A hierarchical decision-making framework for the assessment of the prediction capability of prognostic methods

Zhiguo Zeng, Francesco Di Maio, Enrico Zio, Rui Kang

A hierarchical decision making framework for the assessment of the prediction capability of prognostic methods

Zhiguo Zeng [a], Francesco Di Maio [b], Enrico Zio (*) [a,b], Rui Kang [c]

[a] Chair System Science and the Energy Challenge, Fondation Electricité de France (EDF), CentraleSupélec, Université Paris Saclay Chatenay-Malabry, 92290 France
[b] Energy Department, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy
[c] School of Reliability and Systems Engineering, Beihang University, Beijing, China

Abstract

In Prognostics and Health Management (PHM), the prediction capability of a prognostic method refers to its ability to provide trustable predictions of the Remaining Useful Life (RUL), with the quality characteristics required by the related maintenance decision making. The prediction capability heavily influences the decision maker's attitude towards taking the risk of using the predicted RUL to inform the maintenance decisions. In this paper, a four-layer, top-down, hierarchical decision making framework is proposed to assess the prediction capability of prognostic methods. In the framework, prediction capability is broken down into two criteria (Layer 2), six sub-criteria (Layer 3) and 19 basic sub-criteria (Layer 4). Based on the hierarchical framework, a bottom-up, quantitative approach is developed for the assessment of the prediction capability, using the information and data collected at the Layer-4 basic sub-criteria level. Analytical Hierarchical Process (AHP) is applied for the evaluation and aggregation of the sub-criteria and Support Vector Machine (SVM) is applied to develop a classification-based approach for prediction capability assessment. The framework and quantitative approach are applied on a simulated case study to assess the prediction capabilities of three prognostic methods of literature: fuzzy similarity, feed-forward neural network and hidden semi-Markov model. The results show the feasibility of the practical application of the framework and its quantitative assessment approach, and that the assessed prediction capability can be used to support the selection of the suitable prognostic method for a given application.

Keywords

Prognostic and Health Management (PHM), Remaining Useful Life (RUL), prediction capability, Analytical Hierarchical Process (AHP), fuzzy similarity, feed-forward neural network, hidden semi-Markov model

1. Introduction

Prediction capability of a prognostic method refers to its ability to provide trustable predictions of the Remaining Useful Life (RUL), with the quality characteristics required by the related maintenance decision making. In Prognostics and Health Management (PHM), the predicted RUL (either by model-based prognostic methods [1-4] or data-driven prognostic methods [5-7]) is typically used by decision makers to schedule proper and timely maintenance. Usually, the decision makers choose between Condition-Based Maintenance (CBM) policy [8, 9] or Preventive Maintenance (PM) policy [10]. The choice of maintenance policies, then, depends on the decision makers' attitude to the risk of relying on a predicted RUL to plan maintenance services. Undoubtedly, such attitude is heavily influenced by the prediction capability of the prognostic method used. For instance, a prognostic method with high prediction capability might make the decision maker risk-prone, because he/she feels that he/she can trust the RUL predictions provided by the method. As a result, he/she is willing to take the risk of using them to plan PM. On the other hand, if the prediction capability of the prognostic method is not sufficient, the decision maker might be risk-averse towards using the RUL predictions to support any maintenance decision. Assessment of the prediction capability for a prognostic method is, then, an important task in PHM.

Conventionally, the prediction capability is assessed by calculating some purposely defined Prognostic Performance Indicators (PPIs), based on test or benchmark data [11]. Most commonly used PPIs are related to the accuracy and precision of a prognostic method [12]. Accuracy PPIs quantify the closeness between the model prediction and the true measured values [11, 13]. Precision PPIs measure how confident the model prediction is and the degree to which the prognostic method will yield the same results if repeatedly applied [11, 13, 14]. In general, good values of the PPIs give confidence to the decision makers about the predicted RUL, and make them prone to use the prediction results for supporting maintenance decisions. For example, by calculating some accuracy PPIs, Tobon-Micea et al. [15] compare the prognostic performance of a proposed wavelet-based prognostic method to that of a traditional time-domain method, and conclude that the new method can be applied to support CBM. Using accuracy PPIs, Micea et al. [16] compare the prognostic performances of two prognostic methods for application to Ni-MH-batteries. Hu et al. [17] develop an online assessment method for the PPIs of model-based prognostic methods. In [18], both accuracy and precision PPIs are used to compare the performances of two prognostic methods applied to high-power white Light Emitting

Diodes (LEDs). Other similar examples can be found for Lithium-Ion batteries [19], rotating machinery [20], composite laminates [21], etc., where accuracy and precision PPIs are used to compare the prognostic performances of different prognostic methods.

Although fundamental in practice, the existing PPIs reflect only one dimension of the prediction capability, i.e., the degree to which a prognostic method is able to explain the available data (referred to as the prediction performance in this paper) [22, 23]. Indeed, the prediction capability of a prognostic method is also influenced by the trustworthiness of the method, which is defined in this paper as the confidence that the prognostic method can provide an accurate and precise RUL, with correct and fair quantification of its related uncertainty. Such confidence comes from our knowledge on the prognostic method, such as its proven records of successful applications on similar problems or our knowledge on its inherent methodological characteristics in relation to RUL predictions. Suppose that two prognostic methods, denoted by method A and method B respectively, perform equally well in terms of prediction performances (measured by the PPIs computed on the same available data); while method A has been applied successfully in various scenarios of setting similar to the one of interest, method B is newly developed and has rarely been applied before: it seems reasonable that in this situation, a decision maker would prefer to implement and use method A to support maintenance decisions.

In this view, when evaluating the prediction capability of a prognostic method, both the prognostic performance (in terms of PPIs) and the trustworthiness of the prognostic method should be considered. Whereas the assessment of the prognostic performance is relatively mature through the quantification of PPIs [12, 14, 22], the assessment of the trustworthiness of the prognostic method deserves further consideration. In literature, the trustworthiness of a method or a process is often measured in terms of its maturity [24, 25]. The concept of maturity originated in the 1970s, when a model was developed to assess the maturity of an information system's function [26]. Later, the Software Engineering Institute (SEI) developed the Capability Maturity Model (CMM) to assess the maturity of a process for developing software with desirable quality/reliability/trustfulness characteristics [27]. Based on the CMM, a Prediction Capability Maturity Model (PCMM) has been recently developed to assess the maturity of modeling and simulation efforts [24]. Other approaches of maturity assessment are being developed and applied in different areas, e.g., master data maturity assessment [28], enterprise risk management [29], hospital information system [30], etc. However,

there is no existing maturity assessment methods in the specific context and for the specific aim of prognostics and maintenance decision making.

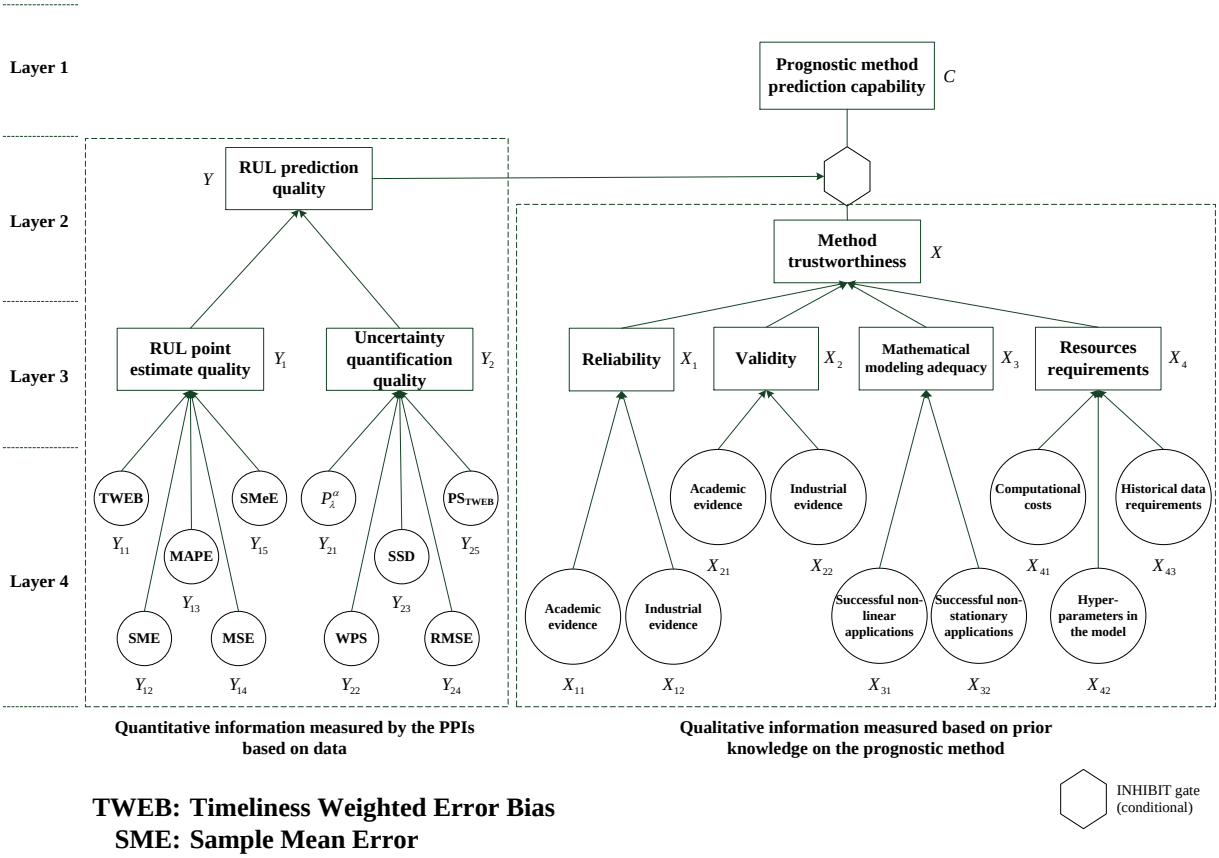
To this aim, in this paper, we consider both the prediction performance and the method trustworthiness to assess the prediction capability of a prognostic method. It should be noted that an initial effort on prediction capability assessment was published by the authors in [31], however without considering the contribution of method trustworthiness and using only a simple weighted average of the PPIs to quantify prediction quality. The rest of the paper is organized as follows. The hierarchical framework is presented in Section 2 and, then, used in Section 3 to assess the prediction capabilities of three data-driven prognostic methods of literature. In Section 4, we draw some conclusions and give some ideas of future research.

