Article

Explainable Artificial Intelligence for Anomaly Detection and Prognostic of Gas Turbines using Uncertainty Quantification with Sensor-Related Data

Ahmad K.M. Nor ¹, Srinivasa Rao Pedapati ¹, Masdi Muhammad ¹, Víctor Leiva ^{2,*}

¹ Mechanical Department, Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia; ahmad_18002773@utp.edu.my (A.K.K.N.), srinivasa.pedapati@utp.edu.my (S.R.P.), masdimuhammad@utp.edu.my (M.M.)

² School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, 2362807 Valparaíso, Chile

* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com (V.L.)

Abstract: Explainable artificial intelligence (XAI) is in its assimilation phase in the prognostic and health management (PHM). The literature on PHM-XAI is deficient with respect to metrics of uncertainty quantification and explanation evaluation. This paper proposes a new method of anomaly detection and prognostic for gas turbines using Bayesian deep learning and Shapley additive explanations (SHAP). The method explains the anomaly detection and prognostic, and improves the performance of the prognostic, aspects that have not been considered in the literature of PHM-XAI. The uncertainty measures considered serve to broaden explanation scope and can also be exploited as anomaly indicators. Real-world gas turbine sensor-related data are tested for the anomaly detection, while NASA commercial modular aero-propulsion system simulation data, related to turbofan sensors, were used for prognostic. The generated explanation is evaluated using two metrics: consistency and local accuracy. All anomalies were successfully detected using the uncertainty indicators. Meanwhile, the turbofan prognostic results showed up to 9% improvement in root mean square error and 43% enhancement in early prognostic due to the SHAP, making it comparable to the best existing methods. The XAI and uncertainty quantification offer a comprehensive explanation for assisting decision-making. Additionally, the SHAP ability to increase PHM performance confirms its value in AI-based reliability research.

Keywords: Artificial intelligence; CMAPSS; consistency and local accuracy; CUSUM chart; deep learning; prognostic and health management; RMSE; sensing and data extraction; SHAP; Uncertainty; XAI

1. Introduction

1.1. Artificial Intelligence

Artificial intelligence (AI) is a marvel of today's technological advancement. AI marks the culmination of a decades-long effort by the technical community in imitating biological reasoning. The expansion of data volume from big data [1], the availability of open-source development tools, the easing of collaboration between AI players and countless unexplored opportunities push AI on a global scale. Backed by a steady flow of investment and enjoying supports from tech-friendly authorities, AI-based projects flourish, replacing the old ways of doing things. AI brings optimization, automation, and efficiency to the table.

Nowadays, AI-powered applications are practically everywhere, whether it is apparent or hidden. AI penetration is not limited to social media, where it is probably more visible to the public, but it reaches far into niche areas. Much progress has been felt in recent years, especially in fields such as healthcare [2], defense [3], manufacturing [4], biology [5], robotics [6], and reliability [7].

Tech firms and external funders define the AI investment landscape now, with machine learning (ML) startups being one of the most funded sectors since 2011 [8]. Approximately 30% augmentation in AI investment was registered from 2010 to 2013 and 40% from 2013 to 2016 [9]. To give an idea of the scale this represents, around \$26 to \$39 billion were invested in 2016.

Price Water Cooper (www.pwc.com, accessed on 24 August 2021) projects an equivalent of \$15.7 trillion or 14% of added gross domestic product value by the 2030, fueled by the growth in productivity and consumer demand due to AI [10]. McKinsey (www.mckinsey.com, accessed on 24 August 2021) estimated an annual increase of 1.2% in the global gross domestic product, or \$13 trillion by 2030, driven by AI substitution of the workforce and AI-driven industrial innovation [11].

1.2. Black Box Obstacle

The most used and the most powerful AI methods are black boxes in nature. Deep learning (DL) [12], for example, is opaque. In other words, the reason the model produces an output is unknown. Naturally, this presents an obstacle, a risk, in integrating AI for high-performance and high-risk markets. Decision-making in these markets depends much on supportive evidence and not a merely point-estimate prediction. A wrong forecast by AI models could prove disastrous in terms of life, health, time, or financially.

Regulation bodies see red in this opacity and started introducing laws to protect users. The General Data Protection Regulation (GDPR) in the European Union went into effect in 2018 [13]. The GDPR is a comprehensive set of regulations governing algorithmic responsibility, requiring openness, procedure, and supervision when computers are used to make a significant decision concerning human being. The year after, the Ethics Guidelines for Trustworthy Artificial Intelligence presented by the European Commission High-Level Expert Group on AI was published [14]. It suggests some essential requirements to make AI trustworthy. These laws and guidelines echo the same idea: AI transparency.

1.2. Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) is a discipline dedicated to making AI models discoverable and more transparent. While the term has existed early on, it recently picked up steam because of rising scrutiny in AI usage [15]. The accumulation of publications and the surge in interest expressed for the search term XAI since 2016, shown in Figure 1, reflect the growing interest in the field [16]. In 2017, Defense Advanced Research Projects Agency (www.darpa.mil, accessed on 24 August 2021)) launched the "XAI initiative", while the Chinese government published "The Development Plan for New Generation of Artificial Intelligence" in the same year, both aiming to increase XAI [15].

It is essential that XAI transcends regulations and is rewarding rather than burdensome to the AI community. Some of the incentives in incorporating XAI are as follows:

- (i) Justify decisions, detect problems, and improve AI models.
- (ii) Comply with the regulations, bias, ethics, reliability, accountability, safety, and security of the AI use.
- (iii) Allow the user to verify the desirable properties of the model, encourage interactivity, gain new insights about the model or data, and increase human intuition.
- (iv) Permit user's task, efforts, and resources to be more optimized and targeted.
- (v) Detect when the cost of error is high or when the AI system is not proven reliable.
- (vi) Foster the collaboration between experts, data scientists, users, and stakeholders.

