Validating Machine-learned Diagnostic Classifiers in Safety Critical Applications with Imbalanced Populations[†]

Daniel Wade[1], Dr. Andrew Wilson[2], Abraham Reddy[3], and Raj Bharadwaj[4]

[1,2]*United States Army AMRDEC, Redstone Arsenal, AL, 35898, United States*

[3,4]*Honeywell Aerospace, Plymouth, MN, 55441, United States*

ABSTRACT

Data science techniques such as machine learning are rapidly becoming available to engineers building models from system data, such as aircraft operations data. These techniques require validation for use in fielded systems providing recommendations to operators or maintainers. The methods for validating and testing machine learned algorithms generally focus on model performance metrics such as accuracy or F1-score. Many aviation datasets are highly imbalanced, which can invalidate some underlying assumptions of machine learning models. Two simulations are performed to show how some common performance metrics respond to imbalanced populations. The results show that each performance metric responds differently to a sample depending on the imbalance ratio between two classes. The results indicate that traditional methods for repairing underlying imbalance in the sample may not provide the rigorous validation necessary in safety critical applications. The two simulations indicate that authorities must be cautious when mandating metrics for model acceptance criteria because they can significantly influence the model parameters.

1. INTRODUCTION

The United States Army Aviation Engineering Directorate (AED) is the Army's airworthiness certification authority; it also provides engineering expertise to the Army's Program Executive Office for Aviation (PEO-AVN). One portion of this service is data analysis and management of Health and Usage Monitoring System (HUMS) data collected on all Army rotorcraft. The AED is leading an Aviation and Missile Research Development and Engineering Center (AMRDEC) Science and Technology effort to investigate the application of data science to aviation data, e.g. HUMS and maintenance log data. As part of this effort, the AED is defining the substantiation required to validate machine-learned diagnostics that could replace existing physics-based diagnostics. This paper investigates the qualities of metrics used to inform the substantiation and qualification of diagnostic classifiers; it identifies aviation specific requirements for the use of less often reported classification metrics (Powers, 2011).

In particular, many aviation data sets are highly imbalanced, i.e. one class (such as healthy or nominal) is 100:1 (or worse) more prevalent than other critical classes (such as faulted or damaging). This population imbalance has significant consequences to the behavior of classifier metrics and the resulting decisions. This paper will investigate the typical metrics reported over classification problems using a hypothetical situation, an oncoming zombie apocalypse, to understand the consequences of using them to make diagnostic decisions. Before investigating the metrics, the paper will discuss how a metric, informedness, was recently used to rank models for a rotorcraft health monitoring application and select a model for deployment to the PEO-AVN customer.

1.1. Definitions

This paper attempts to use industry standard data science nomenclature. In an effort to be absolutely clear, frequently used terms are defined in this section.

True Positive Rate (**TPR**) is the ratio of correct positive identifications to the total number of positive examples in the sample. It is the conditional probability that the classifier will label an example as positive given that it actually was positive. It describes the ability of a classifier to correctly identify a positive example. In this paper, it refers to how often a model correctly identifies zombies in the sample. The sum of the TPR and False Negative Rate (*missed detection rate*) is one. TPR is known in some communities as **recall**.

[†] Distribution Statement A: Approved for Public Release

False Positive Rate (**FPR**) is the ratio of incorrect positive identifications over the total number of negative examples in the sample. It is the conditional probability that the classifier will label an example as positive given that it actually was negative. It describes the ability of a classifier to correctly identify a negative example. In this paper it refers to how often a model errors by calling a human a zombie. The sum of the FPR (*false alarm rate*) and True Negative Rate is one.

Zombies and humans are used in this paper as a fictional way to understand a real world problem – making critical, discrete decisions with noisy, continuous data. The underlying cause of the zombie transformation is unknown and there is no known treatment available.

Imbalance in this paper is the ratio of humans to zombies. In the case of this paper, it is assumed that initially there are only a small number of zombies in the overall population. Imbalance progressing from a very small ratio (500:1) towards a 1:1 ratio is a surrogate for time and widespread prevalence of the condition. From a fictional perspective, it is the desire of the health authority to prevent spread of the disease.

Model performance metrics such as: Accuracy, Weighted Accuracy, Informedness, and F1-Score are properties of a chosen threshold boundary between two (or more) distributions. They are often used in support of model validation for applications in data science and statistics. They are also used as model selection criteria for use in decision making.

