Performance Metrics for Artificial Intelligence (AI) Algorithms Adopted in Prognostics and Health Management (PHM) of Mechanical Systems

Sunday Ochella[1] and Mahmood Shafiee[2,*]

[1] Department of Energy and Power, Cranfield University, College Road, Bedfordshire MK43 0AL, UK
[2] School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NT, UK

E-mail: Sunday.Ochella@cranfield.ac.uk; m.shafiee@kent.ac.uk (corresponding author)

Abstract. Research into the use of artificial intelligence (AI) algorithms within the field of prognostics and health management (PHM), in particular for predicting the remaining useful life (RUL) of mechanical systems that are subject to condition monitoring, has gained widespread attention in recent years. It is important to establish confidence levels for RUL predictions, so as to aid operators as well as regulators in making informed decisions regarding maintenance and asset life-cycle planning. Over the past decade, many researchers have devised indicators or metrics for determining the performance of AI algorithms in RUL prediction. While most of the popularly used metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), etc. were adapted from other applications, some bespoke metrics are designed and intended specifically for use in PHM research. This study provides a synopsis of key performance indicators (KPIs) that are applied to AI-driven PHM technologies of mechanical systems. It presents details of the application scenarios, suitability of using a particular metric in different scenarios, the pros and cons of each metric, the trade-offs that may need to be made in choosing one metric over another, and some other factors that engineers should take into account when applying the metrics.

1. Introduction

Prognostics and Health Management (PHM) involves assessing the health state of systems, sub-systems or components throughout their lifecycle with a view towards avoiding unexpected failures as well as possibly extending their useful life [1]. A mechanical system is considered to be under a normal or healthy state of operation if certain parameters remain above a predetermined threshold [2]. This threshold is often defined based on temperature, pressure, vibration, noise or other measurable parameters. These measurements can be used as an indication of the current health state of the system or to alert any deviation from normal operating condition, which can help determine how much longer the system would run before its condition falls below the threshold. A key activity in PHM, therefore, is the prediction of the remaining useful life (RUL) of systems.

RUL is defined as the duration between the current time and the time when system condition reaches the failure threshold. Up to date, many different approaches have been proposed in the literature to predict the RUL of mechanical systems. In general, the RUL prediction approaches can be categorized into three

types according to their principles: physic-based, data-driven, and hybrid (or fusion) techniques [3]. The data-driven RUL prediction approaches involves the use of artificial intelligence (AI) algorithms along with sensor data from the monitored equipment [4]. Given the recent rapid advances in the field of AI, a plethora of AI algorithms have been applied to predict the RUL of mechanical systems. These algorithms range from conventional techniques such as artificial neural network (ANN), neuro-fuzzy systems, support vector machine (SVM) and Gaussian process regression (GPR) [5] to more recent techniques such as deep learning algorithms [6]. Irrespective of the type of algorithm used, an important step in adopting AI in PHM science is being able to measure the performance of the algorithm.

The performance metrics serve as indicators of the level of confidence one may have in the accuracy of an algorithm and the associated methodology. As the RUL prediction is inherently a regression problem, a performance metric is required to assess the prediction error (as opposed to classification problems that seek to determine right or wrong classification). The majority of key performance indicators (KPIs) have been adapted from the ones used in forecasting to measure forecast error, that are predominantly statistical. Even though the statistical-based metrics are popular and still widely in use, some researchers have developed bespoke performance measures for PHM algorithms. For instance, the readers can refer to [7-9].

The current study provides a synopsis of the KPIs and metrics that are being used for AI-driven PHM of mechanical systems and equipment. It also presents details of the application scenarios, suitability of using a particular metric in different scenarios, the pros and cons of each metric, and the various considerations that should be made when choosing a metric for assessing the performance of different AI algorithms. A broad classification scheme is presented using not only the conventional forecasting metrics but also incorporating other metrics developed by PHM researchers. The results of this study can serve as a useful resource to help researchers select the most suitable AI algorithm for their PHM related research.

The remainder of this paper is organized as follows. Section 2 provides a classification of metrics used to measure the performance of AI-driven PHM algorithms. Section 3 presents details of the key considerations which should be made in choosing one metric over another. Section 4 briefly discusses some challenges that may arise when defining suitable metrics to use for AI algorithms in PHM research. Some concluding remarks are also provided at the end.

