental update to the reference model that the expert could more easily understand and analyze.

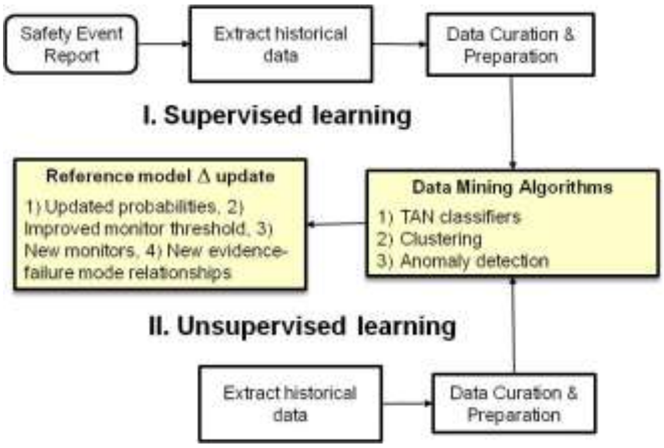
Safety Event Report
Extract historical data
Data Curation & Preparation
I. Supervised learning
Reference model Δ update
1) Updated probabilities, 2) Improved monitor threshold, 3) New monitors, 4) New evidence-failure mode relationships
Data Mining Algorithms
1) TAN classifiers
2) Clustering
3) Anomaly detection
II. Unsupervised learning
Extract historical data
Data Curation & Preparation

Fig. 2: Learning loop

Fig. 2 above shows the resulting learning loop methodology employed. The reference model incremental update is achieved in four ways. First, the conditional probabilities between the evidence and failure modes may be updated. Second, the indict/exonerate region of a condition indicator (CI) monitor may be adjusted. Third, new monitors may be created, specifically new monitors that look at the combined results of previously generated monitors. Finally, new evidence-failure mode relationships may be discovered, resulting in addition of new monitors, new evidence, and creation of connections between the new monitors/evidence with existing failure modes identified in the model.

Two different sources of data were used in this work. First, CMAPS-S (Commercial Modular Aero-Propulsion System Simulation) developed at NASA Glenn Research Center was used to generate data from simulated faults in a commercial turbofan engine. This relatively clean, controlled data set was used to support development of the learning loop methodology designed during this project. The methodology was then testing on a set of data generated from a fleet of aircraft. The fleet consisted of more than 30 identical aircraft, each operating 2-3 flights per day. Data was gathered from this fleet over the course of 3 years. During this time period the airline experienced several safety incidents for which information was available through the Aviation Safety Information Analysis and Sharing (ASIAS) system.

In particular, one of the adverse events that occurred during the data collection time period for which the ASIAS report was available was an in-flight engine shutdown (IFESD), a highly undesirable event that led to the pilots returning to the departure airport. The IEFSD for engine 3 (out of 4 total) was initiated by the aircraft computer. The root cause for the shutdown was determined to be a faulty fuel hydro-mechanical actuator (HMA). This component was replaced after the event, and afterwards indications were that the engine and aircraft resumed normal operation. An initial assessment showed that using the original expert-designed reference model VIPR would have been unable to disambiguate the real fault between several fault candidates. The task, then, was to employee the learning loop using historical data to try to improve the performance of VIPR recognizing this particular fault. This was done through a series of experiments, described below.

A domain expert provided insight into this fault, specifically that it would likely begin to manifest around 50 flights prior to the shutdown as well as what data was relevant to indicate this error. This data was transformed into CI's, which were used in the initial experiment to avoid any effects of overly-conservative thresholds used to generate health indicators (HI). There were 25 such features to be used per flight per engine and a total of 200 samples (50 flights x 4 engines) of each of these features. A discrete Tree Augmented Naive Bayes (TAN) algorithm was used to generate the necessary classifiers for this data.

Classifiers were trained using 10-fold cross-validation. The resulting accuracy of these classifiers was quite high, with an average accuracy of 99.5% with a 0.7% false positive rate and no false negatives.