Accuracy is the ratio of the correct classifications to the total number of samples (scikit-learn, 0.17.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Weighted Accuracy corrects for imbalance in a sample. It weights the underrepresented class' error by the imbalance ratio. Weighted accuracy is the weighted average of the accuracy of the classes (scikit-learn, 0.17.1).

$$Class\ Weight = \frac{1}{N_{classes} * size(Class)} \quad (2)$$

Informedness quantifies how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition (versus chance) (Powers, 2011).

$$\text{Informedness} = \text{TPR} - \text{FPR} \quad (3)$$

Precision is the ratio of True Positives over the sum of True Positives and False Positives. It is also known as the positive predictive accuracy (scikit-learn, 0.17.1).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

F1-Score is the weighted geometric average of precision and recall. It can be expressed in terms of number of True Positives (TP), False Positives (FP), and False Negatives (FN) (scikit-learn, 0.17.1).

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (5)$$

Error Rate is the sum of the errors divided by the sample size (scikit-learn, 0.17.1).

$$Error\ Rate = \frac{FN + FP}{size(Sample)} \quad (6)$$

Methods are ways the health or safety authority decides to use metrics to set diagnostic thresholds. An example of a method would be to choose a diagnostic threshold based on maximum F1-score.

1.2. Prior Applications of Performance Metrics

The Army recently used machine learning methods to investigate and potentially improve vibration based diagnostics for rotorcraft gearboxes. As part of this effort, tens of thousands of models were learned from the data (Wilson, Wade, Albarado, Partain, & Statham, 2016) by a cross organizational (government and industry) team of engineers and data scientists. The models were primarily evaluated using informedness. The Army explored avenues for validating and accrediting models using a train-validate-test data partitioning method with rigid model standards (Wade & Wilson, 2017).

Model selection was broken into two serial procedures starting with a *best of breed* decision (~10,000 models) among similar model types followed by a *best of show* decision among the remaining models (15 models). The *best of show* method used threshold and objective requirements for TPR and FPR. The details of how the problem was setup are described in prior publications by Wade and Wilson (2017) and Wilson et al. (2016). The success criteria for this problem was to outperform the physics based diagnostic

algorithm programmed into the aircraft Health and Usage Monitoring System.

The results of this development applicable to this paper showed that informedness was preferred over F1-score and accuracy for model selection. This was found anecdotally by the engineers prior to accomplishing the simulations shown in this paper, as well as supported by a literature search (Powers, 2011). The gearbox population statistics for which the diagnostic was developed were similar to those used in the simulation for this paper. The data was split into train, validate, and test partitions with a 1:35 imbalance ratio (faulted:healthy). This ratio was based on operation of the aircraft over hundreds of flight hours.

2. SIMULATION PROCEDURES

Let us suppose that there is an oncoming zombie apocalypse for which a diagnostic test is under development to determine if a human will soon turn into a zombie. In this hypothetical case, only a small number of transitioning zombie test subjects are available for diagnostic test threshold determination but many humans are available. Let us also say that the diagnostic is immature and has unknown behaviors that introduce noise into the output such that the two populations are only partially separable. The underlying diagnostic value (D_v) model for the human population is a Rayleigh distribution and for the zombie population is a Gaussian distribution. Finally, assume that it is the objective of the health authorities to quickly identify zombies inside the human population to prevent further spread, and the end of humanity. This hypothetical case is not unlike discovering mechanical faults (e.g. gear or bearing faults) from noisy sensor data (e.g. vibration data) installed on rotorcraft (Wade, Tucker, Davis, Knapp, Hasbroucq, Saporiti, Garrington, & Rudy, 2017).

The proposed problem was translated into a simulation using Python packages NumPy, and scikit-learn (Anaconda, 4.1.1), (Python, 3.5.2), (NumPy, 1.11.1), and (scikit-learn, 0.17.1). A zombie diagnostic value (D_{vz}) is determined by randomly choosing from a Gaussian distribution centered at 10.5 (mean) with a scale of 2.0 (standard deviation). The human diagnostic value (D_{vh}) is determined by randomly choosing from a Rayleigh distribution of scale 3.0. When sampled without imbalance (1:1), the two populations have histograms as shown in Figure 1.

The types of distributions and their setup are chosen so that the problem is interesting to study. The diagnostic is designed to be noisy with occasional outliers reaching the mean of the other population's distribution. The human diagnostic value is based on a Rayleigh distribution so that it has values greater than zero and is skewed toward the zombie population. This distribution is similar to the distribution of aircraft health data collected by HUMS.

