Proposal of Metrics to Visualise Performance of Prognostics Case Studies in Aerospace

MSc Thesis

Sahil Panse

TU Delft

Proposal of Metrics to Visualise Performance of Prognostics Case Studies in Aerospace

MSc Thesis

by

Sahil Panse

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday, 27th January 2021
at 9:00 AM.

Student number:	4996563
Project duration:	February 2021 – January 2022
Thesis Committee:	
Chair	Dr. B.F. Santos
Supervisor	Dr. M.L. Baptista
Examiner	Dr. B. Chen

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Acknowledgements

I present before you the result of my Master Thesis carried out at the Delft University of Technology in The Netherlands. This thesis also marks the end of journey in the graduate studies at TU Delft. This thesis was a journey that I undertook in February 2021, when I had first called my supervisor Marcia to understand what the idea behind this thesis was. Being under restrictions because of the COVID-19 pandemic was certainly far from ideal and most conversations between Marcia and I had to be on video calls. Nevertheless, the support I had received from her and the timely feedback from Bruno held me in great stead going forward with the completion of my thesis. I would like to express my gratitude to dr. Boyang Chen from the Aerospace Structures and Materials faculty for agreeing to be part of my thesis defense committee.

I would also like to recognize the support I had received from friends and family back home to provide the much needed motivation during times of difficulty in this thesis. A special word of thanks to my close friends here in the Netherlands who made Delft feel like home over the course of the last two and a half years.

Sahil Panse
Delft, January 2021

Contents

List of Figures

List of Tables

Introduction

A large portion of the direct costs incurred by an airline and delays can be attributed to maintenance. The major drivers of maintenance related costs are unnecessary, unplanned maintenance checks or unexpected component or system failure. With rise in the availability of sensor data to monitor system health and the ability to leverage that data into meaningful predictions, traditional maintenance practices such as hard time maintenance are beginning to fade away. The meaningful predictions mentioned earlier are remaining useful life (RUL) predictions that are obtained during the life of a system or a component with the aim that key decisions such as the optimal scheduling of maintenance can be taken based on them. This is achieved with the help of prognostics where the systems health data can be used to train ML models, resulting in RUL predictions.

There is however a concern on whether these RUL forecasts are trustworthy and therefore of any significance to maintenance organizations. A bad RUL forecast can trigger uncalled for failure if the forecast is a late prediction or an unnecessary maintenance check if it is an early prediction. Conventional forecasting metrics such as the mean absolute error, the root mean square error lack the depth necessary to judge a RUL forecast from an operational viewpoint and there is hence the necessity to propose meaningful metrics to evaluate and visualize the performance of prognostics models. The main research objective in this thesis is as follows.

To evaluate and visualize the performance of data-driven prognostics models by proposing modifications to currently available performance metrics.

This report is structured as follows. In Part I, a scientific paper is provided entailing the key aspects of the research. This is a concise version of the report and can be therefore read as a standalone document. In addition, section 2 is an in depth literature review focusing on the performance evaluation of prognostics data-driven prognostics models culminating in a research objective. Finally, Part III provides an overview of the initial version of the proposed project plan during the thesis kickoff.

I

Scientific Paper

Proposal of Metrics to Visualise Performance of Prognostics Case Studies in Aerospace

Sahil Panse
Chair: dr. ir. B.F. Santos, Supervisor: dr. M.L. Baptista
Section: Air Transport & Operations, Controls & Operations Department
Faculty of Aerospace Engineering, Delft University of Technology,
Delft, The Netherlands

***Abstract* - With the increasing availability of sensor data to monitor the health of systems and components, maintenance is encountering a shift from the traditional preventive and corrective methodologies to a predictive approach. Given that maintenance, repair and overhaul are a significant operational cost to organizations, it is imperative that the adoption of predictive maintenance techniques such as prognostics comes to the forefront. Prognostics models ensures that remaining useful life (RUL) estimates are available throughout the life of the component by leveraging sensor data. The obstacle faced by prognostics though, is the lack of standardized metrics, performance evaluation and their application to different case studies. Visualization results in understanding the evolution of the performance over time rather than single point estimates, which helps in identifying performance at crucial points in the life of systems. This paper focuses on assessing and visualizing the performance of prognostics models by addressing the shortcomings of the currently available performance metrics, advancing these metrics to suit the requirements and applying these to visualize the performance of two linear regression models over time. The results show that the modification of currently available performance metrics can enhance performance visualization, leading to better interpretation of the performance of prognostics models.**

I. INTRODUCTION

Condition based maintenance (CBM) has been the trend in the maintenance of systems and components in aviation, replacing preventive maintenance techniques such as hard time maintenance where maintenance takes place after certain time intervals resulting in additional maintenance checks and increasing the downtime of systems. The key aspect of CBM is that maintenance takes place when the system reaches certain unacceptable levels, defined for each system. This is however different from predictive maintenance wherein scheduling of maintenance happens farther in advance owing to forecasts about the remaining useful life (RUL) of the system or component. Prognostics, which is a part of the domain of prognostics and health management (PHM) ensures that forecasts of the RUL of a component are constantly made available throughout its useful life so that a decision can be taken towards the optimal scheduling of a maintenance check, which is becoming the trend increasingly as more sensor data is made available to monitor system health [1].

This monitoring of system health in aircraft is realized with the help of the airplane condition monitoring system (ACMS), involving a number of sensors connected throughout the aircraft [2]. For instance, if we consider aircraft such as the Airbus A350, one would find a total of 6000 sensors on various parts of the aircraft to monitor the health of key subsystems [3]. With the rise in computational power and the development of new data technologies, analysis of such quantities of data has been made possible, which then translates into timely RUL forecasts notifying of impending failure through prognostics. Prognostics is therefore a key tool that provides the potential of cost savings by avoiding unnecessary maintenance checks, which is a significant benefit given that maintenance, repair and overhaul activities form 10% - 15% of an airlines direct operational costs [4].

A simple classification of prognostics approaches would be data-driven models and physics-based models. Data-driven prognostics relies on machine learning models that are trained by the input data at hand, which essentially involve the run-to-failure (RtF) data for the system under study. RtF means that the system is utilized until failure, with the aim being to deliberately allow them to fail. The quality of the dataset available impacts the accuracy of the RUL predictions of data-driven prognostics models [5]. Physics-based models build on the concept of describing the failure propagation or degradation in a component with

the help of mathematical models [6]. Hybrid models on the other hand exists of a data-driven ML approach alongside a model based approach. An example of this is a particle filter approach used to determine model parameters implemented in conjunction with a data-driven neural network model [7].

While a significant amount of publications deal with the development of prognostics models (data-driven prognostics models majorly) [6], there is still work to be done in terms of the performance evaluation of these prognostics models. Performance evaluation of prognostics models mainly deals with judging the quality of RUL predictions over time with regards to a variety of aspects, the most crucial of which is the error in RUL prediction. RUL prediction error is simply the difference between the predicted RUL and the true RUL, hereafter referred to as the *ground truth*.

Until quite recently, there had been a lack of standardized metrics to deal with the performance evaluation of prognostics models [8]. This is partly because of the end expectations of the researcher being model development, metrics have often been used without much thought or emphasis and largely based on the convenience of the researcher [9]. This is especially observed in a large number of publications making use of basic conventional forecasting metrics such as the MSE, MAE, RMSE to evaluate complex data-driven prognostics models that will be discussed further. But prognostics will only continue to play a vital role in maintenance if the users can express a certain level of confidence in the prognostics models being used and only then can they realistically be incorporated in the decision making process when it comes to maintenance [9]. Furthermore, performance evaluation plays a role in the design of prognostics systems as well, where appropriate feedback on performance metrics such as the prediction accuracy for instance can help tune the models further.

This paper focuses on advancing the current set of prognostics metrics by broadening focus on the consequences of late predictions over early predictions for the RUL predictions of data-driven prognostics models. Once the performance evaluation metrics are proposed, the visualization of the different performance parameters of a prognostics model such as the prediction accuracy of RUL predictions is considered. This is because currently, the metrics developed for use in prognostics lack a visual perspective [9] needed to interpret the performance over the life of the system and rather estimate single values to judge the performance of the models and compare them. This is especially crucial for metrics that yield a single estimate as a measure of performance such as the MSE, RMSE or MASE. A quantitative assessment from these metrics is not sufficient to judge the performance of prognostics models because most of the aforementioned metrics only consider the average and provide no indication of the performance evolves over time. For model development or model selection purposes for a prognostics case study, it would be beneficial for the performance to provide such kind of feedback. When dealing with a fleet of systems (multiple units), visualization can also provide the user with a schematic representation of the performance of each of the units, which is a part of this research.

The structure of this paper is as follows. Firstly, section II will deep dive into the work carried out by previous researchers pertaining to prognostics performance evaluation techniques, a critical analysis of the same culminating in the formulation of a research question. This will then be followed by section III which will explain the methodology based on which the metrics have been incorporated for the visualization of two prognostics models. Then section IV will focus on actual model performance visualization resulting from the framework of metrics defined in section III along with the interpretations of the results in section VI. Certain limitations of the metrics defined in this paper have been outlined in section V followed by recommendations for future work in section VII.

II. LITERATURE REVIEW

Performance evaluation has consisted of significant gaps in terms of research work in prognostics and health management (PHM) of systems and components. Lack of standardized metrics has been a hindrance in the progress of PHM and the metrics that currently exist have largely not been applied to prognostics algorithms to assess their validity, which is crucial for prognostics to expand in a large operational scenario to aid decision making. There also exists the issue of the lack of visualization of prognostics performance over time, with the main focus currently being on metrics such as RMSE, MSE, MAE, scoring functions providing only a single value at the end of life to assess prognostics performance. It would therefore be prudent to further study the progress of performance metrics that are applicable to prognostics from a research point of view. There have been metrics outside the prognostics domain that have been considered because of their applications in general forecasting techniques. From an engineering research point of view as pointed out by [10], algorithm performance metrics and computational performance metrics are of particular importance to evaluate a prognostics algorithm from a research viewpoint.

The best estimate of the RUL obtained from a prognostics algorithm is then compared to the ground truth RUL. According to [8], the ground truth can be defined as the best belief of the true value of a certain variable, which in this case is the RUL. Publications reviewed in the literature study were found to deal with deterministic RUL predictions and probabilistic RUL predictions. Conventional forecasting metrics such as the MSE, MAE, RMSE and the MASE in prognostics applications have also been implemented in performance evaluation, the drawbacks of which will be discussed here. Hence, the performance metrics in this review will be classified into *prognostics-based metrics, conventional forecasting metrics and probabilistic forecasting metrics*. The scope of this research pertains to deterministic forecasts and hence, only a brief overview is provided for probabilistic forecasting metrics.

A. Prognostics Based Metrics

As is the case in most forecasting scenarios, not all algorithms are ideal. This means that there is often a difference between the parameter estimated from the algorithm and the ground truth. In prognostics, this introduces a key aspect in performance evaluation known as the prediction error. Some algorithms denote this as the difference between the true RUL and the predicted RUL whereas other algorithms assume the other way around. Therefore, error-based metrics are of particular importance when assessing performance. A large number of publications dealing with these metrics often entail prediction horizon (PH), $\alpha - \lambda$ accuracy, relative accuracy and convergence [8][10]. While these metrics are valid for point forecasts, their use in probabilistic RUL predictions is more abundant and hence will be discussed in subsection II.C. An overview of the classification of performance metrics based on the requirements of the researchers has been provided in [6].

The most basic of these metrics which will be discussed in subsection II.B is the RMSE used by [11] to assess their data-driven prognostics model. Another metric used in the same paper is the confidence interval which focuses on the accuracy of predictions as opposed to the error estimate provided by the RMSE. The confidence interval (CI) metric introduced by [11] deals with the accuracy of the predictions as compared to the error estimates that RMSE deals with. A narrower CI would indicate better performance in terms of accuracy with a higher number of RUL forecasts lying within a small interval.

PH and $\alpha - \lambda$ accuracy are the most common prognostics evaluation metrics and have been used in a variety of publications. [6] define PH as the difference between the time when the predictions satisfy an accuracy criterion known as the β criterion and the end of life. Since these two metrics have also been applied to suit probabilistic forecasts, a greater emphasis has been provided in subsection II.C.

An estimate of the RUL can either be a late prediction or an early prediction. If the predicted RUL is greater than that of the ground truth, this is classified as a late prediction. If a greater RUL estimate is predicted, this runs the risk of component failure during its life because its real failure is earlier than the predicted failure. When critical systems are considered in a fleet such as aircraft engines, for instance, this would be a great risk. It would in such a case be beneficial to deal with early predictions in which the RUL prediction indicates failure is earlier than reality. In such a case though, maintenance would have to be scheduled earlier incurring additional costs and downtime for the component. This concept of early and late predictions has been captured by scoring functions in prognostics case studies. Each RUL prediction is awarded a certain value known as the score (or penalty), with the aim being to minimize the overall penalty for all predictions.

The most prominent example of such a scoring function in prognostics literature is Eq. 1 has been highlighted in [12]. Here, s is the score, d is the difference between the true RUL and the estimated RUL and the arbitrary constants are set at $a_1 = 10$ and $a_2 = 13$. The asymmetric scoring function aims to penalize early predictions more severely as compared to late predictions.

$$s = \begin{cases} \sum_{i=1}^{n} e^{-(\frac{d}{a_1})} - 1 \text{ for } d < 0 \\ \sum_{i=1}^{n} e^{(\frac{d}{a_2})} - 1 \text{ for } d \geq 0 \end{cases} \quad \text{(see [12])} \qquad (1)$$

The one shortcoming in most publications making use of scoring functions is the lack of explanations on how the constants in Eq. 1 have been determined. If the argument has to be made that early predictions are worse than late predictions, substituting the above constants in Eq. 1 does not yield ideal results for certain cases. Consider prediction errors of +50 and -50. The above scoring function then awards a higher score to the prediction of -50 than +50. In the evaluation of their deep convolution neural networks model, [13] chose to interchange the constants by assigning a_2 to late predictions and a_1 to early predictions. If the idea is to penalize late predictions more than early predictions, the above scoring function fails.

In another model, [14] produced a scoring function where $a_1 = 10$ and $a_2 = 4.5$ for their prognostics model. A comparison of the two scoring functions used in [12] and

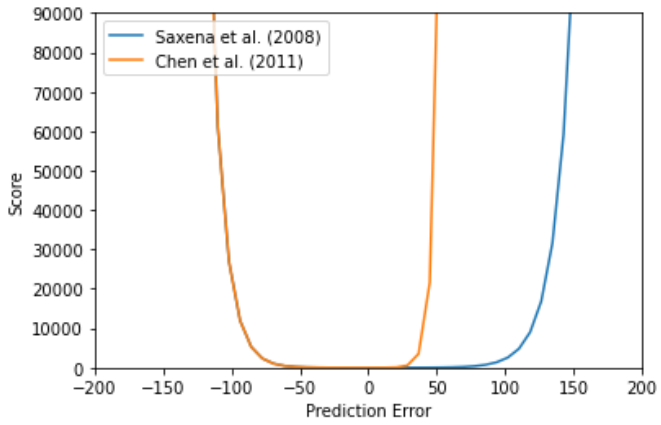


Figure 1. Comparison of scoring functions in [12] and [14].

[14] was plotted against the prediction error in Fig. 1. In Fig. 1, the left hand side of the graph wherein the prediction error is negative indicates a similar response for both the case studies. However, late predictions are penalized far more severely in [12]. This may attributed to the fact that the system under test in [14] was a cutting tool which was a less critical component as opposed to an aircraft engine in [12], resulting in a more conservative approach in [12]. But the no justification has been provided on why the particular parameters (a_1 and a_2) were selected for the two case studies, hence it is only possible to speculate. A suggestion in this case would be to discuss the impact of different constants on the scoring function and how the score then varies with error. This may lead to the selection of optimal constants.

In the case that the scoring function proves indecisive because the models receive similar scores, [12] propose the coefficient of correlation where a high correlation is preferred. However, [15] argue that coefficient of correlation (r) cannot be used to determine the quality of forecasts as this measure only determines the goodness of fit between data.

B. Conventional Forecasting Metrics

Since prognostics is eventually a forecasting technique, it might also be prudent to glance over publications that discuss general forecasting metrics. Given that such metrics do not take into account whether an RUL prediction is an early or late predictions, they cannot be prioritized over the metrics that were discussed in subsection II.A. The metrics reviewed here will include MSE, MASE, MAPE to name a few.

The mean absolute scaled error (MASE) was first proposed by [16] to be used as the standard for the comparison of multiple time-series in forecasting. They compared different forecasting techniques considering three different time series datasets. Metrics that had been used before the publication of the paper had been applied to the time series including MAPE, median absolute percentage error (MdAPE), symmetric MAPE (sMAPE), symmetric MdAPE (sMdAPE), median relative absolute error (MdRAE), geometric relative absolute error (GMRAE) and MASE. The authors also highlight that most publications until that point recommended the use of MAPE for determining forecast accuracy.

A common trend observed in the study was that when a forecast for a particular time index is zero, all the corresponding error metrics (mentioned above) take an undefined or an infinite value, not indicating the quality of the forecast. This shortcoming has also been highlighted by [17] as a gap for their proposal of a new metric. This resulted in the new MASE metric shown in Eq. 2 and Eq. 3, where q_t is the scaled error e_t is the error in the forecast whereas the difference between two successive forecasts is given by $Y_i - Y_{i-1}$.

$$q_t = \frac{e_t}{\frac{1}{N-1}\sum_{i=1}^{N} Y_i - Y_{i-1}} \tag{2}$$

$$MASE = |q_t| \tag{3}$$

The key characteristic that hinders MASE as a metric for prognostic applications is that the mean of the scaled error considers only absolute values of q_t. This means that only positive errors are considered and as has been mentioned previously, we need to classify errors as positive or negative to determine whether an RUL prediction is an over-prediction or an under-prediction to be able to study the extent of how bad a prediction is from an operational viewpoint.

Also, [18] do illustrate the shortcomings of conventional forecasting metrics such as MASE, MSE, etc. in their battery capacity degradation prognostics models. Four prognostics models were used for the analysis, evaluated using PH, $\alpha - \lambda$ accuracy, relative accuracy, cumulative relative accuracy & convergence alongside conventional forecasting metrics such as bias, MAPE, MSE, etc. The results conclude that the latter set of metrics provides an insight into the evolution of the predictions over time and when these predictions are trustworthy, as opposed to the conventional metrics which only provide a measure of deviation from the ground truth. These error metrics, therefore do not account for progressive acceptable accuracy and precision levels such as the prediction horizon and the $\alpha - \lambda$ accuracy introduced in subsection II.A [19].

An interesting feature was observed in [13] when the scoring function discussed in subsection II.A was used in conjunction with the root mean square error to validate their prognostics algorithm. The values of RMSE seem to be more symmetric about the zero error mark than in the case of the scoring function. For example, there seems to be little distinction between an error of -20 and +20 in the RMSE which is inherently incorrect because of the risk of component failure associated with a positive prediction error.

C. Probabilistic Forecast Metrics

In an idealistic scenario, RUL forecasts obtained will be point estimates resulting in a single valued RUL forecast at each time index. However in prognostics, estimation of the remaining useful life of a component is of relatively lesser importance without considering the uncertainties associated with making such a prediction [8]. Prognostics uncertainties have been categorized as those that arise due to lack of understanding of the true state of the system, future uncertainty (lack of knowledge of future loading and operating conditions) and modeling uncertainty by [20]. Therefore, it is likely that an RUL prediction will generally have a PDF associated with it and a confidence bound [21]. What this does, is that there is a need to also rely on metrics that can evaluate a prognostics algorithm based on a probabilistic prediction, rather than a point estimate of RUL. The incorporation of a PDF also allows for propagation of uncertainties for subsequent predictions [22].

This then makes it crucial to enable the forecaster to predict correct probabilities. Therefore if we consider probabilistic forecasting in prognostics, a prediction may be associated with different RULs at that particular time points with different corresponding probabilities. If the most probable outcome matches the ground truth, this would be an example of a good forecast. It might then be prudent to consider the suggestion of [8] who mention that a point estimate can be extracted from the probability distribution of RUL forecasts. Yet even without converting probabilistic forecasts into a point estimate, there are methods and metrics to assess probabilistic forecasts.

First and foremost, the metrics introduced in subsection II.A have been adapted to suit probabilistic forecasts. Hence, this would be an appropriate starting point. Prediction horizon and $\alpha - \lambda$ accuracy have been modified to suit probabilistic forecasts as has been highlighted by [8][10] by altering the criterion to classify a RUL PDF as accurate. When it comes to prediction horizon, the earlier a prediction satisfies a condition known as the β-criterion, the better is the accuracy of the prediction algorithm. Hence, the β-criterion essentially determines the first time instant wherein a significant portion of the probability mass of the RUL prediction PDF falling inside the predetermined shaded area is greater than a threshold value β.

Similarly, the $\alpha - \lambda$ accuracy metric has been adapted to suit probabilistic forecasts as well. The same β-criterion holds when it comes to determining whether a RUL prediction PDF is accurate or not. A prediction at any time instant λ that falls within the cone of accuracy (i.e, satisfying the β-criterion) is classified as accurate [8]. It was pointed out by [12] that the increasing accuracy in RUL predictions should be observed closer to the end of life of the component and $\alpha - \lambda$ accuracy highlights this point. It can be seen that the cone of accuracy shrinks with time leading to increased accuracy of the predictions and precision of the PDF. This is an improvement over the prediction horizon which solely gives an indication of the time point when the β criterion is satisfied. It could have been interesting however, if the author had included an analysis on how these results vary for different values of α or how an appropriate α could be set for the analysis. This is because a higher accuracy level will increase the size of the bounds that are imposed on the forecasts. This would mean that the algorithms that perform relatively poorly for a 20% accuracy for instance may be considered acceptable for a higher value of α. Understandably, the accuracy level that needs to be set is up to the user and that factor needs to be considered when developing prognostics models.

While there have been no contributions towards prognostics, the domain of scoring rules has gained popularity when it comes to evaluating probabilistic forecasts. These may be considered similar to scoring functions, but how a score is assigned is vastly different. An overview of the different scoring rules used to evaluate probabilistic forecasts is presented by [23] & [24], who define a score as shown in Eq. 4. Here, X is used to represent the observed value of a random variable (ground truth) and $p(x)$ represent the probabilities in the forecast for each value x. A lower overall score is preferable in the case of scoring rules.

$$S = \frac{1}{N}\sum_{i}^{N} S[p_i(x), X_i] \text{ (see [23])} \tag{4}$$

The key difference scoring rules have as compared to scoring functions is that instead of considering the prediction error, scoring rules take into account the probability or the confidence with which a set of predictions are made. A good score will be assigned to predictions that predict the ground truth RUL with high probabilities.

Table 1. Overview of reviewed metrics and their limitations

Metrics	Limitations
Scoring function	Current scoring function has certain instances wherein early predictions are penalized more than late predictions
Prediction horizon	No distinction for early predictions and late predictions. Also, it only provides the first instance wherein the accuracy criterion is satisfied.
$\alpha - \lambda$ performance	No distinction for early predictions and late predictions.
MSE, MAE, RMSE	Mean value considered which is not a true representation of performance. Also, only the magnitude of the RUL prediction error is considered.
MASE	Same as for MSE, MAE, RMSE, scaled error also provides no distinction between early and late predictions
Scoring rules	No limitations as such since scoring rules are not present in prognostics literature

From this literature review, it can be established that there have been a number of metrics developed to evaluate the performance of prognostics models. Nevertheless, the gaps discovered in this review pertaining to the current set of metrics from a prognostics point of view are significant and deserve to be worked upon, a summary of which is tabulated in Table 1. Lack of standardized metrics had been brought out as a research gap by [8] in their paper proposing a variety of metrics to assess prognostics models. Currently, there is a scarcity of visualization methods as well to able to effectively compare multiple modeling approaches in addition to the gaps within the current set of metrics themselves [9].

For prognostics to be adapted by maintenance organizations more effectively, the RUL predictions made by prognostics models need to be trustworthy because of the consequences that can occur taking into account a bad prediction. Poor RUL forecasting may also result in suboptimal maintenance check scheduling of systems and the interpretation of feedback from metrics that feature gaps can also contribute to sub optimal prognostics model development. Considering the above remarks, the following research question is proposed.

How can currently available performance evaluation metrics be modified to evaluate and visualize the performance of prognostics algorithms?

III. METHODOLOGY

This section of the paper focuses on answering the first half of the research question. The modification of currently available performance metrics has been discussed in this section, wherein a new scoring function is proposed alongside the $\alpha - \lambda$ performance to assist the performance visualization. Since the visualization is a part of the performance evaluation of the models, the results of visualization will be illustrated in the results in section IV, whereas the methodology used for the visualization will be discussed in this section. The scoring function is chosen because an important area of focus is the preference of having early predictions over late predictions for this case study whereas the $\alpha - \lambda$ accuracy ensures the predictions closer to the end of life are stressed upon. Poor predictions close to the end of life are more critical to the user of the prognostics system because they leave little time to take corrective action on the system before failure.

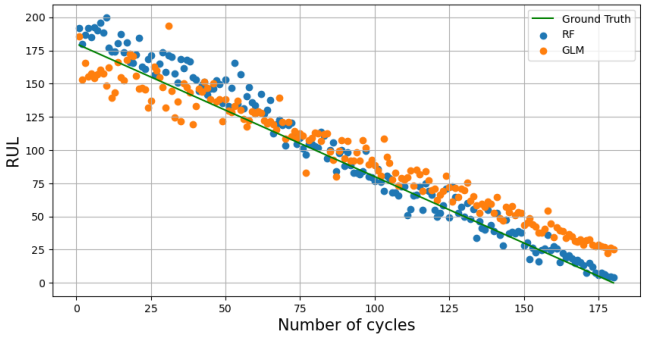
These performance metrics have been developed with the help of deterministic RUL predictions from two linear regression models. The random forest (RF) model was incorporated alongside a generalized linear model (GLM) with a normal distribution function and a log-link to get the required RUL predictions. With the aim being model performance visualization and evaluation rather than model development, these simple models producing a time series of RUL predictions are sufficient for the purposes of this paper.