2. Performance Metrics for AI Algorithms in PHM

Currently, most performance metrics used for PHM have been adapted from other disciplines such as forecasting, meteorology, finance, medicine, etc. A broad classification of the PHM metrics is presented in reference [8]. This study not only uses the conventional forecasting metrics but also incorporates other metrics developed by PHM researchers. The proposed classification framework for PHM performance metrics is shown in Figure 1.

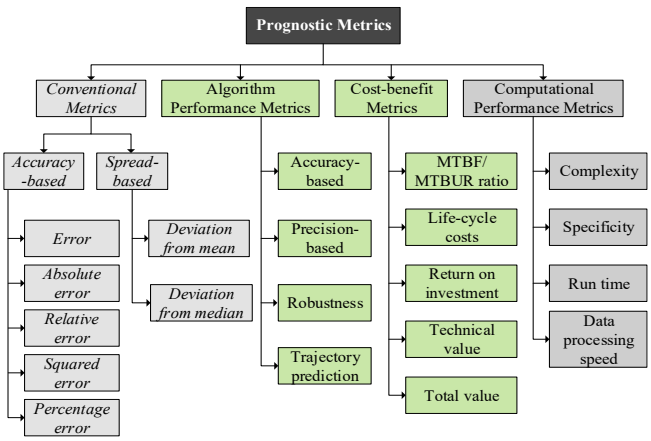

Figure 1. A classification framework for PHM performance metrics.

2.1. Conventional Metrics

Conventional metrics constitute the building block of most PHM performance measures. Given that the fundamental approach used to determine the performance of a PHM algorithm relies on comparing the predicted RUL value with the true value, the statistical-based measures are by far the most common metrics being adopted. Some studies have shown that the mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are the most widely used metrics for measuring the performance of AI algorithms [10]. These "primary" metrics are typically determined by a three-step approach involving (1) calculation of the point distance (between the estimated and true values), (2) normalization, and (3) aggregation of point results over the entire dataset. The conventional performance metrics are based on either accuracy or precision. In what follows, the accuracy-based and precision-based performance measures are briefly reviewed.

2.1.1 Accuracy-based Metrics

Accuracy essentially measures how close the estimated RUL value is to the actual value [11]. Most metrics aggregate the errors in point estimates over the complete dataset, and thus they need to take the mean or median as measures of performance evaluation. The disadvantage of these measures is that they weigh the errors equally, irrespective of when they occurred. Depending on end-user requirements, it may be expedient to give more weights to the errors obtained from predictions made near the end-of-life (EOL) of the system. Directly related to this point is the notion of timeliness, which has also been proposed as a metric. Making accurate predictions early enough is important in order to help with maintenance planning and logistics.

In this study, we define $\tilde{y}_i(t)$ as the predicted RUL value and $y_i(t)$ as the true RUL value, both at time instant t, and at the i^{th} prediction. Some conventional accuracy-based metrics for PHM applications are given below:

$$Mean\ Absolute\ Error\ (MAE) = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y}_i(t) - y_i(t)|\,. \tag{1}$$

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\tilde{y}_i(t) - y_i(t)|}{y_i}\,. \tag{2}$$

$$Sum\ of\ Squared\ Errors\ (SSE) = \sum_{i=1}^{n}(\tilde{y}_i(t) - y_i(t))^2\,. \tag{3}$$

$$Mean\ Squared\ Error\ (MSE) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i(t) - y_i(t))^2\,. \tag{4}$$

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{MSE}\,. \tag{5}$$

$$Symmetric\ Mean\ Absolute\ Percentage\ Error\ (SMAPE) = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\tilde{y}_i(t) - y_i(t)|}{(|\tilde{y}_i(t)| + |y_i(t)|)/2}\,. \tag{6}$$

$$Median\ Absolute\ Error\ (MdAE) = \underset{i=1,\dots,n}{Median}\,|\tilde{y}_i(t) - y_i(t)|\,. \tag{7}$$

$$Median\ Absolute\ Percentage\ Error\ (MdAPE) = 100\% \times \underset{i=1,\dots,n}{Median}\,\frac{|\tilde{y}_i(t) - y_i(t)|}{|y_i(t)|}\,. \tag{8}$$

The merits and demerits of these metrics are discussed later in this paper.