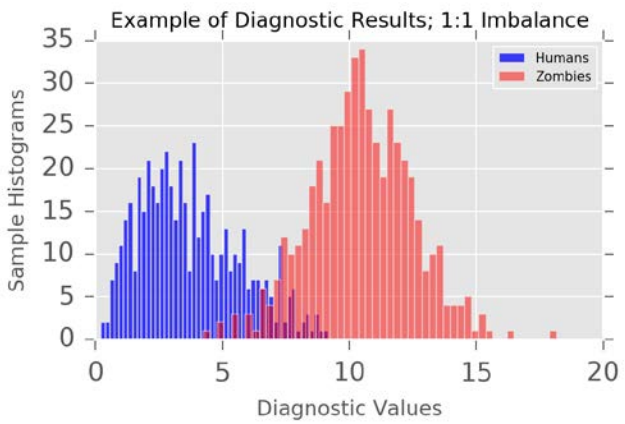

Figure 1. Histogram examples of a single simulation at the 1:1 imbalance ratio.

The simulation follows this procedure:

1. Define the initial imbalance ratio, R_I, humans to zombies, e.g. 250:1.
2. Define the fixed sample size, N_s, the number of total subjects in a sample.
3. Define an imbalance change function to iterate over, e.g. $R_{In} = R_{In-2} + R_{In-1.}$
4. Determine which model validation metrics should be computed, e.g. informedness and weighted accuracy.
5. Call a simulation function that can be iterated.
6. Pull subjects into the sample with random diagnostic values (D_{vz} and D_{vh}) such that the number of zombies (N_z) and humans (N_h) summed equals the sample size (N_s).
7. Find each threshold value that maximizes each of the metrics.
8. Save the threshold and confusion matrix for each of the metrics.
9. Repeat the process 1,000 times for the same imbalance ratio (Step 4).
10. Increment the imbalance ratio (Step 3) until the imbalance ratio goes to 1:1.
11. Plot mean TPR, FPR, Error Rate, and Threshold Value as a function of imbalance.

Results for this simulation are shown in Figures 2-5 and in Appendix A, Figures A1-A4. The x-axis is shown on a log scale for readability at low imbalance ratios. The x-axis is the number of zombies in the sample normalized by $N_s/2$. Some results are shown as mean simulation output with a ±1 standard deviation cloud to show variance across simulations.

A second simulation, to understand the influence of sample size, was also accomplished. It uses the same procedure above with changes to steps 1, 2, 3, 9, 10, and 11 shown underlined for clarity.

1. Define the initial sample size, N_s.
2. Define the fixed imbalance ratio, R_I.
3. Define a sample size increase function to iterate over.
4. Determine which model validation metrics should be computed.
5. Call a simulation function that can be iterated.
6. Pull subjects into the sample with random diagnostic values such that $N_z + N_h = N_s$.
7. Find each threshold value that maximizes each of the metrics.
8. Save the threshold and confusion matrix for each of the metrics.
9. Repeat the process 1,000 times for the same sample size (Step 4).
10. Increment the sample size (Step 3) up to 72,000.
11. Plot mean TPR, FPR, Error Rate, and Threshold Value as a function of sample size.

Results for the simulations are presented below, starting with the first, sample imbalance study, then moving on to the second, sample size study. Discussions specific to each study are presented sequentially with the results. A final discussion section regarding how the simulation results are related to the development of a rotorcraft diagnostic for gearbox health is then offered. The code that produces these simulations is provided for the reader and heavily commented; it is available on github at the following address:

https://github.com/DanielRWade/AMRDEC-aviation-data-science.

3. VARIABLE IMBALANCE RATIO RESULTS

The normalized threshold chosen by the four methods is shown in Figure 2. The threshold is normalized by the mean Zombie Diagnostic Value (10.5), thus some threshold values will be greater than 1. Note that the threshold values generally converge at the 1:1 imbalance ratio but they approach this threshold in very different ways. The threshold at the 1:1 imbalance ratio for the weighted accuracy and informedness methods is achieved at very low imbalance ratios. The accuracy method produces the widest range of threshold values and never achieves steady state. The F1-score method moderates between the others but has more similarities to the accuracy method in that it never truly settles on a threshold value. Note that the standard deviation range of all the methods is similar as a function of imbalance ratio.

The uncertainty in the methods never goes to zero, even when the imbalance ratio is 1:1 due to the overlapped distributions. If the plot was continued beyond the 1:1 ratio, indicating that the zombies outnumbered the humans, the methods would not be perfect mirror images due the use of the Rayleigh distribution.