The case study being considered for this paper is the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset which is used to train these two regression models. This particular dataset is fairly

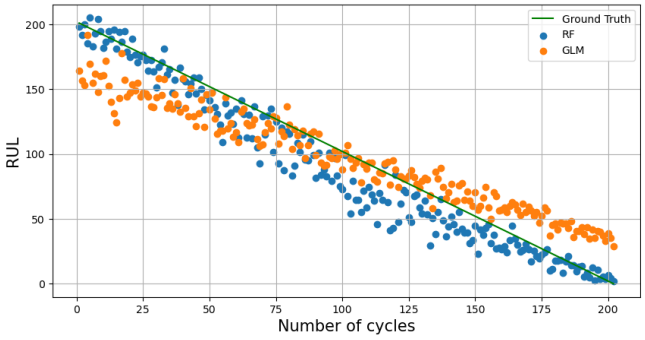
simplistic in nature and is well suited to train basic machine learning models such as the models chosen. Some of the sensors in this dataset exhibit monotonic (non-increasing or non-decreasing) trends with their sensor readings, which is important for better RUL predictions. The availability of the ground truth RUL for each engine was another reason for the selection of the C-MAPSS dataset.

The aim is therefore to assess and visualize the performance of both these models based on the metrics that will be proposed in this section. Before proceeding with the methodology of the performance metrics outlined in subsection III.B & subsection III.C, there needs to be a more detailed understanding of the dataset being used, which has been provided in subsection III.A.

Within the dataset, once the test dataset is chosen, the RUL predictions over time can be studied for all of the test units, or units under test (UUTs) separately for each simulation. As an example for one of the simulations, RUL predictions over the life of two UUTs are illustrated in Fig. 2. These plots were obtained by separating the data based on each test unit after both models had been trained using the C-MAPSS training data. The section will now be



(a) RUL predictions varying with number of cycles for UUT 14



(b) RUL predictions varying with number of cycles for UUT 22

Figure 2. Remaining useful life predictions varying with number of cycles

divided into subsections discussing the metrics that have been developed to visualize the performance of the two models, with subsection III.B discussing the formulation of the case-study specific scoring and subsection III.C discussing the $\alpha - \lambda$ accuracy.

A. Case Study

C-MAPSS is a high fidelity simulation system based on which the datasets introduced in section III are created. The dataset consists of multiple multivariate time series, which are subsequently divided into four different sub-datasets, each with different operational conditions and fault conditions. The first of these sub-datasets ($FD001$) is considered is sufficient for this paper. Each time series in the $FD001$ dataset relates to the same engineering system which is an aircraft turbofan engine. Each engine begins with varying degrees of initial wear, which is different from a fault condition. A fault is then considered to be developed at some point during the time series which grows in magnitude until failure in the training set. In the test set, the time series ends before failure and the goal is to then predict the RUL at each time point in the test set.

The training dataset includes operational data of a certain number of the turbofan engines, which can be varied using a different train-test split. Next, the FD001 sub dataset also makes use of three engine operational settings (namely the altitude, Mach number and power setting) and 21 sensors that monitor quantities such as the engine fan speed, pressure ratios, bypass ratios, etc. Each of the 100 units in the FD001 sub dataset possesses readings corresponding to these sensors, with the final column of the dataset being the ground truth RUL.

In each dataset, the engine is run for a different number of cycles until failure takes place. Since the length of the runs for each engine are different with a maximum number of cycles of 362 corresponding to the 69^{th} engine in the dataset, the number of cycles corresponding to each test unit will need to be normalized to be able to compare different prognostics models using performance visualization techniques. This has been achieved by converting the number of cycles of each UUT in the test dataset to percentage so that the cycles corresponding to each unit vary from 0% to 100% of the unit life. For the i^{th} observation corresponding to each engine in the dataset, this conversion takes place as shown in Eq. 5, where t_i is the number of cycles elapsed at time point i and N is the total number of cycles for the particular UUT.

$$[\text{Time (in \%)}]_i = \frac{t_i}{\text{N}} \cdot 100 \tag{5}$$

B. Scoring Function

To get an overview of how better or worse deterministic RUL predictions are to each other, the first metric proposed is the scoring function. A scoring function can be considered as a cost function wherein the cost has to be minimized for all the test units in the simulation. The cost is a penalty (or score) applied to each RUL prediction at each instant of time (or cycle) where a low overall score is preferred. For the C-MAPSS case study, the preference is to score late predictions more severely than early predictions for the reason that overestimating the RUL brings the possibility of failure to the aircraft engine as opposed to an early prediction in which case, the integrity is not compromised. As an improvement to the currently used scoring function

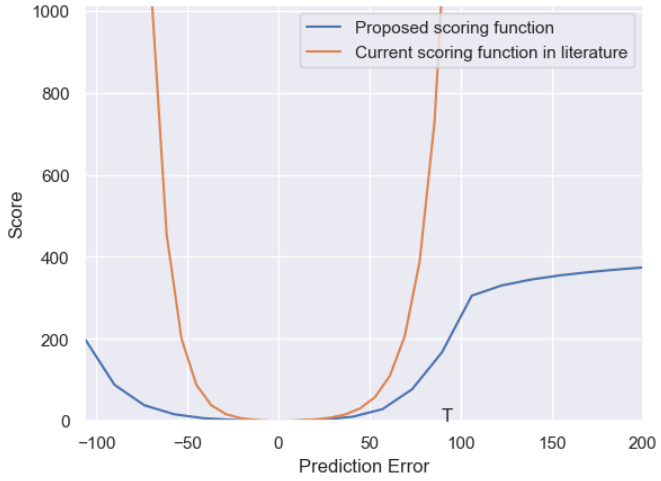


Figure 3. Proposed scoring function with a tolerance of 90 cycles in comparison to the current scoring function

in prognostics literature [12] (Eq. 1), this paper proposes the use of the scoring function highlighted in Fig. 3. An exponential profile is used to score early and late predictions as can be observed. Note that the parameters set for the proposed scoring function illustrated in Fig. 3 are only an example. The proposed methodology for the determination of these parameters will be discussed further.

One major difference in this paper however is the concept of *tolerance*, which is shown on the horizontal axis (T). This is a buffer that has been provided for late predictions. From an error of zero to an error equivalent to tolerance, the score rises exponentially. To understand the reason for proposing tolerance and its significance, consider a small late prediction of $+y$ flight cycles and assume a tolerance of x flight cycles. In such a scenario, a maintenance check could be scheduled at $(y - x)$ flight cycles because $(y - x)$ would be a point before the zero error mark in Fig. 3, which averts the risk of failure of the system as long as $y < x$. This ensures that the component avoids failure and a state of functionality is maintained, which is particularly crucial for critical systems in a fleet. This enables certain late RUL predictions to be considered as acceptable given the fact that the maintenance would be scheduled accordingly.

Beyond the tolerance (T) in Fig. 3, there is no reason to exponentially increase the score further. This is because given a tolerance of +50 as observed in Fig. 3 an error of +80 for example, is not too different from an error of +180. From the point of view of a maintenance organization, trusting either of those two late predictions will lead to the component failing in any case and hence the score should not be too different for both such cases.

Hence beyond tolerance, instead of an exponential rise in the score, a logarithmic trend is proposed which results in only a marginal increase in score with respect to increasing error. A constant score beyond tolerance however is not considered because the algorithms need to be evaluated not only from the point of view of a maintenance organization but from a research viewpoint focusing on how close the model predicts to reality. So there still needs to be a distinction between two RUL predictions at different stages beyond tolerance.

From multiple simulations, it was observed that the range of the prediction errors kept varying significantly. Hence, the limits for the prediction error (x axis) in Fig. 3 were considered as the least error and the maximum error observed for each simulation. From the above discussion, the scoring function with S_i being the score at time index i is as follows with d_i being the prediction error at time index i, T being the tolerance and p_T being the score at the tolerance point.

$$S_i = \begin{cases} e^{-(\frac{d_i}{a_1})} - 1, d_i \leq 0 \\ e^{(\frac{d_i}{a_2})} - 1, 0 < d_i \leq T \\ p_T + 40 \cdot log_4(d_i - T), d_i > T \end{cases} \tag{6}$$

Here, the constants a_1 and a_2 determine the extent to how much early and late predictions are penalized and this depends on the case study being considered. There is therefore need to find the optimal constants for each unique simulation as such a methodology does not exist currently. This is also the case with the tolerance buffer, which is not a preset value but rather depend on the case study being considered. Hence, subsubsection III.B.1 and subsubsection III.B.2 will present a methodology to estimate these parameters.

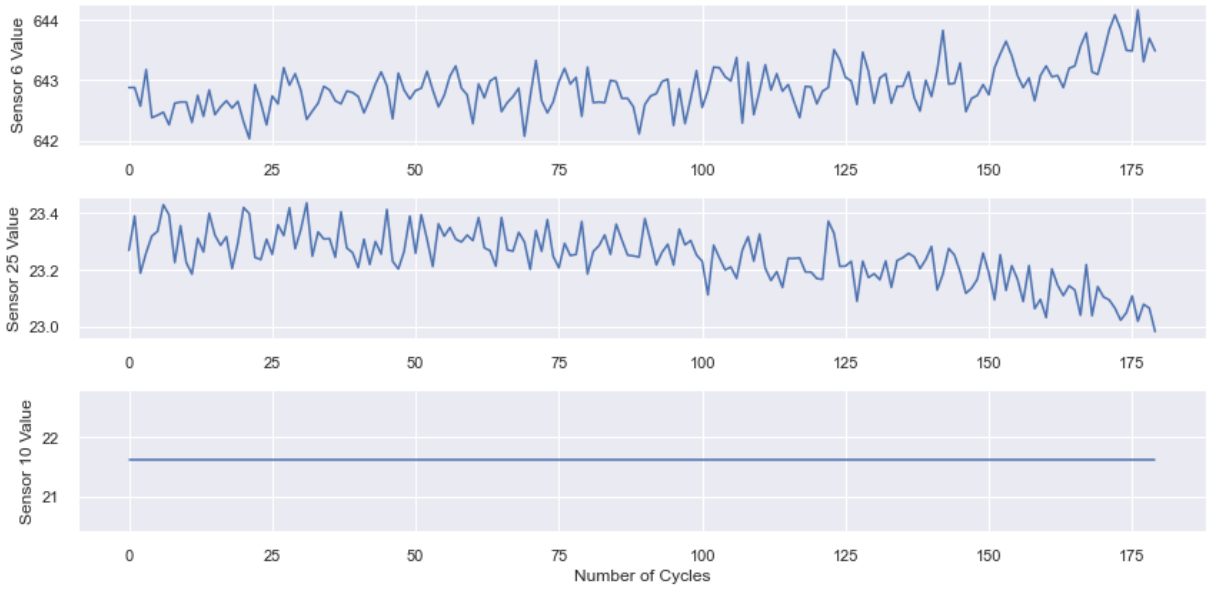


Figure 4. Non-monotonic trends of sensors 6 & 25 with monotonic trend of sensor 10 in C-MAPSS

1. Determination of Constants

The parameters a_1 & a_2 in Eq. 6 quantify the exponential nature of the rise in scores for early predictions and late predictions until the tolerance. Parameter a_1 controls the rise of the exponential penalty in the case of early predictions and a_2 does the same for late predictions until tolerance. The idea here is to fix a value for a_1. Consider that $a_1 = 20$ has been used as a fixed value to obtain a_2 with the assumption that the maximum score assigned to the early predictions *equals* the maximum score assigned to the late predictions at the tolerance point. This is a reasonable assumption because the rise in penalty for the early predictions can be a bit more gradual, but has to be sudden for late predictions until the tolerance point, given the risks involved with late predictions.

If d_{min} is the highest early prediction in magnitude, this results in Eq. 7 when Eq. 6 is evaluated at $d_i = d_{min}$ and $d_i = T$. Then, a_2 can be calculated from Eq. 8.

$$e^{-\frac{d_{min}}{a_1}} - 1 = e^{\frac{T}{a_2}} - 1 \tag{7}$$

$$a_2 = \frac{20 \cdot T}{-d_{min}} \tag{8}$$

2. Determination of Tolerance

To determine the tolerance for a particular simulation, this paper considers the training data available for model training. Given that this case study deals with run to failure datasets where the system is allowed to fail, the system health cannot exhibit self healing properties and can only degrade. Therefore treatment of these non-monotonic trends as observed for sensors 6 and 25 in Fig. 4 by converting this trend into an entirely non-increasing or non-decreasing trend (as seen in sensor 10) will make the data much more suited for training the models thereby resulting in more accurate RUL predictions. Note that Fig. 4 shows the sensor readings for only one particular training unit. This theory motivated the use of a metric that could quantify a degradation trend or a non-increasing trend for the sensor data. This resulted in choosing monotonicity as the metric (m), which quantifies the non-increasing or non-decreasing nature of a trend wherein a monotonicity of 0 shows a perfectly non-monotonic data and vice versa. A prognostics model that is trained with largely non-monotonic data finds more difficulty in producing good RUL predictions as opposed to monotonic data because monotonicity of the input data is an important criterion in prognostics for model training when it comes to run to failure datasets.

This then provided the motivation, to pair RUL predictions wherein the models have been trained with monotonic data with a lower value of tolerance for the scoring function than for non-monotonic data. Therefore, the purpose of this section is to establish a feasible method to determine the tolerance on the basis of the input sensor data by using monotonicity as the metric to assess the quality of data, which is defined in Eq. 9 and is calculated for each of the sensors in the C-MAPSS dataset [19].

Here, x_j refers to the sensor readings which are observed at two successive time points k and $k + 1$. The derivatives are computed as the difference between these two successive

observations in each sensor per UUT.

$$m = \frac{1}{M}\sum_{j=1}^{M}\left|\sum_{k=1}^{N_j-1}\frac{sgn(x_j(k+1)-x_j(k))}{N_j-1}\right| \quad (9)$$

The sign function in Eq. 9 is then used to quantify the non-decreasing or non-increasing trend at that point as is shown in Eq. 10.

$$sgn(x_j(k+1)-x_j(k)) = \begin{cases} 1, (x_j(k+1)-x_j(k)) \geq 0 \\ -1, (x_j(k+1)-x_j(k)) < 0 \end{cases} \quad (10)$$

For each simulation, the maximum positive prediction error that is observed in either of the two models is considered to be the tolerance for that particular simulation provided the data is perfectly non-monotonic as is shown in Eq. 11. This assumption is made because a non-monotonic dataset that is contaminated with noise gives the least chance for the models to be trained accurately for RUL prediction.

$$T_{m=0} = max(\text{max error RF, max error GLM}) \quad (11)$$

From that point onwards, a linearly decreasing trend is proposed for the variation of tolerance with increasing monotonicity as shown in Fig. 5. As an example, the plot in Fig. 5 is a result of the fact that a maximum prediction error of 87 cycles was obtained from both models which is considered as the corresponding tolerance for a perfectly non-monotonic trend. The early RUL predictions are not considered in this process since tolerance accounts only for the late predictions. The following section now describes the proposed methodology to estimate the monotonicity for the entire C-MAPSS dataset using the training data from which tolerance can be determined.

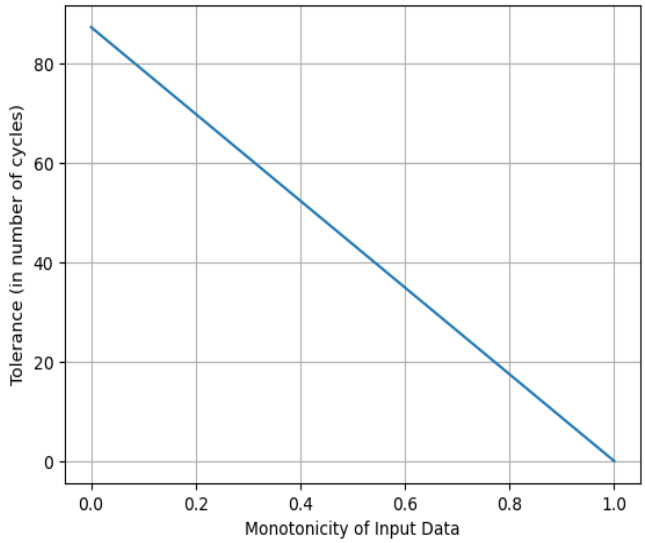


Figure 5. Variation of Tolerance with respect to Monotonicity of Input Data

3. Computation of Monotonicity of the C-MAPSS Dataset

To compute the monotonicity of the training set, the process outlined in Fig. 7 has been proposed, where the monotonicity for each train unit is computed using Eq. 9. The single valued discrete sensors with constant sensor measurements throughout do not provide any information for prognostics and are hence not considered in this process. The first step in Fig. 7 estimates the monotonicity at each time k. This depicts the non-increasing or non-decreasing nature of the trend at every k as is shown in Fig. 6 for the third sensor in the dataset for a particular train unit. As mentioned in subsubsection III.B.2, the monotonicity is computed with a window size of one, i.e, for every successive sensor reading. For the first two readings, the monotonicity will as expected, be 1. From the third observation ($k = 2$ & $k + 1 = 3$), the monotonicity estimates reveal more about the nature of the trend. Subsequently, the average monotonicity for a singular train unit is estimated. The monotonicity obtained from the

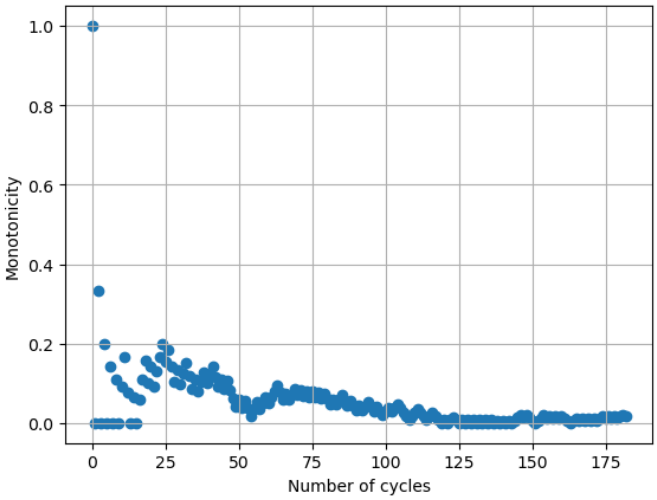


Figure 6. Variation of monotonicity for sensor 3, train unit 99

process outlined is then used to obtain the corresponding tolerance from Fig. 5, which is then used to determine the profile of the scoring function.

4. Performance Visualization

Once the scoring function formulated in this section is applied to the RUL predictions for the particular simulation, there are two methods used in this paper to visualize the performance, illustrated in subsubsection IV..1. Firstly, a more general way is provided to visualize performance where the concentration of errors is illustrated along side its contribution to the overall score. Since the RUL prediction errors for each simulation have a fairly large range, the errors have first been converted into a percentage scale where early predictions vary from 0% to -100%

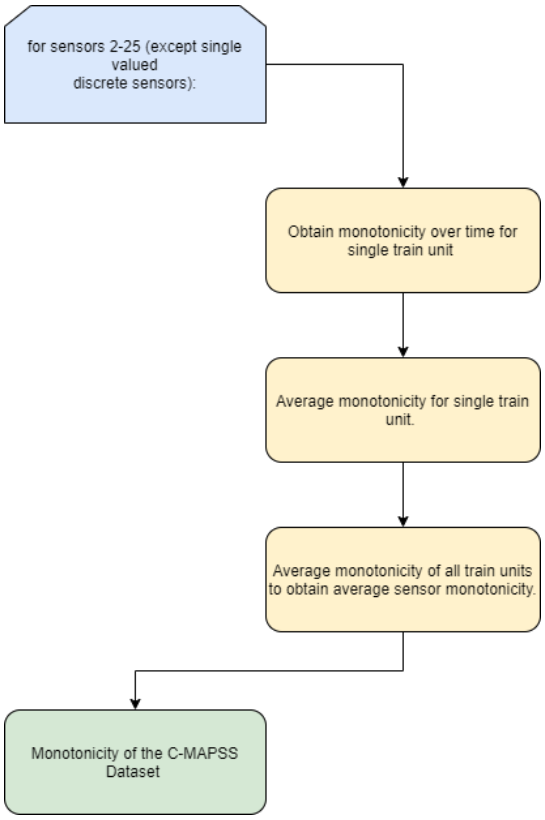


Figure 7. Methodology used for calculating the monotonicity of the C-MAPSS dataset

of the largest early prediction in magnitude whereas late predictions vary from 0% to 100% of the largest late prediction. This method then groups the percentage of average errors and the corresponding score into these intervals, over time. The reason to propose this method is to observe which type of errors result in a larger score.

Secondly, the score is grouped according to the time interval of the predictions (time converted to percentage using Eq. 5) and the corresponding score is observed for each UUT. This enables the user to understand which unit in the fleet is under performing for the particular prognostics model and whether the score is increasing at a critical period or not. If there is a high score observed towards the end of life of the unit, this then leaves no chance to avoid failure. Bad predictions early in the life of the unit are considered relatively better in the sense that the predictions may get better with time, also leaving enough time to take corrective action if confidence is shown in such predictions.

C. The $\alpha - \lambda$ accuracy

As has been discussed, the $\alpha - \lambda$ accuracy imposes bounds on each set of RUL predictions over time wherein any prediction found to stay within the bounds can be classified as accurate. These bounds take the shape of a cone, with a vertex con There is a need to evaluate the performance towards the end of life of the units because this is the most critical time period in their useful lives. The significance of end of life predictions is that a poor prediction very late in the life of a system is dangerous because it leaves little time to take any corrective action before failure. Since this concept is not explored by the scoring function proposed in subsection III.B, the $\alpha - \lambda$ accuracy is used as the other performance metric in this paper. The motivation for this step is because while the current $\alpha - \lambda$ metric can classify whether a prediction is accurate or not, it does not estimate the accuracy of a set of predictions.

The methodology in this paper first aims to determine the $\alpha - \lambda$ accuracy rather than simply determine whether a prediction is classified as accurate or inaccurate, which is described in subsubsection III.C.1. Next, subsubsection III.C.2 discusses the tools proposed for performance visualization of this metric once accuracy has been determined. Here, the accuracy is determined taking into account each prediction made at every cycle of each UUT, instead of taking RUL predictions only from well defined time points λ. This allows all sets of RUL predictions to be taken into account when determining accuracy.

1. Determination of Accuracy

The first step here will be to focus on firstly imposing the bounds on the RUL forecasts over time resulting in the cone of accuracy for different values of cone sizes or accuracy levels (α). The bounds are calculated with the help of Eq. 12 and Eq. 13, where α is a predefined accuracy value, $r_*(i_\lambda)$ is the ground truth RUL at each cycle i within the FD001 dataset. Fig. 8 has been provided as a reference to illustrate this concept.

$$\alpha^+(i) = r_*(i_\lambda) \cdot (1 + \alpha) \tag{12}$$

$$\alpha^-(i) = r_*(i_\lambda) \cdot (1 - \alpha) \tag{13}$$

Firstly, to classify any RUL prediction at a time instant as accurate, the condition needs to be met that the prediction lies within the cone of accuracy. Mathematically, this is described by Eq. 14 for an instant i where $r(i_\lambda)$ is the predicted RUL at i. Next, this paper takes into account the accuracy of RUL predictions for each UUT. This can be defined as shown in Eq. 15, with N being the number of cycles for each UUT.

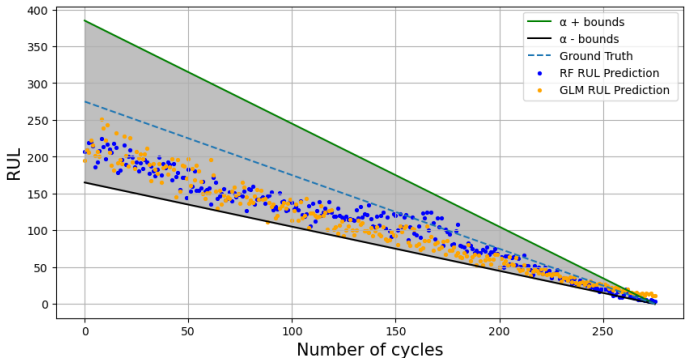
$$(\alpha - \lambda)_i = \begin{cases} 1, & \alpha^+(i) \leq r(i_\lambda) \leq \alpha^-(i) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$Accuracy = \frac{\sum_{i=1}^{N} (\alpha - \lambda)_i}{N} \cdot 100 \tag{15}$$

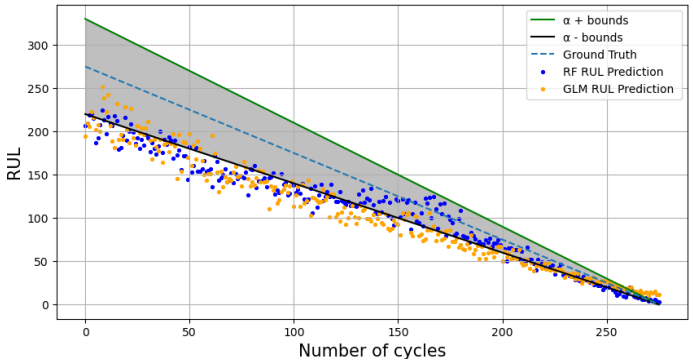
The accuracy therefore gives the number of RUL forecasts for each UUT that lie inside the cone of accuracy for each simulation. Now, this concept needs to be coupled with the one drawback of the $\alpha - \lambda$ performance pointed out in subsection II.C i.e, setting a value for α and solely judging the performance of an algorithm through its RUL predictions for that one particular α value.