The TAN created in the first experiment showed a complex relationship between some of the CI's from the engine shutdown phase and some of the CI's from the engine start-up phase. A second experiment was devised to look at evolution of the classifier scheme as the fault progressed. The 50 flights were divided into 5 bins of 10 flights chronologically so bin 5 was from earliest before the event and bin 1 was from right before the event. Classifiers were trained using each bin individually as the training set and using the remaining 4 bins as the test set. The results of the experiments are shown in Table 1.

Table 1: Results from second experiment

Bin	**Training Flights**	**Accuracy on Holdout Set**	**False Positive %**	**Observation Root Node**	**Children of ORN**
1	1 to 10	97.65%	2.30%	IdleSpeed	StartTime
2	11 to 20	93.90%	5.70%	peakEGTC	liteOff,dipEGTC
3	21 to 30	94.65%	5.30%	peakEGTC	liteOff,dipEGTC
4	31 to 40	96.62%	3.50%	startTime	peakEGTC
5	41 to 50	96.06%	4.10%	liteOff	phaseTwo,RollTime

In the second experiment, the accuracy and false positive rates were the best for Bins 1 and 4. Upon closer examination of the TAN structures generated for these two bins, the expert noted the apparent causal relationship between the startTime and peakEGTC CI's, corresponding to the time it took the engine to reach stoichiometry and the peak exhaust gas temperature during engine start-up, respectively.

The results from experiments 1 and 2 were examined more closely to determine how to update the system reference model to include the discovered knowledge. For all existing health indicators (HI's) related to these CI's, the thresholds were compared to the values in the conditional probability tables (CPT) for the classifiers. Expert examination of these values and comparison with the existing thresholds prompted modification of the thresholds to potentially improve accuracy without becoming too noisy. In addition, the structure of the TANs showed that, in addition to the existing HI for a slow engine start, for this particular fault the engine start time was shorter than nominal cases. A new HI was created to fire when the start time was below a specific threshold suggested by the TAN. Also, as alluded to previously, there was a causal relationship noted between start time and peak exhaust gas temperature during start-up. This prompted the creation of a "super monitor" that would fire in case the existing HighTemp HI and the new fastStart HI both fired at the same time.

These changes were incorporated into the system reference model, and simulations were run for the 50 flights considered here in addition to 10 "nominal" flights after the component was replaced. Without this system reference model update, VIPR could not disambiguate between the faulty fuel HMA and a set of other faults. With the updates VIPR was able to establish the fuel HMA fault 20-30 flights before the IFESD event, which would allow the operator to avoid the safety incident entirely.

Another aspect of the VIPR project was to include anomaly detection. The anomaly detection monitors in VIPR collect and process on-board data and continuously seek emerging patterns. The full process is shown in Fig. 3.