2. Hierarchical framework and assessment approach

We present the hierarchical framework developed to assess the prediction capability of prognostic methods in Subsection 2.1, considering two main attributes, RUL prediction quality and method trustworthiness. Prediction quality is assessed in Subsection 2.2 and AHP method is applied in Subsection 2.3 to assess the method trustworthiness. In Subsection 2.4, a classification-based method is developed to determine the prediction capability based on the prediction the quality and method trustworthiness.

2.1. Framework of prediction capability assessment

We present a four-layer hierarchical model to support the assessment of prediction capability, as shown in Figure 1. The prediction capability represented by C (Layer 1 in Figure 1) is characterized in terms of RUL prediction quality and prognostic method trustworthiness (Layer 2 in Figure 1). The former, represented by Y, measures the performance of the prognostic method with respect to the specific application and data, while the latter, represented by X, measures the confidence based on knowledge related to the fact that the prognostic method provides trustworthy predictions, in terms of point estimates and uncertainty quantifications. The inhibit (conditional) gate indicates the logical relationship between X and Y in determining the prediction capability: to have a good prediction capability, the prognostic method should at least satisfy a minimum requirement of prediction quality; once this minimum requirement is satisfied, the prediction capability is determined jointly by the prediction quality and the method trustworthiness.



TWEB: Timeliness Weighted Error Bias
SME: Sample Mean Error
MAPE: Mean Absolute Percentage Error
MSE: Mean Square Error
SMeE: Sample Median Error
P_λ^α: α - λ Performance
WPS Weighted Prediction Spread
SSD: Sample Standard Deviation
RMSE: Root Mean Square Error
PS_{TWEB}: Prediction Spread

Figure 1 Hierarchical framework for prediction capability assessment

The two attributes in Layer 2 are further broken down into factors that influence them, leading to the six criteria in Layer 3: RUL point estimate quality (Y_1) and uncertainty quantification quality (Y_2), which contribute to the RUL prediction quality and reliability (X_1), validity (X_2), mathematical modeling adequacy (X_3) and resources requirements (X_4), which influence the method trustworthiness. Detailed descriptions of the criteria are given in Table 1.

Table 1 Descriptions of the Layer-3 characteristics

Notation	Meaning	Description
Y_1	RUL point estimate quality	Measures the distance of the RUL point estimates from the true RUL values, i.e., the accuracy of the prognostic method. An accurate prognostic method is obviously preferred.
Y_2	Uncertainty quantification quality	Measures the spread and variability of the RUL predictions, i.e., the precision of the prognostic method. A precise prognostic method is preferred.
X_1	Reliability	Measures the capability of the method to yield the same RUL prediction quality, when different analysts apply it on similar sets of data related to similar problems: the larger the reliability, the more trustworthy the method is.
X_2	Validity	Measures the capability of the method to achieve the same RUL prediction quality, when applied to solve different problems with similar characteristics: the larger the validity, the more trustworthy the method is for use in different problems of similar characteristics.
X_3	Mathematical modeling adequacy	Measures the capability of the method to deal with problems of given complexity: a less advanced method may handle well linear and simplified problems, whereas a more complex and advanced method is needed to deal with more realistic problems, e.g., nonlinear and non-stationary problems. In these situations, such methods would be more adequate and trustworthy compared to less mathematically complex and advanced methods.
X_4	Resources requirements	Measures the required resources by the prognostic methods, e.g, the data requirements, the computational costs, the number of hyper-parameters, etc. A prognostic method with lower resource requirements is more controllable and verifiable during the training phase under the realistic available data, and therefore, is more trustworthy for such settings.

The six criteria in Layer 3 are further decomposed into a layer of 19 basic sub-criteria (Layer 4 in Figure 1), where data and information can be used to support the assessment of prediction capability. Detailed descriptions of all the 19 basic sub-criteria can be found in the Appendix. Depending on the nature of the basic sub-criteria, they might take either numerical or linguistic values. The basic sub-criteria used to evaluate the RUL prediction quality are, in fact, quantitative PPIs related to accuracy and precision of a prognostic method. All of them take numerical values, e.g., the Timeless Weighted Error Bias (TWEB, Y_{11}) in Figure 1. The basic sub-criteria used to evaluate the method trustworthiness, on the other hand, represent evidence on various aspects of the trustworthiness of the prognostic method. Some of them are objective in nature, and, therefore, can be measured by numerical indicators, e.g., the number of academic evidence (X_{11}) in Figure 1. Others are qualitative in nature and can only be represented by linguistic or non-numerical values: evaluation of these basic sub-criteria requires the involvement of subjective judgements.

To assess the prediction capability, data and information are collected to support the evaluation of the Layer 4 basic sub-criteria (X_{ij} and Y_{ij}). Then, the basic sub-criteria are aggregated to assess the criteria in Level-3, and

further aggregated to assess the Level-2 attributes of prediction quality and method trustworthiness. Finally, the prediction capability of the prognostic method is determined based on the joint contributions of the two Level-2 attributes, as shown in Figure 2. The obtained prediction capability incorporates the influences from both prediction quality and method trustworthiness, and, therefore, can be used to support the selection of appropriate prognostic methods for given maintenance planning requirements.

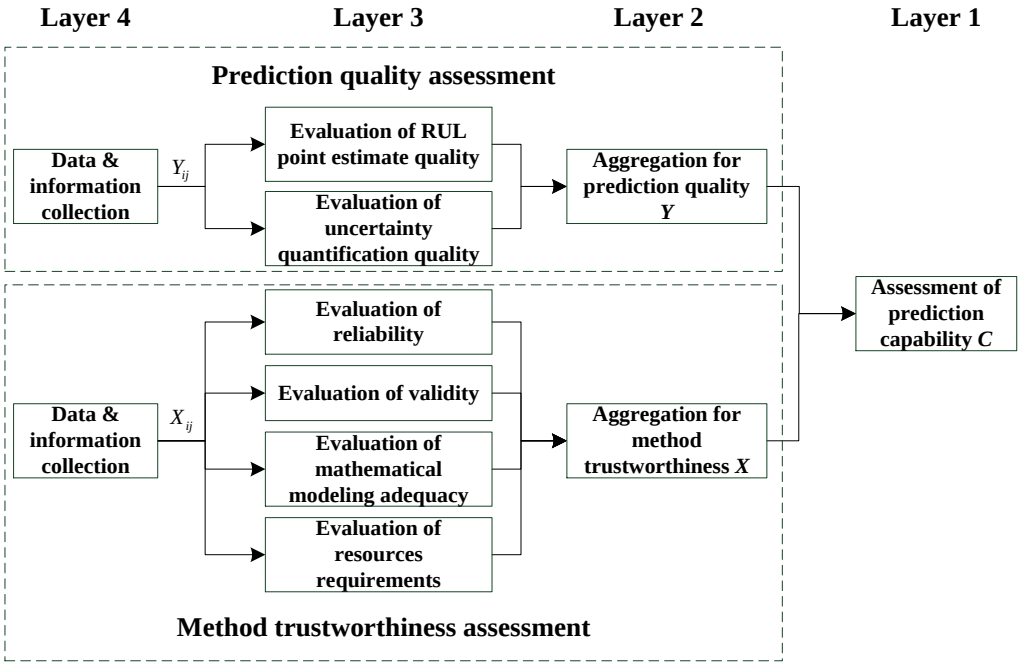


Figure 2 Procedures for prediction capability assessment

2.2. Prediction quality assessment method

As shown in Figure 2, the assessment of prediction quality starts from collecting data and information for the Level-4 basic sub-criteria related to prediction quality. Various numerical indicators, referred to as Prognostic Performance Indicators (PPIs), have been defined in literature to assess the performance of a RUL prediction method with respect to both point estimates and uncertainty quantifications. In this work, the PPIs listed in Table A.1 are adopted as the Layer-4 basic sub-criteria; their values, denoted by Y_{ij}, $i=1,2,\ j=1,2,\ldots,5$, are calculated based on the formula listed in Table A.1.