To discern the effect of an incremental α on the overall prediction accuracy of an algorithm, a sensitivity analysis is proposed for the parameter wherein α is varied from 10% to 90% in increments of 10% which will give a broader picture of how the algorithms fare over a range of values of α. A large α will no doubt be more lenient on poor RUL predictions whereas a narrower cone formed by small values of α will be more stricter on the these predictions. For algorithms that do not predict RUL accurately enough, a large value of α will perceive these predictions as accurate.

To illustrate the above with an example, the $\alpha - \lambda$ performance metric has been applied a particular UUT for a $\alpha = 40\%$ and $\alpha = 20\%$ as shown in Fig. 8. An accuracy



(a) RF Accuracy = 97.46% and GLM Accuracy = 94.93%



(b) RF Accuracy = 43.11% and GLM Accuracy = 28.62%

Figure 8. Changes in prediction accuracy with variation of α for UUT 5

of 97.46% was observed for the RF model and 94.93% was observed for the GLM model for $\alpha = 40\%$. However, when α was reduced to 20%, the accuracy levels of both algorithms dropped by 55.76% for the RF model and by a significant 69.85% for the GLM model. The question that arises here is whether such a large jump in accuracy is justified for the applied change to α. This necessitates the need to observe how accuracy varies over all values of α which is illustrated in this paper by iterating each simulation from $\alpha = 0.1$ to $\alpha = 0.9$ and analysing the prediction accuracy at each α.

2. Performance Visualization

Once the accuracies are obtained with the help of the methodology outlined above, the performance needs to be visualized. The three parameters that influence the interpretation of performance for $\alpha - \lambda$ accuracy are the accuracy level α, the prediction accuracy defined in Eq. 15 and time. This motivated the following three ways to visualize the results for the $\alpha - \lambda$ performance, which are illustrated in subsubsection IV..2.

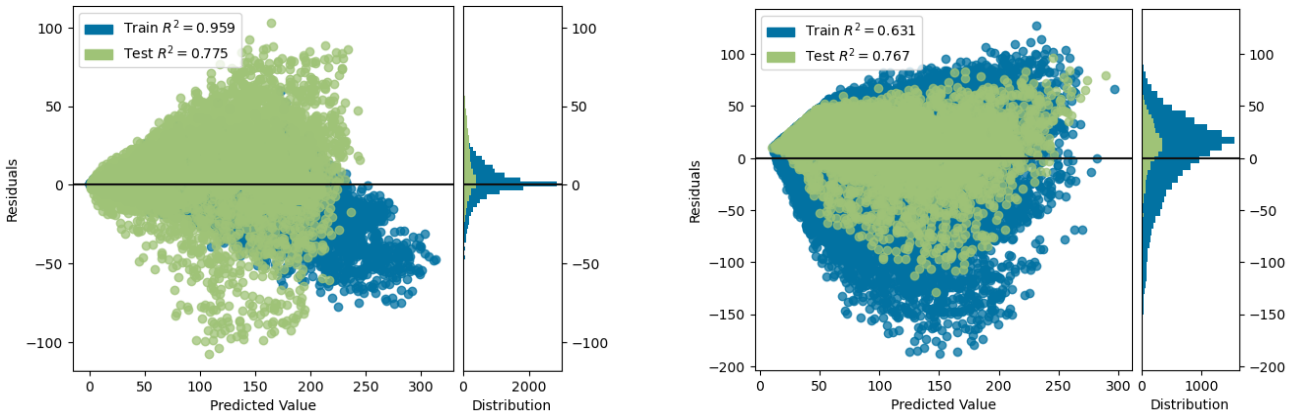
If the conventional path is followed by selecting a single value for α over an incremental α for the analysis, the prediction accuracy obtained is illustrated for all the test units at once. The time domain is converted from number of cycles and discretized into bins from 0% - 10%, 10% - 20% and so on until the end of life (100%) using Eq. 5. This is the solution to visualize the performance of multiple test units in a fleet.

Secondly, a visualization of accuracy over the life of the test units is provided for each incremental value of α with the aid of a bar plot. In this case, the average accuracy for all test units is considered within the time intervals rather than each test unit individually. This is done to observe the overall model performance in a single visual tool, albeit for multiple values of accuracy levels.

Finally, a combination of all the three factors that influence the interpretation of performance evaluation are considered. The variation of the prediction accuracy is illustrated along with the variation in accuracy levels α over the duration of the life of all test units. Here as well, the overall model performance is considered with all UUTs considered together. To also provide a measure of the errors of the RUL predictions, a heatmap illustrating the magnitude of errors over time is proposed because this would provide an ideal representation of the type of prediction errors obtained within each interval.

IV. RESULTS

The focus of this section will be on the visualization of the performance that has been obtained by implementing the



(a) RF Model

(b) GLM Model

Figure 9. Residual plots for the prognostics models

performance metrics and visualization techniques outlined in section III. To reiterate, the goal here is not to compare the two models but rather focus on how visualization can contribute to the interpretation of model performance with the C-MAPSS dataset only being used as a case study.

For the simulation used to obtain these results, an 80-20 train-test split is used for the C-MAPSS dataset. The residual plots obtained illustrate the distribution of the prediction errors about the zero error mark for both models is shown in Fig. 9. Residuals on the vertical axis in Fig. 9 represent the RUL prediction errors for the model and the horizontal axis depicts the corresponding RUL predictions. From Fig. 9, the RF model features a greater number of very early predictions. Late predictions large in magnitude seem to be more prevalent in the RF model as well. In the GLM model in Fig. 9b, a denser concentration of late predictions that are within the tolerance number of cycles are observed. To discuss the performance visualization, this section will further be subdivided into two sections to discuss the scoring function and the $\alpha - \lambda$ accuracy.

1. Scoring Function

The scoring function defined in Eq. 6 is firstly used to penalize the RUL predictions according to the severity of their corresponding errors. Based on the computed monotonicity (m = 0.12) for the C-MAPSS dataset and the maximum error obtained from the test data for the two models, a tolerance of 91 cycles has been applied for late predictions to the proposed scoring function. This means that any late predictions made up to an error of 91 cycles may not necessarily result in failure if the maintenance check is scheduled 91 cycles in advance. The profile of the scoring function with respect to the prediction error for this simulation has been illustrated in Fig. 10. The results

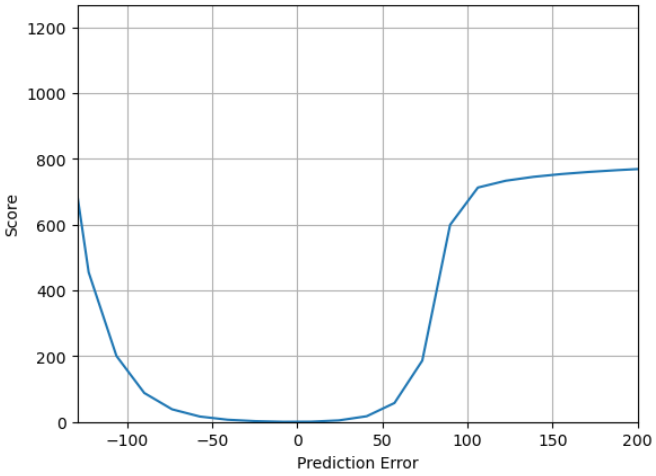


Figure 10. Scoring function as a function of error with $a_1 = 20$, $a_2 = 14.03$ and $T = 91$ cycles

of the scoring function for a particular test unit (UUT 42) out of the 20 UUTs have been shown in Fig. 11. In the case of UUT 42, the average score observed was 4.30 for the RF model and 4.24 for the GLM model. A sharp increase in score from 80% - 100% is observable for the GLM model, whereas the average scores are almost similar. The difference in the two average scores may be negligible and this can induce a false sense of security, especially when visualizing end of life predictions. There is hence the issue of the average score per unit not providing the best comparison between the RUL predictions made by different models. This necessitated the development of a more visual approach to display the performance results of the test units on the basis of the scoring function. Each RUL prediction consists of an error associated with it and the scoring function assigns a particular penalty to these predictions. However, this does not translate into the fact that the highest

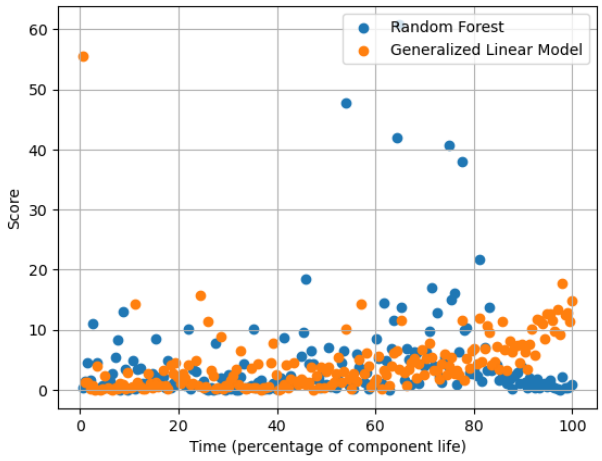
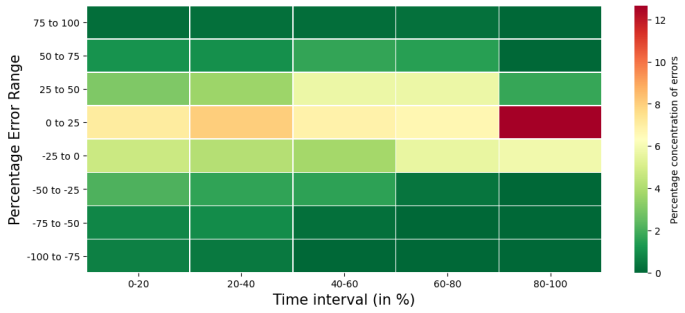
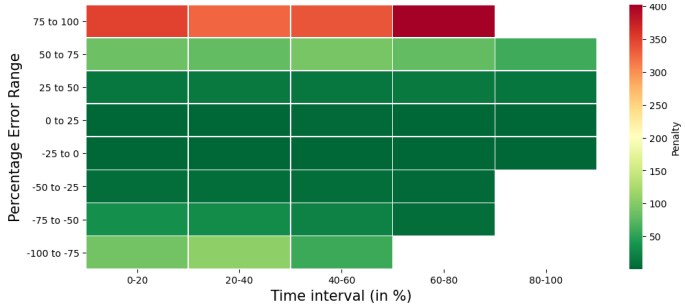


Figure 11. Score throughout the life of UUT 42

penalty will be observed where the concentration of errors in the model is the highest. To illustrate this, the errors observed in either model over time have been converted into a percentage scale as was discussed in subsubsection III.B.4. Early predictions vary from 0% to -100% of the least early prediction error whereas late predictions vary from 0% to 100% of the maximum late prediction error. In Fig. 12a, the highest concentration of errors observed are late predictions (0% - 25%) towards the end of life of the UUTs.



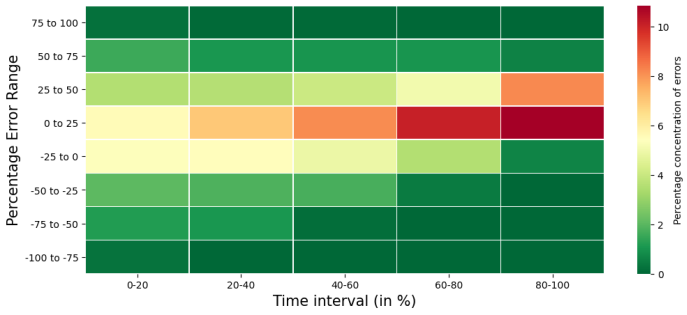
(a) Concentration of RUL prediction errors for the RF model



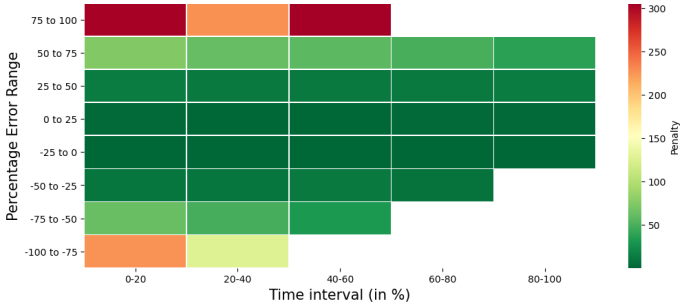
(b) Scoring function penalty for the RF model over time

Figure 12. Dispersion of RUL prediction errors of the RF model over time and the corresponding penalty

Looking at the RUL predictions made by the two models, the errors made for each prediction are clustered into the



(a) Concentration of RUL prediction errors for the GLM model



(b) Scoring function penalty for the GLM model over time

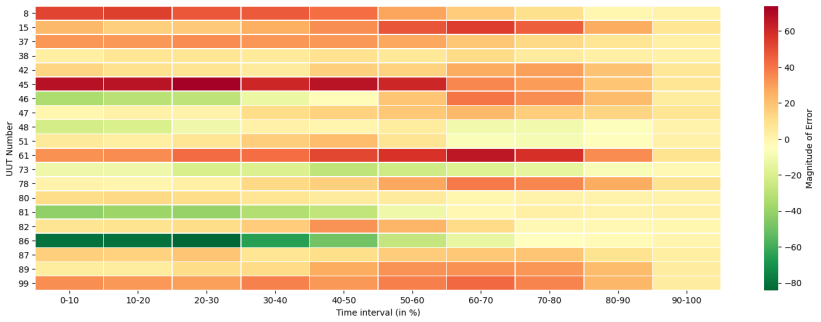
Figure 13. Dispersion of GLM prediction errors of the RF model over time and the corresponding penalty

corresponding time interval the predictions are made. This is done to observe at which point of time the errors get significantly worse. This visualization also validates the goal of the scoring function to penalize very late predictions more severely because while both models have fewer errors present that are late predictions, the maximum penalty is observed in those intervals.

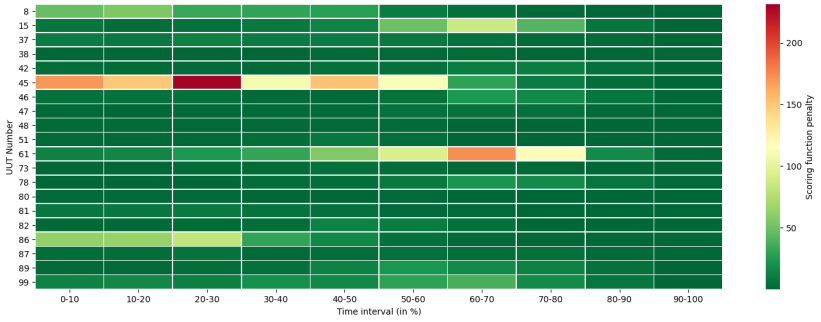
When observing the results for the RF model in Fig. 14a, the majority of the test units perform well given the errors are within a small range around the zero error mark, be it early or late predictions. Test units 45 and 61 however find some bad late predictions throughout their life but these predictions seem to get better towards the end of life of the engine. Only the late predictions of unit 45 get worse in the 20% - 30% time interval. Looking at the extremes for the late predictions in Fig. 14c in the color bar, the maximum positive error is below the tolerance value for this simulation. A larger score is still observed towards the end of life because a greater number of RUL predictions lie in this interval.

2. *The $\alpha - \lambda$ accuracy*

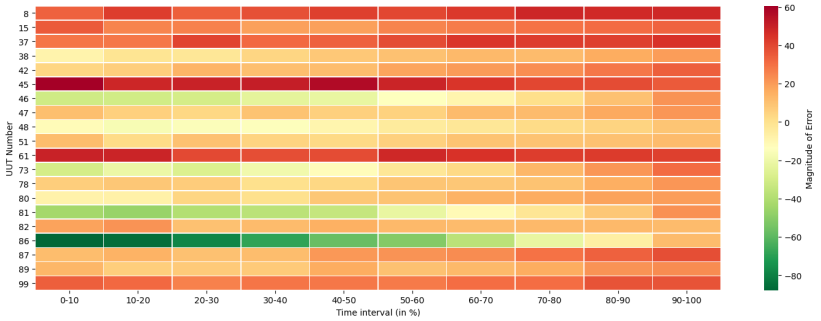
The methodology for the $\alpha - \lambda$ accuracy implemented in subsection III.C yields results for $\alpha = 0.1, 0.2, ..., 0.9$. The variation of α with respect to the prediction accuracy gives a rise in the RUL prediction accuracy with respect to an increasing α as is to be expected because as the cone size keeps increasing, there is a greater chance of more RUL



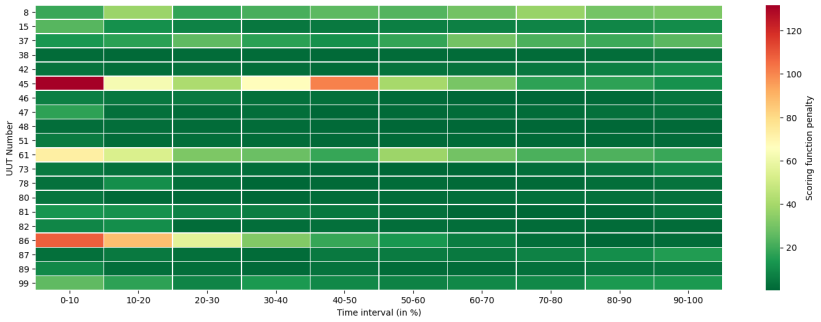
(a) Magnitude of errors throughout the life of the UUTs (random forest)



(b) Scoring function results for the UUTs (random forest)



(c) Magnitude of errors throughout the life of the UUTs (generalized linear model)



(d) Scoring function results for the UUTs (generalized linear model)

Figure 14. Performance on the basis of scoring function for each UUT

predictions to fall within the cone. The variation of the overall $\alpha - \lambda$ accuracy with varying α proposed is illustrated in Fig. 15. At the highest accuracy level (α = 90%), the RF model was found to have an overall accuracy of around 90% while the GLM model was found to have an overall accuracy of around 85%, which means that both models seem to perform similarly in this case. This result more importantly

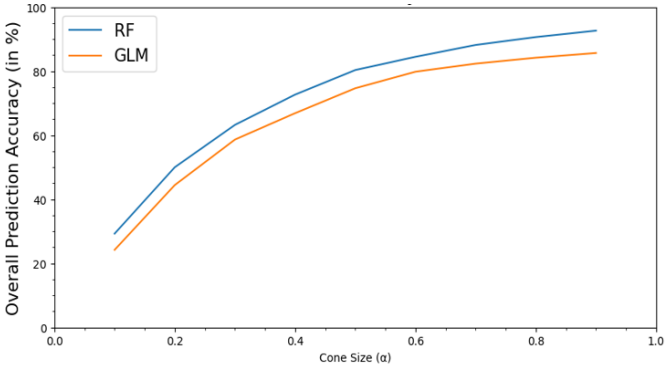
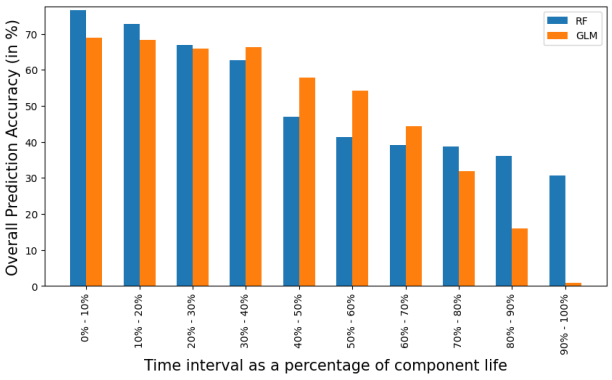


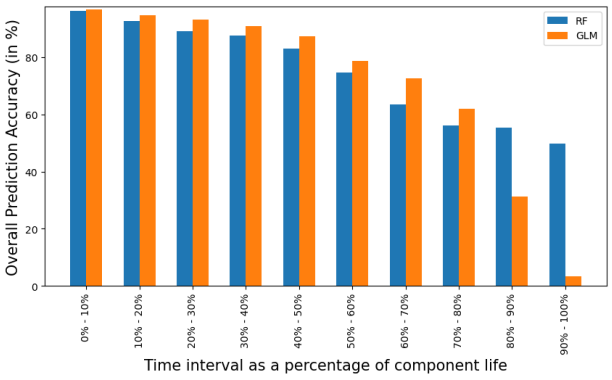
Figure 15. Variation of overall $\alpha - \lambda$ accuracy with cone size α

illustrates the need for a tradeoff to be established for the selection of an optimal α depending on the case study being assessed. A low accuracy level α results in a lower accuracy but intentionally selecting a high α would be far too lenient on poor RUL predictions, crucially towards the end of life. To illustrate this point, the average RUL prediction accuracies obtained above were also analyzed over time for varying levels of accuracy α as is observed in Fig. 16. This enables the user to observe where the loss of accuracy takes place for a particular prognostics model over the life of the test units. Reduced accuracy towards the end of life is expected but the drop off was found to be rapid in the case of the GLM model. From Fig. 16, it is clear that the GLM model accuracy also decreases rapidly towards the end of life. This overlooks the fact that during the time intervals 30% – 70%, the GLM model was found to perform notably better in Fig. 16a. For α = 40%, the GLM model was observed to have a higher accuracy until 80% of the unit life, after which the drop-off in accuracy was significant. This does put into context what an overall accuracy suggests. It is evident from Fig. 15 that the two models perform neck and neck for a range of values of α. But some stark differences are observed when the time factor is introduced. The number of trustworthy RUL predictions reduces drastically for the GLM model over time as opposed to the RF model.

Finally, an illustration of the $\alpha - \lambda$ performance of all the UUTs has been illustrated in Fig. 19.
Looking at the overall $\alpha - \lambda$ performance of both the models,

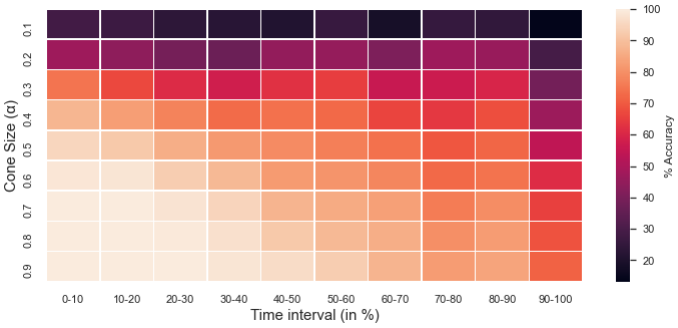


(a) α = 20% accuracy level

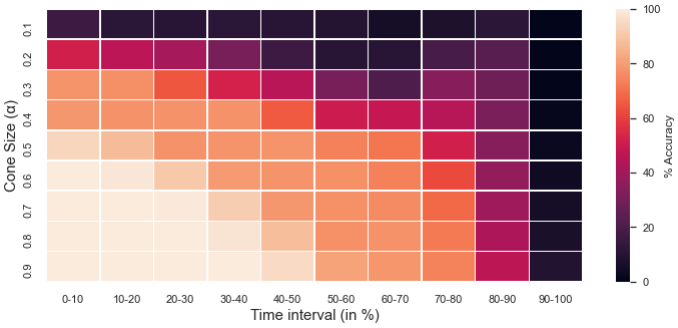


(b) α = 40% accuracy level

Figure 16. Variation of average accuracies for α = 40% & α = 60% accuracy level over the life of the UUTs



(a) RF Model

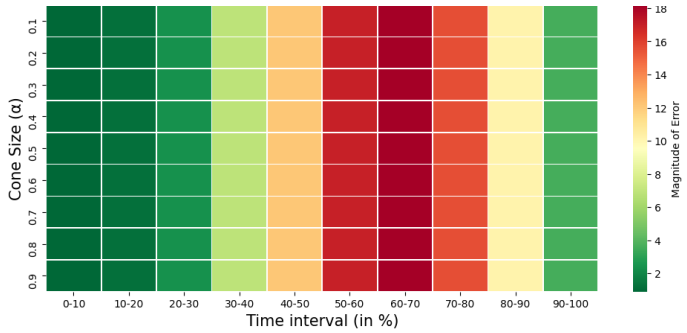


(b) GLM Model

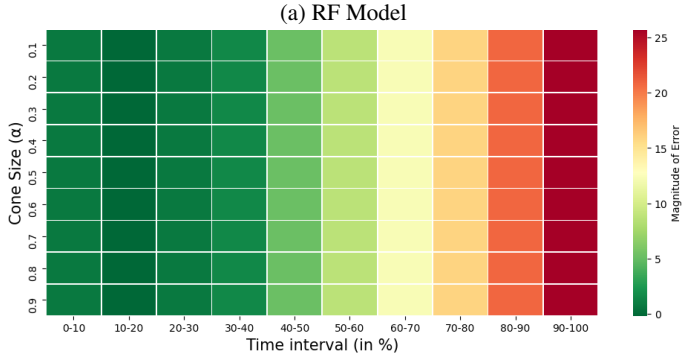
Figure 17. Variation of RUL prediction accuracy with varying α over UUT life

an accurate comparison can be established by considering the overall prediction accuracy of all UUTs over the entire range of $\alpha = 10\%$ to $\alpha = 90\%$ accuracy level as is shown in Fig. 17a and Fig. 17b. It is clear that the overall accuracy is lost towards the end of life for the GLM model while the performance is not too different from the RF model early on in the life of the UUTs. As was deduced earlier, the accuracy does increase w.r.t α but the RF model yields a higher overall accuracy. This sets a benchmark for the GLM model in that RUL predictions towards the $90\% - 100\%$ time interval mark cannot be trusted in comparison to the RF model. In fact, a significant decrease in accuracy is observed for the GLM model at time interval $80\% - 90\%$. In fact, the GLM model was found to lose 98% accuracy from time intervals 80%-90% to 90%-100% when $\alpha = 10\%$.