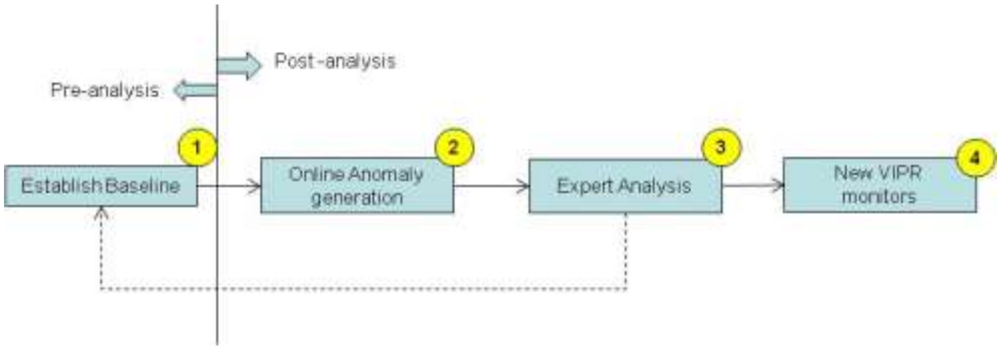

Fig. 3: Operational steps in VIPR anomaly detection

The first step (step 1 in Fig. 3) to incorporating anomaly detection into VIPR was to establish a nominal model to serve as a baseline for the expected behavior during flight. This was achieved using an unsupervised learning approach. The data set consisted of P flight segments, each associated with N features (here, an aircraft sensor parametric value), and each feature associated with M samples defining its time-varying characteristics. Each flight segment defines a unique data point. Pair-wise feature distances between each data point were calculated using the Kolmogorov complexity measure. A matrix was generated containing the pair-wise distances between flight segments, using the Euclidean metric to calculate the pair-wise distances based on the pair-wise feature dissimilarities. A hierarchical clustering approach (here, the complete link clustering algorithm) was used to generate the dendrogram defining clusters of nominal flight segments as well as the outliers and anomalous clusters. At this point, the pre-analysis ends by both extracting a nominal model to employ for on-aircraft anomaly detection with VIPR and analyzing the anomalous clusters to generate new VIPR monitors directly. This full process is illustrated in Fig. 4.

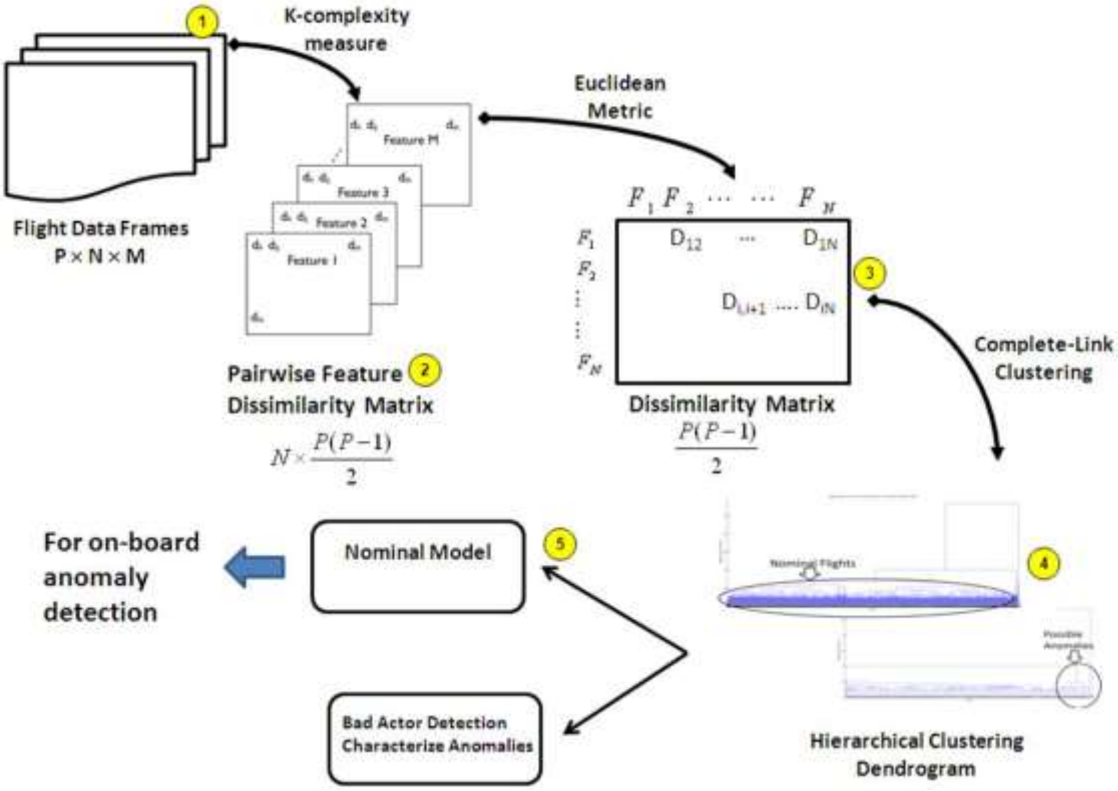

Fig. 4: Establishing the baseline through offline unsupervised analysis

The remaining 3 steps shown in Fig. 3 are labelled the "Post-Analysis" phase. The nominal model discovered in the baseline establishment process is used to create an anomaly monitor in VIPR. VIPR then uses this monitor to capture potentially anomalous data for later download and analysis by an expert. The analysis may result in discovering an emerging fault and/or may result in the addition of a new diagnostic or prognostic monitor to VIPR to explicitly watch for future occurrences of such a fault.