2.1.2 Precision-based Metrics

Precision refers to the spread or narrowness of the interval within which the estimates variate. Precision-based metrics provide an indication of the spread of RUL predictions, given the same set of inputs. An emphasis is made here on the difference between sensitivity and precision. Sensitivity gives an indication of how the predictions from an algorithm would change with the changes in inputs. Thus, the sensitivity is a measure of robustness. Some conventional precision-based metrics for PHM applications are given below:

$$Standard\ Deviation\ (SD) = \left(\frac{\sum_{i=1}^{n}(\tilde{y}_i(t) - \overline{y_i})^2}{n-1}\right)^{1/2}, \tag{9}$$

where $\overline{y_i}$ is the mean of the errors.

$$Mean\ Absolute\ Deviation\ (MAD) = \frac{1}{n}\sum_{i=1}^{n}|\tilde{y}_i(t) - \overline{y_i}|\ . \tag{10}$$

$$Median\ Absolute\ Deviation\ (MdAD) = \underset{i=1,\dots,n}{Median}\ |\tilde{y}_i(t) - M|\ , \tag{11}$$

where M is the median of the errors.

2.2 PHM-specific Algorithms

Engel *et al.* [11] proposed the use of confidence intervals, along with accuracy and precision, to determine whether the RUL estimates are within acceptable bounds. In addition to these metrics, Vachtsevanos *et al.* [12] proposed also some other metrics for fault diagnosis and prognosis, including: timeliness, similarity, sensitivity, incorporation of uncertainty, as well as cost-benefit metrics such as technical value and total value. Leao *et al.* [7] extended the conventional metrics and proposed some new metrics which are discussed in section 2.2.1. Other researchers, e.g. Saxena *et al.* [13], proposed a number of hierarchical metrics for PHM. Sharp [9] argued that the hierarchical metrics proposed in [13] are complicated to use. Therefore, he proposed a set of metrics which could be used in a hierarchical manner or integrated with other metrics.

In an attempt to develop metrics for assessing the performance of PHM algorithms, there has been a somewhat unintended multiplicity of proposed metrics. Thus, most researchers resort to the use of metrics such as MAE, RMSE, etc., instead of those designed for PHM. The PHM metrics can be classified into three groups: (1) the metrics that directly measure the algorithm's performance, (2) the metrics that are based on a cost-benefit criterion, and (3) the metrics that can be used to measure the computational performance. There three groups are described in more detail below.

2.2.1 Algorithm Performance Metrics

Leao *et al.* [7] proposed a number of metrics to help with defining user requirements as well as verifying algorithm performance. Some of these performance metrics include:

a. *Prognostic Hits Score (PHS)* – this is defined as the number of correct prognostics estimates divided by the total number of estimates, where the alert time is greater than or equal to actual time to failure. This metric gives an indication of number of useful predictions (NuP).
b. *False Alarm Rate (FAR)* – this is defined as the number of false alarms due to prognostics estimates divided by NuP. "False alarm" in prognostics implies the occurrence of actual failure of an equipment later than the RUL predicted by the algorithm.
c. *Correct Rejection Rate (CRR)* – this is defined as the number of correct rejections divided by total number of prognostics estimates that meet rejection criterion. Rejection criterion is met when alert time plus confidence interval is less than the ground truth RUL. Alert time in this context means minimum lead time required to take an action, while correct rejection implies rejecting the prediction when not enough time is available to take an action before a failure occurs.
d. *Imprecise Correct Estimation Rate (ICER)* – this is defined as the number of correct predictions that do not provide enough precision in order to be useful to the user, divided by the total number of correct prognostics estimates.
e. *Prognostic effectivity* – this is defined as the capacity of prognostics algorithm to avoid unscheduled maintenance. This metric is calculated by dividing the number of avoided unscheduled maintenance events by total number of unscheduled maintenance events caused by the failure mode of interest.
f. *Average Bias (AB)* – this metric is given by:

$$AB = \frac{1}{n}\sum_{i=0}^{n}(\tilde{y}_i(t) - y_i(t))\ , \tag{12}$$

where $y_i(t)$ is the ground truth RUL at time *t* and *n* is the total number of predictions that helped to avoid unplanned maintenance.

g. *Average Absolute Bias (AAB)* – this metric is similar to *AB* but uses absolute difference. This is an accuracy measure and is given by:

$$AAB = \frac{1}{n}\sum_{i=0}^{n}|\tilde{y}_i(t) - y_i(t)|\ . \tag{13}$$

h. *Coverage* – this is defined as the relative frequency of occurrence of the failure mode of interest, which is calculated by dividing the number of failures caused by the failure mode of interest by the total number of recorded failures for a component. It does not directly measure the algorithm performance but may be used as a weighting factor when considering all failure modes of the component.