However, there is still one aspect that needs to be considered. While the loss of accuracy is prevalent in the GLM model more so towards the end of life, these predictions may still hold an advantage in one aspect. Fig. 18a and Fig. 18b illustrate the magnitude of the



(a) RF Model



(b) GLM Model

Figure 18. Magnitude of RUL prediction errors over UUT life

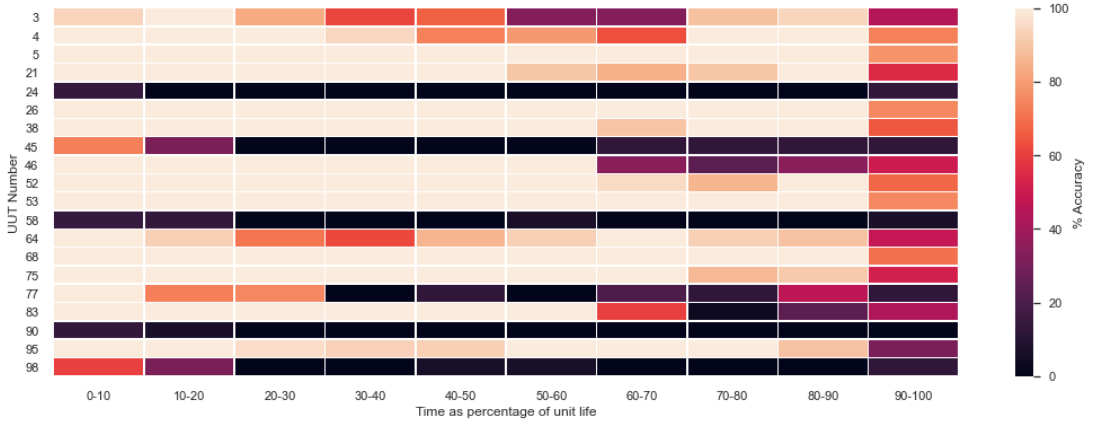
RUL prediction errors over the life of all the UUTs. This information provides details on whether the predictions are early predictions or late predictions at each time interval. The reason for the uniform distribution for all the rows in Fig. 18a and Fig. 18b is that the magnitude of errors is independent of α. Such a representation is however preferred over a line plot, to compare the two heatmaps for both models simultaneously to draw accurate comparisons. It was established that the reduced RUL prediction accuracy towards the end of life in Fig. 17b was a concern for the GLM model. However, the predictions observed within the 90% - 100% time interval are also late predictions (Fig. 18b) as opposed to late predictions of a lower magnitude in the case of the RF model as observed in Fig. 18a for the same time interval. To some extent, it can be inferred that the presence of late predictions in the time intervals where accuracy is high is a concern for the RF model. But the end of life performance as can be observed in Fig. 17a and Fig. 18a is better for the RF model. It is therefore a lot easier to take corrective action for the RF model close to the actual failure because of greater reliability of the RUL forecasts.

V. LIMITATIONS

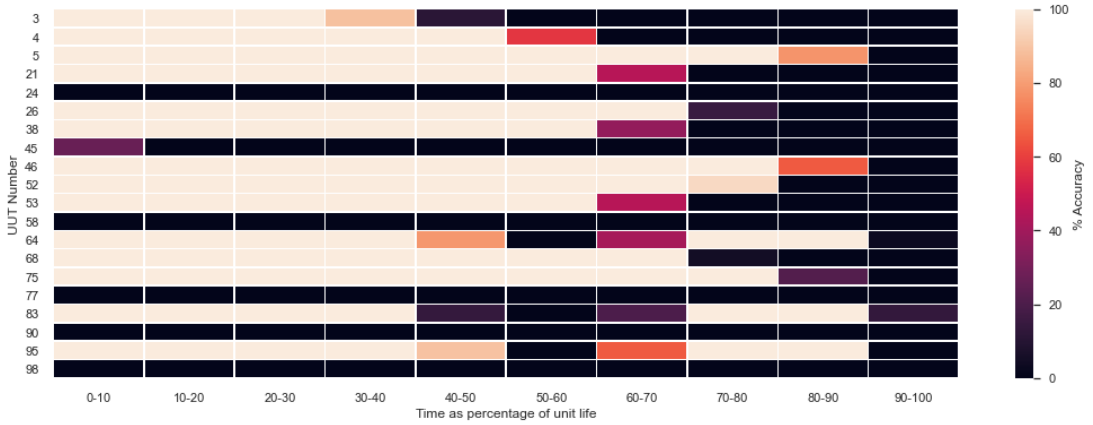
While satisfactory results have been obtained using the methodology outlined in section III, there need to be some limitations that need to be pointed out. These also serve as questions for those wishing to build on the metrics discussed further. Since this research paper focuses on deterministic RUL predictions from the outset, certain performance metrics such as the scoring function cannot be applied when the prognostics models output a probability distribution. This may be possible if the PDF is converted into a deterministic measure of RUL, but alternative metrics need to be explored.

Secondly, there is merit in the argument that both the current and the proposed scoring functions do not take into account that higher weights must be assigned to poor RUL predictions closer to the end of life, which is discussed by [12]. This is especially true when the average score of the test units is considered because a single value should be able to capture such a feature. This is also the reason why a more visual approach was taken in this paper to study the scoring function performance for each of the test units without considering the final average score. Also, the $\alpha - \lambda$ accuracy metric was specifically chosen to counter this drawback of the scoring function.

Lastly, the input sensor data from the C-MAPSS case study has not been treated any with preprocessing steps such as denoising for example since this was outside the scope of this paper. Sensor noise in this case may have hampered the learning process of the RF and the GLM models in order to make RUL predictions. This was also observed when the low monotonicity was calculated in



(a) RF Model



(b) GLM model

Figure 19. $\alpha - \lambda$ performance of each UUT for $\alpha = 40\%$

section IV. More emphasis was laid to evaluate the models at hand than the preprocessing of the input data resulting in a highly non-monotonic dataset. Pre-processing would have resulted in a lower tolerance value for the scoring function thereby resulting in more strict scoring of late RUL predictions.

VI. CONCLUSIONS

In this paper, a framework of metrics and visualization tools have been proposed to visualize performance of the C-MAPSS case study, taking into consideration two linear regression models. The random forest model and the generalized linear models were trained using the C-MAPSS FD001 sub-dataset to generate RUL predictions at each cycle for all test units.

Firstly by developing a scoring function along with a methodology to determine the constants that quantify the growth of score w.r.t the prediction error, it is ensured that each case study is associated with a unique scoring function for evaluation. This ensures that there are no situations wherein late predictions are favored over early predictions, given the assumptions used to determine the constants. Secondly, the impact of a significant percentage of late predictions could be limited by scheduling a maintenance check with the introduction of the tolerance buffer, which has been estimated based on the quality of the training data used for model training. This also enables the use of suitable pre-processing techniques by researchers to achieve a lower value of tolerance, since it may be more feasible to schedule maintenance fewer cycles in advance. Owing to the extremely low monotonicity of the input data pertaining to this case study, a high tolerance value was applied to the scoring function.

The results obtained from visualization of the scoring function do in fact show that while the score obtained might be similar in certain cases for any two models (Fig. 11), the performance can still be vastly different for crucial stages during the life of the unit. The end of life has been chosen to be the focus for the performance visualization because an inaccurate RUL prediction towards the end of life leaves little room for any corrective action before

actual failure. In this particular case study, late predictions towards the end of life were found to be more prevalent for the GLM model and hence resulted in higher penalties. The corresponding magnitude of errors chart provided illustrated what types of errors (early or late predictions) were contributing to the corresponding penalties. Such a visual perspective also enables the user to determine which specific unit is contributing to a bad performance. The user can then for instance trace back the input data metrics such as the monotonicity, prognosability and trendability corresponding to this particular unit used for training the model and improve the data accordingly. An example shown in the results would be UUT 45 featuring significant errors for both the models nearly throughout its life.

When looking at the $\alpha - \lambda$ accuracy, the accuracy level (α) was initially set to a pre defined value of 40% as shown in Fig. 19 to analyze individual UUT performance. The motivation behind this approach was similar to the one highlighted for the scoring function. The training data corresponding to each poorly performing UUT can be improved based on the input data metrics such as monotonicity, prognosability and trendability for better model training. From Fig. 19 for instance, UUTs 24, 58, 90 and 98 perform poorly for both the models and hence the input data corresponding to these test units can be observed for improvement.

Next, the average $\alpha - \lambda$ accuracies were found to be similar when considering all values of α, with the RF model having only a marginally higher accuracy (Fig. 15). But when visualizing the accuracy over well defined time intervals, there was a substantial drop off in accuracy observed towards the time intervals 80% - 100% for the GLM model for all values of α. This may be attributed to the converging cone of accuracy in these intervals, but the performance is significantly better for the RF model. This means that any RUL prediction obtained within these intervals can not be trusted for decision making, and the GLM model only possessed a high overall accuracy because of relatively better performance early during the life. This visualization therefore provides a tool to interpret end of life RUL predictions and in this case, a clear difference in performance can be observed when compared to an overall accuracy value.

Considering the limitations of the $\alpha - \lambda$ performance in distinguishing between the early and late predictions, a secondary visual perspective has also been provided to visualize the magnitude of errors at each stage during the useful life of the UUTs. The novelty of such an approach lies in that the two charts (Fig. 17 & Fig. 18) can be compared side by side to find areas of overlap where the critical regions are. These regions would be time intervals of a low accuracy and very late predictions. A case in point here was the end of life performance of the GLM model wherein large magnitude late predictions coincide with regions of low accuracy, which is unfavorable. Although even for the RF model, regions of higher accuracy earlier in the life (50% to 70%) were found to coincide with late predictions. However since this is not towards the end of life, a higher accuracy can be considered as the advantage.

These visualization tools therefore give useful information to users who rely on prognostics to schedule maintenance checks as well as researchers involved in prognostics model development. These findings also form the foundation for future researchers to work upon. The importance of a visual perspective has been outlined throughout this paper and therefore, section VII now outlines a few recommendations that can be followed for future steps in this research area.

VII. FUTURE WORK

In this paper, several visualization tools have been proposed by the modification of existing performance metrics in prognostics. Deterministic RUL predictions had been considered as the scope for this paper, but there are be situations where prognostics models provide probabilistic RUL predictions when uncertainties are taken into account. While the mean and variance of the RUL PDF can provide feasible deterministic forecasts from a PDF [8], the framework of metrics in this paper can account for uncertainties with minor alterations. This is also where the domain of scoring rules introduced in subsection II.C may find application since scoring rules evaluate forecasts on the basis of the prediction and the confidence associated with such a prediction.

There also needs to be a greater focus on online prognostics performance evaluation methods and not many models are under development if reviewing most publications on prognostics. Currently, offline prognostics uses run to failure datasets meaning that the failure has already occurred and any analysis pertaining to the performance evaluation would serve as a benchmark for future test units of the same type. Online prognostics on the other hand would therefore assist in real time decision making, thereby enhancing the potential of performance visualization even further.

Lastly the performance metrics and visualization tools proposed in the paper were intended to work on the gaps that arose when conventional forecasting metrics are used.

Therefore, an ideal path forward would be to incorporate these metrics and the visualization techniques provided here to evaluate the performance of other data-driven prognostics models, even outside the aerospace domain. Another interesting progression would be to further propose methods to estimate the tolerance based on whether a system is critical to the operation or not. In addition, the impact of the tolerance buffer on the optimal scheduling of maintenance would be a noteworthy addition to this work in the future.

References

[1] Sun, B., Zeng, S., Kang, R., and Pecht, M., "Benefits and Challenges of System Prognostics," *IEEE Transactions on Reliability - TR*, Vol. 61, 2012, pp. 323–335. doi: 10.1109/TR.2012.2194173.

[2] SUN, J., WANG, F., and NING, S., "Aircraft air conditioning system health state estimation and prediction for predictive maintenance," *Chinese Journal of Aeronautics*, Vol. 33, No. 3, 2020, pp. 947–955. doi: https://doi.org/10.1016/j.cja.2019.03.039, URL https://www.sciencedirect.com/science/article/pii/S1000936119302055.

[3] Rajaraman, V., "Big data analytics," *Resonance*, Vol. 21, 2016, pp. 695–716. doi: 10.1007/s12045-016-0376-7.

[4] Sprong, J., Jiang, X., and Polinder, H., "Deployment of Prognostics to Optimize Aircraft Maintenance - A Literature Review: A Literature Review," *Annual Conference of the PHM Society*, Vol. 11, 2019. doi: 10.36001/phmconf.2019.v11i1.776.

[5] Aizpurua Unanue, J., and Catterson, V., "Towards a methodology for design of prognostics systems," 2015.

[6] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., and Lin, J., "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mechanical Systems and Signal Processing*, Vol. 104, 2018, pp. 799–834. doi: https://doi.org/10.1016/j.ymssp.2017.11.016, URL https://www.sciencedirect.com/science/article/pii/S0888327017305988.

[7] Li, T., Sbarufatti, C., Cadini, F., Chen, J., and Yuan, S., "Particle filterbased hybrid damage prognosis considering measurement bias," *Structural Control and Health Monitoring*, 2021. doi: 10.1002/stc.2914.

[8] Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K., "Metrics for Offline Evaluation of Prognostic Performance," *International Journal of Prognostics and Health Management*, Vol. 1, 2010, pp. 2153–2648.

[9] "Prognostics Performance Evaluation," , ???? URL https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostics-performance-evaluation/.

[10] Goebel, K., Saxena, A., Saha, S., Saha, B., and Celaya, J., "Prognostic Performance Metrics," *Machine Learning and Knowledge Discovery for Engineering Systems Health Management*, Vol. 22, 2011, p. 147. doi: 10.1201/b11580-7.

[11] Yang, F., Habibullah, M., Zhang, T., Xu, Z., Pin, L., and Nadarajan, S., "Health Index-based Prognostics for Remaining Useful Life Predictions in Electrical Machines," *IEEE Transactions on Industrial Electronics*, Vol. 63, 2016, pp. 1–1. doi: 10.1109/TIE.2016.2515054.

[12] Saxena, A., Goebel, K., Simon, D., and Eklund, N., "Damage propagation modeling for aircraft engine run-to-failure simulation," *International Conference on Prognostics and Health Management*, 2008. doi: 10.1109/PHM.2008.4711414.

[13] Li, X., Ding, Q., and Sun, J. Q., "Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks," *Reliability Engineering & System Safety*, Vol. 172, 2017. doi: 10.1016/j.ress.2017.11.021.

[14] Chen, H., "A multiple model prediction algorithm for CNC machinewear PHM," *International Journal of Prognostics and Health Management*, Vol. 2, 2011.

[15] Li, J., "Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what?" *PLOS ONE*, Vol. 12, 2017, p. e0183250. doi: 10.1371/journal.pone.0183250.

[16] Hyndman, R. J., and Koehler, A. B., "Another look at measures of forecast accuracy," *International Journal of Forecasting*, Vol. 22, No. 4, 2006, pp. 679–688. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001, URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.

[17] Kim, S., and Kim, H., "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, Vol. 32, No. 3, 2016, pp. 669–679. doi: https://doi.org/10.1016/j.ijforecast.2015.12.003, URL https://www.sciencedirect.com/science/article/pii/S0169207016000121.

[18] Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K., "Evaluating algorithm performance metrics tailored for prognostics," *IEEE Aerospace Conference Proceedings*, 2009, pp. 1 – 13. doi: 10.1109/AERO.2009.4839666.

[19] Coble, J., "Merging Data Sources to Predict Remaining Useful Life An Automated Method to Identify Prognostic Parameters," 2010.

[20] Sankararaman, S., and Goebel, K., "Uncertainty in Prognostics and Systems Health Management," *International Journal of Prognostics and Health Management*, Vol. 6, 2015. doi: 10.36001/ijphm.2015.v6i4.2319.

[21] Uckun, S., Goebel, K., and Lucas, P. J., "Standardizing research methods for prognostics," *Physical Review Letters - PHYS REV LETT*, 2008, pp. 1 – 10. doi: 10.1109/PHM.2008.4711437.

[22] Orchard, M. E., and Vachtsevanos, G. J., "A particle-filtering approach for on-line fault diagnosis and failure prognosis," *Transactions of the Institute of Measurement and Control*, Vol. 31, No. 3-4, 2009, pp. 221–246. doi: 10.1177/0142331208092026, URL https://doi.org/10.1177/0142331208092026, _eprint: https://doi.org/10.1177/0142331208092026.

[23] Bröcker, J., and Smith, L., "Scoring Probabilistic Forecasts: The Importance of Being Proper," *Weather and Forecasting - WEATHER FORECAST*, Vol. 22, 2007. doi: 10.1175/WAF966.1.

[24] Roulston, M. S., and Smith, L. A., "Evaluating Probabilistic Forecasts Using Information Theory," *Monthly Weather Review*, Vol. 130, No. 6, 2002, pp. 1653 – 1660. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml, place: Boston MA, USA Publisher: American Meteorological Society.

II

Literature Study
previously graded under AE4020

1. Introduction

This project relates to the domain of PHM, whereby corrective action is taken with the help of a forecast for the RUL of a system. Any system or a system, especially in the case of non-repairables exhibit a certain failure degradation pattern during their lifetime when in operation. In an idealistic scenario, this degradation can be seen as a linear decrease of health of the system followed by a non-linear degradation over time till the system crosses the EoL, post which it enters a state of failure. As the number of cycles of the system increase, the machine health reduces. Non-repairables cannot be operated on beyond the state of failure and if these are systems that may be deemed 'critical', then prognostics is all the more crucial when it comes to maintenance. In time-based maintenance techniques, the system undergoes maintenance checks at frequent intervals to ensure that a state of functionality is restored for future use. Being far from ideal, RUL forecasts from prognostics models are laden with errors and hence, there is a need to validate the predictions made by these models.

Prognostic maintenance is becoming increasingly adopted in maintenance organizations with the increasing availability of sensor data, leveraging that data to monitor the health of systems and systems culminating in the a forecast of the RUL. This is from an aerospace point of view. With increasing acceptance of prognostics by maintenance organizations, there is the added benefit of potential cost savings since prognostics ensures that additional maintenance checks for systems can be avoided, parallelly ensuring there are advance warnings before the system fails. These savings are no doubt significant given that maintenance, repair and overhaul activities form 10-15% of an airline's direct operational costs [26]. Diagnostics on the other hand is the act of knowing when a problem is taking place, identifying and isolating the fault [14].

Most prognostics models can be classified as physics-based models, data-driven models, or hybrid models. Physics based models consider mathematical models developed solely on the failure degradation understanding of systems, which require greater knowledge of failure mechanisms whereas data-driven models rely on sensor measurements and previous events such as maintenance checks or repairs that had been carried out on the system or part [1]. Statistical, or data-driven prognostics are the most widely used approaches, followed by AI models and then physics-based models [10].

While data-driven prognostics algorithms are widely used for predictive maintenance purposes, the algorithms themselves may have errors in their RUL predictions. These may be early predictions or late predictions or just inaccurate predictions. A variety of metrics are available that can judge a prognostics algorithm based on certain properties. Some common metrics such as those discussed in [22][5] are the $\alpha - \lambda$ accuracy, prognostics horizon, convergence and the relative accuracy. A metric known as the scoring function focusing solely on early predictions and late predictions has been discussed in [20] aiming to penalize bad predictions depending upon the nature of the system. For some systems, early predictions may be bad predictions but in the case of most critical systems, they are preferred over late predictions. The $\alpha - \lambda$ accuracy metric discussed in subsection 2.2 and Equation 2.2 focuses on the accuracy of RUL forecasts overtime with the aim of classifying forecasts as inaccurate more strictly towards the end of life of the system. Conventional forecasting metrics such as RMSE, MSE, MASE, MAE and bias are not accurate enough to capture the required features of a RUL prediction, yet they have been discussed in the literature review. Therefore the main goal of the following section is to identify research gaps in the currently available performance evaluation techniques of data-driven prognostics algorithms, to then define a problem and research questions based on the conclusions of the review.

2. Literature Review

To come to terms with the domain of PHM and understand the prognostic models that deal with RUL estimation and to explore performance evaluation metrics, a wide variety of literature was studied. This section therefore goes further into the essential literature dealing with prognostics. As was discussed in section 1, PHM is a very diverse field and contributions have been made by fellow researchers in each branch of the field, be it data preprocessing and fault detection, or prognostics itself where innovative algorithms are developed for RUL estimation. The final aspect of PHM before the decision making process refers to performance evaluation, and this is where there has been the least contributions in the PHM field. Hence, the idea is to review and critique the past work done by contributors, to understand the gaps and to finally formulate a research question and objective(s).

2.1. Review of Prognostic Approaches

While the core subject of this thesis is performance evaluation of prognostics algorithms, it might be a good start to understand the different approaches used in prognostics. This section will review literature involving the different types of models currently used for the purpose of prognostics.

A simple classification of prognostics approaches would be data-driven and physical. Physics-based models build on the concept of describing the failure propagation or degradation in a component with the help of mathematical models, with the Paris-Erdogan model being one of the most widely used prognostic models in machinery. Data-driven models use machine learning models to establish a relationship between the input data available indicating the degradation behavior and the observed end observations. Hybrid models on the other hand exists of a data-driven ML approach alongside a model based approach. An example of this is a particle filter approach used to determine model parameters implemented in conjunction with a data-driven neural network model [12]. It is pointed out by [1] that such approaches are more accurate than data-driven and hybrid methods over a longer time interval in terms of RUL prediction but require expert knowledge, as is validated by [10] who mention that one needs a complete understanding of failure mechanisms and effective estimation of model parameters. Yet as observed in Figure 1, physics based model approaches form only a small percentage of the prognostic methods with regards to the number of publications that were considered in a study [10]. On this front, [1] argue that this may be attributed to the high implementation cost and the fact that some simplified assumptions lower the range of applicability of physics-based algorithms.

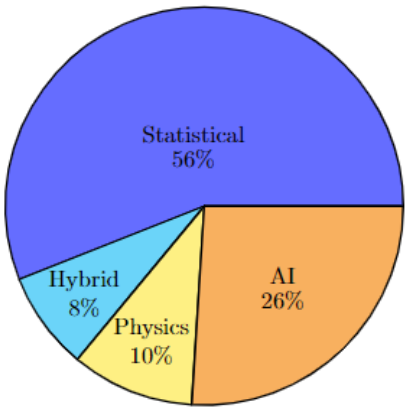


Figure 1: Distribution of prognostics models among different publications

However, another obstacle faced by the development of data-driven prognostic methods is the lack of run-to-failure (RtF) datasets that capture the failure evolution throughout the failure of a system [15]. It is also highlighted by [1] that one drawback of data-driven approaches is that they require a lot of data, and the lack of RtF data will therefore be a hindrance. Since they form the highest concentration of publications in prognostic algorithms, it will be of interest to delve further. A classification on data-driven approaches on whether it is a directly observed state process or not is provided by [25]. Directly observed state processes are cases where event data, such as past maintenance checks of a particular component is not available and one must rather rely on current observations to make an estimation of the RUL. They argue that such cases often have an advantage over scenarios where one must rely on event data to make RUL predictions.

2.2. Review of Performance Metrics

Performance evaluation has consisted of significant gaps in terms of research work in the field of PHM. Lack of standardized metrics has been a hindrance in the progress of PHM and the metrics that currently exist have largely not been applied to prognostics algorithms to assess their validity, which is crucial for prognostics to expand in a large operational scenario to aid decision making. Perhaps it would be prudent to study the progress of performance metrics that are applicable to prognostics purely from a research point of view. From an engineering research point of view as pointed out by [5], algorithm performance metrics and computational performance metrics are of particular importance. It is now known that prognostics deals with the estimation of the remaining useful life of a machine or a component with the help of a physics based

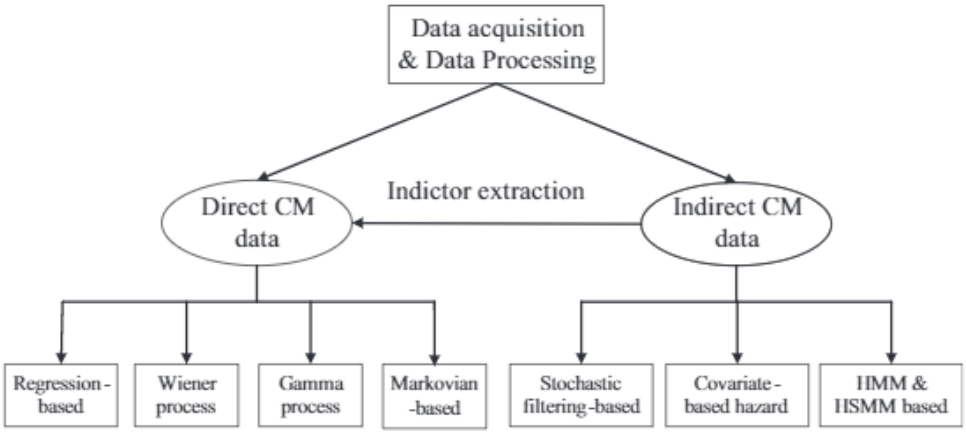


Figure 2: Classification of data-driven approaches [25]

model, data-driven model or even a hybrid model. The best estimate of RUL obtained from a prognostics algorithm is then comparable to the actual reality, hereafter referred to as ground truth. The ground truth can hence be defined as the best estimate of a certain feature or a parameter, which in this case can be considered to be the true RUL. This section will be structured as follows.

- *Metrics for Prognostic Applications:* These performance evaluation metrics have been majorly used in prognostics applications and they focus on accuracy of the RUL predictions and not solely the deviation from ground truth. Most importantly, the metrics discussed here will be only for deterministic forecasts and not forecasts with uncertainty. Scoring functions will be majorly discussed here, along with prediction horizon (PH), relative accuracy (RA), $\alpha - \lambda$ accuracy. An insight into the approaches used in few data competitions held by the PHM Society will be assessed.
- *Conventional Metrics:* These metrics may be applicable to prognostics applications but are majorly used in conventional forecast evaluations. These include MAPE, MASE, bias, etc.
- *Metrics for Probabilistic Forecasting:* Finally, this section will focus on literature involving methods for evaluating probabilistic forecasts because, most RUL forecasts will have a degree of uncertainty involved due to a variety of factors that will be mentioned later. Some methods from the first subsection can be applied to probabilistic forecasts, but majority of this section will be dedicated to scoring rules.