Next, the PPIs are aggregated to evaluate the two Layer-3 criteria related to prediction quality, i.e., the RUL point estimate quality and uncertainty quantification quality. As shown in Table A.1, the values of all the Y_{ij} s are bounded in the range $(-\infty,1]$. A weighted-average method is used to aggregate the Layer-4 basic

sub-criteria:

$$Y_i = \sum_{j=1}^{n_i} \omega_{ij} \cdot Y_{ij}, i = 1,2,\ j = 1,2,\ldots,n_i, \tag{1}$$

where $n_i, i = 1,2$ denotes the number of the Layer-4 basic sub-criteria associated with the i th Layer-3 sub-criteria and $\omega_{ij}, j = 1,2,\ldots,n_i$ are the weights of the Layer-4 basic sub-criteria. In this paper, we have $n_1 = n_2 = 5$, and $\sum_{j=1}^{n_i} \omega_{ij} = 1, i = 1,2$. The weights represent the relative contribution of a basic sub-criteria to the corresponding Layer-3 criterion. In practice, the weights can be obtained by expert assessments or through some structured analysis method, e.g., the Analytical Hierarchical Processes (AHP) method [32]. It is easy to verify from (1) that both Y_1 and Y_2 take values in $(-\infty,1]$, where a value close to 1 indicates good performance.

The two Layer-3 criteria are again aggregated to yield the prediction quality Y by the weighted average:

$$Y = \exp\left\{\sum_{i=1}^{2} \omega_i \cdot Y_i - 1\right\}, \tag{2}$$

where ω_i is the weight for Y_i and $\sum_{i=1}^{2} \omega_i = 1$. As for the weights in the Layer-3 calculations, the ω_i can also be determined by experts assessments based on structured analysis methods such as the AHP method [32]. The exponential function in (2) is used for normalization: since $Y_i \in (-\infty,1]$, it is easy to verify that $Y \in (0,1]$ and a value close to 1 indicates good prediction quality.

2.3. Method trustworthiness assessment method

Since the assessment of method trustworthiness involves multiple quantitative (i.e., the $X_{11} - X_{32}$ in Figure 1) and qualitative sub-criteria (i.e., the $X_{41} - X_{43}$ in Figure 1), it is formulated as a Multi-Criteria Decision Analysis (MCDA) problem [33]. As a widely applied MCDA method [34], AHP is selected for the assessment. AHP, first introduced in 1977 [32], is a hierarchical framework to support multi-criteria decision analysis, where the decision problem considered (the first, top, layer in the hierarchy) is decomposed into several layers of criteria and, eventually, the last, bottom layer containing the alternatives available for the solution of the decision problem. Through pairwise comparisons among elements in the same layer, the alternative solutions in the bottom layer can be ranked with respect to the decision problem in the top layer [32]. For a detailed discussion on the implementation procedures of AHP, readers might refer to [35, 36]. The AHP model for method trustworthiness assessment is illustrated in Figure 3.

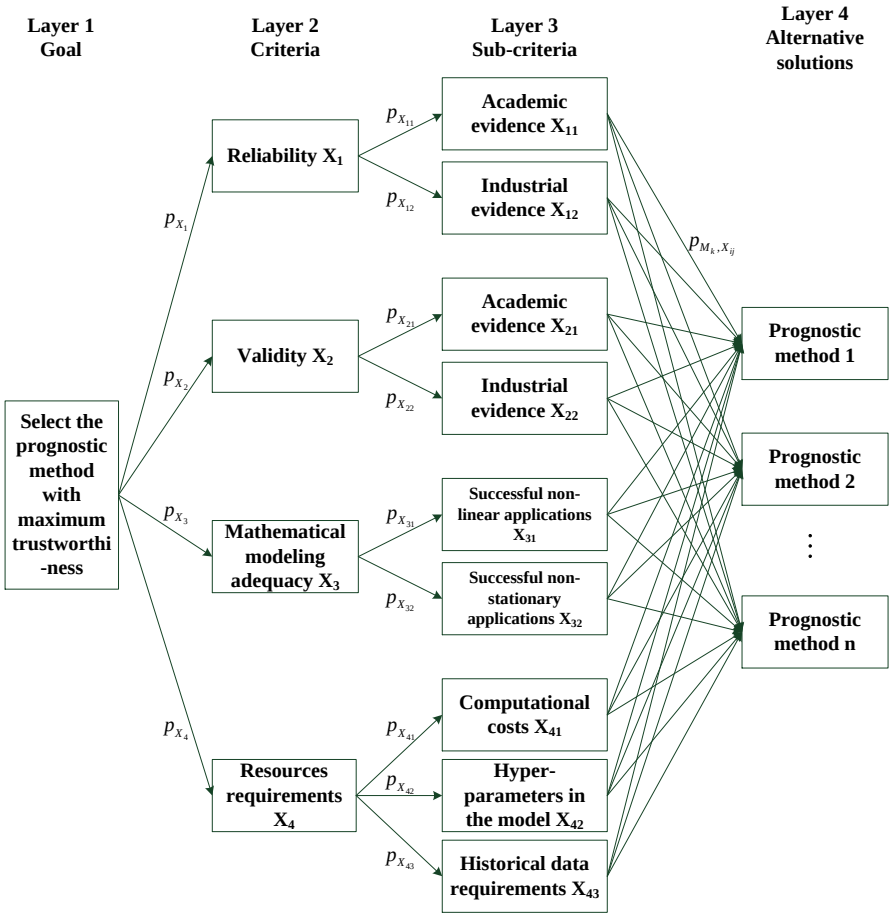


Figure 3 The AHP model for trustworthiness assessment

Based on the AHP model of Figure 3, the assessment of method trustworthiness involves three steps:

Step 1: Determine the inter-level priorities for the criteria (p_{X_i}), sub-criteria ($p_{X_{ij}}$) and alternative solutions ($p_{M_k,X_{ij}}$). The inter-level priorities quantify the relative importance of the lower-level elements with respect to the corresponding high-level element. For the qualitative sub-criteria, experts compare their relative importance using the 1-9 scaling system defined in [37], where scale 9 represents "i is extremely more important than j", scale 1 represents "equally important" and scale $1/9$ represents "j is extremely more important than i". Pairwise comparison matrices, indicated with the symbol A in this paper, are constructed by filling out each element a_{ij} with the numerical value of relative importance and considering the reciprocity property, which indicates that $a_{ij} = 1/a_{ji}$.

For the quantitative basic sub-criteria $X_{11} - X_{32}$, their inter-level priorities can be determined by calculating priority weights as:

$$p_{M_k,X_{ij}} = \frac{X_{M_k,X_{ij}}}{\sum_{k=1}^{n} X_{M_k,X_{ij}}}, \tag{3}$$

where $p_{M_k,X_{ij}}$ is the inter-level priority of the k th prognostic method with respect to the basic sub-criteria X_{ij}, and $X_{M_k,X_{ij}}$ is the numerical value that the k th prognostic method takes with respect to the basic sub-criteria X_{ij}.

Once the comparison matrix for a given level of the hierarchy has been constructed, the eigenvalue method is used to calculate the inter-level priorities [32]. Suppose the priorities associated with a comparison matrix A are denoted by $\mathbf{p} = \left[p_{L,1}, p_{L,2}, \cdots, p_{L,n} \right]^T$. The eigenvalue method first calculates the eigenvector of A that corresponds to the largest eigenvalue, denoted by $\mathbf{p}_M$ and λ_M, respectively. The priority vector $\mathbf{p}$ is, then, calculated by normalizing the vector $\mathbf{p}_M$, as in (4) below, where $\mathbf{p}_{(i)}$ and $\mathbf{p}_{M,(i)}$ represent the i th component in $\mathbf{p}$ and $\mathbf{p}_M$, respectively:

$$\begin{aligned} &\mathbf{p}_{(i)} = \frac{\mathbf{p}_{M,(i)}}{\mathbf{p}_M^T \cdot \mathbf{p}_M}, i = 1, 2, \ldots, n, \\ &A\mathbf{p}_M = \lambda_M \mathbf{p}_M . \end{aligned} \tag{4}$$

Finally, the consistency of the comparison matrix is checked to see if the calculated priority vector makes sense. A comparison matrix A is consistent if it satisfies both the reciprocity rule [33]:

$$a_{ij} = \frac{1}{a_{ji}} \tag{5}$$

and the transitivity rule [33]:

$$a_{ij} = a_{ik} \times a_{kj}, \tag{6}$$

where a_{ij} is the element in the i th row and j th column of A and i, j, k are indexes for the criteria or alternative solutions in A. The consistency can be checked following the procedure in Figure 4, where RI is the CI (Confidence Index) of a randomly generated $n \times n$ matrix whose values can be found in [32, 37]. The three-step procedures are repeated for each criteria, sub-criteria and alternative solutions, until all the p_{X_i}, $p_{X_{ij}}$ and $p_{M_k,X_{ij}}$ are determined.