The following set of hierarchical metrics were proposed in [13], with additional guidance in [14, 15] on how to apply them. The metrics need to be applied in a logical manner in order to make any meaningful deductions from them.

i. *Prognostic Horizon* – this metric gives the difference between the time when prediction first meets the specified performance criteria (i.e. $\pm\alpha\%$ error on RUL) and the EOL, i.e. the time when prediction crosses the failure threshold.
j. *α-λ performance* – this metric gives an indication of the prediction accuracy at specific time instances, i.e., it checks if prediction is within acceptable bounds ($\pm\alpha\%$ of RUL) at a given time fraction λ, between first prediction and EoL ($\lambda = 0$ at time of first prediction; $\lambda = 1$ at EoL).
k. *Relative Accuracy* – this is an instantaneous measure of error in RUL prediction relative to ground truth RUL.
l. *Cumulative Relative Accuracy* – this is a normalised weighted sum of instantaneous relative accuracies over the lifetime of the prediction (i.e. from first prediction to EoL). Weights are assigned such that predictions at critical times, such as near the EoL, are more important than earlier predictions.
m. *Convergence* – it measures the manner in which any metric improves with time, e.g. how quickly a prediction converges towards the actual RUL as it progresses towards EoL.
n. *Robustness* – it attempts to quantify the sensitivity of an algorithm with respect to its parameters, like the size of the training data or choice of prior distributions. Confidence bounds of a robust algorithm are not expected to vary wildly with changes in the input parameters.

Sharp [9] proposed four metrics intended to measure the performance of algorithms and possibly compare different prognostics sets, independent of the unit of RUL (e.g. time, number of cycles). These metrics capture fundamental aspects of accuracy, precision and timeliness. They are briefly explained below:

1) Weighted Error Bias (WEB) – this is defined as the effective bias in all predictions as a percentage of total equipment lifetime. Therefore,

$$WEB = \frac{100}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} w_i(t) \times \frac{(\tilde{y}_i(t) - y_i(t))}{TotalUnitLifetime_i}\ , \tag{14}$$

where $\tilde{y}_i(t)$ is the predicted RUL for unit *i* at time instant *t;* $w_i(t)$ is the importance weight of unit *i* at time *t*; *T* is the total number of times that RUL prediction is made; and *N* is the number of units. The optimal value for the *WEB* is zero, indicating that the average prediction is centred on the true RUL value.

2) Weighted Prediction Spread (WPS) – this metric is designed to capture prediction uncertainties and, simultaneously, apply weights to prediction importance across various points of the equipment lifetime. First, instantaneous percentage errors in RUL prediction are allocated into bins, across the lifetime of an equipment, with the percentage error computed by the following equation:

$$\%Er = 100\% \times \frac{(\tilde{y}_i(t) - y_i(t))}{Unit\_Lifetime_i}\ . \tag{15}$$

Instantaneous percentage errors can then be placed in bins divided either equally between 0% and 100% of total unit lifetime or by points centred around $(t/Total\ Lifetime)$. The *WPS* can then be computed as:

$$WPS = 100\% \times \frac{\sum_{bi=1}^{B}(W_{bi} * CI_{bi})}{\sum_{bi=1}^{B} W_{bi}}\ . \tag{16}$$

where B is the number of bins, W_{bi} is a weighting function based on the centre value of each reference bin. *WPS* values give an indication of the level of uncertainty in the prediction, with higher values indicating larger uncertainty.

3) *Confidence Interval Coverage (CIC)* – this metric helps to check whether or not the true RUL value lies within the confidence interval of the prediction. This is given by:

$$CIC = 100\% \times \frac{\sum_{bi=1}^{B}(RUL_{bi} \in B_{bi})}{B}. \quad (17)$$

CIC is therefore interpreted as the sum of the percentage of true RUL values contained within their corresponding error bin sets divided by the number of bins. A 100% score implies that all predictions fall within the true RUL values.

4) *Confidence Convergence Horizon (CCH)* – this metric identifies the predicted RUL value that once reached, all remaining predictions would fall within no more than ±10% of the true RUL, 95% of the time (assuming a 95% confidence level). This metric is somewhat similar to the *α-λ performance* metric as proposed in [13], however it is more focused on the quality of prediction towards EOL.