Metrics for Prognostic Applications

As is the case in most forecasting scenarios, not all algorithms are ideal. This means that there is often a difference between the parameter estimated from the algorithm and the ground truth. This introduces a key aspect in performance evaluation, that is known as the prediction error. Some algorithms denote this as the difference between the true RUL and the predicted RUL whereas other algorithms assume the other way around. Therefore in the category of algorithm performance metrics, error-based metrics are of particular importance.

A review of the classification of performance evaluation metrics regarding RUL predictions based on the different requirements of previous researchers and operators is tabulated in [10]. The most basic of these is the root mean square error (RMSE), used by [31] to assess their data-driven prognostics model, involving RUL prediction methodology for an electric motor case. An important consideration in the most conventional metrics such as RMSE is that while the difference in the estimated RUL and the true RUL is considered, it does not have any significance on early and late predictions as will be discussed in Equation 2.2. The confidence interval (CI) metric introduced in [31] deals with the accuracy of the predictions as compared to the error estimates that RMSE deals with. If the confidence is set at a high percentage (95% for example) and if the width of the confidence interval is small, this would be a case of a good prognostics algorithm since the concentration of bulk of the RUL estimates lie within a small interval.

Four main performance metrics, namely the prediction horizon (PH), $\alpha - \lambda$ performance, relative accuracy

and convergence are discussed in [22]. Let us explore the first three metrics first. This section concerns itself with only point forecasts and not probabilistic forecasts for performance evaluation. PH and $\alpha-\lambda$ accuracy are the most common prognostics evaluation metrics and have been used in a variety of publications. PH has been defined as the difference between the time when the predictions satisfy an accuracy criterion known as the β criterion and the EoL in [10]. Thus the foundations of this metrics are vastly different to that of a conventional metric such as RMSE, even though both depend on the ground truth RUL, where PH depends on the accuracy of the predictions and the RMSE measures the deviation from the ground truth.

Since PH is more effective when uncertainties are taken into consideration when making a RUL prediction, this will be discussed further in Equation 2.2 along with the β criterion. However, it might be interesting to look at the aforementioned $\alpha-\lambda$ metric even if it is for a deterministic forecast. $\alpha-\lambda$ accuracy distinguishes between predictions that lie within a certain accuracy levels and those that do not. The accuracy levels vary over time and are obtained at each time point as a function of the ground truth at that time. The bounds are calculated as follows, where α is a predefined accuracy value, $r_*(i_\lambda)$ is the ground truth RUL at each time point i.

$$\alpha^+ = r_*(i_\lambda)\cdot(1+\alpha) \tag{1}$$

$$\alpha^- = r_*(i_\lambda)\cdot(1-\alpha) \tag{2}$$

To assess their data-driven prognostics model dealing with the prediction of end of discharge for Li-ion batteries, [23] evaluated the predictions with the help of the $\alpha-\lambda$ metric [23]. Three different algorithms were proposed in their paper including two learning algorithms and one empirical model based prediction. The following results were illustrated for $\alpha = 20\%$ and $\lambda = 0.5$.

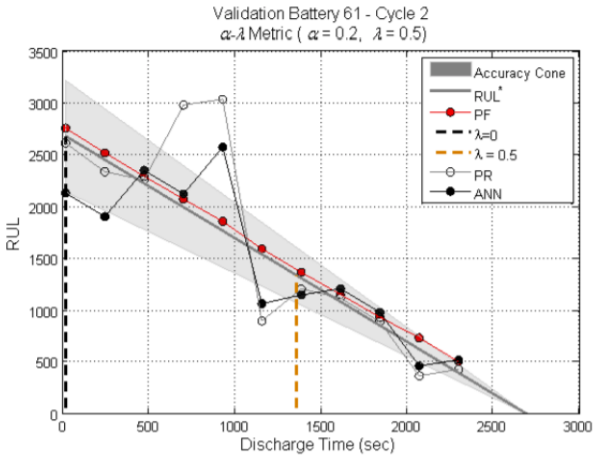


Figure 3: $\alpha-\lambda$ metric plot for comparison of algorithmic performance [23]

In Figure 3 artificial neural network (ANN), polynomial regression (PR) and particle filter (PF) are the three algorithms that have made their corresponding RUL forecasts whereas RUL' is the ground truth. As can be observed, the essence of this metric is its capability to distinguish algorithms on the basis of their accuracy levels with the highest accuracy appearing towards the EoL of the component. It could have been interesting however, if the author had included an analysis on how these results vary for different values of α. This is because a higher accuracy level will increase the size of the bounds that are imposed on the forecasts. This would mean that the two algorithms that perform relatively poorly for a 20% accuracy (namely ANN and PR) in Figure 3 may be considered acceptable for a higher accuracy. Understandably, the accuracy level that needs to be set is up to the user and that factor needs to be considered when developing prognostics models.

Before explaining the limitations of relative accuracy, this would be a good point to introduce late predictions, early predictions and their significance from an operational viewpoint. A negative prediction error or an underestimate hereafter referred to as an early prediction means that the estimate of the RUL is lesser than the true RUL, resulting in the remaining useful life of the component not being maximized to its capability. Such an estimate often leads to an early maintenance check when it is not needed, contributing an additional cost to the operator. On the contrary a positive prediction error or an overestimate, hereafter referred to as a

late prediction poses the risk of the component crossing the failure threshold. Depending upon how critical the component is to the system, an early prediction may or may not be suitable. Hence when dealing with error based metrics, considering the sign of the error may itself translate into meaningful advice and potential cost savings for the organization.

To highlight the above explanation, consider a very simplified example in Figure 4 where a series of predictions are made for a certain system at a moment in time t_i. The ground truth corresponds to a RUL of 40 cycles and hence a prediction of 30 is determined to be an underestimate with a prediction error of $(30-40) = -10$ cycles, whereas a RUL prediction of 50 is an overestimate with a prediction error of $(50-40) = 10$ cycles.

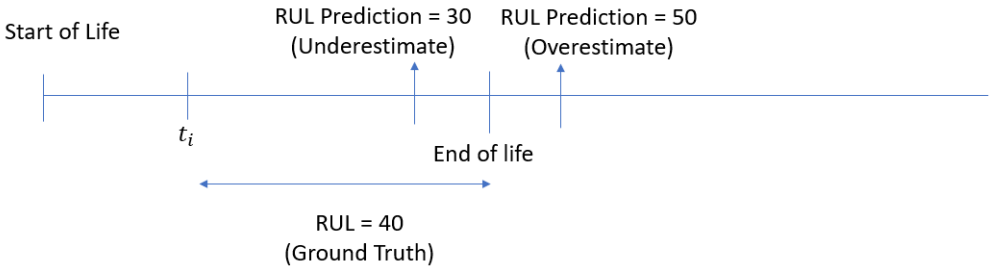


Figure 4: Prediction error terminology

The relative accuracy deals with the ratio of the prediction RUL error to that of the ground truth RUL at the given time index [22]. Since the estimated RUL is often a distribution rather than a point estimate, the author has taken into account the central tendency point estimate for the estimated RUL. Let us take a closer look at the relation as defined in [22].

$$RA_{\lambda}^{l} = 1 - \frac{|r_{*}^{l}(i_{\lambda}) - \langle r^{l}(i_{\lambda})\rangle|}{r_{*}^{l}(i_{\lambda})} \text{ (see [22])} \tag{3}$$

The term in the modulus sign is the prediction error with the term $r_{*}^{l}(i_{\lambda})$ denoting the ground truth for the RUL estimate. λ here is a time window modifier relating to the time index during prediction whereas the superscript l denotes the index of the particular system under consideration, or the unit under test (UUT). The key here lies in the fact that only the magnitude of the error is considered, as can be observed in the numerator. Therefore, the drawback of the relative accuracy metric in most applications is that there is little distinction between a negative prediction error and a positive prediction error. This is a significant gap in a prognostics case study because of some components being critical, by virtue of which an overestimate is not preferred and vice versa for non-critical components.

In addition to the limitations of metrics mentioned previously, [20] in modeling the damage propagation for an aircraft engine with a RtF simulation using the C-MAPSS tool for aircraft engine simulation discussed the prospect of an error scoring function. The scoring function was fully dependent on the sign of the error overcoming the shortcomings of relative accuracy described in Equation 3. It was therefore interesting that this metric was not mentioned in [22]. One would attribute this to the concept being relatively in its infant stages for prognostics applications. The system model discussed here was used to generate RtF data which has been a challenge in prognostics where a system needs to be made to fail deliberately.

Now the scoring function specifically developed for the application by [20] aims to penalize late predictions with a higher score owing to the risk of failure associated with a late prediction. The following scoring function was also made use of by [32] in their proposal of a convolution neural network prognostics model to predict the RUL. It would seem like a rather conservative approach from the author but it is understandable given that the case study being considered is an aircraft engine, a critical component in an aircraft. Some gaps arise from the scoring function itself, shown in Equation 4.

$$s = \begin{cases} \sum_{i=1}^{n} e^{-(\frac{d}{a_1})} - 1 \text{ for } d < 0 \\ \sum_{i=1}^{n} e^{(\frac{d}{a_2})} - 1 \text{ for } d \geq 0 \end{cases} \quad \text{(see [20])} \tag{4}$$

where s is the score, d is the difference between the true RUL and the estimated RUL and the arbitrary constants are set at $a_1 = 10$ and $a_2 = 13$ [20]. While this scoring function does achieve its task of penalizing late predictions, the author does not state how the constants have been estimated. It would perhaps have been ideal if the variation of the results of the scoring function had been illustrated with differing constants to portray their significance on the end score. Given that the concept of the scoring function is to introduce bias in terms of predictions by penalizing late RUL predictions, the results seem to indicate a peculiarity as is explained after Figure 5.

On the other hand, the scoring function as mentioned takes into account the shortcomings of conventional performance metrics such as RMSE as has been mentioned before. RMSE gives no indication on singling out single bad estimates and only focus on averaging the errors which are a deviation from the ground truth. The scoring function enables the user to exponentially penalize the predictions, thereby ensuring that the outliers are easily spotted in a series of RUL forecasts. These outliers can then be excluded from the time series.

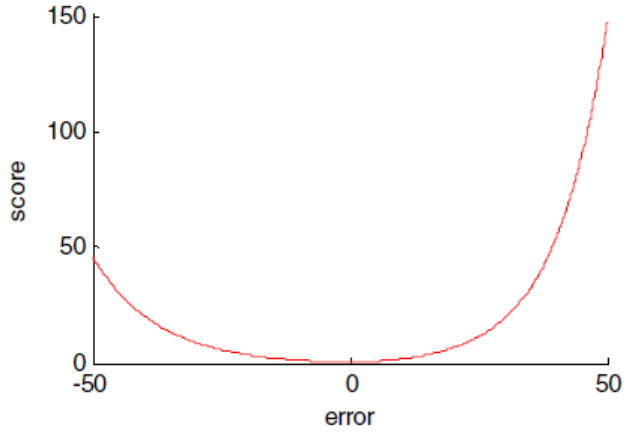


Figure 5: Score as a function of error [20].

When the score is analysed in terms of the prediction error as illustrated in Figure 5, an error of -20 for example seems to have a similar score to an error of +20. The scoring function seems to show symmetric results for such ranges when there should clearly be a distinction between the two, as the author has mentioned that the scoring function is asymmetric. Furthermore, if the values of $d = -50$ and $d = +50$ are substituted in Equation 4, it appears that a higher penalty has been provided for an early prediction of 50 than a late prediction of 50. The authors do not explain this aspect of the results, and this may have been due to the concept of scoring functions in prognostics being relatively new and untested. More likely, the scoring function may have been adapted to a scenario wherein a late prediction may have been favored over early predictions where the risk of failure of the system is not as significant as an additional cost that may result due to an earlier maintenance check. While this setback is not explained, a correlation metric is introduced for algorithms with multiple UUTs in a scenario when the performance cannot be judged by a scoring function in the event that the score is the same for multiple cases. While the exact correlation metric being used is not specified, the results do seem to show how this metric overcomes the limitations of the scoring function for this application, as shown in Figure 6. The predictions that are closest to the ground truth as expected have a higher correlation coefficient and vice versa. The key objective here is to distinguish between scenarios that have the same score, but varying closeness to ground truth. From Figure 6, it is therefore likely that the correlation coefficient would not be used as a standalone metric in the absence of the scoring function because there may be certain scenarios where a series of late predictions and a series of early predictions would have the same correlation coefficient, leading to no distinguishing feature between the two. The reason this could happen is because the data may be symmetric around the ground truth resulting in the same correlation coefficient regardless of late or early predictions. This would therefore lead to the same drawback that relative accuracy mentioned by [22] had in terms of prognostics applications for a variety of components. Secondly, [11] also argues that the correlation coefficient cannot be used as a metric to assess the accuracy of a prediction in forecast models because correlation only determines the goodness of fit

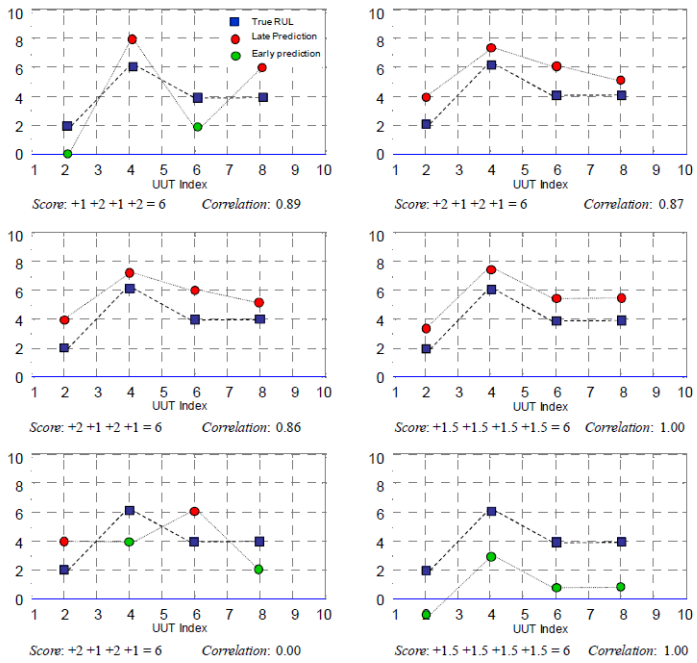


Figure 6: Scenarios illustrating various cases where error based aggregated scores may be same but the correlation score distinguishes further between different algorithms [20].

between the data, which is the estimated and true RULs in this case. It might therefore be a better idea to let correlation coefficient to be used as a backup metric in the event that the scoring function proves indecisive.

Perhaps another approach that [20] and [13] could have considered was to penalize late predictions close to the zero error point a little lesser than the ones farther away. In reality, this could be logical as the decision to conduct a maintenance check is made before the estimated EoL. If a tolerance period of x flight cycles is considered and if the estimated positive prediction error is $+y$ flight cycles, a maintenance check could be scheduled at $(y-x)$ flight cycles. Subtracting this tolerance from a small positive error will mostly lead to a point in time before the EoL of the component, which corresponds to the negative error part of Figure 5 ($error < 0$). This ensures that the component never fails and a state of functionality is restored, which is particularly crucial for critical components in a fleet. Considering these aspects can be treated as a research gap from a prognostics algorithm evaluation point of view.

A similar scoring function from Equation 4 along with RMSE was incorporated by [13] for their deep convolution neural networks prognostics model which forecast RUL for the C-MAPSS data-set used by [20]. Here, [13] chose to interchange their exponential terms resulting in the following scoring function shown in Equation 5. The notation used to represent the scoring function will be borrowed from that of Equation 4 for comparison.

$$s = \begin{cases} \sum_{i=1}^{n} e^{-(\frac{d}{a_2})} - 1 \text{ for } d < 0 \\ \sum_{i=1}^{n} e^{(\frac{d}{a_1})} - 1 \text{ for } d \geq 0, \end{cases} \quad \text{(see [13])} \tag{5}$$

When comparing the two scoring functions, the following profile is obtained as shown in Figure 7. It needs to be known that such plots need not be plotted in practice since the prediction errors will not necessarily vary with time as shown in the horizontal axis of Figure 7. For comparison with the corresponding scoring function, the root mean square error (RMSE) metric incorporated by [13] and [32] is shown by Equation 6. As can be seen, the RMSE takes into consideration only the magnitude of the prediction error since the error is eventually squared, and hence we can express a symmetric response when RMSE is plotted with respect to the prediction error. Such a plot has been illustrated in Figure 8.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \text{ (see [13])} \tag{6}$$

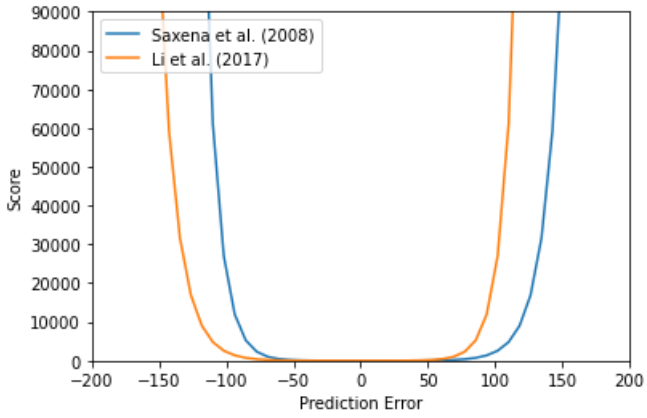


Figure 7: Comparison of scoring functions in [13]

and [20]

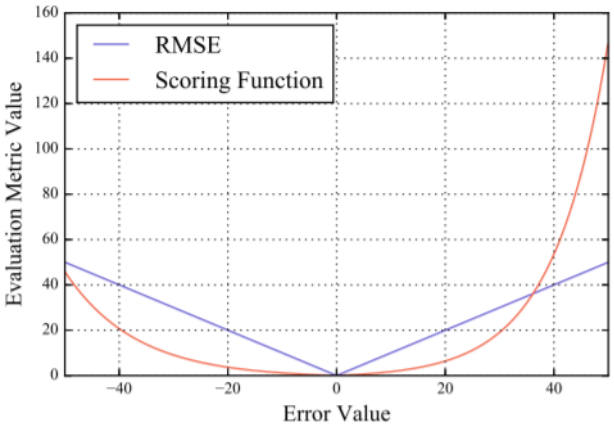


Figure 8: Comparison of RMSE with scoring function for CMAPSS data-set [13].

This comparison of the RMSE with the scoring function highlights the gap that is present between conventional forecasting metrics and a scoring function. A symmetric profile around the zero error mark indicates that overestimates and underestimates are penalized in a similar manner, which should not be all that an error metric captures in a prognostics assessment. A discussion on conventional metrics will follow in Equation 2.2.

Specifically on the C-MAPSS simulation data-set of the turbofan engine, a set of performance metrics was applied by various publications in their respective papers. A significant proportion of publications made use of the scoring function mentioned in Equation 4. This would make sense since an accuracy performance evaluation metric takes into account the evolution of prediction errors over time rather than an averaging of the errors procedure. In fact, 31 out of 40 publications that were considered for a review focused on accuracy based metrics including the aforementioned scoring function, MAPE, MAE, MSE, FPR and FNR [15]. Precision based metrics such as ME, MAD and conventional prognostics metrics including $\alpha - \lambda$ accuracy, PH, RA, CV, AB as mentioned by [5] form only a negligible portion of the publications considered in [15]. The less number of publications on $\alpha - \lambda$ accuracy and PH may be due to the fact that these metrics focus on getting accurate predictions earlier in the forecast rather than evaluating the incorrect forecasts and penalizing them.

With the advancement of data-driven prognostics, the PHM Society and IEEE had conducted a variety of competitions over the last 12 years (2008-2020) with focus on various aspects of PHM including supervised and unsupervised fault detection, prognosis and diagnosis, risk assessment and health assessment [8]. The following table summarizes the competitions annually hosted by the PHM Society and IEEE from the year 2008 to 2017.

Table 1: Research tasks for the data competition 2008 2017 [8]

Host & Year	System	Tasks
PHMS 2017	Bogie	Supervised fault detection & diagnosis
PHMS 2016	Semiconductor CMP	Virtual metrology
PHMS 2015	Power plant	Supervised fault detection & diagnosis
PHMS 2014	Unknown	Supervised risk assessment & fault detection
IEEE 2014	Fuel cell	Prognosis and health assessment
PHMS 2013	Unknown	Supervised fault detection & diagnosis
PHMS 2012	Bearing	Prognosis
PHMS 2011	Anemometer	Unsupervised fault detection
PHMS 2010	Milling machine	Prognosis
PHMS 2009	Gearbox	Unsupervised fault detection & diagnosis
PHMS 2008	Aircraft engine	Prognosis

It would be prudent to consider from Table 1 the challenges that focus on prognostics as there is a greater scope for the application of performance metrics. For example, a submission of PHMS 2008 data challenge by [20] which was applied on the aircraft turbofan engine with the help of the C-MAPSS aircraft engine simulation tool has already been discussed earlier in this review. A scoring function was used to highlight the significance of the prediction error which was found to distinguish between early and late predictions.

While Table 1 summarizes the data competitions hosted by the two societies from the year 2008 to 2017, the PHM Society had hosted another data challenge in 2020, wherein a prognostics algorithm was to be developed and evaluated using an error metric. One of the submissions for the same will be considered in this review.

The PHM Data Challenge conducted in 2010 saw the implementation of a scoring function to evaluate the prognostics algorithms developed for estimating the maximum number of cuts that could be made by a CNC milling machine cutter for a certain wear limit. The problem could be viewed as an analogy to the tradition RUL estimation problems that have been discussed so far. The number of cuts that could be made is analogous to the remaining useful life whereas the wear is analogous to the failure growth or component health degradation profile. There is of course, a ground truth profile for the maximum wear for each value of the number of cuts with which the estimated number of cuts could be compared with. The concepts of overestimates and underestimates then come into play here. One submission to the competition that will be reviewed here is that of [4].

Six data-sets are considered by [4], three of which are training data (c1, c4 and c6) and the other three are test data (c2, c3 and c5) each containing data for the cutter with which the cuts were going to be made. The RUL to be estimated was the maximum number of cuts that could be made, given a certain wear limit. From the training data and the sensor measurements, the maximum wear for the corresponding number of cuts was estimated as shown in Figure 9.

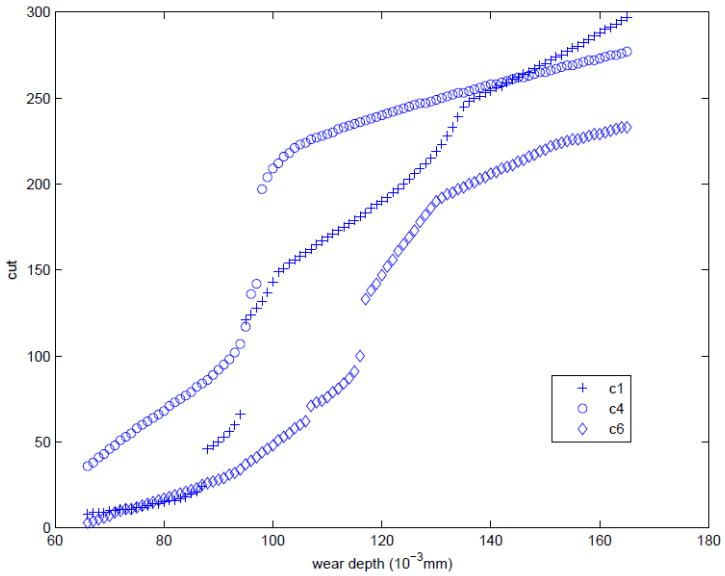


Figure 9: Maximum safe cuts for c1, c4 and c6 [4]

The data obtained in Figure 9 can be considered as the ground truth with which the estimated RULs or the number of cuts can be made for the test data-sets c2, c3 and c5. It is to be noted that the initial wear on each cutter was unavailable to the authors. Upon the application of the RUL algorithm to estimate the RULs, the results that were obtained by [4] are illustrated in Figure 10.

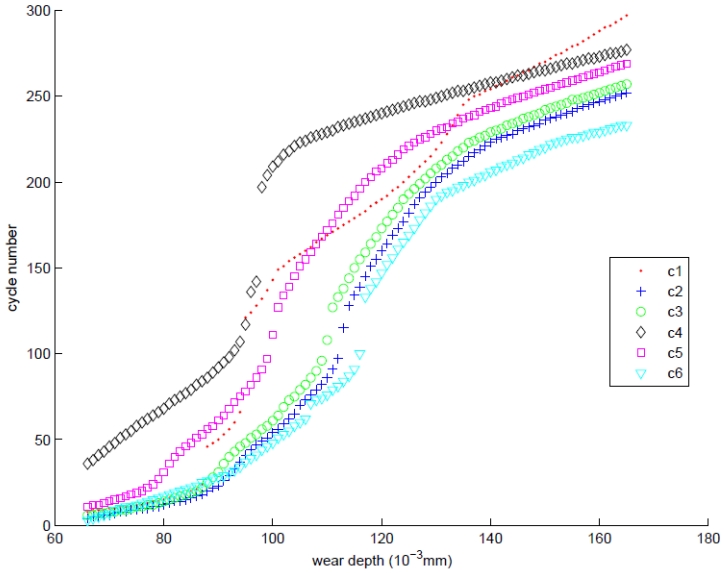


Figure 10: First submission to PHM data challenge for cutter c2, c3, c5 (maximum safe cuts of c1, c4, c6 included for comparison) [4].