Case II: Development of HUMS Super-CI's

A common problem of interest is to update thresholds of HUMS CI's to improve performance in detecting incipient faults. Generally this amounts to improving the sensitivity (true positive rate) and/or the specificity (true negative rate) associated with the CI. HUMS systems generate a great deal of data over time and throughout a fleet. In the Standoff Approach for Drive Systems Prognostics program sponsored by the Army Aviation Applied Technology Directorate (AATD), Honeywell led an effort to explore data-driven methods to update HUMS CI's using historical HUMS data stored in the Honeywell Intelligent Machinery Diagnostics System (iMDS) [2]. This data was supplemented by fault information from The Army Maintenance Management System-Aviation (TAMMS-A) DA Form 2410 – Component Removal and Repair/Overhaul Record.

Initially the TAMMS-A removal records were examined for removal for cause events related to the aft hanger bearing (AHB), intermediate gearbox (IGB), and tail gearbox (TGB). HUMS survey data from the iMDS was extracted from 90 days before and after each of the events. The analysis proceeded in two main directions. First, improvement of thresholds of individual CI's related to the identified components was examined based on the historical removal and CI data. This information could be utilized to update fleet-wide thresholds, achieved by deployment through the iMDS system to the HUMS systems. The second experiment involved looking at using combinations of CI's to create "Super-CI's" that could better identify incipient faults. Such "Super-CI's" could be deployed to the fleet through a HUMS database update.

The approach to improving thresholds on CI's individually produced limited results. This was a relatively straightforward analysis in which each combination of CI and fault type (e.g. bearing failure, beyond specified tolerance, corroded, leaking, excessively worn, etc.) was examined individually, breaking the data into "before" data gathered prior to the removal event and "after" data gathered after the removal event. Receiver operating characteristic (ROC) curves were plotted to explore the relationship between sensitivity and specificity for various threshold values for a given CI. The information in the ROC curves was compared to the existing CI thresholds to look for potential improvements. A primary challenge with this approach is the relatively few occurrences of exceedances included in the iMDS data surrounding these particular events. A few thresholds could be adjusted, but real meaningful improvement in CI performance in diagnosing the fault was not observed. The established CI thresholds either performed well or not enough information was present in the data to establish otherwise.

In light of the results of the single CI investigation the team explored ways of combining information from multiple CI's to create more informative "Super-CI's" based on the historical data and removal records. For this investigation, 43 IGB removal for cause events were utilized. For these events, 21 related CI's were extractions for 90 days before and after the event. To try to limit bias, in cases where a specific CI on a specific platform associated with an event was collected multiple times on a given day, that CI value was averaged for that day. Similarly, for those CI's that were collected multiple times during the 90 days before or after the event, the values were averaged to give one data point before or after the event.

The goal of this analysis was to produce a "Super-CI" that served as a binary classifier, essentially an anomaly detector, giving a classification of nominal or anomalous. To obtain this classifier, linear discriminant analysis (LDA) was employed. LDA generates a linear combination of variables that maximizes classification effectiveness. Generally, LDA transforms input variables (here, the 21 CI's) to canonical variables. The classifier is then obtained by transforming input points onto the first canonical variable axis to obtain a score. A binary classification is achieved when an appropriate threshold is associated with the score. The results of the LDA analysis are shown in Fig. 5, where the blue dots represent the "before" data and the red dots represent the "after" data. Fig. 6 shows the ROC curve for thresholds on the first canonical variable. This indicates, for example, that it is possible to achieve a 68% detection rate with only 10% false positives. This represents a 25% increase in detection accuracy over the established single-CI threshold capabilities.