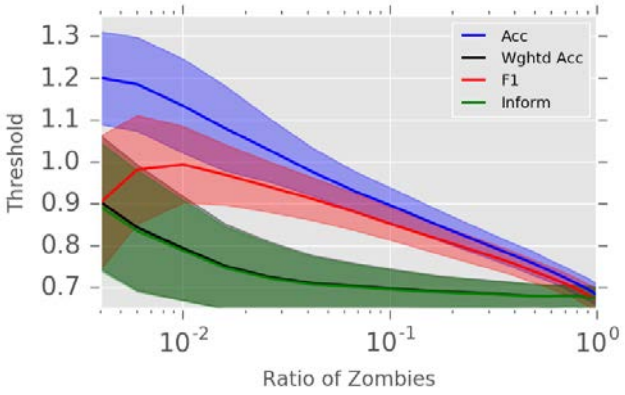

Figure 2. Mean (line) and one standard deviation (cloud) normalized thresholds chosen by each of the four methods after one full simulation run from 499:1 to 1:1 imbalance ratio.

The error rate of each threshold method is shown in Figure 3. For the sake of simplicity and legibility, only the mean error rate is shown. The error rate generally follows the behavior of the threshold plot, except that it shows the error rates specific to the way the problem is setup. Note that the methods converge to an error rate of 0.05, with the exception of the informedness method which has slightly lower error, 0.049, at the 1:1 ratio. The plot of the error rate shows how the different methods, which are choosing slightly different thresholds (Figure 1) at the 1:1 ratio, still result in the same amount of error, thus making a trade-off that can only be seen by plotting the associated TPRs and FPRs.

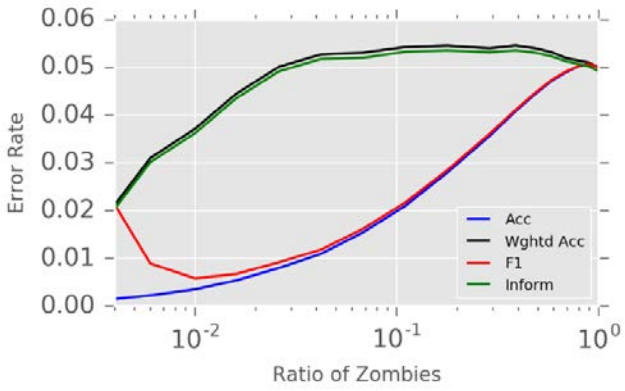

Figure 3. Mean Error rate for each of the four methods after one full simulation run from 499:1 to 1:1 imbalance ratio.

Figures 4 and 5 plot the TPRs and FPRs of each method respectively. The standard deviation is included in these plots. Figure 4 is possibly the most interesting plot shown yet because it is the first plot that demonstrates a difference between all of the methods for choosing a threshold value. The TPR varies significantly, from as low as 0.3 to 1.0, at low imbalance ratios but then converges to nearly the same value at the 1:1 ratio (0.96). The standard deviation of the methods

is also quite different with the smallest deviation in the informedness method, followed by the weighted accuracy method, the F1 method, and finally the largest deviations in the accuracy method. Notice that the TPR for the informedness method is mostly steady for all imbalance ratios and that the weighted accuracy method quickly converges with the informedness TPR. Conversely, F1 and accuracy have similar behavior but significantly different values for low imbalance ratios; over the range of imbalance ratios, they converge at the 1:1 ratio.

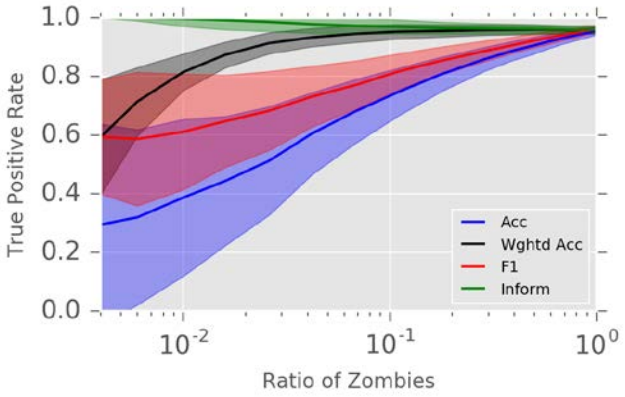

Figure 4. Mean TPR (line) and one standard deviation (cloud) for each of the four methods after one full simulation run from 499:1 to 1:1 imbalance ratio.

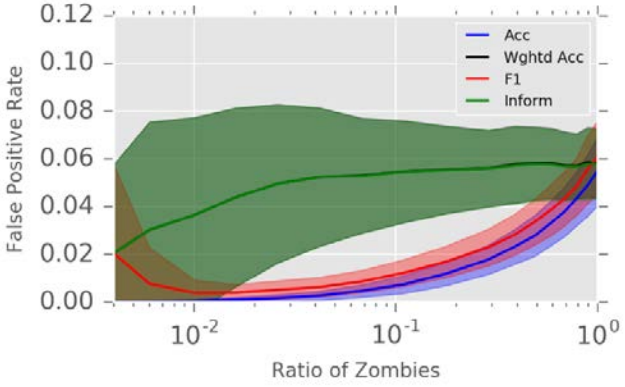

Figure 5. Mean FPR (line) and one standard deviation (cloud) for each of the four methods after one full simulation run from 499:1 to 1:1 imbalance ratio.