From Figure 10, the wear after each cut for the test data-sets can be compared to the ground truth from the training data c1, c4 and c6 which is shown in Figure 9. The ground truth is illustrated with a different legend in Figure 10 as compared to Figure 9. It seems clear from the results obtained by [4] that the trajectory of their RULs followed a profile similar to that of the ground truth and the correlation between the two seems high. The error metric as discussed earlier in this competition was a scoring function. The scoring function defined by the rules of the competition is as follows, where δ is the RUL prediction error with respect to the number

of cuts and S is the score which is to be minimized.

$$S(\delta) = \begin{cases} e^{-(\frac{\delta}{10})} - 1 \text{ for } \delta < 0 \\ e^{(\frac{\delta}{4.5})} - 1 \text{ for } \delta \geq 0, \end{cases} \quad \text{(see [4])} \tag{7}$$

The scoring function defined in Equation 7 on first glance seems similar to that defined in [20] which penalizes over predictions as opposed to early predictions. But the lack of emphasis on the variation of δ with respect to the score as shown in Figure 5 or variation of δ with respect to the time intervals means that there is no clear way to determine a certain tolerance level before which a maintenance check can be performed on the machine, be it lubrication or some other procedure. Secondly, the authors provide only a point estimate of the final score ($9 \cdot 10^5$) from which no conclusions can be drawn unless a range of data is available in terms of the prediction error. Finally, when more components in the CNC machine are considered in a prognostics algorithm, it would be interesting to see the approach of [20] where a correlation metric is used after the scoring function estimate for different algorithms to compare their results with that of the ground truth [20].

To compare the scoring function of [20] given by Equation 4 to that of [4] given by Equation 7, both of them have been plotted in Figure 11. The horizontal axis illustrates a range of prediction errors while the vertical axis represents the penalty (or score) given by the scoring function.

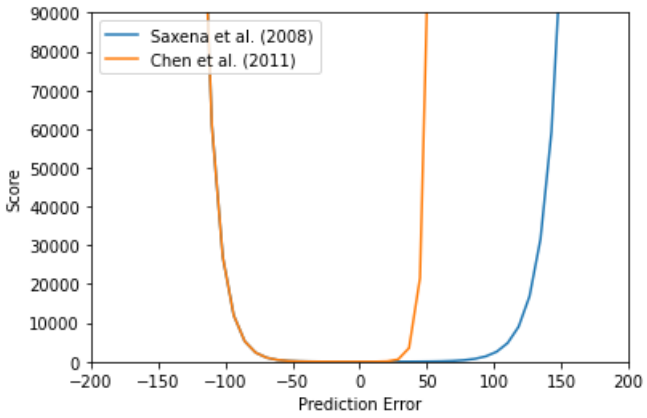


Figure 11: Comparison of scoring functions in [20] and [4].

The left hand side of the graph indicates a similar response in terms of the scoring function since the function for a negative prediction error is the same in Equation 4 and Equation 7, which shows the overlap between the two lines in Figure 11. The key difference lies in the area of a positive prediction error. A small positive prediction is penalized in the case of the CNC milling machine cutter [4] by an amount similar to the case of a much larger positive prediction in the case of an aircraft engine [20]. This is an indication of a slightly more conservative approach in the case of the cutting tool where even a small positive error is penalized, as compared to the aircraft engine. Hence we can conclude that depending upon how critical a component is, the variation of the terms inside the scoring function can penalize predictions accordingly.

Before proceeding to another example, let us finally consider all three scoring functions discussed so far by [20], [4] and [13] together as shown in Figure 12. This shows how particular components need different scoring functions based on how critical they are to the fleet. Expanding on the previous argument regarding the symmetric nature of the score versus error profile, it would be understandable if the score would be higher for a positive prediction error beyond the preset tolerance value. In an industry application, it may not be wise to argue that a negative prediction error of 150 cycles be considered the same as a positive prediction error of 100 cycles. This is because a negative prediction error is not a bad forecast since this would result only in an additional maintenance check at its worst but a positive prediction error of 100 cycles would mean that the component would have failed. Let us assume a tolerance of 50 cycles which means that if a positive prediction error is beyond 50 cycles, that is an example of a bad prediction in which failure will have occurred. It then does not make sense to exponentially increase the penalty beyond that point because it would not mean anything. Whether the error is +70 or +1000, it would really make no difference in a real life situation, even

from a monetary standpoint to the organization. Therefore, it could also be considered to have a 'jump' in the penalty value from the point when the tolerance period is over and maintain a constant penalty beyond that point. This constant penalty should also be greater than that of the peak penalty observed when the prediction error is negative.

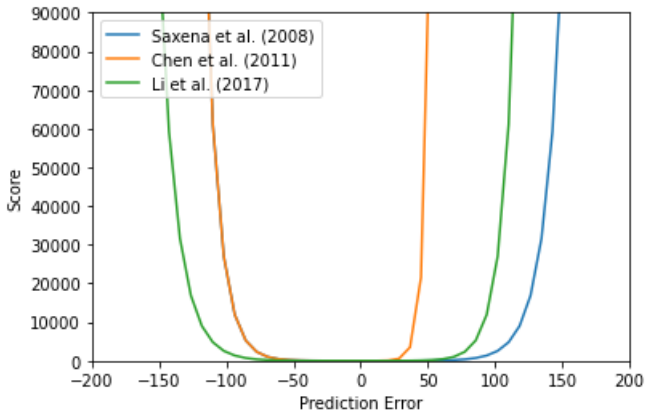


Figure 12: Comparison of scoring functions defined in [20], [4] and [13].

Another approach to performance evaluation of a prognostics algorithm was highlighted by [2] for the PHM Data Challenge 2020. The task here was to determine the RUL of an experimental filtration system with the help of a data-driven prognostics model, mainly when the filter is clogged [2]. There were 24 run to failure experiments available for the model training data-sets and eight available for the model validation data-sets and the performance was then judged upon how the model could be built using firstly all the experiments and secondly, a varying percentage (25%, 50%, 75%, 100%) of all the experiments available in the training and validation data-sets.

Finally, an error function was used to assess the performance of each of the above models for the training, validation data-sets and an independent test set. The error function deployed was the mean absolute error (MAE). The MAE was estimated for the combination of the training and validation data-set, as well as the independent test set. The penalty function, or the scoring function as referred to by [20], was determined in this case as follows.

$$\text{Penalty} = \sum_{i \in 25,50,75,100} (1.5 \cdot MAE(M_i(Te)) + MAE(M_i(Tv))) \text{ (see [2])} \tag{8}$$

In Equation 8, the term M_i refers to the model generated with i% experiments Tv refers to the combination of the training and validation data-set and Te is the test set [2].

Upon the application of the mean absolute error to the three proposed models for RUL estimation of the experimental filtration system, one of the models was selected to undergo validation based on its penalty score applied to only the training and validation set. After applying Equation 8, the penalty for the chosen model was obtained to be 57.24. Now, the issue seems from here to be the same as that of the relative accuracy metric. There is no distinction between the errors as far as a late or an early prediction goes, rather only focusing on a direct score. From a readers point of view, this gives no indication to any health features over the life of the filtration system. The mean can only capture a certain value which in no way is a true representation of all the data, especially the outliers in terms of errors. This also means that the reader should first go through the rules and regulations of the competition to understand the need for such a penalty function to evaluate the algorithm. While such an approach may seem to be sufficient for a clogged filtration system, it perhaps falls short when it comes to critical components in aerospace applications. It may therefore be prudent to implement with slight modifications, the scoring function as defined by [20] in Equation 4 [20].

A Note on Conventional Metrics

The idea behind this thesis is to discuss and implement innovative methods to assess the quality of a prognostics algorithm. These metrics have been discussed until this point. But it might also be interesting to focus on a slightly broader picture where we do not only consider prognostics and RUL estimation, but forecasting error and how we can evaluated the quality of a prediction in a forecasting model. Prognostics can then

be considered as a subset of the many forecasting techniques. This enables us to consider a broader range of metrics to evaluate prediction error. Given that such metrics in most cases do not distinguish between overestimates and underestimates, they will usually be second in priority to the metrics discussed in subsection 2.2. RUL forecasts at individual time points can be considered as a univariate time series with the only varying quantity being the RUL. As such, there are metrics available such as mean absolute percentage error (MAPE), mean absolute scaled error (MASE), etc. that will be discussed here.

Hyndman & Koehler (2006) proposed that the mean absolute scaled error (MASE) be used as the standard for the comparison of multiple time series in forecasting [6]. They compared different forecasting techniques including *historical mean upto the most recent observation, naive forecasting, simple exponential smoothing* and *Holt's method* taking into account three different time series. Metrics that had been used before the publication of the paper had been applied to the time series including MAPE, median absolute percentage error (MdAPE), symmetric MAPE (sMAPE), symmetric MdAPE (sMdAPE), median relative absolute error (MdRAE), geometric relative absolute error (GMRAE) and MASE. The authors also highlight that most publications until that point recommended the use of MAPE for determining forecast accuracy.

A common trend observed in the study was that when a forecast for a particular time index is zero, all of the corresponding error metrics (mentioned above) take an undefined or an infinite value, giving no indication on the quality of the forecast. This particular shortcoming has also been highlighted by Kim & Kim (2016) in their paper focusing on the proposal of a new metric [9]. Thus, Hyndman & Koehler (2006) propose a scale free error metric known as the scaled error, shown in Equation 9 and the mean of the scaled error is the MASE. In the equation, e_t is the forecast error between the forecast and the ground truth ($e_t = Y_t - F_t$) whereas the observations at two successive time points is denoted by Y_{i-1} and Y_i.

$$Scaled\ Error\ (q_t) = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|} \text{ (see [6])} \tag{9}$$

If MASE is applied to a prognostics algorithm, this will seem similar to the relative accuracy metric proposed by [22] given in Equation 3. The key similarity with relative accuracy that hinders MASE as a metric for prognostic applications is that the the mean of the scaled error considers only absolute values of q_t. This means that only positive errors are considered and as has been mentioned previously, we need to classify errors as positive or negative to then penalize them depending upon the case study being considered. One option from Equation 9 is that we do not proceed to average the scaled error to calculate MASE. If only the scaled errors can be used as a metric, this would be beneficial in obtaining a positive and a negative error from the time series. Whether or not this can be applied to relative accuracy is not clear.

Since MASE is a relatively new metric only developed by [6], there are no limitations in metrics literature as such. However [21] do mention that the other conventional metrics mentioned above had been proposed from a general forecasting point of view and had shortcomings, which is in agreement with [6], albeit from a prognostics viewpoint. This was illustrated by [21] in modeling battery capacity degradation between a certain time frame. They consider modeling the RUL prediction using four algorithms each assessed by four conventional metrics (bias, SSD, MAPE, MSE) and four newer metrics including prediction horizon, relative accuracy (Equation 3), cummulative relative accuracy and convergence. The values obtained for the metrics conclude that the latter set of metrics provide an insight into the evolution of the predictions over time and when these predictions are trustworthy, unlike bias, SSD, MAPE and MSE which only providing a measure of deviation from the ground truth [21].

Uncertainties in Prognostics and Probabilistic Forecasts

In prognostics, estimation of the remaining useful life of a component is of little importance without taking into account the uncertainties associated with making such a prediction. In an ideal scenario, the RUL forecasts obtained at each time points will be point estimates resulting in a univariate time series as mentioned in Equation 2.2. In practice, this is hardly ever true because of the many uncertainties that accompany forecasts including noise in sensor measurements, modeling errors, input data uncertainties, etc [22]. Sanakaraman & Goebel (2015) classify the uncertainties that arise in prognostics as present uncertainty which deals with lack of understanding of the true state of the system at an instant in time, future uncertainty which deals with the lack of knowledge of future loading and operating conditions and modeling uncertainty [18]. Therefore, it is likely that a RUL prediction will generally have a PDF associated with it and a confidence bound. This

is also in concurrence with the point made by [29] regarding uncertainty management in prognostics. What this does, is that there is a need to also rely on metrics that are capable of evaluating a prognostics algorithm on the basis of a probability prediction, rather than a point estimate of RUL.

This then makes it crucial to enable the forecaster to predict correct probabilities. Therefore if we consider probabilistic forecasting in prognostics, a prediction may be associated with different RULs at that particular time points with different corresponding probabilities. If the most probable outcome matches the ground truth, this would be an example of a good forecast. It might then be prudent to consider the suggestion of [22] who mention that a point estimate can be extracted from the probability distribution of RUL forecasts. Yet even without converting probabilistic forecasts into a point estimate, there are methods and metrics to assess such forecasting techniques.

Perhaps it would be wise to clear a misconception between precision and accuracy before advancing to metrics. While we can consider accuracy to be a measure of deviation from the ground truth, precision is rather dependent on the nature of the probability distribution of the forecast [19]. The narrower the distribution, the more precise it is. A quantification of this can be provided with the help of the standard deviation of the distribution wherein a lower standard deviation leads to higher precision. While accuracy and precision may be considered by the novice as desired characteristics, there is a strange irony that exists. A high precision probability distribution of the RUL forecast is not always desirable as highlighted by [17]. They argue that it is rarely possible to obtain an RUL estimate with complete precision without computing the associated uncertainties. This results in the thought that a highly precise forecast may be considered to be inaccurate. The stochastic nature of the EoL point is also highlighted by [17], which illustrates the point made earlier regarding a higher preference for probabilistic forecast metrics. Accuracy and precision of a probabilistic forecast is shown in Figure 13.

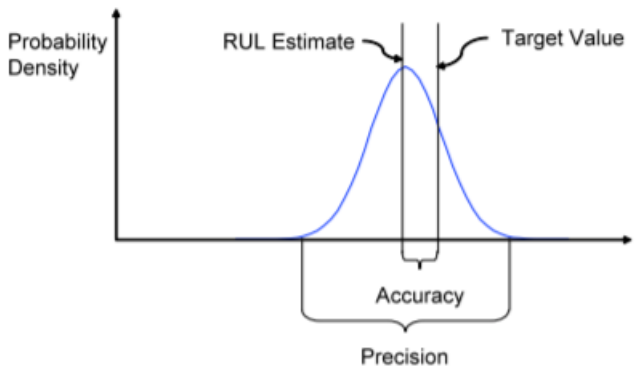


Figure 13: Illustration of accuracy and precision for RUL [29]

It should made me made clear here that metrics discussed in Equation 2.2 lack the depth necessary to assess features of a probabilistic forecast, whose dynamics are very much different from that of a deterministic forecast. This can be highlighted by a very simple example. Consider a certain repeatable event with a probability of positive outcome being 15%, which is the ground truth. Consider now two forecasting models attempting to estimate the above positive outcome probability. If Model A estimates the probability as 0% and Model B estimates the probability as 30%, it is more likely that Model B will be considered as the one with a better forecast since Model A says that event will not occur at all, while it does in reality with a 15% probability. Both the above models according to a conventional metric such as MSE only give an error of 0.15, which provides nothing to distinguish between the two models. This is in a way, similar to the limitation to using correlation coefficient as a stand alone metric as observed from Figure 6. Fortunately for the researcher, there is enough literature focusing on metrics that are capable on evaluating probabilistic forecasts. One such concept or domain rather, is known as scoring rules.

Before getting to scoring rules, prediction horizon, α - λ accuracy mentioned in subsection 2.2 have also been modified to consider probabilistic forecasts as highlighted in [22]. As discussed earlier, these metrics rely on accuracy of predictions (for point forecasts and probabilistic forecasts) and not the evolution of predictions

over time. The earlier a prediction satisfies a particular criterion known as the β criterion (resulting in a large PH), the better is the accuracy of the prediction algorithm. For example, the length of the PH for a certain probabilistic prediction at k is shown in Figure 14. The author has mentioned that the first time index where a prediction satisfies the criterion is given by Equation 10.

$$i_{\alpha\beta} = \min\{j|(j \epsilon p) \wedge (\pi[r(j)]|_{-\alpha}^{+\alpha}) \geq \beta\} \text{ (see [22])} \tag{10}$$

where,
p is the set of time points involving a prediction, of which j is an instant. In Figure 14, j is replaced by k.
$r(j)$ is a RUL distribution at time t_j instead of a point forecast.
β is the minimum acceptable probability mass.
$\pi[r(j)]$ is the probability mass of the prediction PDF.
$-\alpha, +\alpha$ are the bounds given by the grey shaded region in Figure 14. While Figure 14 illustrates the length of

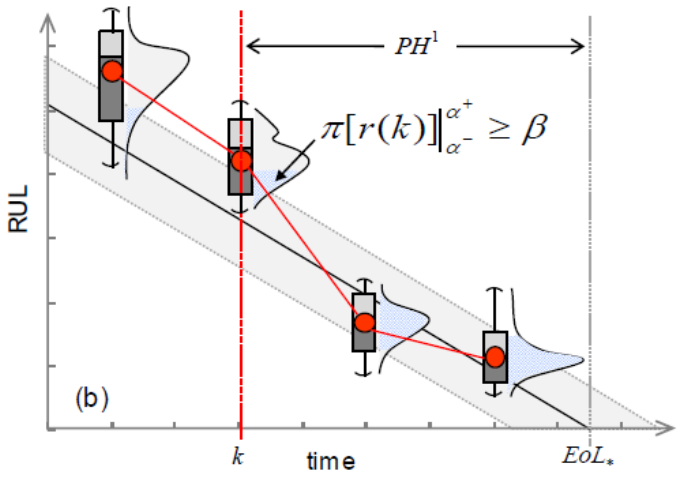


Figure 14: PH based on β criterion for probabilistic forecasts [22]

the prediction horizon for a certain RUL prediction forecast, it does not show why the first prediction in the figure does not fulfil the criterion. The second prediction that fulfils the criterion seems to be similar to the first prediction in terms of area lying outside the α bounds.

Another metric suggested by [22] and [10]) is the $\alpha - \lambda$ accuracy, whereby probabilistic forecasts are assessed upon whether they fall in a desired 'cone of accuracy' levels. This cone is formed by imposing bounds on α, dependent on the ground truth at each time point. This metric enables to study accuracy of forecasts on multiple levels and is quantified as a binary metric that takes the value of 1 if the desired prediction accuracy is satisfied at a specific time. At a glance, this metric can be seen as an extension of the prediction horizon but instead of solely giving an indication of the time point when the β criterion is satisfied, $\alpha - \lambda$ accuracy gives us an indication of how many predictions lie in the desired accuracy levels set by the user. An illustration of the desired cone of accuracy and the corresponding probabilistic RUL predictions is illustrated in Figure 15. The notations used are similar to that of prediction horizon shown in Figure 14.

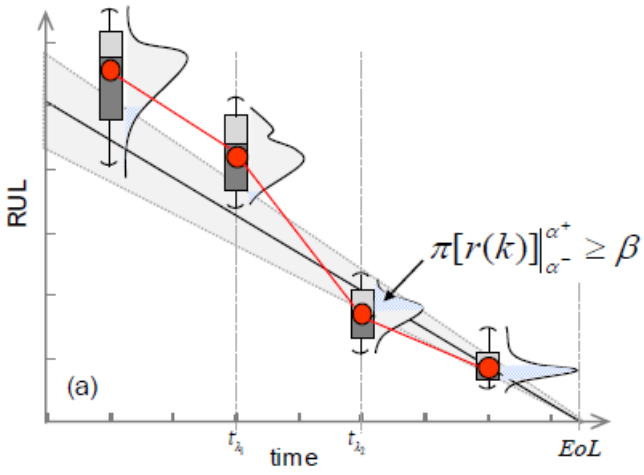


Figure 15: $\alpha - \lambda$ accuracy with the accuracy cone shrinking with time on RUL vs. time plot [22]

It was pointed out by [20] that the increasing accuracy in RUL predictions should be observed closer to the EoL of the component and this example seems to highlight that point. It can be seen that the cone of accuracy shrinks with time leading to increased accuracy of the predictions and precision of the PDF. This is an improvement over the prediction horizon which solely gives an indication of the time point when the β criterion is satisfied.

While it is of significant importance to discern which predictions fall into the desired accuracy levels set by the user, it also shows a certain amount of untapped potential that the metric possesses. For instance, a scoring function can be used in conjunction with the $\alpha - \lambda$ accuracy to penalize predictions that lie outside the bounds of the cone of accuracy. This would also help in countering the limitation of the scoring functions, which so far have been shown to be applicable only to point forecasts dependent on the sign of the prediction error. By dividing each probabilistic forecast PDF into regions, it can be observed what percentage of the forecast lies outside the bounds of the cone and the error can be obtained with respect to the ground truth estimate. When the ground truth estimate is superimposed on the forecast PDF, it can be determined what percentage of the PDF indicates an overestimate and what percentage of the PDF indicates an underestimate.

Scoring Rules

While there have been no contributions towards prognostics, the domain of scoring rules has gained popularity when it comes to evaluating probabilistic forecasts. This section aims to highlight the untapped potential that scoring rules possess when it comes to applicability to performance evaluation in prognostics. These may be considered similar to scoring functions, but the way in which a score is assigned is vastly different. An overview of the different scoring rules used to evaluate probabilistic forecasts is presented by [3]. If X is used to represent the observed value of a random variable (ground truth) and $p(x)$ represent the probabilities in the forecast for each value x, a score can be computed as $S[p(x), X]$, which can then be aggregated over all the N forecasts as follows. A lower overall score is preferred.

$$S = \frac{1}{N}\sum_{i}^{N} S[p_i(x), X_i] \text{ (see [3])} \tag{11}$$

Many publications have covered different types of scores such as ignorance, Brier score, mean square error, etc. [16] provide the relationship for the ignorance score, which depends only on the probability of the ground truth as follows. Considering the probability of the forecast which corresponds to the ground truth value shows that the confidence with which the correct forecast is made matters in the forecast evaluation.

$$S[p(x), X] = -\log[p(X)] \text{ (see [16])} \tag{12}$$

The main difference such an evaluation metric has with respect to scoring functions discussed in subsection 2.2 is that the prediction error is not considered, which was considered to be a limitation in most other metrics,

especially those discussed in Equation 2.2. Equation 12 however considers the confidence of the forecast associated with the true observed value, which means that the RUL prediction error is not the basis of such a metric and is hence, not required. A similar scoring rule to this is the logarithmic scoring rule, which scores on the basis of the likelihood of an outcome. If there is a positive outcome with 80% probability, the score assigned is ln(0.8) and a score of ln(0.2) is assigned to the other outcome. The goal in this case is to maximise the score, since the sign is opposite to that of Equation 12. Thus if the forecaster predicts an event with a high level of confidence and it turns out to be incorrect, this is an example of a poor forecast in which case the penalty can rise exponentially, asymptotically reaching negative infinity. This property is known as the asymptotic property.

There are two properties of scoring rules as described by [24] that are worth mentioning here - hypersensitivity and insensitivity. Insensitivity corresponds to the fact that if a particular forecast receives a score of infinity, then the rest of the predictions made by the algorithm do not matter. The logarithmic nature of the scoring rule in Equation 12 means that this is a possibility and enables the forecaster to accurately make predictions. While it is not possible to compare this to a deterministic forecast, this property might hold some value when considering scoring functions from Figure 12. An exponential rise in penalty is observed for overestimates and underestimates in the figure and if a prediction there is penalized by infinity, then even a good prognostics model may be deemed as a bad one. According to hypersensitivity, the asymptotes can become extreme very quickly for a small difference in probabilities. In terms of a deterministic forecast made by the scoring functions, the exponential rise observed there is observed for a small difference in RUL prediction error. The author believes that this difference should not be too extreme, which may well be the case with the logarithmic scoring rule as well.

Another interesting approach would be the incorporation of a conventional deterministic forecasting metric to deal with a probabilistic forecast. The mean square error measures the square of the difference between the forecast value and the ground truth. Many publications have illustrated the Brier Score in research papers that essentially performs the same task for a probabilistic forecast. If the probability interval [0,1] is subdivided into m mutually exclusive sub-intervals, labelled by $k = 1,2,...,m$ and n_k number of forecasts falling in the k^{th} bin are denoted by f_{kj} with $j = 1,2,...,n_k$ and o_{kj} are the corresponding ground truth values, the Brier Score according to [27] can be determined as follows. As it was with the ignorance score, a lower score is the indication of a better forecast.

$$BS = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj} - o_{kj})^2 \text{ (see [27])} \tag{13}$$

Also, [3] simplify the Brier Score given Equation 13 as follows. The notation for the variables is borrowed from Equation 11.

$$BS = S(p,X) = (X-p)^2 \text{ (see [3])} \tag{14}$$

The one reason why such a metric lacks in a prognostics application is the binary nature of the random variable X. Probabilistic RUL forecasts usually contain a wider range of random variables rather than just $X = 0$ or $X = 1$. While a RUL forecast can be divided into a binary forecast by introducing some sort of threshold (0 if the RUL is below a certain value and 1 if it is above a certain value), the ignorance score would be more preferable. The binary nature of the random variable aside, [7] was also particularly critical on the use of the Brier Score as a metric to evaluate a probabilistic forecast. In his paper, the two forecasts were considered for the same data set and the differences in the Brier Scores for both the forecasts (b_1 and b_2) were estimated by first expanding Equation 14. It was observed that the differences in Brier Score was positive which meant that $b_2 > b_1$, indicating that the forecast corresponding to b_1 (f_1) is a better forecast. But it was observed that the better forecast f_1 had even predicted a zero probability, which is a contradiction in itself [7]. Hence, the author recommends the use of scores that use the concept of likelihood as a replacement for the Brier Score. Even in Brier's original research paper, there was no backing provided as to why the mean square error should be used as a scoring rule over the other available ones.

3. Problem Definition

This section deals with the problem definition for this thesis project along with a few research objectives. As was observed in section 2, several gaps were established and arguments were given regarding the research work in prognostics performance evaluation. To try and fill these gaps, the following research question is proposed.