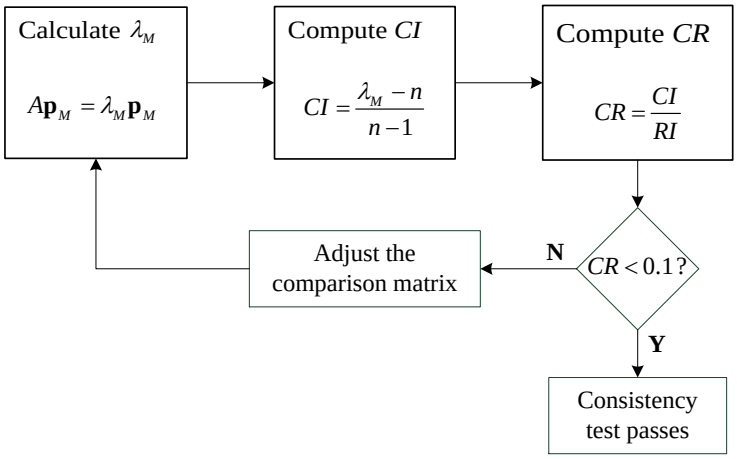


Figure 4 Procedures for consistency tests [32, 37]

Step 2: Calculate the global priority for each alternative solution.

A bottom-up synthesis process is used to calculate the global priority for each alternative solution, with respect to the top goal of the hierarchy:

$$p_{M_k} = \sum_{i=1}^{4} \sum_{j=1}^{n_{X_i}} p_{X_i} \cdot p_{X_{ij}} \cdot p_{M_k, X_{ij}}, \tag{7}$$

where p_{M_k} is the global priority of the k th prognostic method and n_{X_i} is the number of sub-criteria under the criterion X_i. Note that the global priorities should sum up to 1, i.e., $\sum_{i=1}^{n} p_{M_i} = 1$.

Step 3: Determine the method trustworthiness.

The method trustworthiness for each prognostic method, denoted by $X_{M_i}, i = 1, 2, \cdots, n$, is then determined based on the global priorities:

$$X_{M_i} = \frac{p_{M_i} \cdot X_{\max}}{\max_{i=1}^{n} \left(p_{M_i} \right)}, \tag{8}$$

where $X_{\max}$ is the method trustworthiness of the prognostic method with the largest global priority, which is evaluated based on expert judgements. The value of $X_{\max}$ ranges in $[0,1]$, where a value closer to 1 indicates that the prognostic method is more trustworthy.

2.4. Prediction capability assessment and prognostic method selection

Prediction capability C is an integrated metric that supports the selection of appropriate prognostic methods for a given application scenario. Depending on the role that the predicted RUL plays in maintenance

planning, three typical application scenarios are usually distinguished: fully supportive, where the predicted RUL is used to support Predictive Maintenance (PM) planning; partially supportive, where the predicted RUL is used to support Condition-Based Maintenance (CBM) planning; and non-supportive, where the predicted RUL is not directly applicable in maintenance planning. Therefore, the prediction capability is assumed to take three discrete values, $C \in \{C_0, C_1, C_2\}$, where C_0, C_1, C_2 correspond to the required prediction capability for the non-supportive, partially supportive and fully supportive application scenarios, respectively. The issue of prediction capability assessment, is, then, formulated within a classification framework: given a prognostic method, which is characterized in terms of prediction quality and method trustworthiness, select among the above three candidates a proper value for its prediction capability.

In this paper, we assume that training data are available to construct a classifier for prediction capability assessment using supervised learning algorithms. The training data comprise of prognostic methods with known prediction quality, method trustworthiness and prediction capability. In Figure 5, we present 200 training data, which are constructed by randomly generating 200 samples of X and Y, and then, inviting decision makers to assess the prediction capability for each combination of X and Y. Support Vector Machine (SVM) is used to construct a classifier for prediction capability assessment. We directly apply the SVM algorithm in MATLAB® R2015b and the result is shown in **Figure 6**. A 10-fold cross validation is conducted to validate the classifier. The average misclassification rate of the classifier is $\epsilon_1 = 0.04$, which indicates good classification performance. The $X-Y$ plane is partitioned in non-supportive, partially supportive and fully supportive regions, corresponding to $C = C_0$, $C = C_1$ and $C = C_2$, respectively. The prediction capability of a prognostic method can, then, be determined based on its position in the $X-Y$ plane of **Figure 6**.

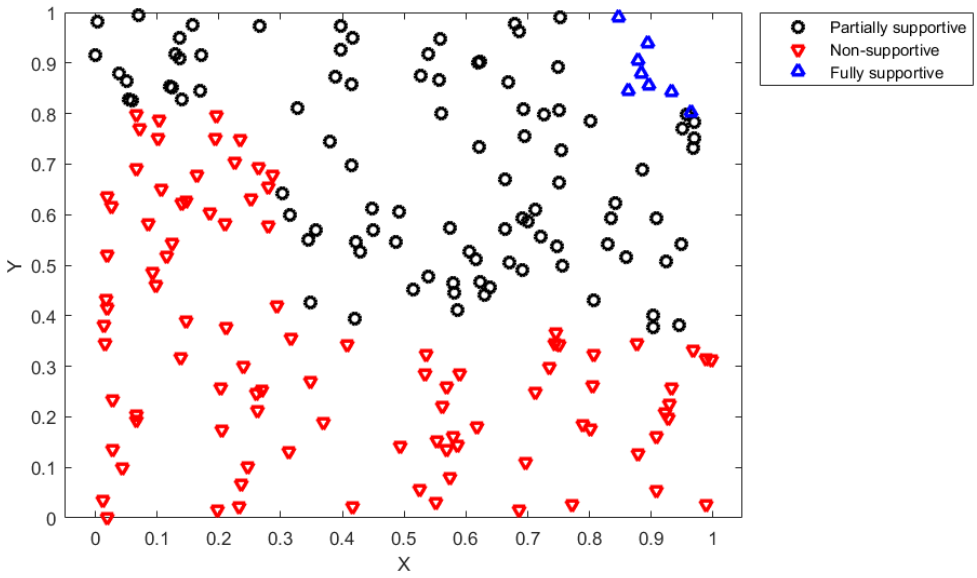


Figure 5 Training data for prediction capability assessment

Figure 6 reflects, based on the training data, the decision makers attitude to the risk of relying on a predicted RUL to plan maintenance services. It can be seen from Figure 6 that to be qualified to support PM, the decision maker thinks that a prognostic method needs to have both high prediction quality and high trustworthiness (fully supportive region). Also, when $Y \le e^{-1}$ (roughly speaking, it means that the average prediction error between the predicted and true RUL is higher than the total life, see Table A.1 and Eq. (2)), the decision maker is not willing to apply the prognostic method to support any kind of maintenance decisions (non-supportive region), regardless of how well the method trustworthiness is. This fact is also reflected by the conditional gate in Figure 1. If the minimum requirement of Y is satisfied ($Y > e^{-1}$), the prediction capability further depends on the value of method trustworthiness: if the method trustworthiness is medium or high (roughly speaking, $X \ge 0.3$), the decision maker would apply the method to support CBM (partially supportive region); otherwise, only with higher prediction quality (roughly speaking, $Y > 0.8$), the prognostic method can be qualified to support CBM.

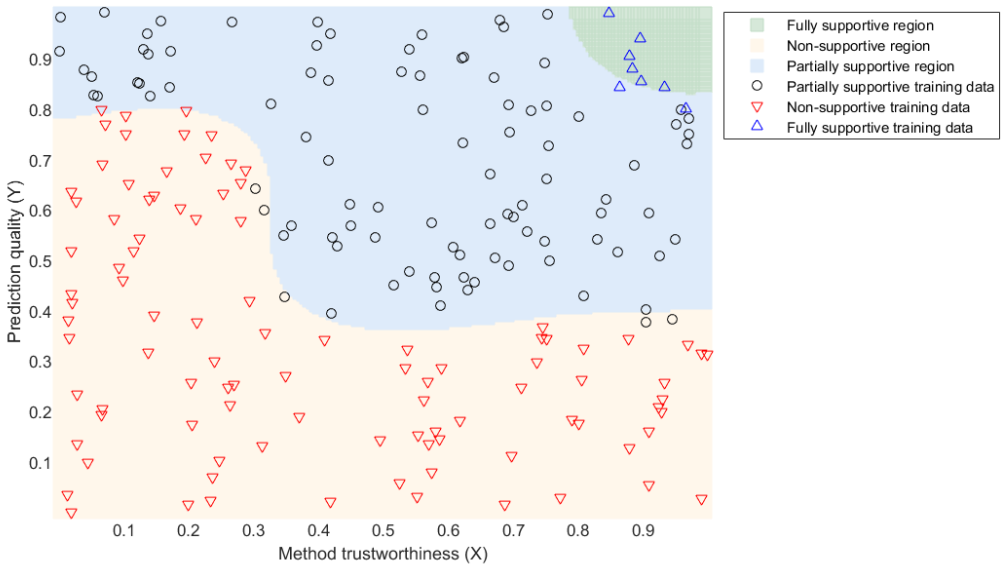


Figure 6 A classifier constructed for prediction capability assessment

A major strength of the developed prediction capability assessment framework is that it integrates both prediction quality and method trustworthiness of the prognostic method, while existing frameworks, such as those in [12] or [14], often neglect method trustworthiness. To demonstrate the strength of the developed framework, we also apply the prediction quality based framework on the training data in Figure 5. Since only the prediction quality is considered, we only use Y to construct the classifier. We again use SVM to construct the classifier and the result is given **Figure 7**. A 10-fold cross validation is conducted. The average misclassification rate for this classifier is $\epsilon_2 = 0.22$, which is much larger than that of the developed framework ($\epsilon_1 = 0.04$). The comparison shows that by considering the method trustworthiness, the developed assessment framework provides a more comprehensive description of the prediction capability.