Sharp [9] further proposed a "*Total Score*" metric that aggregates the above-mentioned four metrics, namely, *WEB, WPS, CIC* and *CCH*. This metric is calculated as below:

$$TotalScore = \vec{N} \times \begin{bmatrix} 100 - |WEB| \\ 100 - WPS \\ CIC \\ CCH \end{bmatrix}, \quad (18)$$

where $\vec{N}$ is a normalized vector representing the importance weight of the four metrics. For example, $\vec{N}$ = [0.25, 0.25, 0.25, 0.25] means equal weights.

Three metrics were proposed in Zemouri and Gouriveau [16] to adopt for AI-driven PHM, in a scenario where *M* different prediction algorithms were used to make *n* different RUL predictions. The three metrics are:

a. *Overall Average Bias (OAB)* – this gives the average of the absolute value of the prediction errors. It is calculated by:

$$OAB = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{n=1}^{n}|\tilde{y}_i(t) - y_i(t)|. \quad (19)$$

b. *Overall Average Variability (OAV)* – this is computed as the mean of the standard deviations. It is calculated by:

$$OAV = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n}(\tilde{y}_i(t) - \bar{y}_i)^2}{n-1}\right)^{1/2}. \quad (20)$$

c. *Reproducibility* – this represents the mean distance between RUL predictions of the *M* different algorithms. It is calculated by:

$$Rep = \left(\frac{2}{M(M-1)}\sum_{i<j}(d_{ij})^2\right)^{1/2}, \quad (21)$$

where d_{ij} is the Euclidean distance between the i^{th} and j^{th} prediction algorithms and is given by:

$$(d_{ij})^2 = (E_j - E_i)^2 + (StdDev_j - StdDev_i)^2, \quad (22)$$

where E is the error $(\tilde{y}_i(t) - \bar{y}_i)$, and E and $StdDev$ both are *n*-dimensional.

2.2.2 Cost-Benefit Metrics

The metrics discussed so far are meant to measure the quality of RUL predictions. However, the real benefit of making correct predictions is to record less number of unexpected failures and minimize the hassles associated with unplanned interventions. Cost-benefit metrics measure the anticipated benefits of adopting PHM in business, such as life-cycle cost savings or risk reduction. Some cost-benefit metrics are:

a. *Life Cycle Cost* – It calculates the total cost of acquisition, operation and maintenance under a PHM system and compares with the costs when there is no PHM decision system. In order to justify the adoption of PHM, the costs with PHM should be lower.

b. MTBF-to-MTBUR Ratio – It is defined as the ratio of mean time between failures (which is estimated by conventional reliability methods) to mean time between unit replacement (after PHM implementation). This metric gives an indication of the effectiveness of predictions. Lower MTBF-to-MTBUR ratio indicates the efficiency of the PHM decision system.

c. *Return-on-Investment (ROI)* – this is defined as the average annual profit as a percentage of the initial investment made for PHM implementation [17].

d. Technical Value and Total Value – technical value measures the benefits of correct predictions for critical failure modes against the cost of wrong predictions and the associated resource requirements. Total value, on the other hand, looks at the benefits across all the failure modes that a PHM system can effectively cover, less all costs associated with the PHM implementation.

Details of the formulae associated with the above cost-benefit metrics can be found in [12, 13]. Luna [18] analyzed the cost implications of accurate and timely estimates of RUL on four logistics support scenarios: (*i*) lead times for ordering the spare parts required for maintenance actions; (*ii*) mitigation of consequences of failures; (*iii*) extension of useful operational lifetime; and (*iv*) reduction in maintenance cost. Tang *et al.* [19] proposed two metrics of 'skill' and 'value', which were adapted from meteorology literature. Skill measures how much better a prediction model is than the reference prediction; for example, whether the prediction of RUL using AI algorithms can help make more accurate decisions about maintenance. The measure of "value" represents whether the RUL estimates actually lead to lower maintenance expenditure, compared to the reference case.

An important note on cost-benefit metrics is the fact that the analysis is based on historical figures about lifetimes of similar equipment or experimental run-to-failure data. Consequently, evaluation of the actual performance of AI algorithms using these cost-benefit metrics can only be correctly performed after PHM implementation and is thus not immediately or directly applicable to newly introduced equipment. This further underscores the limitation of offline PHM metrics because the actual RUL is required, which is not available in most cases in real practice. The concept of online PHM performance metrics is discussed in section 2.3 below.