3.1. Research Objective

The following research objective has been established after considering the research gaps observed in section 2.

To evaluate the accuracy of RUL predictions of a prognostics algorithm by proposing modifications to currently available performance evaluation metrics.

A key aspect to remember is that error scoring functions and other performance metrics need to be specified to validate RUL predictions of an algorithm so that they overcome the limitations of conventional metrics such as root mean square error, mean absolute error, etc. These metrics fail to capture essential features such as late predictions may be more prejudicial than early predictions, that accuracy of predictions should be particularly higher closer to the end of life of the component and that RUL estimations should exhibit desirable properties such as monotonicity and trendability.

When validating the metrics developed on a real prognostics case, only the performance evaluation of algorithms is focused on, with the limitations being that no part of the existing algorithms will be modified. This will provide an unbiased assessment of the algorithms. Input measurements from sensors for prognostic model development may contain noise and non-monotonic trends and if this has not been treated by the code developer, it is out of scope of the assignment. This thesis can therefore be considered to be part of post-processing of the results (RUL forecasts) and not pre-processing of the input features.

3.2. Research Question

How can currently available performance evaluation metrics be modified to evaluate RUL forecasts of a prognostics algorithm and visualize the corresponding performance?

To break down the research question into smaller fragments and to help with the overall project planning, the following sub-questions are established, along with a few additional questions that need to be answered alongside.

1. Which performance evaluation metrics can be applied to a RUL prediction algorithm to quantify prediction error?
    - Which metrics consider prediction accuracy when it comes to RUL estimation, and how is prediction accuracy quantified by the metric?
    - What other aspects need to be addressed by the metric when it comes to assessing a prognostics algorithm?
2. What are the limitations of each performance evaluation metric in the context of the application?
    - What parameters do each of the studied metrics deal with? For instance, RMSE deals with the deviation from the ground truth whereas prediction horizon deals with the accuracy of predictions.
    - What are the hybrid solutions, if they already exist, wherein limitations of two or more metrics are solved by working in combination?
3. How can a new scoring function be proposed that can deal with the shortcomings from the currently available scoring function in literature?
    - How can the constants be determined so that the RUL predictions are penalized in the appropriate manner for a case study?
    - What should be the nature of the curve when observing the plot between score versus prediction error?
    - Which forecasts should be penalized more depending upon the nature of the system being tested?
    - How worse are inaccurate predictions close to the end of life than at the beginning?
    - What assumptions should be made in order to arrive at this scoring function without any bias when formulating it for a specific case study?
4. How can the $\alpha - \lambda$ accuracy metric be amended to overcome its current limitations?

- How can prediction accuracy be computed for a particular test unit and a certain value of α?
- What effect does a variation in the α parameter have over the prediction accuracy?
- How can the changing prediction accuracy be visualized over the life of the unit under test?
- Will this metric be used in conjunction with the scoring function or is there a way to incorporate early and late predictions in this metric?

5. How can this framework of metrics be incorporated into an existing prognostics case study?

- How do the set of constants in the scoring function need to be modified to suit the case study?
- How does the algorithm in the case study perform in terms of accuracy of predictions as compared to the algorithms used for metric development?
- What set of conclusions can be drawn when these metrics are applied to the case study?

4. Methodology

From section 2, an overview of the literature involving prognostics models, forecasting error metrics such as MAE, MAPE, RMSE, etc and specific metrics applied to prognostics such as PH, $\alpha - \lambda$ accuracy, convergence and relative accuracy have been discussed. Furthermore, there has also been sufficient emphasis laid on scoring rules in the case that the algorithm prediction works on probabilistic forecasting. The research gap that currently exists between defining performance evaluation metrics of prognostics algorithms and the selection of the right metrics for an application has been substantiated upon in the literature review. Hence, the purpose of this section will be to see how the work of previous researchers can be improved upon taking into account the research questions laid out in section 3.

The main idea would be to work with an existing prognostics algorithm that computes remaining useful life predictions with the help of data-driven approaches. In the initial phase, a sample code that performs RUL estimation for components (or units) will be used. Two different machine learning models in this code will be used to compute RUL - namely random forest (RF) regression and multilayer perceptron (MLP). The most ideal data-set used in the development of prognostic models is the one generated by the C-MAPSS simulation tool used to simulate an aircraft turbofan engine and this will be the dataset the code makes use of. A framework of metrics will then be developed keeping in mind the limitations discussed in the gap of existing performance evaluation metrics. Depending upon the nature of the forecast, these metrics as discussed earlier can be categorized into deterministic and probabilistic. The main idea for the deterministic metric would be an improved scoring function, which accounts for tolerance as discussed in subsection 2.2. The idea is to then modify the currently existing $\alpha - \lambda$ accuracy metric to compute the prediction accuracy for each test unit, upon which further analyses will be performed such as establishing the relation between the α parameter and the prediction accuracy which was discussed as a limitation of the currently available metric. Variation of the α parameter will then be a sensitivity analysis which will affect the accuracy. Further ways to visualize performance of the prognostics algorithms will be taken into consideration. Post the mid-term review, the framework of metrics developed on the sample code will be expanded to a real publicly available prognostics algorithm to prove the validity of the metrics.

The expected results should be plots showing the evolution of prognostics prediction errors over each time point a prediction is made by an algorithm. Validation on a real prognostics case study should show that the essential features are captured by the metrics such as penalizing predictions correctly on the basis of accuracy and whether they are early or late predictions. The publicly available algorithms for the estimation of remaining useful life of the UUT are mostly available in Python, and hence this will be the programming language that will be used for the remainder of the project. Being an open-source software with many packages available, Python aids in the addition of performance evaluation code to the existing code.

5. Conclusions

The overview of the literature provided has been discussed according to whether the RUL forecasts generated are deterministic or probabilistic in nature. This affects the kind of performance metrics that will be used to assess the RUL predictions made by a prognostics algorithm. Relative accuracy, prediction horizon, confidence interval, $\alpha - \lambda$ accuracy and scoring functions were discussed upon in the case of a deterministic forecasts.

Most of these metrics did not take into account the prediction error between the true RUL and the ground truth itself, but rather focused on the accuracy of the forecast. Scoring functions, used to penalize late predictions over early predictions, were studied from three different case studies and compared. The results showed an odd peculiarity where the penalty for late predictions and early predictions were observed to be similar for a few values of prediction error given the symmetric nature of the profile. This of course does not illustrate the importance of penalizing late predictions more significantly than early predictions. One particular case study also illustrated the use of a correlation metric to determine the correlation between the true RUL and the forecast RUL, only when the scoring function proved indecisive [20].

Conventional metrics discussed in Equation 2.2 covered metrics such as MAPE, MdAPE, sMAPE, sMdAPE, etc. with similar limitations wherein a majority of these metrics take a value of infinity, giving no indication on the quality of the forecast. One such improvement was proposed by [6] with the introduction of the mean absolute scaled error which improved upon the earlier conventional forecasting metrics. Their use in prognostics however was found to be limited as they do not highlight the evolution of prediction errors over time.

Several uncertainties that are observed when developing prognostics algorithms were subsequently highlighted in Equation 2.2. This ensures that a probabilistic forecast would be favorable over a deterministic forecast, thereby enabling a different set of metrics to be considered for performance evaluation. Prediction horizon and $\alpha - \lambda$ accuracy were modified to account for forecasts with uncertainty, which is already available in literature. A suggestion was also provided wherein a scoring function could be incorporated in conjunction with $\alpha - \lambda$ accuracy. Lastly, the section concluded by an emphasis on scoring rules which is a domain dealing with evaluation of probabilistic forecasts by awarding a score to each forecast. This score needs to be minimized. Scores such as ignorance, logarithmic loss and the Brier score were explored. The scoring functions discussed in subsection 2.2 were studied on the basis of a few properties dedicated to scoring rules such as hypersensitivity and insensitivity. The general limitations of the Brier score were finally highlighted.

6. Project Planning

The project timeline, that has been planned during the literature review phase has been illustrated in Figure 4 the form of a Gantt chart. The literature review phase occurs before the kickoff and therefore has not been highlighted in the chart. Note that currently, all meetings with the thesis supervisor are scheduled over video conferencing because of the COVID-19 pandemic. These meetings take place biweekly before the kickoff phase. The frequency of these meetings may increase if needed post kickoff.

The priority will be given to maintaining the schedule but if a certain task takes more time than expected, the Gantt chart will be updated accordingly. Note that this is the initial subdivision of tasks and if more tasks need to be added with dependencies, this will again update the plan. Usually the mid term review should take place three months after the kickoff meeting. In the initial version of the project plan, this is set at four months, since the remaining tasks mostly involving the final thesis report documentation and presentation are deemed possible to be completed in the remaining two months for the green light review.

III

Research Methodologies
previously graded under AE4010

1. Executive Summary

Condition based maintenance (CBM), or predictive maintenance has been the trend in the maintenance of systems and components in aviation, replacing preventive maintenance techniques such as hard time maintenance where maintenance takes place after certain time intervals. Preventive maintenance is a more conservative approach where additional maintenance costs are compromised over failure of the component. Prognostics is the type of CBM wherein forecasts of remaining useful life (RUL) of a component are constantly made available throughout its useful life so that a decision can be taken whether to schedule a maintenance check or not, which is becoming the trend increasingly as more sensor data is made available to monitor system health. This form of maintenance is in opposition to diagnostics where the cause of failure is determined once the end of life (EoL) threshold is crossed, and the component enters a state of failure. The lack of standardized performance metrics that deal with the verification and validation of these RUL forecasts has been the Achilles Heel in terms of progress in prognostics applications. The project proposal focuses on the modification of these performance evaluation metrics by developing a solution that captures essential features of a RUL forecast. One such example is that predictions in which the predicted RUL is greater than the true RUL will be penalized more than the other way around because an overprediction, as will be discussed later, runs the risk of failure of the system. Developing a set of metrics that consider wider aspects of RUL forecasts such as the sign of the prediction error and its significance in terms of the useful life of the component, will no doubt lead to more reliance on prognosis in maintenance organizations. With increasing acceptance of prognostics by maintenance organizations, there is the added benefit of potential cost savings since prognostics ensures that additional maintenance checks for systems can be avoided, parallelly ensuring there are advance warnings before the system fails. The same has been highlighted by [28].

2. Introduction

This project relates to the domain of prognostics and health management (PHM), whereby corrective action is taken with the help of a forecast for the RUL of a component. Any system or a component, especially in the case of non-repairables exhibit a certain failure degradation pattern during their lifetime when in operation. In an idealistic scenario, this degradation can be seen as a linear decrease of remaining life of the component over time till the component crosses the EoL, post which it enters a state of failure. Non-repairables cannot be operated on beyond the state of failure. In time-based maintenance techniques, the component undergoes maintenance checks at frequent intervals to ensure that a state of functionality is restored for future use. Being far from ideal, RUL forecasts from prognostics models are laden with errors and hence, there is a need to validate the predictions made by these models. Prognostic maintenance is becoming increasingly adopted in maintenance organizations with the increasing availability of sensor data, leveraging that data to monitor the health of systems and components culminating in the a forecast of the RUL. Diagnostics on the other hand according to [14], is the act of knowing when a problem is taking place, identifying and isolating the fault. Most applications of prognostics, or predictive maintenance use different datasets for the estimation of RUL. Most prognostics models can be classified as physics-based models, data-driven models, or hybrid approaches. Physics based models consider mathematical models developed solely on the failure degradation understanding of components, which require greater knowledge of failure mechanisms whereas data-driven models rely on sensor measurements and previous events such as maintenance checks or repairs that had been carried out on the system or part [1]. Statistical, or data-driven prognostics are the most widely used approaches, followed by AI models and then physics-based models as has been illustrated by [10]. An argument is made by [1] that this may be attributed to the high implementation cost and the fact that some simplified assumptions lower the range of applicability of physics-based algorithms.

3. State-of-the-art/Literature Review

Performance evaluation has consisted of significant gaps in terms of research work in prognostics and health management (PHM) of systems and components. Lack of standardized metrics has been a hindrance in the progress of PHM and the metrics that currently exist have largely not been applied to prognostics algorithms to assess their validity, which is crucial for prognostics to expand in a large operational scenario to aid decision making. It would therefore be prudent to study the progress of performance metrics that are applicable to prognostics purely from a research point of view. There have been metrics outside the prognostics domain that have been considered because of their applications in general forecasting techniques. From an engineering research point of view as pointed out by [5], algorithm performance metrics and computational performance

metrics are of particular importance to evaluate a prognostics algorithm from a research viewpoint. The best estimate of the RUL obtained from a prognostics algorithm is then comparable to the actual reality, hereafter referred to as ground truth. The ground truth can hence be defined as the best estimate of a certain feature or a parameter, which in this case can is the true RUL. The available metrics that will be discussed in this state-of-the-art can broadly be classified as prognostics-based metrics, conventional forecasting metrics, metrics dedicated to probabilistic forecasting.

3.1. Prognostics Based Metrics

As is the case in most forecasting scenarios, not all algorithms are ideal. This means that there is often a difference between the parameter estimated from the algorithm and the ground truth. This introduces a key aspect in performance evaluation, that is known as the prediction error. Some algorithms denote this as the difference between the true RUL and the predicted RUL whereas other algorithms assume the other way around. Therefore, in the category of algorithm performance metrics, error-based metrics are of particular importance. Most of the prognostics based metrics used to validate have been proposed by various authors often feature prediction horizon (PH), - accuracy, relative accuracy and convergence [22][5]. Lastly an emphasis has been laid on another metric known as a scoring function, which captures the key features of a prediction error in terms of an overestimate and an underestimate. Scoring functions have been studied from three separate case studies mentioned in [20], [4] and [13]. PH is used to determine the time index when the RUL prediction made satisfies a particular criterion known as the -criterion, whereas - accuracy assume a certain accuracy level for each time index that varies with time and the ground truth value at each instant. A value of 1 is assigned to those RUL predictions which fall into the accuracy levels and a value of 0 is assigned to those that do not. Both these metrics come under accuracy-based metrics. To be critical, it would not be possible to extend PH to consider overestimates and underestimates in terms of prediction errors as that metric only gives the time point at which the first prediction satisfies the -criterion. While these metrics are valid for point forecasts, their use in probabilistic forecasts is much more important, as is discussed in subsection 3.3.

The most important metric that needs to be discussed under this section is the scoring function. A scoring function either penalizes overestimates or underestimates depending upon how critical the component under consideration is to the operator. The goal in performance evaluation with the help of a scoring function is to aggregate the score at every time point and minimize the overall score. As mentioned earlier, three separate case studies will be considered to highlight the significance of the scoring function. In modeling the damage propagation for an aircraft engine with a run to failure (RTF) simulation using the C-MAPSS tool for aircraft engine, [20] developed the following scoring function to penalize late predictions. The prediction error in the following equation is denoted by d.

$$s = \begin{cases} \sum_{i=1}^{n} e^{-\left(\frac{d}{a_1}\right)} - 1 \text{ for } d < 0 \\ \sum_{i=1}^{n} e^{\left(\frac{d}{a_2}\right)} - 1 \text{ for } d \geq 0 \end{cases} \quad \text{(see [20])} \tag{1}$$

The parameters have been set to a1 = 10 and a2 = 13 by [20]. The one limitation here is that the author has not states as to how the constants came into consideration. Perhaps it would have been ideal to illustrate the effect of variation of the arbitrary constants on the performance evaluation. Nevertheless, as can be seen in the above scoring function, the higher score indicating a higher penalty is reserved for predictions in which the prediction error is positive. This is an RUL overestimate, and this would lead to the operator thinking that there is still more useful life that remains whereas the actual failure is due much earlier. It may be a conservative approach, but the author has managed to strike a tradeoff between additional maintenance checks/costs and the risk of failure. The author has also mentioned that the accuracy of predictions should be greater towards the EoL of the component, which is a claim that requires further work. Currently, the penalties awarded to RUL predictions do not take into account that the EOL errors are more crucial. The $\alpha - \lambda$ accuracy metric that will be described in subsection 3.3 takes into account EOL prediction accuracy and hence solves the limitation of the scoring function.

The following scoring function was incorporated by [4] (Equation 2) in their submission to the PHM Data Challenge 2010 to model the wear after each cut made by the cutting tool of a CNC milling machine. The number of cuts that could made be by the cutting tool of the CNC after a certain wear on the cutting tool was estimated in this prognostics model. This problem can be considered as an analogy to the traditional

RUL estimation prognostics model where the RUL is analogous to the remaining number of cuts that could be made whereas the wear on the cutting tool can be considered being analogous to failure degradation of a component. In this case, the prediction error is denoted by . The one change here is that the exponential term has been modified with the denominator changing from 13 in [20] to 4.5. This changes the way over predictions are penalized. It is to be noted here that the following equation calculates the scores for each time point and is not an aggregate. Summing these scores over all time intervals will give the overall penalty.

$$S(\delta) = \begin{cases} e^{-(\frac{\delta}{10})} - 1 \text{ for } \delta < 0 \\ e^{(\frac{\delta}{4.5})} - 1 \text{ for } \delta \geq 0, \end{cases} \quad \text{(see [4])} \tag{2}$$

This review will also make use of the scoring function discussed by [13] for a deep convolution neural networks prognostics model which forecast RUL for the same C-MAPSS data-set used by [20]. The values of the constants used in the scoring function were borrowed from that of the scoring function in [20] for comparison. The only difference was that the constants were interchanged to change the magnitude of penalty in this case. To compare the different approaches undertaken by the various authors, a resulting plot of score as a function of prediction error was plotted on Python using the matplotlib and numpy packages. This resulted in the following profile. This shows how particular components need different scoring functions based on how

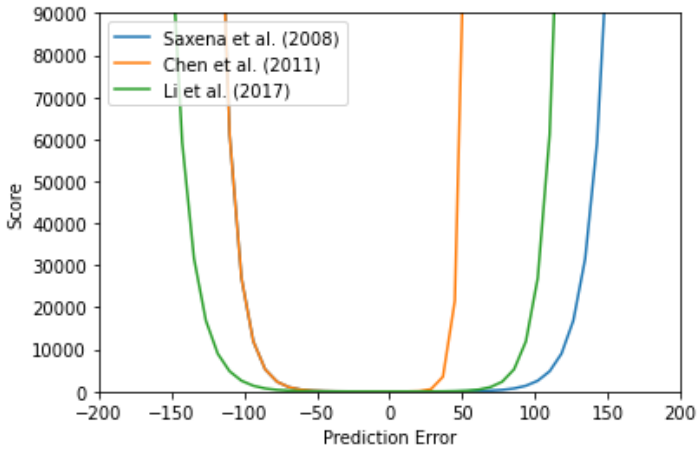


Figure 1: Score as a function of prediction errors for scoring functions highlighted in [20], [4] [13]

critical they are in terms of needing maintenance or quick repair. For instance if we consider the scoring function illustrated by [4] and [20], both penalize early predictions in a similar manner since their plots coincide for negative prediction error but because of a more conservative approach that was required in the case of the cutting tool from the CNC milling machine, positive error predictions are penalized much earlier than they are for an aircraft engine scenario as is shown by [20]. Using the same data set for RUL predictions as [13], [20] penalize very early predictions as well as significantly late predictions in a similar manner. A correlation coefficient metric was also proposed if the scoring function proved to be indecisive in the case of fleet of components, all ending up with the same score [20]. The correlation of the predictions with respect to the ground truth where a high correlation would be example of a good forecast was suggested. There is however an understanding that even this has limitations. The correlation coefficient could also end up being the same for underestimates and overestimates if those predictions are symmetric about the ground truth. Secondly, [11] argue that coefficient of correlation (r) cannot be used to determine the quality of forecasts as this measure only determines the goodness of fit between data (predicted RUL and ground truth in this case).

3.2. Conventional Metrics

It is also practical to focus on a slightly broader picture where we do not only consider prognostics and RUL estimation, but forecasting error and how we can evaluate the quality of a prediction in any generalized forecasting model. Prognostics can then be considered as a subset of the many forecasting techniques. Given that such metrics in most cases do not distinguish between overestimates and underestimates, they will usually be second in priority to the metrics discussed in subsection 3.1. Assuming that RUL estimations are point forecasts, forecasts at individual time points can be considered as a univariate time series with the

only varying quantity being the RUL. As such, there are metrics available such as mean absolute percentage error (MAPE), mean absolute scaled error (MASE), etc. that will be discussed here.

It was proposed by [6] that the mean absolute scaled error (MASE) be used as the standard for the comparison of multiple time series in forecasting. They compared different forecasting techniques considering three different time series datasets. Metrics that had been used before the publication of the paper had been applied to the time series including MAPE, median absolute percentage error (MdAPE), symmetric MAPE (sMAPE), symmetric MdAPE (sMdAPE), median relative absolute error (MdRAE), geometric relative absolute error (GMRAE) and MASE. The authors also highlight that most publications until that point recommended the use of MAPE for determining forecast accuracy.

A common trend observed in the study was that when a forecast for a particular time index is zero, all the corresponding error metrics (mentioned above) take an undefined or an infinite value, giving no indication on the quality of the forecast. This shortcoming has also been highlighted by [9] as a gap for their proposal of a new metric. Thus, [6] propose a scale free error metric known as the scaled error, shown below and the mean of the scaled error is the MASE . In the equation, e_t is the forecast error between the forecast and the ground truth ($e_t = Y_t - F_t$) whereas the observations at two successive time points is denoted by $Y(i)$ and $Y(i-1)$. MASE is then calculated as the mean of all values of q_t as shown in Equation 3.

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|} \text{ (see [6])} \tag{3}$$

The key characteristic that hinders MASE as a metric for prognostic applications is that the mean of the scaled error considers only absolute values of qt. This means that only positive errors are considered and as has been mentioned previously, we need to classify errors as positive or negative to then penalize them depending upon the case study being considered.

Since MASE is a relatively new metric, there are no limitations in metrics literature as such. However, [21] do mention that the other conventional metrics mentioned above had been proposed from a general forecasting point of view and had shortcomings, which is in agreement with [6], albeit from a prognostics viewpoint. This was illustrated by [21] in modeling the battery capacity degradation between a certain time frame. They consider modeling the RUL prediction using four algorithms each assessed by four conventional metrics (bias, SSD, MAPE, MSE) and four newer metrics including prediction horizon, relative accuracy, cumulative relative accuracy convergence. The values obtained for the metrics conclude that the latter set of metrics provide an insight into the evolution of the predictions over time and when these predictions are trustworthy, unlike conventional metrics which only providing a measure of deviation from the ground truth [21].

Another interesting feature from [13] was observed when the scoring function discussed in subsection 3.1 was used in conjunction with the root mean square error to validate their prognostics algorithm. The root mean square error is a well-known error metric that squares the prediction errors over all the time points, averages them and then the square root is taken. The results are illustrated in Figure 2. In fact, the RMSE was

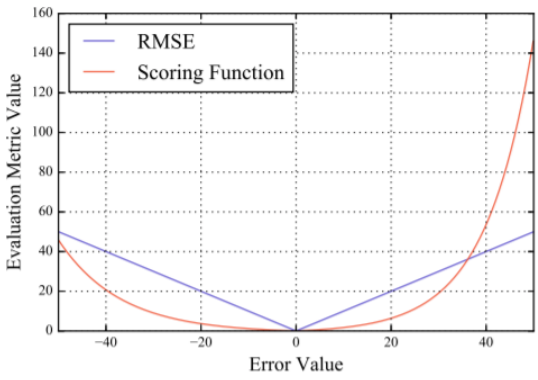


Figure 2: Scoring function used in conjunction with the RMSE [13]

also used by [31] to assess their data-driven prognostics model, involving RUL prediction methodology for

an electric motor case. The limitation of root mean square error as observed in Figure 2 is that it gives a very symmetric profile when plotted with respect to the prediction error. If one wishes to study the significance of predicting an overestimated RUL forecast and an underestimated RUL forecast, the limitations of RMSE are clear. Both overestimates and underestimates are penalized in a similar manner, whereas as can be seen for the scoring function, the penalty is awarded in an asymmetric manner to distinguish between late and early RUL predictions. This limitation has been observed in most conventional metrics discussed in this section and is perfectly in agreement with the argument put forward by [21].

3.3. Uncertainties in Prognostics and Probabilistic Forecasts

In prognostics, estimation of the remaining useful life of a component is of little importance without considering the uncertainties associated with making such a prediction. In an ideal scenario, the RUL forecasts obtained at each time point will be point estimates resulting in a univariate time series as mentioned in subsection 3.2. In practice, this is hardly ever true because of the many uncertainties that accompany forecasts including noise in sensor measurements, modeling errors, input data uncertainties, etc. Therefore, it is likely that a RUL prediction will generally have a PDF associated with it and a confidence bound. This is also in concurrence with the point made by [29] regarding uncertainty management in prognostics. What this does, is that there is a need to also rely on metrics that can evaluate a prognostics algorithm based on a probability prediction, rather than a point estimate of RUL. This also makes it difficult to compare metrics that have been discussed for point estimates. Fortunately for the researcher, there is enough literature focusing on metrics that are capable on evaluating probabilistic forecasts.