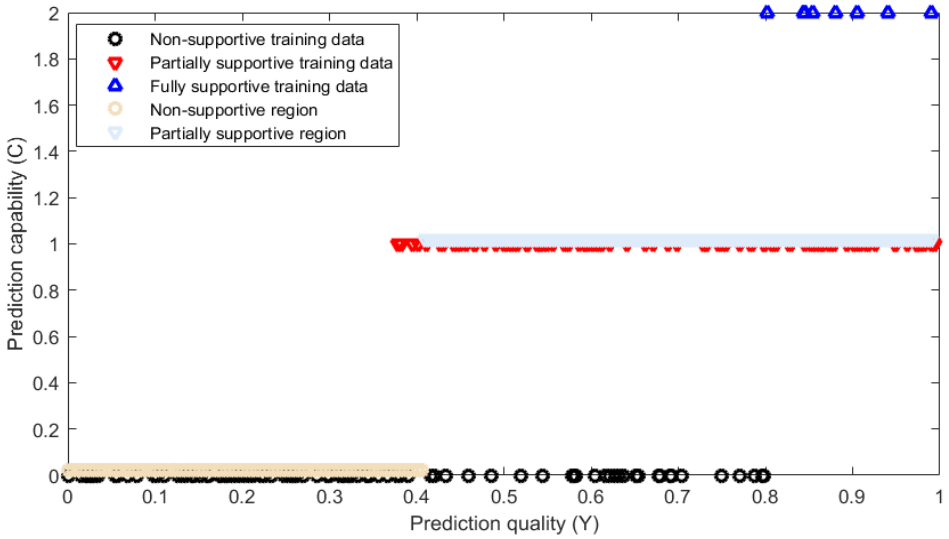


Figure 7 Training data and classifier when only Y is considered

3. Application

In this section, the framework developed in Section 2 is implemented to assess the prediction capabilities of three prognostic methods of literature, i.e., Fuzzy Similarity (FS) [38, 39], Feed-forward Artificial Neural Networks (FANN) [40] and Hidden Semi-Markov Model (HSMM) [41-43]. A simulation case study of nine run-to-failure trajectories is considered as data, as shown in Figure 8 [31]. These data represent the failure trajectories that can be extracted based on simulated mono-dimensional signal (e.g., temperature, pressure, or vibration signal) of a generic component. The three methods are applied to predict the RUL, and their prediction quality and method trustworthiness are assessed in Subsections 3.1 and 3.2, respectively. The prediction capabilities of the three methods are determined in Subsection 3.3 by combining the prediction quality and method trustworthiness.

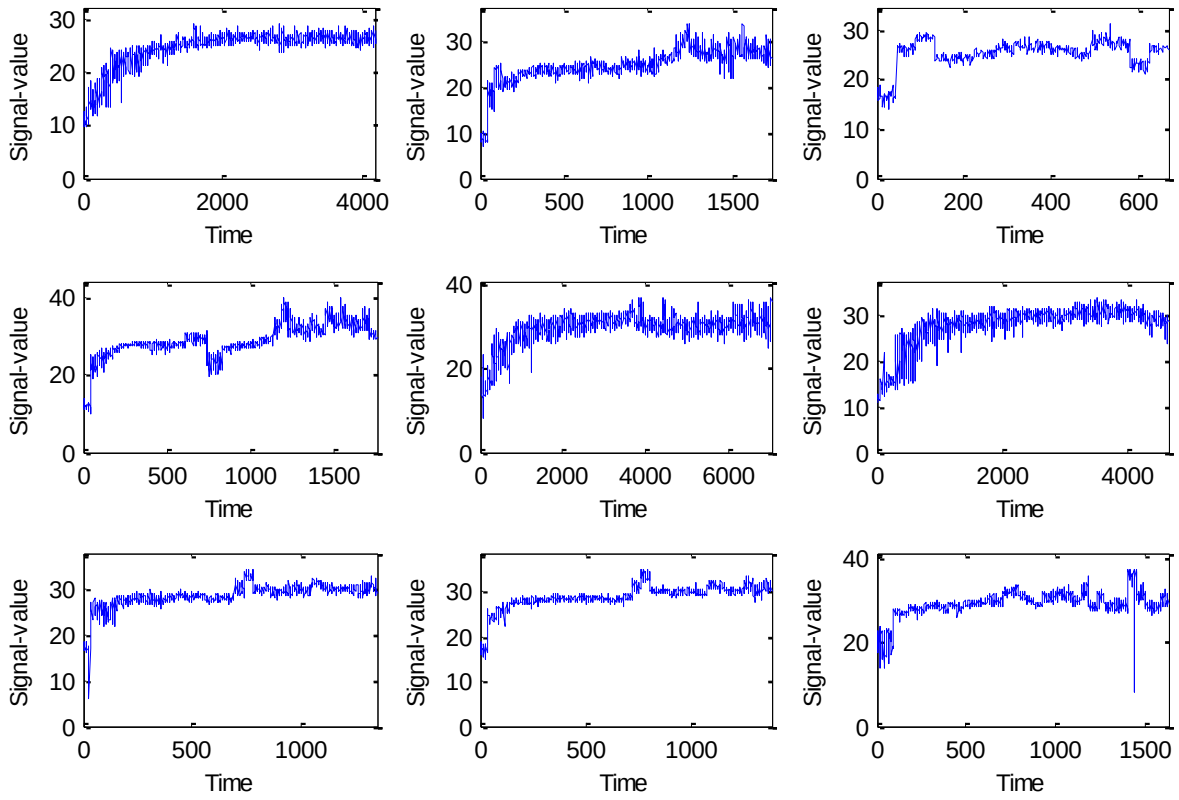


Figure 8 The nine simulated run-to-failure degradation trajectories

3.1. Prediction quality assessment

The three prognostic methods are applied to predict the RUL of the case study in Figure 8. Leave-one-out cross validations are used to compare the prediction quality of the three methods, where for each validation, one of the nine samples is left out while the rest eight are used as training samples. The RUL of the left-out sample is regarded as the true RUL so that the PPIs in Table A.1 can be calculated. Empirical Mode Decomposition (see [44] for details) is used for the three methods to pre-process the raw signal and construct health indicators (HIs). The RUL prediction from each method is given in Figure 9. The accuracy and precision PPIs calculated based on Table A.1 are listed in Table 2 and Table 3.

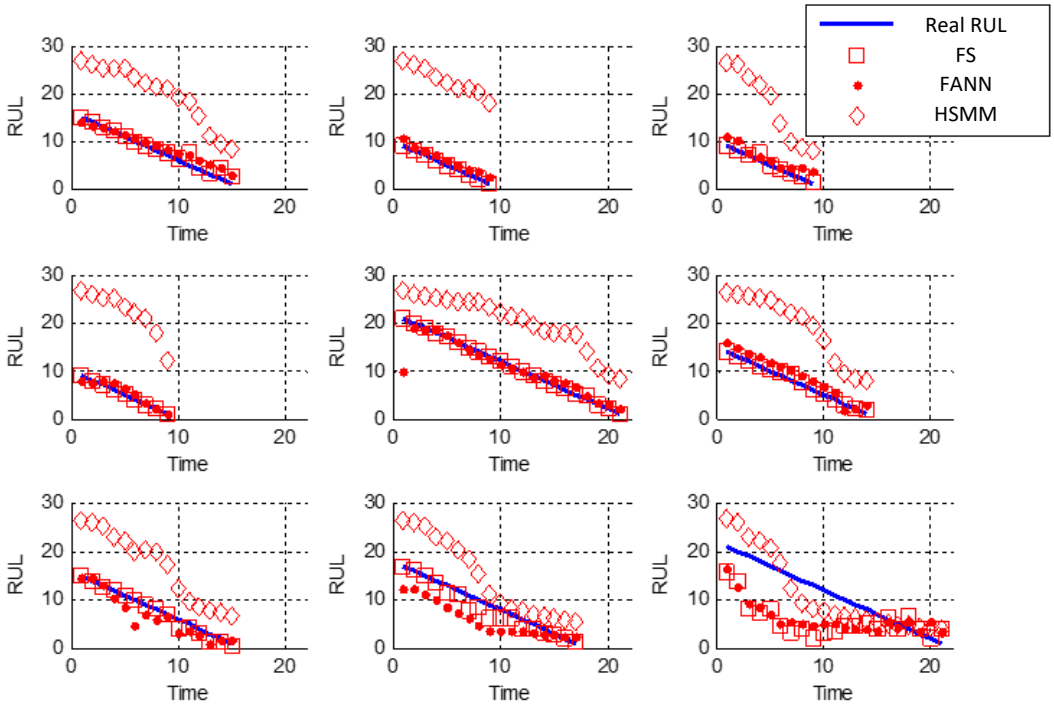


Figure 9 The predicted RUL of FS, FANN and HSMM [31]