The plot of FPRs in Figure 5 shows how the informedness and weighted accuracy methods perform equally from an FPR perspective at all imbalance ratios. This should immediately be cross referenced against the TPR plot however, where there are significant differences in the methods for low imbalance ratios. The very slight threshold differences at low imbalance (Figure 2) result in significant difference in overall performance of the metric to maximize overall performance. As has been seen in previous plots, the F1 and accuracy methods have similar behavior except at low imbalance ratios where accuracy has a nearly zero FPR, while the F1 method starts at a value similar to the weighted accuracy and informedness methods.

3.1. Variable Imbalance Ratio Discussion

When reviewed as a full set, the four figures tell an important story. It can be seen that the accuracy method, as expected, reduces overall error, which results in an overwhelming preference for reduction of error by assuming all subjects are humans. For a health official that is interested in discovering the zombies quickly and possibly preventing the introduction of a disease into the healthy population, accuracy fails.

Accuracy provides an excellent baseline for comparison of the three other methods. The weighted accuracy and informedness methods should be considered together because they have so much in common. In general, when maximized, they result in nearly the same threshold value, with informedness choosing a threshold slightly closer to the healthy population. This slight difference between the two thresholds results in a barely perceptible difference in overall error (Figure 3), but results in a dramatic difference between the two methods when the imbalance between the populations is maximum (low imbalance). This is quite evident in Figures 4 and 5. This appears to support a conclusion that using both of them, especially in situations when the practitioner knows that imbalance exists but does not know the extent to which it occurs in the overall population, is necessary. While they result in very similar models being chosen, they are still different enough that they provide unique information to the practitioner trying to make the best choice and inform other decision makers about model performance.

The F1 method chooses thresholds in a hybrid fashion. At the lowest imbalance values, it settles on the same threshold as weighted accuracy, but quickly diverts away, focusing on moving the threshold toward the zombie population. This results in excellent control of the FPR for the majority of imbalance values, but this is of course at the cost of allowing TPR to fall quite significantly. The F1 method does not result in models that have even reasonable discriminatory power (TPR above 0.9) until the imbalance ratio is greater than 5:4 (humans:zombies).

In machine learning, metrics are not often used to drive model behavior, cost functions are used; the performance of a set of models is shown on validation and test data and represented by any number of metrics. The model builders utilize these metrics to understand which models are best. The metrics used to rank and choose models will influence the outcome and how a metric is biased by the underlying population characteristics is influencing the chosen model. This must be well understood by the safety or health authority.

Often times, in cases where there is a known imbalance between populations, minority oversampling (or majority

decimation) is used to prevent bias in the metrics and decision making. For the purposes of validation and model selection, this is equivalent to using the weighted accuracy. The plots shown in this paper show that none of the metrics is immune to sample size imbalance, including weighted accuracy. The metrics offer a trade space which is not trivial in applications where models are used to determine which patients should receive treatment or which aircraft is fit to fly a mission.

Based on the evidence presented, it can be stated that there is actually a correct threshold for the diagnostic that is shown in Figure 2, at the 1:1 imbalance ratio. There is convergence of the methods on a threshold zone represented by the standard deviation clouds. All of the methods do not start at the optimal threshold when imbalance is significant. All of the methods exhibit a reduction in uncertainty as imbalance ratio approaches 1:1. One additional point can be made, the weighted accuracy and informedness methods achieve the optimal threshold for lower imbalance ratios than F1-score and accuracy.

4. VARIABLE SAMPLE SIZE RESULTS

This study offers an understanding of the influence sample size has on the metrics. Fewer plots are shown here than in the previous section for the sake of brevity. The plots show mean values with a cloud representing the standard deviation across the simulations.

Figure 6 shows how the chosen threshold changes as a function of the sample size. Using the maximum accuracy method for setting a threshold results in the least mean variance for the sample sizes shown here which vary from 1000 subjects, up to 72,000 subjects. The other three methods choose thresholds that are nearly all the same, and then diverge towards a final value determined by the underlying population distributions. In the case of informedness and weighted accuracy, the thresholds converge and have the same value for very large, but still imbalanced samples. This simulation uses the initial imbalance (499:1) from the previous simulation for consistency.