One such concept or domain rather, is known as scoring rules. Only the Brier score will be looked at here. Before getting to scoring rules, prediction horizon, accuracy mentioned in subsection 3.1 have also been modified to consider probabilistic forecasts as shown by [22]. The -criterion (illustrated in Figure 3) ensures that the sufficient part of the PDF lies inside the accuracy levels, where is the minimum acceptable probability and the left hand side of the inequality is the probability mass of the PDF of the prediction within the bounds [22]. This counters the limitation of applying such a metric to solely point forecasts, yet there is scope for extending this concept further to include scoring functions as will be discussed later in this proposal. Scoring

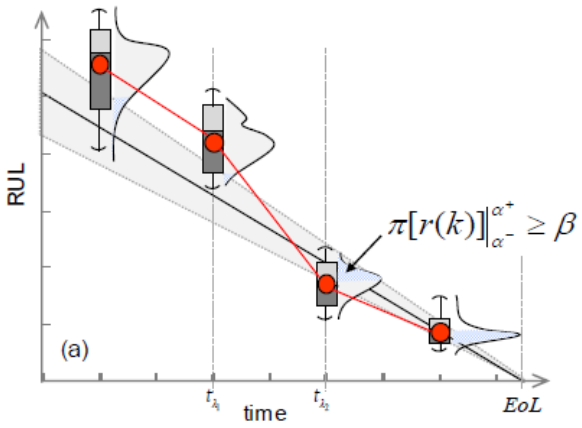


Figure 3: $\alpha - \lambda$ accuracy cone with the cone shrinking over time [22]

rules compare the ground truth to each of the probabilistic forecasts in the PDF. Therefore if X is the ground truth and $p(x)$ is a single probability in the PDF of the forecast, a score is computed as a function of X and $p(x)$ denoted by $S[X, p(x)]$ [3]. There are many types of scores available that can be computed to evaluate a probabilistic forecast, but one such score that requires emphasis here is the Brier score, which is defined for only a binary random variable X. The score is then computed as shown in Equation 4 [3].

$$S[X, p(x)] = (X - p(x))^2 \tag{4}$$

This metric has also been extended to a series of forecasts by considering the average over the series of forecasts, where a lower overall Brier score is desirable [30]. The limitation here is that a probability mass function is considered in the case of the Brier score, which means the metric has not yet been explored for a

continuous random variable series forecast and there is no clarity whether the Brier score can be extended to forecasts beyond binary which will be the case in the selected case studies. Gaps therefore have been identified in this state-of-the-art regarding the use of metrics in prognostics algorithms performance evaluation and each metric has its shortcoming when used individually, especially in the case of conventional metrics. With the research questions and objectives laid out in the next section, a solution is then proposed at for developing a set of metrics in section 5.

4. Research Question, Aim/Objectives and Sub-Goals

The research questions and the objectives have been discussed in subsection 4.1 and subsection 4.2.

4.1. Research Question

The main research question for this thesis assignment will be as follows.

How can currently available performance evaluation metrics be modified to evaluate and visualize performance of prognostics algorithms?

This research question will be answered with the help of the following sub-questions that have been developed keeping in mind the SMART technique.

1. Which performance evaluation metrics can be applied to a RUL prediction algorithm to quantify prediction error?
   (a) Which metrics consider prediction accuracy when it comes to RUL estimation, and how is prediction accuracy quantified by the metric?
   (b) What other aspects need to be addressed by the metric when it comes to assessing a prognostics algorithm?
2. What are the limitations of each performance evaluation metric in the context of the application?
   (a) What parameters do each of the studied metrics deal with? For instance, RMSE deals with the deviation from the ground truth whereas prediction horizon deals with the accuracy of predictions.
   (b) What are the hybrid solutions, if they already exist, wherein limitations of two or more metrics are solved by working in combination?
3. How can the currently available performance evaluation metrics be amended to suit the probabilistic RUL forecasts made by a prognostics algorithm?
   (a) How is the ground truth of the RUL forecasts presented by the algorithm?
   (b) What are the uncertainties involved in the prognostics algorithm?
   (c) How do the currently available metrics relate to early predictions or late predictions?
   (d) Which specific probabilistic forecasting metrics can be assigned to prognostics applications?
4. How can the modified metrics deal with evaluation of different prognostics algorithms?
   (a) Does the scoring function achieve its task of penalizing late predictions in the algorithms?
   (b) Do the metrics give the evolution of prediction errors over time?
   (c) Is a clear distinction provided between forecasts that lie within the given accuracy levels or not?

4.2. Research Objective

The following research objective is established from the gaps outlined in the literature review.

To evaluate and visualize the performance of prognostics models by proposing modifications to currently available performance metrics.

This research objective or goal of evaluating the RUL predictions of a prognostics algorithm first needs to be decomposed into several research sub-goals to achieve the larger picture. These sub-goals are also set up

to assist the project planning stage as is shown in the Gantt Chart attached in section 8.

Firstly, it is imperative to consider how the research objective is going to be achieved. As mentioned above, the accuracy of RUL predictions will be judged by the help of proposal of performance evaluation metrics. Thereby, it would be prudent to conduct a theoretical analysis of the performance evaluation requirements of different prognostics case studies in the aerospace domain. This step will help capture the features that need to be studied during performance evaluation such as accuracy and precision of RUL predictions.

Once a foundation is laid on what the performance evaluations requirements of a prognostics algorithm are, it will be crucial to delve deeper into the specifics of different performance metrics themselves. For instance, metrics that focus on precision will be different from those that focus on accuracy of predictions. Then a variety of scoring functions are available, which would be extended to include more ways to distinguish prediction error instead of just positive and negative error. This would be a key aspect in the proposal since current available metrics will need to be modified to assess a prognostics algorithm as has been outlined in subsection 4.1. From section 3, it can be seen that the different performance metrics focus on different aspects of evaluation such as precision, accuracy, computation time of predictions whereas some metrics not discussed focus on life cycle costs, technical value, total value and ROI for the organization as well [5]. Hence, an understanding of the different performance metrics will be instrumental in contributing to the answer to the research question. This also forms the foundation for the literature review carried out for the thesis, a highlight of which is illustrated in the state-of-the-art in section 3.

A key aspect to remember is that error scoring functions and other performance metrics need to be specified to validate RUL predictions of an algorithm so that they overcome the limitations of conventional metrics such as root mean square error, mean absolute error, etc. These metrics fail to capture essential features such as late predictions may be more prejudicial than early predictions, that accuracy of predictions should be particularly higher closer to the end of life of the component and that RUL estimations should exhibit desirable properties such as monotonicity and trendability. A simple scoring function that can be applied to a point forecast penalizes late predictions (error > 0) as compared to early predictions (error < 0), depending upon the sign of the prediction error. Late predictions are penalized since it risks the component crossing the failure threshold into state of failure whereas early predictions induce unwarranted maintenance checks on the component incurring additional costs to the operator, which is not as risky as failure itself.

The research goal is scoped in such a way that only the performance evaluation of algorithms is focused on, with the limitations being that no part of the existing algorithms will be modified. This will provide an unbiased assessment of the algorithms. Input measurements from sensors for prognostic model development may contain noise and non-monotonic trends and if this has not been treated by the code developer, it is out of scope of the assignment. This thesis can therefore be considered to be part of post-processing of the results (RUL forecasts) and not pre-processing of the input features.

5. Methodology

Of the methods proposed in the state-of-the-art, this thesis will focus on the methods discussed specifically in subsection 3.3 since the chosen case study and the C-MAPSS dataset may yield a probabilistic forecast for the remaining useful life of the component at every time index and can be adapted to point forecasts as well. This will ensure that metrics linked to point forecasts be deemed insufficient to capture the essential features of an evaluation metric for probabilistic forecasts that yield a PDF instead of a single value (see subsection 3.2). When it comes to probabilistic forecasts, there is a need to modify these metrics to extract a wider range of features when it comes to performance evaluation. There will therefore have to be a hybrid approach to formulate a metric that can fill the gap in terms of accuracy of predictions and penalizing late predictions and early predictions accordingly. It is to be mentioned here that this is an initial proposed methodology and may be subject to modifications as the thesis progresses.

The $\alpha - \lambda$ accuracy metric that has been discussed for probabilistic forecasts will be the key metric that will be focused on in the methodology. Currently, this metric only considers the forecasts that lie within a certain accuracy level. A value of 1 is currently assigned to forecasts that lie within the pre-determined accuracy level and 0 is assigned when the forecast do not meet the required accuracy level [10][22]. According to the

current method, the accuracy levels are determined by imposing bounds on α : α^+ and α^-, resulting in a cone of accuracy as observed in the grey shaded region in Figure 3. These two values are calculated at every time point and therefore form a cone. If the predicted RUL forecasts lie within this cone, the predictions are considered accurate and are assigned a score of 1 and 0 otherwise.

But the goal of the thesis is to extend the same metric so that it can incorporate the corresponding ground truth estimate of the RUL at every time index. The forecast PDF can then be categorized into a region of overestimate and an underestimate. This overestimate and an underestimate will run in parallel with the accuracy levels that the metric already focuses on. The region of the PDF that lies above the ground truth value will be considered as an overestimate or a late prediction (positive prediction error) and that region of the PDF that lies below the ground truth value will be considered as an underestimate or an early prediction (negative prediction error). Remember that if we were dealing with point forecasts, the difference between the true and predicted RUL would have been easily computed but since it is a probability distribution, the PDF will have to be divided into two regions. Once this is achieved, the following classification will be made.

1. Late predictions lying outside the cone of accuracy.
2. Late predictions lying inside the cone of accuracy.
3. Early predictions lying inside the cone of accuracy.
4. Early predictions lying outside the cone of accuracy.

Classifying the PDF of the forecasts at each time point into the above four classes will aid in assessing what percentage of the forecast was a good forecast. For instance, if 80% of the PDF is an underestimate lying inside the cone of accuracy, this would be an example of a very good forecast whereas if 80% of the forecast is an overestimate lying outside the cone of accuracy, this would be an example of a bad forecast. This is when the hybrid approach mentioned earlier would come into play. Scoring functions discussed in subsection 3.1 have only considered positive and negative errors to penalize predictions. The scoring function that will be developed in this project will be developed into the four classes mentioned above. Classes 1 and 2 will be penalized for being overestimates, but Class 1 will have a higher penalty since the predictions are overestimates and at the same time lie outside the cone of accuracy whereas Class 3 will have the least penalty for being the best forecasts. Class 4 will have a higher penalty than Class 3 but less than 1 and 2. This is how the project aims to bridge the gap between a prognostics metric designed to determine accuracy and a scoring function that penalizes predictions based on their error.

Next, as a secondary metric to evaluate the prognosis of RUL forecasts, the aim is to amend the currently available scoring rules discussed in subsection 3.3. The most important out of these scoring rules would be the Brier score which is like a scoring function in that a lower Brier score is ideal for forecasting. Currently, the Brier score is only available for binary forecasts where the random variable only takes the value of 1 or 0. This project will aim to extend this to deal with RUL forecasts.

6. Experimental Setup

In terms of the experimental set-up, a significant part of the project will be performed solely from home, with regular meetings with the thesis supervisor and the chair from TU Delft who will provide expertise. As such, there will not be any assistance from any industry. Because of the COVID-19 pandemic that is currently disabling the opportunity to work at the university, no work is likely to take place at the university campus. This is of course not a major concern because of the theoretical and conceptual nature of the subject. From an experimental point of view, performance evaluation of an RUL estimation algorithm falls under the category of modifying a pre-existing computer program, thereby not requiring any visits to a laboratory, testing sites or the university itself (due to the pandemic). The exact planning has been detailed in section 8.

The algorithms for the estimation of remaining useful life of the UUT are mostly available in Python, and hence this will be the programming language that will be used for the remainder of the project. The dataset used in these algorithms will be the publicly available C-MAPSS simulation tool dataset that simulates the failure degradation of a turbofan engine so that algorithms can be compared with a standardized dataset. This is the same dataset used by [20] and [13]. Being an open-source software with many packages available,

Python aids in the addition of performance evaluation code to the existing code. Packages such as *numpy, matplotlib* and *scikitlearn* provided by the Anaconda distribution will largely be used for the performance evaluation code.

7. Expected Results

In section 5, the theory and methodology has been laid out by which this assignment aims to amend currently available performance assessment techniques of prognostics algorithms to suit the demands of the research gap. Firstly, the - accuracy metric will be considered to subdivide each RUL probabilistic forecast of different case studies into four regions and each region will be penalized with a score relative to each other, using a scoring function. The following are the key results that will be analyzed.

Firstly, the percentage of the PDF that lies in each of the four categories will be analyzed over time. For instance, at time index 1, 40% of the forecast could belong to Class I and at time index 2, 50% of the forecast could belong to Class I. This increase in forecasts belonging to Class I would be a sign of poor predictions by the algorithm. The evolution of these forecasts over time will provide an accurate representation of the accuracy of the predictions. This plot will also indicate at which time index, predictions start getting better and when they begin to get poor. Evolution of Class I forecasts over time for example should ideally be a plot that reduces over time since that is the worst category out of the four as discussed in subsection 4. This will therefore be extended to all four classes to see how each categories predictions evolve over time. Similarly, to be more critical in the assessment of RUL forecasts, the aim would be to vary to observe how this shift affects the evolution of each category of forecasts. For example, if a PDF has a higher standard deviation (less precision), with a reduction in , there is a good chance that the percentage of the forecast that was in Class II earlier would fall into Class I. Therefore, another key result would be to vary the cone of accuracy by varying and analyzing the impact on the quality of forecasts.

From the state-of-the-art documented in section 3, a few gaps were evident in performance evaluation and metric selection for certain cases. It was pointed out that not all cases were accounted for when it comes to understanding the significance of prediction error and the impacts that it can have on the organization when forecasts are overestimates or underestimates. The main motivation for this study is that if the metrics developed in this thesis can evaluate a prognostics algorithm considering the accuracy of predictions as well as the difference between an RUL overestimate and an RUL underestimate, this will fill a significant gap in performance evaluation research for prognostics and make it viable for organizations to implement prognostics in their framework with the knowledge that their forecasts can be trustworthy. As mentioned in the introduction, lack of standardized metrics for performance evaluation of prognostics algorithms has been a hindrance in the progress of this field. Implementing prognostics will no doubt bring the added potential of cost savings wherein non-essential maintenance checks can be scrapped and the component failure can be predicted.

8. Project Planning

The project timeline, that has been planned during the literature review phase has been illustrated in Figure 4 in the form of a Gantt chart. The literature review phase occurs before the kickoff and therefore has not been highlighted in the chart. Note that currently, all meetings with the thesis supervisor are scheduled over video conferencing because of the COVID-19 pandemic. These meetings take place biweekly before the kickoff phase. The frequency of these meetings may increase if needed post kickoff.

9. Conclusions

The incorporation of CBM and specifically prognostics is only going to increase as more sensor data to monitor system health is available. Continuing with a time-based maintenance approach such as hard time maintenance today would be considered as a missed opportunity with the cost savings that CBM can bring to an organization. This method also ensures maximum utilization ensuring reduced downtimes and larger availability of systems and components. This is especially crucial when dealing with a large fleet in maintenance. Even with these benefits, comes added responsibility for the researcher. Prognostics will only progress as a field when one can rely on the RUL forecasts. RUL forecasts are undoubtedly of high importance to maintenance organizations that incorporate prognostics for their fleet.

Performance evaluation metrics for prognostics algorithms are important to validate the accuracy of the RUL forecasts made by prognostics models. The lack of standardized metrics for prognostics models and the over-reliance of conventional metrics is a situation that needs to be looked at because of the many shortcomings of such evaluation tools. With added cost-benefits and wider fleet availability, the last thing such organizations would then want is to have a component enter a state of failure when a prediction had been made that the RUL was still a fair time away. Thus, with the extension of currently available performance metrics to integrate early and late RUL predictions, a critical gap will potentially be countered by amending the evaluation metrics currently designed to deal with prognostics applications and probabilistic forecasts in general. This would make a stronger case for the implementation of prognostics with increased operational reliability, at the same time motivating researchers to explore prognostics forecast validation further.

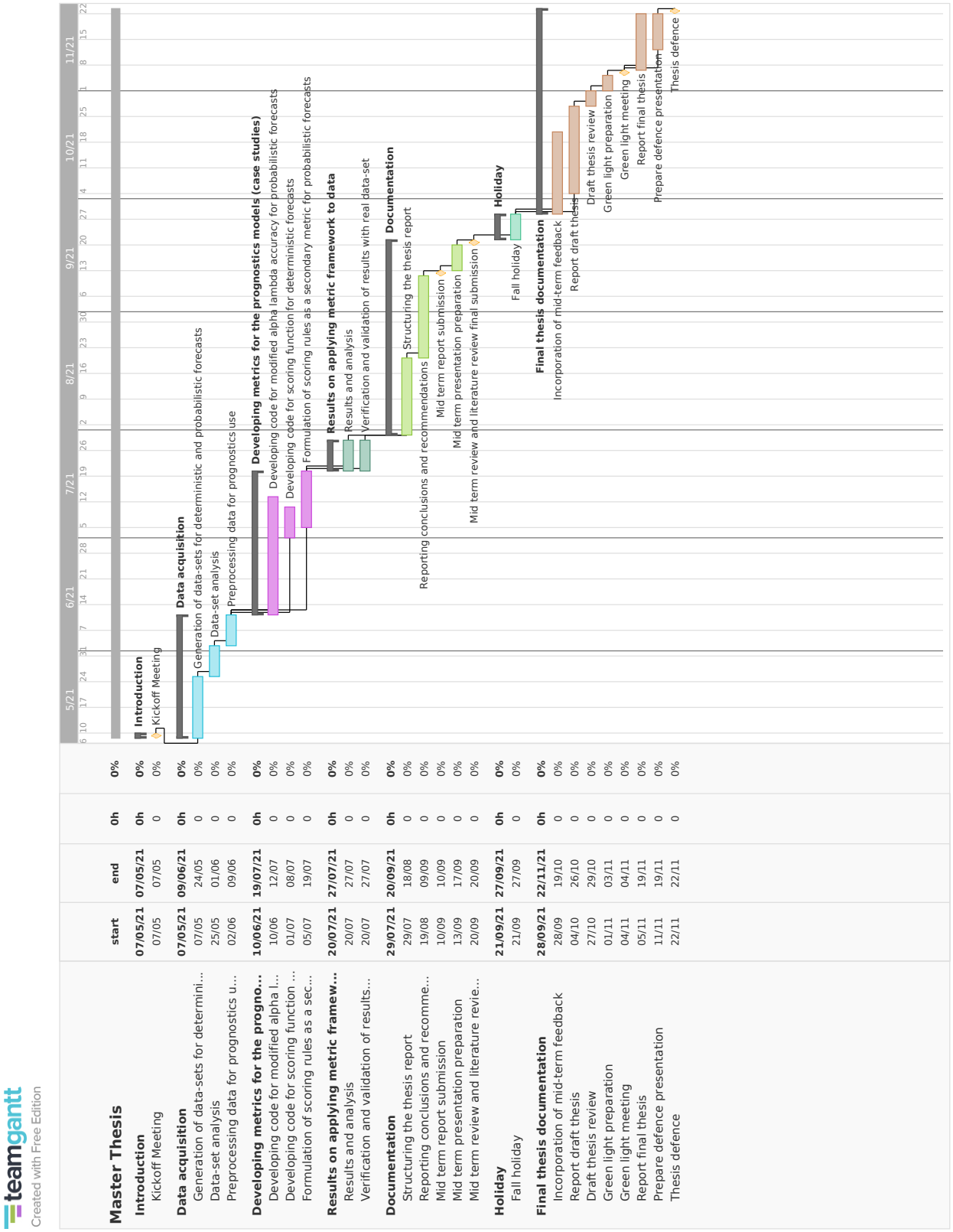


Figure 4: Master thesis plan

Bibliography

[1] Vepa Atamuradov, Kamal Medjaher, Pierre Dersin, Benjamin Lamoureux, and Noureddine Zerhouni. Prognostics and Health Management for Maintenance Practitioners-Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management*, 8:31, 2017.

[2] H. Beirami, D. Calza, A. Cimatti, M. Islam, M. Roveri, and P. Svaizer. A Data-driven Approach for RUL Prediction of an Experimental Filtration System. 2020.

[3] Jochen Bröcker and Leonard Smith. Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather and Forecasting - WEATHER FORECAST*, 22, 2007. doi: 10.1175/WAF966.1.

[4] Huimin Chen. A multiple model prediction algorithm for CNC machinewear PHM. *International Journal of Prognostics and Health Management*, 2, 2011.

[5] Kai Goebel, Abhinav Saxena, Sankalita Saha, Bhaskar Saha, and Jose Celaya. Prognostic Performance Metrics. *Machine Learning and Knowledge Discovery for Engineering Systems Health Management*, 22: 147, 2011. doi: 10.1201/b11580-7.

[6] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.

[7] Stephen Jewson. The problem with the Brier score. 2004.

[8] Xiaodong Jia, Bin Huang, Jianshe Feng, Haoshu Cai, and Jay Lee. Review of PHM Data Competitions from 2008 to 2017: Methodologies and Analytics. *Annual Conference of the PHM Society*, 10, 2018. doi: 10.36001/phmconf.2018.v10i1.462.

[9] Sungil Kim and Heeyoung Kim. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679, 2016. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2015.12.003. URL https://www.sciencedirect.com/science/article/pii/S0169207016000121.

[10] Yaguo Lei, Naipeng Li, Liang Guo, Ningbo Li, Tao Yan, and Jing Lin. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018. ISSN 0888-3270. doi: https://doi.org/10.1016/j.ymssp.2017.11.016. URL https://www.sciencedirect.com/science/article/pii/S0888327017305988.

[11] Jin Li. Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PLOS ONE*, 12:e0183250, 2017. doi: 10.1371/journal.pone.0183250.

[12] Tianzhi Li, Claudio Sbarufatti, Francesco Cadini, Jian Chen, and Shenfang Yuan. Particle filterbased hybrid damage prognosis considering measurement bias. *Structural Control and Health Monitoring*, 12 2021. doi: 10.1002/stc.2914.

[13] Xiang Li, Qian Ding, and J. Q. Sun. Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks. *Reliability Engineering & System Safety*, 172, 2017. doi: 10.1016/j.ress.2017.11.021.

[14] Ying Peng, Ming Dong, and Ming Zuo. Current status of machine prognostics in condition-based maintenance: A review. *International Journal of Advanced Manufacturing Technology*, 50:297–313, 2010. doi: 10.1007/s00170-009-2482-0.

[15] Emmanuel Ramasso and Abhinav Saxena. Review and Analysis of Algorithmic Approaches Developed for Prognostics on CMAPSS Dataset. *PHM 2014 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014*, 2014.

[16] Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653 – 1660, June 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml. Place: Boston MA, USA Publisher: American Meteorological Society.

[17] Shankar Sankararaman and Kai Goebel. Why is the remaining useful life prediction uncertain? *PHM 2013 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2013*, pages 337–349, 2013.

[18] Shankar Sankararaman and Kai Goebel. Uncertainty in Prognostics and Systems Health Management. *International Journal of Prognostics and Health Management*, 6, 2015. doi: 10.36001/ijphm.2015.v6i4.2319.

[19] A. Saxena, S. Sankararaman, and K. Goebel. Performance Evaluation for Fleet-based and Unit-based Prognostic Methods. 2014.

[20] Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. *International Conference on Prognostics and Health Management*, 2008. doi: 10.1109/PHM.2008.4711414.

[21] Abhinav Saxena, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. Evaluating algorithm performance metrics tailored for prognostics. In *IEEE Aerospace Conference Proceedings*, pages 1 – 13, 2009. doi: 10.1109/AERO.2009.4839666.

[22] Abhinav Saxena, Jose Celaya, Bhaskar Saha, Sankalita Saha, and Kai Goebel. Metrics for Offline Evaluation of Prognostic Performance. *International Journal of Prognostics and Health Management*, 1:2153–2648, 2010.

[23] Abhinav Saxena, Jose Celaya, Indranil Roychoudhury, Bhaskar Saha, Sankalita Saha, and Kai Goebel. Designing Data-Driven Battery Prognostic Approaches for Variable Loading Profiles: Some Lessons Learned. page 10, 2012.

[24] Reinhard Selten. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, 1(1):43–61, June 1998. ISSN 1573-6938. doi: 10.1023/A:1009957816843. URL https://doi.org/10.1023/A:1009957816843.

[25] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation A review on the statistical data driven approaches. *European Journal of Operational Research*, 213 (1):1–14, 2011. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2010.11.018. URL https://www.sciencedirect.com/science/article/pii/S0377221710007903.

[26] Jorben Sprong, Xiaoli Jiang, and Henk Polinder. Deployment of prognostics to optimize aircraft maintenance - a literature review: A literature review. *Annual Conference of the PHM Society*, 11, 09 2019. doi: 10.36001/phmconf.2019.v11i1.776.

[27] D. Stephenson, Sandra Coelho, and Ian Jolliffe. Two Extra Components in the Brier Score Decomposition. *Weather and Forecasting*, 23, 2008. doi: 10.1175/2007WAF2006116.1.

[28] Bo Sun, Shengkui Zeng, Rui Kang, and Michael Pecht. Benefits and Challenges of System Prognostics. *IEEE Transactions on Reliability - TR*, 61:323–335, 2012. doi: 10.1109/TR.2012.2194173.

[29] Serdar Uckun, Kai Goebel, and Peter J. Lucas. Standardizing research methods for prognostics. In *Physical Review Letters - PHYS REV LETT*, pages 1 – 10, 2008. doi: 10.1109/PHM.2008.4711437.

[30] S. Weijs, Schoups G, and Nick van de Giesen. Why hydrological forecasts should be evaluated using information theory. *Hydrology and Earth System Sciences Discussions*, 7, 2010. doi: 10.5194/hessd-7-4657-2010.

[31] Feng Yang, Mohamed Habibullah, Tianyou Zhang, Zhao Xu, Lim Pin, and Sivakumar Nadarajan. Health Index-based Prognostics for Remaining Useful Life Predictions in Electrical Machines. *IEEE Transactions on Industrial Electronics*, 63:1–1, 2016. doi: 10.1109/TIE.2016.2515054.

[32] Hanbo Yang, Fei Zhao, Gedong Jiang, Zheng Sun, and Xuesong Mei. A Novel Deep Learning Approach for Machinery Prognostics Based on Time Windows. *Applied Sciences*, 9(22), 2019. ISSN 2076-3417. doi: 10.3390/app9224813. URL https://www.mdpi.com/2076-3417/9/22/4813.