Table 2 Accuracy PPIs for FS, FANN and HSMM

PPIs	FS	FANN	HSMM
Y_{11}	0.98	0.94	0.11
Y_{12}	0.37	0.56	-9.47
Y_{13}	0.85	0.63	-1.44
Y_{14}	-5.14	-7.56	-143.79
Y_{15}	0.98	0.57	-10.65

Table 3 Precision PPIs for FS, FANN and HSMM

PPIs	FS	FANN	HSMM
Y_{21}	0.61	0.34	0.02
Y_{22}	0.97	0.94	0.67
Y_{23}	-0.71	-1.11	-4.16
Y_{24}	-0.28	-1.42	-10.20
Y_{25}	0.98	0.97	0.14

To assess the values of Y_1 and Y_2, the weights of each Layer-4 basic sub-criteria should be determined first. In this case study, experts are invited to rank all the PPIs in terms of their relative importance in affecting the corresponding prognostic performance. Then, the weight of each PPI can be calculated by:

$$\omega^{(i)} = \frac{1-\frac{i-1}{n}}{\sum_{i=1}^{n}\left(1-\frac{i-1}{n}\right)} = \frac{2(n-i+1)}{n(n+1)}, \tag{9}$$

where i is the ranking of the PPI (in descending order of importance) and n is the total number of the PPIs in the same category. Suppose that $Y_{11}, Y_{12}, \cdots, Y_{15}$ and $Y_{21}, Y_{22}, \cdots, Y_{25}$ are in descending order of importance, respectively. According to (9), their weights are calculated and listed in Table 4.

Table 4 Weights of the PPIs

	Accuracy PPIs					Precision PPIs				
PPIs	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
Weights	0.333	0.267	0.200	0.133	0.067	0.333	0.267	0.200	0.133	0.067

Then, the values of Y_1 and Y_2 are calculated based on (1) and the results are given in Table 5.

Table 5 Evaluation results of Y_1 and Y_2

RUL point estimate prediction quality			Uncertainty quantification quality		
$Y_{1,FS}$	$Y_{1,FANN}$	$Y_{1,HSMM}$	$Y_{2,FS}$	$Y_{2,FANN}$	$Y_{2,HSMM}$
-0.025	-0.381	-22.659	0.348	0.017	-1.997

The prediction quality Y is, then, calculated based on (2) and Table 5, where the RUL point estimate quality and uncertainty quantification quality are assumed to have equal weights, $\omega_1 = \omega_2 = 0.5$. The results are tabulated in Table 6. The results in Table 6 show that considering both the point estimate and uncertainty quantification quality, Fuzzy Similarity performs the best among the three prognostic methods in terms of prediction quality, whereas the prediction quality of Hidden-Semi Markov Model is the worst among the three methods.

Table 6 Prediction quality of the three prognostic methods

Y_{FS}	Y_{FANN}	Y_{HSMM}
0.433	0.307	1.63×10^{-6}

3.2. Method trustworthiness assessment

3.2.1. Step 1: Determine the inter-level priorities

For the quantitative basic sub-criteria $X_{11} - X_{32}$, the numerical values for the criteria are collected in Table 7, where M_1, M_2 and M_3 correspond to FS, FANN and HSMM, respectively. Based on (3), the local priorities are calculated and given in Table 8. It should be noted that the numerical values in Table 7 are simulated for illustrative purposes. In practice, these values should be collected based on actual data extracted

from literature and engineering applications.

Table 7 Numerical values for the basic sub-criteria

	X_{11}	X_{12}	X_{21}	X_{22}	X_{31}	X_{32}
M_1	24	18	22	16	22	18
M_2	39	31	41	38	38	31
M_3	38	32	35	24	28	22

Table 8 Inter-level priorities of the alternative solutions with respect to $X_{11} - X_{32}$

	X_{11}	X_{12}	X_{21}	X_{22}	X_{31}	X_{32}
$p_{M_1,X_{ij}}$	0.238	0.222	0.224	0.205	0.250	0.254
$p_{M_2,X_{ij}}$	0.386	0.383	0.418	0.487	0.432	0.437
$p_{M_3,X_{ij}}$	0.376	0.395	0.357	0.308	0.318	0.310

For the qualitative sub-criteria $X_{41} - X_{43}$, the local priorities are obtained by constructing pairwise comparison matrices. Altogether, there are eight pairwise comparison matrices that need to be constructed: one for the criteria in Layer 2, four for the sub-criteria in Layer 3 and three for the alternative solutions in Layer 4. For simplicity and illustrative purposes, we assume that all the criteria and sub-criteria in Layer 2 and Layer 3 are indifferent, so that all the elements in these pairwise comparison matrices are 1 and $p_{X_1} = \cdots = p_{X_4} = p_{X_{11}} = \cdots = p_{X_{43}} = 0.5$. For the methods in Layer 4, experts are invited to make pairwise comparisons among the three methods in terms of computational costs, numbers of hyper-parameters and historical data requirements, respectively. The pairwise comparison matrices are constructed following the 1-9 scaling system introduced in Section 2.3. The resulted pairwise comparison matrices are

$$A_{X_{41}} = \begin{bmatrix} 1 & 4 & 2 \\ 1/4 & 1 & 1/3 \\ 1/2 & 3 & 1 \end{bmatrix}, A_{X_{42}} = \begin{bmatrix} 1 & 6 & 4 \\ 1/6 & 1 & 1/3 \\ 1/4 & 3 & 1 \end{bmatrix}, A_{X_{43}} = \begin{bmatrix} 1 & 1/3 & 1/2 \\ 3 & 1 & 2 \\ 2 & 1/2 & 1 \end{bmatrix}.$$

The inter-level priorities are calculated using (4) and listed in Table 9. The value of CR for each comparison matrix is also calculated following the procedures in Figure 4 to check the consistency. It can be seen from Table 9 that all the three CR are less than the threshold value 0.1: therefore, all the three comparison matrices are consistent.

Table 9 Inter-level priorities of the alternative solutions with respect to $X_{41} - X_{43}$

	X_{41}	X_{42}	X_{43}
$p_{M_1,X_{ij}}$	0.558	0.691	0.163
$p_{M_2,X_{ij}}$	0.122	0.091	0.540
$p_{M_3,X_{ij}}$	0.320	0.218	0.297
CR	0.009	0.027	0.005

3.2.2. Step 2: Calculate the global priority for each alternative solution

Equation (7) is, then, used to determine the global priority for each alternative solution, where the local priorities involved have been determined in Section 3.2.1 (see Table 8 and Table 9). The obtained global priorities are given in Table 10.

Table 10 Global priorities for the three prognostic methods

p_{M_1}	p_{M_2}	p_{M_3}
0.312	0.366	0.322

3.2.3. Step 3: Determine the method trustworthiness

It can be seen from Table 10 that FANN (M_2) is the most trustworthy one among the three prognostic methods. Its method trustworthiness is, then, evaluated by expert judgements and serves as benchmark for the other two methods. Suppose the experts judge that the trustworthiness of M_2 is $X_{M_2} = 0.85$; then, the trustworthiness of the other two methods can be determined using (8), as shown in Table 11.

Table 11 Method trustworthiness for the three prognostic methods

X_{M_1}	X_{M_2}	X_{M_3}
0.72	0.85	0.75

3.3. Prediction capability assessment and method selection

The prediction capabilities of the three prognostic methods are assessed using the classifier in **Figure 6**, where the values of X and Y are given in Table 11 and Table 6, respectively. The result is given in Table 12. Based on the assessment results, FS can be used to support CBM decision making for this specific application, while FANN and HSMM should not be used to support maintenance decisions due to their relative poor prediction capabilities in this case study.

Table 12 Prediction capabilities for the three prognostic methods

Prognostic methods	FS (M_1)	FANN (M_2)	HSMM (M_3)
C_{M_i}	C_1	C_0	C_0

4. Conclusions

In this paper, a hierarchical framework is developed to assess the prediction capability of prognostic methods. The framework considers the joint contributions from prediction quality and method trustworthiness (Layer 2). The prediction quality and method trustworthiness are further decomposed into six sub criteria (Layer 3) and 19 basic sub-criteria (Layer 4), where information and data can be collected to support the prediction capability assessment. A bottom-up method is developed to determine the prediction capability based on the information and data collected in the Layer-4 basic sub-criteria, in which the AHP method is applied for the aggregation of qualitative sub-criteria. A classification-based method is developed for the assessment of prediction capability. Based on the assessed prediction capability, the appropriateness of the prognostic method for supporting maintenance decisions can be determined, i.e., labelling it as qualified to support predictive maintenance, qualified to support condition-based maintenance or not qualified to support any maintenance decision.