Figures 7 and 8 show the associated underlying TPR and FPR, respectively, as a function of sample size. All of the methods are influenced by the sample size with initial values being either optimistic (informedness, F1-score, and accuracy), or pessimistic (weighted accuracy). Informedness has an initial TPR at 1.0 which eventually settles to 0.96 for, unrealistically, large sample sizes. Weighted accuracy and F1-score start at 0.59; weighted accuracy then rapidly moves toward the informedness TPR ending at 0.96. F1-Score settles on its own final TPR at 0.42. Accuracy behaves similarly to F1-score, but with lower TPR values, initially 0.32, settling to 0.17.

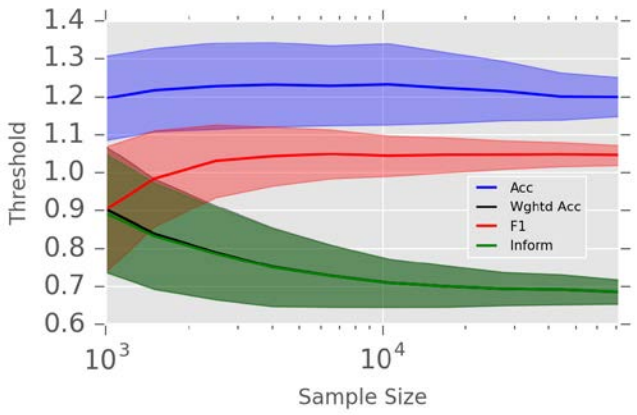

Figure 6. Mean (line) and one standard deviation (cloud) normalized thresholds chosen by each of the four methods as sample size increases.

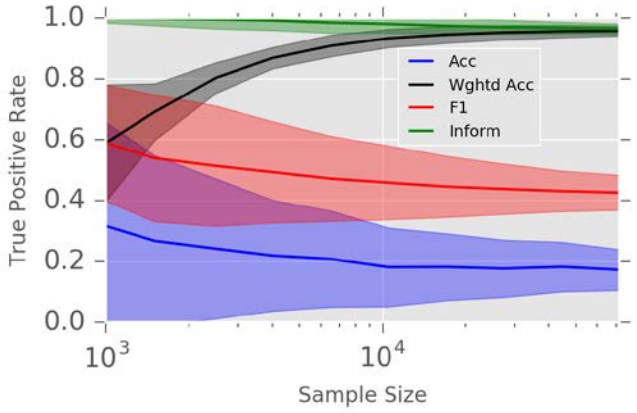

Figure 7. Mean TPR (line) and one standard deviation (cloud) for each of the four methods as sample size increases.

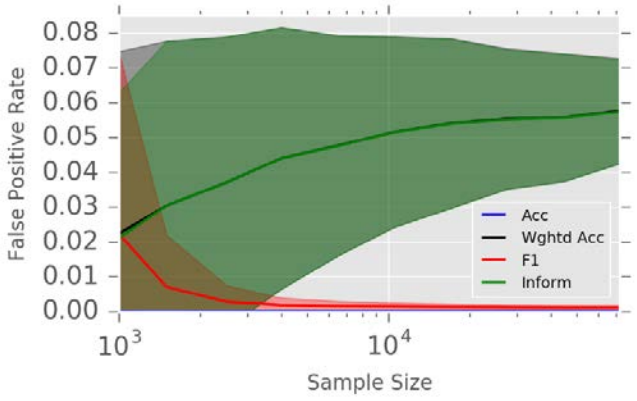

Figure 8. Mean FPR (line) and one standard deviation (cloud) for each of the four methods as sample size increases.

Expectedly, FPRs for the metrics trend in the opposing directions. Informedness and weighted accuracy show almost

identical FPRs over all sample sizes, ending near 0.06. F1-Score starts at the same value as the informedness and weighted accuracy (0.02) and ending at 0.001 which is only non-zero due to the unrealistic sample size. Accuracy maintains the same FPR for all possible samples at nearly zero. This is caused by a small number of the simulations occasionally choosing thresholds with non-zero FPR.

4.1. Variable Sample Size Discussion

The plots that show how the metrics respond as the sample increases but maintain the proportion of zombies are significant. As the sample size increases, the standard deviation around each method generally decreases showing that the methods are more confident in the choice. The F1, informedness and weighted accuracy methods change the threshold decision to some degree based on the sample size. Accuracy, meanders slightly but never changes significantly. As happened in the imbalance study, in these plots, it is seen that the methods **do not converge on the same threshold**; three distinct final thresholds are reached.