2.3 Other Performance Metrics

Two other categories of metrics worth discussing are computational metrics and online evaluation metrics. Computational metrics measure the performance of an algorithm in terms of *run time* and *processing capabilities* of the hardware and software used to run the algorithm. Another broad categorization of PHM performance metrics is to distinguish between offline and online evaluations. Offline metrics assume a priori that run-to-failure data is sufficient to predict the RUL and perform an evaluation. All the metrics discussed so far make this assumption and are thus suitable for offline evaluations. Online performance metrics, on the other hand, involve making RUL estimates based on data available up to the present time, with regular updating as more data becomes available in (near) real-time. Liu and Sun [20] proposed two metrics, namely, *relative accuracy (RA)* and *relative precision (RP)* for online PHM performance evaluation. The metrics were based on the probability of predictions falling within a user-defined acceptance zone, the level of confidence of the predictions, and the actual data measured in (near) real-time during operation (as against previously collected run-to-failure data). Other studies have proposed online, real-time parameter tuning and updating as more operational data become available. For instance, Zhou *et al.* [21] proposed a method using long short-term memory (LSTM) algorithm for updating the RUL prediction model parameters. However, the performance of the algorithm was evaluated using the MSE metric.

3. Considerations and Selection Criteria

Defining user requirements and developing algorithms for PHM are processes that feed into each other. The choice of the cost function, the optimization objective, and performance metrics for use with AI algorithms in PHM will therefore necessarily depend on key factors, some of which are shown in Figure 2. The factors have been broadly grouped into the ones related to user requirements and those that are necessary for algorithm design.

3.1. User Requirements

Timeliness – the time of first prediction should trigger maintenance planning and determine the usefulness of the RUL prediction algorithm.

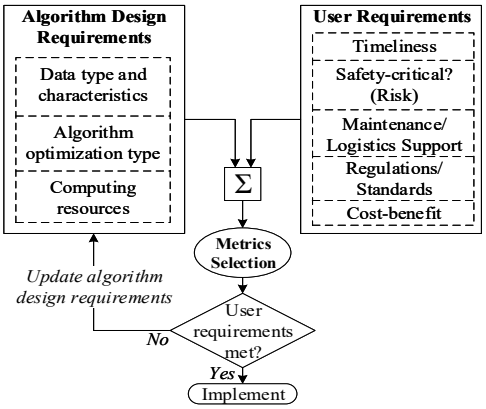

Figure 2. Considerations for PHM metrics selection.

- Criticality – components whose failures result in severe consequences should have stricter performance requirements. For instance, the lead time required to take maintenance action should be longer for safety-critical equipment, along with narrow confidence bounds. Furthermore, defining the failure threshold is a key consideration for critical equipment.
- Maintenance logistics support – the lead time required to order spare parts would influence the choice of PHM metrics. For example, the *prognostic horizon* metric.
- Regulation/standards – extant regulations, or lack thereof, contribute to user requirement specifications, since standards and regulations will have to be complied with.
- Cost-benefit – this is perhaps of utmost importance in PHM research as it determines whether or not a PHM decision system is worth it.

3.2. Algorithm Design Requirements

- Data type and characteristics – although ground truth RUL values obtained from run-to-failure data attempt to simulate real life scenarios, such data will always differ from reality. Algorithm design must therefore factor in noise in sensor data along with other uncertainties associated with health state estimation and future loading conditions.
- Algorithm optimization type – essentially, AI algorithms perform optimization by minimizing a loss function designed around a performance metric, e.g. MSE.
- Computing resources – these must be compatible with data type and size, as well as choice of algorithm; thus influencing algorithm design.

3.3. Other Considerations

An algorithm that penalizes large errors may be rejected even though it makes good predictions towards EOL. This is because, typically, during early life, algorithms are trained using minimal data and predictions could result in large errors. However, the errors would tend to become smaller as the system approaches its EOL because more data becomes available. Typically, metrics that take an average over lifetime as against breaking down the lifetime into different parts exhibit this trait (e.g. MAE). However, in cases where the accuracy of early predictions is very important, penalizing early errors can be a performance requirement. PHM effectivity can also be a very useful input for maintenance planning.

3.4. Pros and Cons of Some Selected Metrics

In addition to discussions in section 2 on each of the metrics, Table 1 below gives the strengths and weaknesses of some selected metrics. As a general note regarding the units of the metrics discussed; accuracy-based and precision-based metrics, along with their other derivatives, are typically measured in the same unit as the RUL – which is either number of cycles or running hours. Weighted metrics and

relative accuracy metrics are devoid of units and are more amenable to easy comparison of results across different simulations and algorithms.