The framework proposed in this paper does not pretend to be exhaustive in the criteria and factors considered, nor rigidly prescriptive in the methods used for their evaluation. In the end, the prediction capability assessment is framed as a process of classification: given all the available information and knowledge, classify the prognostic methods based on their prediction capabilities. Therefore, in the future research, classification algorithms, e.g., Naïve Bayes classifier, majority rule sorting, etc., will also be investigated to develop efficient prediction capability assessment methods. Furthermore, various uncertainties exist in the process of prediction capability assessment. For example, the number of evidence in Figure 1 is often estimated based on sampling approaches. Hence, uncertainty arises from sampling errors. Also, the qualitative basic sub-criteria are evaluated based on pairwise comparisons and, therefore, subjected to uncertainty resulting from incomplete knowledge. How to address the effect of uncertainty in prediction capability assessment deserves further investigations too.

Acknowledgement

This work has been performed within the initiative of the Center for Resilience and Safety of Critical

Infrastructures (CRESCI, http://cresci.cn). The major part of this paper was conducted during Dr. Zhiguo Zeng's visit to Politecnico di Milano. The authors would also like to thank the editor and the two anonymous reviewers for their help in improving the quality of this paper.

Appendix Detailed definitions of the basic sub-criteria

Table A.1 Descriptions of the Layer-4 basic sub-criteria related to prediction quality

Notation	Name	Formula	Description	Range	Category
Y_{11}	Timeliness Weighted Error Bias	$Y_{11}=1-\frac{1}{N}\sum_{i=1}^{N}\varphi\left(\sum_{t=1}^{T_i}\omega_i(t)\cdot\frac{\left(RUL_i^*(t)-RUL_i(t)\right)}{T_i}\right)$, $\varphi(z)=\begin{cases}\exp\left(\frac{\lvert z\rvert}{a_1}\right)-1 & \text{for } z<0,\\ \exp\left(\frac{\lvert z\rvert}{a_2}\right)-1 & \text{for } z>0,\end{cases}$ $a_1>a_2>0.$	Calculate the penalized weighted prediction error over the entire lifetime T_i, with a penalty function $\varphi(z)$ to penalize late predictions ($z\geq 0$) against early predictions ($z<0$). The weighting function $\omega_i(t)$ is a Gaussian Kernel Function with a mean value T_i and a standard deviation $0.5T_i$, which puts more weights on errors made at the end of lifetime. The optimal value for the TWEB is 1, indicating that the average penalized weighted prediction value is centered on the true RUL. Values smaller than 1 indicate that the predictions dispersion is above, or under, the true RUL.	$(-\infty,1]$	Accuracy PPI
Y_{12}	Sample Mean Error	$Y_{12}=1-\left\lvert\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t)-RUL_i(t)\right)\right)\right\rvert$	Measure the average sum of errors over all sample points up to T_i. The optimal value of Y_{12} is 1, indicating that the sum of prediction errors of all the sample points is 0. Therefore, the predicted RUL is equally distributed to both sides of the true RUL. Low values of Y_{12} indicate greater discrepancy between the predicted and true RUL.	$(-\infty,1]$	Accuracy PPI
Y_{13}	Mean Absolute Percentage Error	$Y_{13}=1-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left\lvert\frac{RUL_i^*(t)-RUL_i(t)}{RUL_i(t)}\right\rvert\right)$	Exploits the average absolute percentage error of all N units throughout their lifetime T_i. The optimum value for Y_{13} is 1, indicating that the average absolute percentage error for all units throughout their lifetime T_i is small. A low value tells the user that a discrepancy between the estimated RUL and the true one occurs.	$(-\infty,1]$	Accuracy PPI
Y_{14}	Mean Square Error	$Y_{14}=1-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t)-RUL_i(t)\right)^2\right)$	Takes into account the average for all N units of the average quadratic error of the RULs estimated during the lifetime T_i. The optimum value for the Y_{14} is 1,	$(-\infty,1]$	Accuracy PPI

			indicating that the estimated RULs are equal to the real ones for all units i. A low value indicates that, during the lifetime of the N components, the errors in the RUL estimates are high.		
Y_{15}	Sample Median Error	$Y_{15} = 1 - \left\| \underset{i=1,2,\ldots,N}{\text{Median}} \left(\frac{1}{T_i} \sum_{t=1}^{T_i} \left(RUL_i^*(t) - RUL_i(t) \right) \right) \right\|$	Exploits the absolute value of the median of all mean errors, for all N units, over their lifetime T_i. An optimum value for Y_{15} is 1, indicating that the modulus of the median error is zero. A low Y_{15} indicates that most RUL estimates are wrong.	$(-\infty, 1]$	Accuracy PPI
Y_{21}	$\alpha - \lambda$ performance	$Y_{21} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T_i} \sum_{t=1}^{T_i} b(t) \right)$, where $b(t) = \begin{cases} 1 & \text{if } (1-\lambda) RUL_i^*(t) \in (1 \pm \alpha) RUL_i\left(t + \lambda(EOP_t - t)\right) \\ 0 & \text{otherwise} \end{cases}$	Measures the average fraction of points, during the lifetime T_i over all N units, for which the prediction of the RUL estimated at a specific time t before failure is, with α confidence, the true RUL at $t + \lambda(EOP_t - t)$. The optimum value for Y_{21} is 1, indicating that all estimated RULs have still an accuracy at least of α at a relative distance λ from the current prediction time t. Low values indicate that the prediction made at time t is not reliable in the future time window defined by λ. The parameter α is the confidence modifier and λ is the time window modifier.	$[0,1]$	Precision PPI
Y_{22}	Weighted Prediction Spread	$Y_{21} = 1 - \sigma_{1,2,\ldots,N} \left(\sum_{i=1}^{T_i} \omega_i(t) \frac{\left(RUL_i^*(t) - RUL_i(t) \right)}{T_i} \right)$	Considers the standard deviation of the weighted prediction error during the entire lifetime T_i for all N units. The optimum value for Y_{22} is 1, indicating that all units either share a similar average weighted prediction error or that it is small. A low value of Y_{22} indicates a high dispersion, and thus, a low precision.	$(-\infty, 1]$	Precision PPI
Y_{23}	Sample Standard Deviation	$Y_{23} = 1 - \sqrt{\frac{\sum_{i=1}^{N} \left(ME_i - Y_{12} \right)^2}{N-1}}$, where $ME_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \left(RUL_i^*(t) - RUL_i(t) \right)$	Considers the standard deviation of the average error over the lifetime T_i for all N units. The optimum Y_{23} value is 1, indicating that all errors for all units are closely similar. A low value of Y_{23} indicates that the dispersion of the errors within the N units is high.	$(-\infty, 1]$	Precision PPI

Y_{24}	Root Mean Square Error	$Y_{24} = 1 - \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}\left(RUL_i^*(t) - RUL_i(t)^2\right)}$	Considers the average of the Root Mean Squared Error of the N units during the entire lifetime T_i. The optimum value of Y_{24} is 1, indicating that the error between the estimated RUL and the true RUL is consistent in the model. A low value indicates that the discrepancy between the estimated and the true RUL is inherently stochastic.	$(-\infty, 1]$	Precision PPI
Y_{25}	Prediction Spread	$Y_{25} = 1 - \sigma_{1,2,\ldots,N}\left(M_i(t)\right)$	Considers the standard deviation of the Indicator M for all *N* units. The Y_{25} measures how the indicator M varies through all N units. The optimum value of Y_{25} is 1, indicating that the standard deviation of the indicator is 0: thus, the indicator is concentrated on one value, reducing the variability of the performance throughout the units. A low value indicates that the indicators behavior varies between units.	$(-\infty, 1]$	Precision PPI

Notations:

- i: index for the identification of the unit under test (e.g., the equipment).
- N: total number of units under test.
- t: index for the time instant.
- T: failure time of the unit. Note that each unit has a different T_i value.
- EOP: End-Of-Prediction, time at which the unit is expected to fail, as predicted by the prognostic model.
- n: number of total measurements.
- $RUL_i^*(t)$: Estimated Remaining Useful Life (RUL) for the unit i, at time index t.
- $RUL_i(t)$: Real RUL value for the unit i, at time index t.
- $M_i(t)$: PPI calculated for the unit i, at time t.

Table A.2 Descriptions of the Layer-4 basic sub-criteria related to method trustworthiness

Notation	Meaning	Sub-criterion	Type
X_{11}	Number of academic evidence that supports the method's reliability	Reliability	Quantitative
X_{12}	Number of industrial evidence that supports the method's reliability		Quantitative
X_{21}	Number of academic evidence that supports the method's validity	Validity	Quantitative
X_{22}	Number of industrial evidence that supports the method's validity		Quantitative
X_{31}	Number of successful applications dealing with non-linear problems	Mathematical modeling adequacy	Quantitative
X_{32}	Number of successful applications dealing with non-stationary problems		Quantitative
X_{41}	Requirements on computational costs	Resources requirements	Qualitative
X_{42}	Number of hyper-parameters that needs to be tuned		Qualitative
X_{43}	Requirements on historical data		Qualitative

References

[1]. Pecht, M. and R. Jaai, A Prognostics and Health Management Roadmap for Information and Electronics-Rich Systems. Microelectronics Reliability, 2010. 50(3): p. 317-323.