5. GENERAL DISCUSSION

The results from both studies say something very significant about the four methods, none of them are immune to the class imbalance and underlying distributions. Oversampling the minority population, represented by using the weighted accuracy, is not immune to this effect. This is especially pronounced when sample sizes are between 1,000 and 10,000.

Based on the information from the two simulations, the Aviation Engineering Directorate has come to the conclusion that for diagnostics that drive aviation maintenance, especially for safety critical components, metrics such as weighted accuracy and informedness are superior to F1-score and accuracy. One particularly important fact is that often, it is unknown what the true imbalance ratio is between two populations. Thus choosing metrics that are less sensitive to the imbalance for decision making, such as model validation and selection for fielding to the aircraft, are preferred. Furthermore, metrics that are known to have conservative behavior, such as informedness, are preferred.

Thinking back on the prior work discussed in the introduction, what is now better understood by the authors is that all of the metrics are influenced by the underlying imbalance in the populations. Oversampling of the faulted population would not result in alleviating the bias. Based on the results of the simulation shown in this paper, informedness was a good choice for evaluating overall model performance, especially considering the sensitivity to detecting faults in aviation systems, i.e. maintaining a high TPR.

Additional work to determine how these metrics, including many others not studied, could be evaluated using the same codes to determine the effects on multi-class problems, and different underlying distributions. For example, the multi-class problem is being actively investigated as it relates to propagating labels into the HUMS data from the aircraft maintenance logbook. The authors plan to move forward with investigating when Markedness, positive predictive value (precision), and negative predictive value could be used.

6. CONCLUSIONS

This paper presented two simulations that demonstrate how underlying population imbalance influences metrics used for validation of diagnostic models. The simulations show that accuracy, weighted accuracy, F1-score, and informedness are affected by class imbalance. Informedness and weighted accuracy are less biased by imbalance and are also more applicable to diagnostic models used in safety critical applications. Informedness has slightly better performance than weighted accuracy for significantly imbalanced classes.

The authors of this paper spot checked other ways to mix the underlying models, i.e. changing the mean and standard deviation. The rates of convergence shown in the figures as a function of imbalance ratio do change based on the overlap of the two populations. It is recommended that these additional simulation parameters be added to another study. Furthermore, there are more metrics that could be studied, e.g. Markedness, as relates to the topic of sample imbalance.

ACKNOWLEDGMENT

The authors would like to acknowledge the hard work of the men and women of the US Army. It is the objective of this paper and the work described to make their aircraft safe and reliable. We thank them for their service.

REFERENCES

Powers, David (2011). Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* Vol. 2 (No. 1), pp.37-63. DOI 10.9735/2229-3981

Wade, Daniel, Tucker, Brian, Davis, Mark, Knapp, Doug, Hasbroucq, Sophie, Saporiti, Moreno, Garrington, Malcom, & Rudy, Alexander. (2017). Joint military and commercial rotorcraft mechanical diagnostics gap analysis. *Proceedings of the American Helicopter Society 73rd Annual Forum* (Paper 36), May 9-11, Fort Worth, TX.

Wade, Daniel & Wilson, Andrew. (2017). Applying machine learning-based diagnostic functions to rotorcraft safety. *Proceedings of the Tenth DST Group International Conference on Health and Usage Monitoring Systems 17th Australian Aerospace Congress* (Paper 071), February 26-28, Melbourne, VIC.

Wilson, Andrew, Wade, Daniel, Albarado, Kevin, Partain, Jeremy & Statham, Matthew. (2016). A classifier development process for mechanical diagnostics on US

Army rotorcraft. *Proceedings of the 1st ML and PHM Workshop, SIGKDD 2016,* August 8-13, San Francisco, CA. DOI 10.475/123_4

APPENDIX A: ADDITIONAL IMBALANCE RATIO STUDY RESULTS

It is also interesting to review how the metrics respond to each method. Each of the following figures shows an individual metric across each of the methods. The y-axes of the plots are kept the same for comparison purposes, from 0.1 to 1.05, and the x-axes are maintained from the previous figures, ratio of zombies plotted on a log scale.

Figure A1 shows the accuracy as a function of imbalance for each of the methods. Note in this figure that the accuracy plots for maximum informedness and weighted accuracy, which are not numerically equivalent, are nearly indistinguishable and that maximizing F1 Score only differs from maximizing accuracy for very small imbalance ratios. All accuracy scores eventually settle on nearly the same value, 0.95. It should be noted that this figure is the inverse of Figure 3, which is the overall absolute error.