Table 1. Merits and demerits of AI-driven PHM performance metrics.

Metric	Pros	Cons
MAE; Overall Average Bias	a. Easy to compute and understand. b. Unit is same as unit of RUL. c. Equal weighting for individual errors.	a. Susceptible to outliers. b. Does not reveal bias. c. Requires ground truth RUL. d. Unsuitable for multiple datasets with varying scales.
SSE; MSE	a. Applies weighting to magnitude of error. b. Good for gradient-based algorithms (amenable to optimization).	a. Requires ground truth RUL. b. Sensitive to outliers. c. Unit differs from unit of RUL (i.e. scale-dependent).
RMSE	a. Applies weighting to magnitude of error. b. Unit is same as unit of RUL.	a. Requires ground truth RUL. b. Sensitive to outliers. c. Unsuitable for sparse data.
MAPE; sMAPE	a. No unit; good for comparison across different datasets. b. Easy to compute and understand	a. Does not reveal bias. b. Sensitive to outliers. c. Requires true RUL.
MdAE	a. Less sensitive to outliers (than *MAE*).	a. May not work well with very large datasets.
MdAPE	a. Handles outliers well. b. Not scale-dependent.	a. Not intuitive or directly informative.
Std. Deviation; Overall Average Variability	a. Handles outliers well. b. It is a good indication of spread.	a. Assumes a distribution for RUL. b. Affected by weighting of errors.
MAD (Mean Absolute Deviation)	a. Good for sparse data. b. Easy to compute and understand.	a. May not work well for a large data set with lots of outliers.
MdAD (Median Absolute Deviation)	a. Handles outliers well. b. Good for sparse data.	a. May not work well with very large data sets.
Prognostic Horizon	a. Easy to compute and understand. b. Amenable to user definition.	a. May be confusing to use for multiple predictions.
α-λ performance	a. Flexibility to define user requirements. b. Provides a visual graph of performance.	a. Requires ground truth RUL. b. Requires prediction to remain within α-bounds.
Relative accuracy (RA); Cumulative RA	a. Useful for comparing multiple algorithms.	a. Requires ground truth RUL.
Convergence	a. Good indicator of EOL predictions.	a. Requires ground truth RUL. b. Difficult to measure for predictions with large spread.
WEB, WPS, CIS, CCH and Total score	a. Assigns weights as a function of operational life. b. Mostly scale-independent. c. Incorporates uncertainties.	a. Not easy to compute or understand. b. Requires true RUL.

4. Conclusion and Future Work

A significant amount of effort has been put into the attempt to develop performance metrics for AI algorithms used in PHM research. This drive has led to a multiplicity of metrics to measure the accuracy and precision of RUL estimates. The following are important observations and findings:

a) There is a need to unify performance metrics for PHM applications, thereby narrowing down the list. This is a daunting proposal as different application scenarios, data types, algorithms, etc. pose different sets of challenges.
b) The area of online PHM performance evaluation, which indeed applies to most real-life systems, still remains somewhat under-researched.
c) Incorporating uncertainties into PHM remains a challenge. Even though some AI algorithms now incorporate Bayesian techniques to minimize uncertainty, the RUL predictions are still ultimately evaluated using accuracy-based measures such as MAE, MSE and RMSE.
d) As a consequence of the foregoing points, conventional performance metrics remain popular. It will be of interest to see how these metrics evolve as more PHM solutions become adopted in fielded systems, thereby serving as sources of feedback for the suitability of the metrics.

This work provided a synopsis of a synopsis of the performance metrics used for AI-driven PHM of mechanical systems, by the proposition of a comprehensive classification scheme. Conventional as well as PHM-specific metrics were covered, along with discussion on the key factors that should guide engineers in selecting metrics during algorithm design. A key finding is that although efforts have been made to develop bespoke metrics for use in PHM, with the recent resurgence in the use of AI algorithms for RUL prediction, these bespoke algorithms have not yet found wide acceptance and application. This work therefore serves as a good reference material to help in making a choice between the conventional

performance metrics, which remain popular, and PHM-specific metrics, which give more insight but require a specialized understanding.

Acknowledgments
The first author would like to acknowledge the funding support provided by the Petroleum Technology Development Fund (PTDF) in Nigeria. The second author would like to acknowledge the funding support provided by the Industrial Strategy Challenge Fund 2020 through Kent Innovation & Enterprise.