[2]. Vichare, N.M. and M.G. Pecht, Prognostics and Health Management of Electronics. Components and Packaging Technologies, IEEE Transactions on, 2006. 29(1): p. 222-229.

[3]. Wang, H., Y. Li, Y. Liu, Y. Yang, and H. Huang, Remaining Useful Life Estimation Under Degradation and Shock Damage. Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability, 2015. 229(3): p. 200-208.

[4]. Dalal, M., J. Ma and D. He, Lithium-Ion Battery Life Prognostic Health Management System Using Particle Filtering Framework. Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability, 2011. 225(O1): p. 81-90.

[5]. Zhang, Z.X., X.S. Si, C.H. Hu, and X.Y. Kong, Degradation Modeling-Based Remaining Useful Life Estimation: A Review On Approaches for Systems with Heterogeneity. Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability, 2015. 229(4SI): p. 343-355.

[6]. Tsui, K.L., N. Chen, Q. Zhou, Y.Z. Hai, and W.B. Wang, Prognostics and Health Management: A Review on Data Driven Approaches. Mathematical Problems in Engineering, 2015(793161).

[7]. Si, X.S., W.B. Wang, C.H. Hu, and D.H. Zhou, Remaining Useful Life Estimation - a Review On the Statistical Data Driven Approaches. European Journal of Operational Research, 2011. 213(1): p. 1-14.

[8]. Peng, Y., M. Dong and M.J. Zuo, Current Status of Machine Prognostics in Condition-Based Maintenance: A Review. The International Journal of Advanced Manufacturing Technology, 2010. 50(1-4): p. 297-313.

[9]. Lipi, T.F., J.H. Lim, M.J. Zuo, and W. Wang, A Condition-And Age-Based Replacement Model Using Delay Time Modelling. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2011: p. 1748006X11421265.

[10]. Yun, W.Y. and A.J. Endharta, A Preventive Replacement Policy Based On System Critical Condition. Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability, 2016. 230(1SI): p. 93-100.

[11]. Saxena, A., S. Sankararaman and K. Goebel. Performance Evaluation for Fleet-Based and Unit-Based Prognostic Methods. in Second european conference of the prognostics and health management society. 2014.

[12]. Saxena, A., J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher. Metrics for Evaluating Performance of Prognostic Techniques. in Prognostics and health management, 2008. phm 2008. international conference on. 2008: IEEE.

[13]. Walther, B.A. and J.L. Moore, The Concepts of Bias, Precision and Accuracy, and their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance. Ecography, 2005. 28(6): p. 815-829.

[14]. Saxena, A., J. Celaya, B. Saha, S. Saha, and K. Goebel. On Applying the Prognostics Performance Metrics. in Annual Conference of the PHM Society. 2009.

[15]. Tobon-Mejia, D.A., K. Medjaher, N. Zerhouni, and G. Tripot, A Data-Driven Failure Prognostics Method Based On Mixture of Gaussians Hidden Markov Models. Reliability, IEEE Transactions on, 2012. 61(2): p. 491-503.

[16]. Micea, M.V., L. Ungurean, G.N. Cârstoiu, and V. Groza, Online State-Of-Health Assessment for Battery Management Systems. Instrumentation and Measurement, IEEE Transactions on, 2011. 60(6): p. 1997-2006.

[17]. Hu, Y., P. Baraldi, F. Di Maio, and E. Zio, Online Performance Assessment Method for a Model-Based Prognostic Approach. IEEE Transactions On Reliability, 2016. 2(65): p. 718-735.

[18]. Fan, J., K. Yung and M. Pecht, Prognostics of Lumen Maintenance for High Power White Light Emitting Diodes Using a Nonlinear Filter-Based Approach. Reliability Engineering & System Safety, 2014. 123: p. 63-72.

[19]. Hu, Y., P. Baraldi, F. Di Maio, and E. Zio, A Particle Filtering and Kernel Smoothing-Based Approach for New Design Component Prognostics. Reliability Engineering & System Safety, 2015. 134: p. 19-31.

[20]. Miao, Q., H. Huang and X. Fan, A Comparison Study of Support Vector Machines and Hidden Markov Models in Machinery Condition Monitoring. Journal of Mechanical Science and Technology, 2007. 21(4): p. 607-615.

[21]. Peng, T., Y. Liu, A. Saxena, and K. Goebel, In-Situ Fatigue Life Prognosis for Composite Laminates Based On Stiffness Degradation. Composite Structures, 2015. 132: p. 155-165.

[22]. Saxena, A., J. Celaya, B. Saha, S. Saha, and K. Goebel, Metrics for Offline Evaluation of Prognostic Performance. International Journal of Prognostics and Health Management, 2010. 1(1): p. 4-23.

[23]. Goebel, K., B. Saha and A. Saxena. A Comparison of Three Data-Driven Techniques for Prognostics. in 62nd meeting of the society for machinery failure prevention technology (mfpt). 2008.

[24]. Oberkampf, W.L., M. Pilch and T.G. Trucano, Predictive Capability Maturity Model for Computational Modeling and Simulation. 2007: Sandia National Laboratories.

[25]. Paulk, M.C., B. Curtis, M.B. Chrissis, and C.V. Weber, Capability Maturity Model, Version 1.1. Software, IEEE, 1993. 10(4): p. 18-27.

[26]. Gibson, C.F. and R.L. Nolan, Managing the Four Stages of EDP Growth. 1974: Harvard Business Review.

[27]. Herbsleb, J., D. Zubrow, D. Goldenson, W. Hayes, and M. Paulk, Software Quality and the Capability Maturity

Model. Communications of the ACM, 1997. 40(6): p. 30-40.
[28]. Spruit, M. and K. Pietzka, MD3M: The Master Data Management Maturity Model. Computers in Human Behavior, 2015. 51(SIB): p. 1068-1076.
[29]. Farrell, M. and R. Gallagher, The Valuation Implications of Enterprise Risk Management Maturity. Journal of Risk and Insurance, 2015. 82(3): p. 625-657.
[30]. de Carvalho, J.V., A. Rocha and J. Vasconcelos, Towards an Encompassing Maturity Model for the Management of Hospital Information Systems. Journal of Medical Systems, 2015. 39(999).
[31]. Di Maio, F., P. Turati and E. Zio. Prediction Capability Assessment of Data-Driven Prognostic Methods for Railway Applications. in Third European conference of the prognostic and health management society. 2016. At Bilbao, Spain: PHM Society.
[32]. Saaty, T.L., A Scaling Method for Priorities in Hierarchical Structures. Journal of Mathematical Psychology, 1977. 15(3): p. 234-281.
[33]. Streimikiene, D., J. Sliogeriene and Z. Turskis, Multi-Criteria Analysis of Electricity Generation Technologies in Lithuania. Renewable Energy, 2016. 85: p. 148-156.
[34]. Vaidya, O.S. and S. Kumar, Analytic Hierarchy Process: An Overview of Applications. European Journal of Operational Research, 2006. 169(1): p. 1-29.
[35]. Huguenin, J., Data Envelopment Analysis and Non-Discretionary Inputs: How to Select the Most Suitable Model Using Multi-Criteria Decision Analysis. Expert Systems with Applications, 2015. 42(5): p. 2570-2581.
[36]. Dožić, S. and M. Kalić, An AHP Approach to Aircraft Selection Process. Transportation Research Procedia, 2014. 3: p. 165-174.
[37]. Saaty, T.L., Decision Making with the Analytic Hierarchy Process. International journal of services sciences, 2008. 1(1): p. 83-98.
[38]. Di Maio, F. and E. Zio, Failure Prognostics by a Data-Driven Similarity-Based Approach. International Journal of Reliability, Quality and Safety Engineering, 2013. 20(1).
[39]. Zio, E. and F. Di Maio, A Data-Driven Fuzzy Approach for Predicting the Remaining Useful Life in Dynamic Failure Scenarios of a Nuclear System. Reliability Engineering and System Safety, 2010. 95(1): p. 49-57.
[40].Bai, G., P. Wang and C. Hu, A Self-Cognizant Dynamic System Approach for Prognostics and Health Management. Journal of Power Sources, 2014. 278: p. 163-174.
[41]. Dong, M. and D. He, A Segmental Hidden semi-Markov Model (HSMM)-based Diagnostics and Prognostics Framework and Methodology. Mechanical Systems and Signal Processing, 2007. 21(5): p. 2248-2266.
[42]. Moghaddass, R. and M.J. Zuo, An Integrated Framework for Online Diagnostic and Prognostic Health Monitoring Using a Multistate Deterioration Process. Reliability Engineering & System Safety, 2014. 124: p. 92-104.
[43]. Miao, Q. and V. Makis, Condition Monitoring and Classification of Rotating Machinery Using Wavelets and Hidden Markov Models. Mechanical Systems and Signal Processing, 2007. 21(2): p. 840-855.
[44]. Huang, N.E., Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng,... H.H. Liu. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. in Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 1998: The Royal Society.