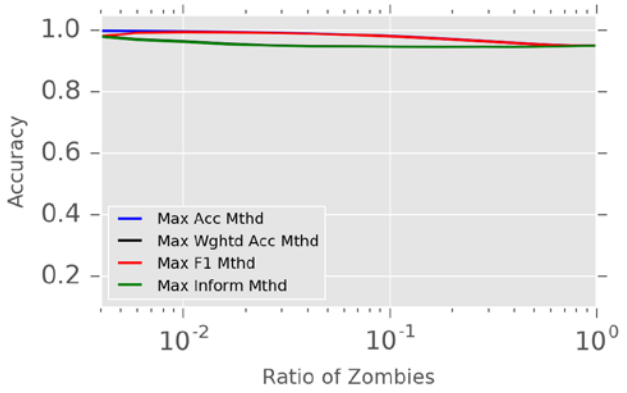

Figure A1. This plot shows the accuracy metric evaluated for each of the four methods to choose a threshold.

Figure A2 shows the weighted accuracy as a function of imbalance for each of the methods. Maximizing weighted accuracy and informedness has the same result on weighted accuracy. Maximizing F1-score results in initially the same weighted accuracy. Finally, maximizing accuracy results in low weighted accuracy until the convergence at the 1:1 ratio.

Figure A3 shows the F1-score as a function of imbalance for each of the methods. In this case, similar shape functions exist between the maximum F1-score and accuracy methods. F1-Score values for all methods are low and continuously approach the same value (0.95) at the 1:1 ratio. The F1-score for the maximum informedness and weighted accuracy methods are nearly identical, except at very low imbalance ratios.

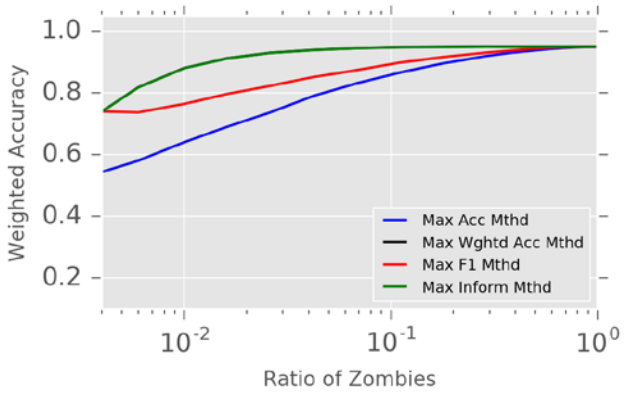

Figure A2. This plot shows the weighted accuracy metric evaluated for each of the four methods to choose a threshold.

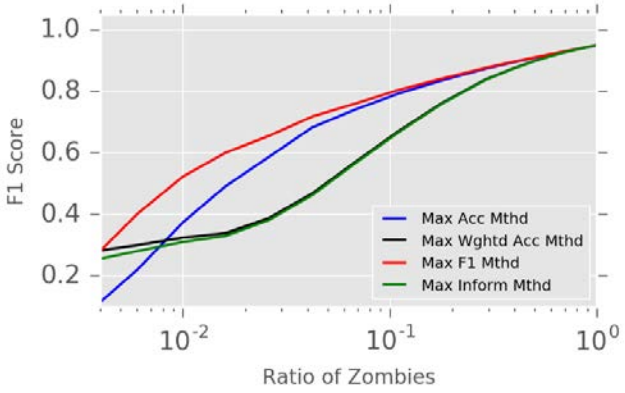

Figure A3.This plot shows the F1-score metric evaluated for each of the four methods to choose a threshold.

Figure A4 shows the informedness as a function of imbalance for each of the methods. Comparing the methods at very low imbalance results in a pronounced difference, with maximum accuracy near 0.3 and maximum informedness at 0.98. The maximum weighted accuracy method is the first to join up with the maximum informedness which started at the same value as the maximum F1-score method.

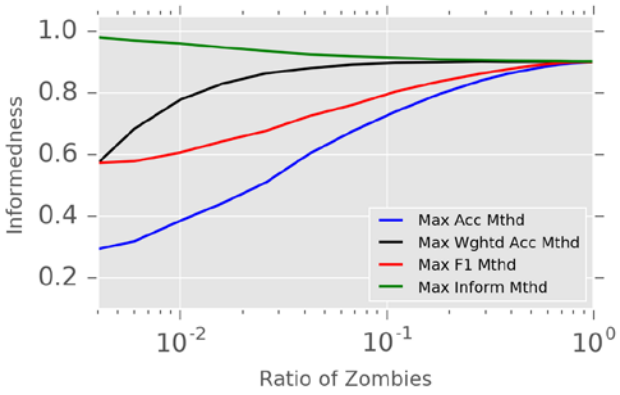

Figure A4. This plot shows the informedness metric evaluated for each of the four methods to choose a threshold.