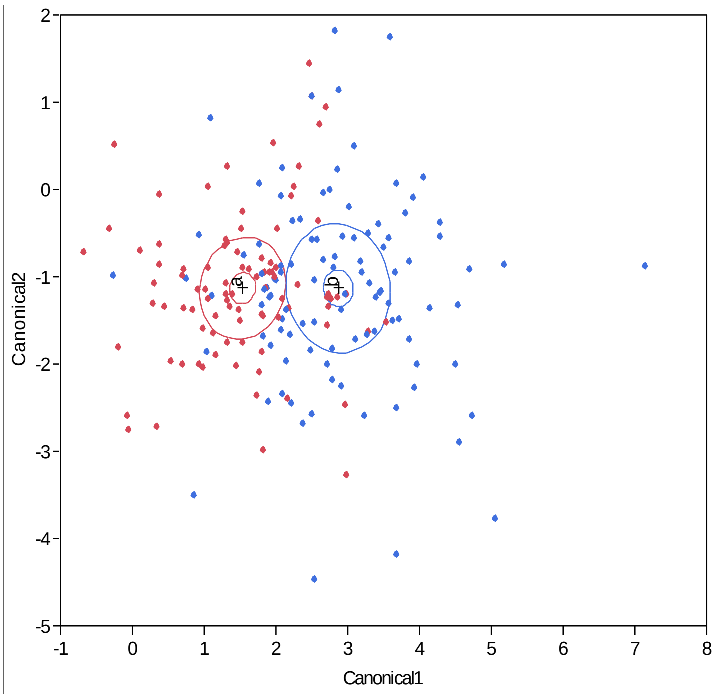

Fig. 5: Results of the LDA

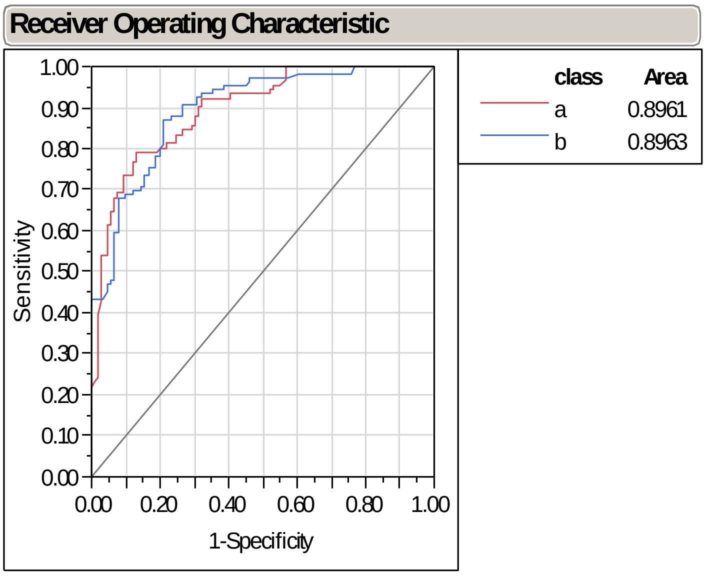

Fig. 6: ROC curve for IGB Super-CI

Conclusion

When big data is discussed, generally reference is made to the “4V” characteristics: volume, velocity, variety, and veracity. From the two case studies presented above for data-driven methods supporting condition-based maintenance, the V’s are present. While the VIPR program operated specifically on a static historical data set, the general methodology developed involved continuous data analysis on a rich, complex, extensive data stream generated at a very high rate (volume and velocity). In the Standoff program, the iMDS provides a great deal of data (volume), and it was married with TAMMS-A 2410 reports (variety) to achieve the goals. Both of these examples lend themselves readily to a big data framework, and through these case studies it has been demonstrated how data-driven methods can effectively and methodically improve upon current condition-based maintenance technologies.

Acknowledgements

The work presented here is in part funded by National Aeronautics and Space Administration (NASA) under contract NNL09AA08B on VLRS and U.S. Army Aviation Applied Technology Directorate (AATD) Ft. Eustis, Virginia, contract number W911W6-11-C-0030. The authors would also like to acknowledge Mr. Eric Cooper (NASA) and Mr. Jason Fetty (AATD) for their valuable feedback through the course of these projects.

References

1. Bharadwaj, R., Mylaraswamy, D., Cornhill, D., Biswas, G., Koutsoukos, X., Mack, D., “Vehicle Integrated Prognostic Reasoner (VIPR)”, Final Report, NASA NNL09AD44T.

2. Bharadwaj, R., “Standoff Approach for Drive Systems Prognostics (6.2 OSST Program)”, Final Report, AATD W911W6-11-C-0030.