References

[1] Kim N-H, An D and Choi J-H 2017 Prognostics and health management of engineering systems: An introduction (Switzerland: Springer International Publishing).
[2] Shafiee M and Finkelstein M 2015 A proactive group maintenance policy for continuously monitored deteriorating systems: Application to offshore wind turbines *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.* **229** 373-384.
[3] Animah I and Shafiee M 2018 Condition assessment, remaining useful life prediction and life extension decision making for offshore oil and gas assets *J. Loss Prev. Process Ind.* **53** 17-28.
[4] Ochella S and Shafiee M 2019 Artificial intelligence in prognostic maintenance *Proc. 29th Eur. Saf. Reliab. Conf. ESREL 2019* 3424–3431.
[5] Lei Y, Li N, Guo L, Li N, Yan T and Lin J 2018 Machinery health prognostics: A systematic review from data acquisition to RUL prediction *Mech. Syst. Signal Process.* **104** 799–834.
[6] Zhao R, Yan R, Chen Z, Mao K, Wang P and Gao R X 2019 Deep learning and its applications to machine health monitoring *Mech. Syst. Signal Process.* **115** 213–37.
[7] Leão B P, Yoneyama T, Rocha G C and Fitzgibbon K T 2008 Prognostics performance metrics and their relation to requirements, design, verification and cost-benefit *Proc. Int. Conf. Progn. Heal. Manag. PHM 2008* 1–8.
[8] Saxena A, Celaya J, Saha B, Saha S and Goebel K 2009 Evaluating algorithm performance metrics tailored for prognostics *IEEE Aerosp. Conf. Proc.* pp 1–13.
[9] Sharp M E 2013 Simple metrics for evaluating and conveying prognostic model performance to users with varied backgrounds *PHM 2013 – Proc. Annu. Conf. Progn. Heal. Manag. Soc. 2013* 556–65.
[10] Botchkarev A 2019 A new typology design of performance metrics to measure errors in machine learning regression algorithms *Interdiscip. J. Information, Knowledge, Manag.* **14** 45–76.
[11] Engel S J, Gilmartin B J, Bongort K and Hess A 2000 Prognostics, the real issues involved with predicting life remaining *IEEE Aerosp. Conf. Proc.* **6** 457–69.
[12] Vachtsevanos G, Lewis F, Roemer M, Hess A and Wu B 2007 Fault Diagnosis and Prognosis Performance Metrics *Intelligent Fault Diagnosis and Prognosis for Engineering Systems* (USA: John Wiley & Sons, Inc.) pp 355–99.
[13] Saxena A, Celaya J, Balaban E, Goebel K, Saha B, Saha S and Schwabacher M 2008 Metrics for evaluating performance of prognostic techniques *Int. Conf. Progn. Heal. Manag. PHM 2008.*
[14] Saxena A, Celaya J, Saha B, Saha S and Goebel K 2009 On applying the prognostic performance metrics *Annu. Conf. Progn. Heal. Manag. Soc. PHM 2009* 1–16.
[15] Goebel K, Saxena A, Saha S, Saha B and Celaya J 2011 Prognostic Performance Metrics *Machine Learning and Knowledge Discovery for Engineering Systems Health Management* eds A N Srivastava and J Han (USA: CRC Press) 147–78.
[16] Zemouri R and Gouriveau R 2010 Towards accurate and reproducible predictions for prognostic: An approach combining a RRBF Network and an AutoRegressive model *IFAC Proc.* **43** 140–5.
[17] Shafiee M, Animah I, Alkali B and Baglee D 2019 Decision support methods and applications in the upstream oil and gas sector *J. Pet. Sci. Eng.* **173** 1173-1186.
[18] Luna J J 2009 Metrics, models, and scenarios for evaluating PHM effects on logistics support *Annu. Conf. Progn. Heal. Manag. Soc. PHM 2009* 1-9.
[19] Tang L, Orchard M E, Goebel K and Vachtsevanos G 2011 Novel metrics and methodologies for the verification and validation of prognostic algorithms *IEEE Aerosp. Conf. Proc.* 1–8.
[20] Liu S and Sun B 2012 A novel method for online prognostics performance evaluation *Proc. IEEE 2012 Progn. Syst. Heal. Manag. Conf. PHM 2012* 1–6.
[21] Zhou F, Hu P and Yang X 2018 RUL prognostics method based on real time updating of LSTM parameters *Proc. 30th Chinese Control and Decision Conf. (2018 CCDC)* 3966–3971.