Figure 1. Interest shown for XAI term in Google search (source: the authors).

1.3. The State XAI in Prognostic and Health Management

Prognostic and health management (PHM) is a maintenance and asset management strategy that exploits signals, measurements, models, and algorithms to anticipate, analyze, and track health deterioration in industrial assets [17]. The PHM provides standards and protocols to ensure that assets are in good working order. It reduces hazards, maintenance costs, and workload, allowing maintenance operations to be optimized.

Failure prognostic, diagnostic, and anomaly detection are the three categories of PHM activities. Prognostic is the process of determining an asset's remaining useful life (RUL) or spare operating time before breakdown. Anomaly detection is the action of identifying unusual patterns going against the norm of operational indicators. In contrast, diagnostic is the action of classifying failure and discovering the detailed root cause of failure. AI-based methods occupy a crucial position in PHM research, as shown in [7]. In addition, the XAI is somewhat a novelty in PHM. A systematic review conducted in [18] summarizes the current state of XAI in PHM where:

- (i) The XAI assimilation in PHM is still in its early years. Nevertheless, it is gaining interest, with a spike in published works in 2020.
- (ii) Interpretable model, rule and knowledge-based model, and attention mechanism are the most commonly used XAI approach in PHM at the moment, as presented in Figure 2.
- (iii) As seen in most diagnostic and anomaly detection works, XAI is becoming vital to PHM since it can be adapted as a tool to execute PHM tasks.
- (iv) The XAI unalters PHM performance.
- (v) Identified gaps in PHM-XAI research comprises of lack of human participation, explanation metrics, and uncertainty management.
- (vi) Mostly real, industrial case studies were tested in previous works to demonstrate the effectiveness of XAI in the PHM domain.

Figure 2. XAI approach in PHM works (source: the authors).

1.4. Related works

Next, we elaborate some literature on PHM-XAI. In presentation order: Interpretable model [19], tree-based [20], knowledge and rule-based [21], logic analysis of data (LAD) [22], feature extraction-based [23], filter-based [24], cluster-based [23], attention-based [26], model-agnostic explainability [27], and layer-wise relevance propagation (LRP) [28].

An interpretable logistic regression model with elastic net regularization is employed for high-pressure plunger pump anomaly detection [19]. Data are first equally divided, and the statistic measures are calculated on each division. A rolling window operation is then applied on the extracted features where a flag is associated, indicating if a failure will occur or not based on the statistical measure calculated before. The flagged representations, having the most relevant features related to failure, serve as input to the regularized logistic regression. The relevance order of features to be included from the flagged representations is determined by considering the non-failure/failure feature distributions and measuring their Kolmogorov–Smirnov distance.

A graphical diagnostic technique based on convolutional neural network (CNN) and extreme gradient boosting (XGBoost), applied on gas turbine failure problems, is presented in [20]. It replaces portions of the CNN architecture with XG-boost, an ML approach for classification and regression, and makes the CNN training model interpretable. XGBoost is a boosting method that combines several weak classifiers into a single robust classifier. The classification and regression tree (CART) is the weak classifier utilized by XGBoost. CART is a binary tree that splits by looking for the best segmentation feature and cut point using the Gini coefficient as a criterion. The time-series data are fed into the CNN. When comparable signals are clustered together, the local features are improved, allowing the CNN to be more accurate. These signals may be sorted with XGBoost, improving feature order interpretability. The original raw data obtained from the gas turbine are first fed into the CNN to determine the accuracy. Then, the signal rankings from the initial raw data, as well as the accuracy gained by CNN, are trained in XGBoost to produce tree models that can choose the optimal features-accuracy sorting combinations.

A K-margin-based interpretable learning (KEEN) is presented in [21] for interpretable aircraft structural damage diagnostic. This framework consists of a residual convolution recurrent neural network (RCR-net), a K-margin diagnostic method and a knowledge-directed interpretation approach. The RCR-net is a DL model that can automatically obtain features and deal with class skewness issues. As input, it accepts augmented data segments. After that, it divides the augmented segments into small fragments and outputs the segment's health-condition prognosis. The K-margin-based diagnostic model is robust against noise. It focuses on the RCR-net's most relevant segments automatically. Its health-condition detector employs segments with top-K confident to

estimate the health status. Simultaneously, a knowledge-based interpretation approach automatically extracts features from the RCR-net responsible for the fault.

A process diagnostic-explanation structure consisting of the knowledge discovery in database (KDD) method [29] and failure tree analysis (FTA) is proposed in [22]. The KDD method, specifically the LAD, extracts patterns from the process dataset and produces rule-based explanation describing the root cause of failure. This explanation is later translated into the FTA logic reasoning. The ability of this KDD method is demonstrated in an actuator system failure diagnostic.

The spectrum anomaly detector with interpretable feature (SAIFE) is an AAE-based model, where AAE is adversarial autoencoders in short. The SAIFE is applied to the problem of wireless spectrum anomaly detection [23]. The long short-term memory (LSTM) is an artificial recurrent neural network. The LSTM is used in DL and acts as the encoder for extracting interpretable features such as signal bandwidth, class, and center frequency via a linear layer and classifying signal via a softmax layer. A CNN acts as a decoder for reconstructing the input data from the extracted features. The AAE architecture is trained in a semi-supervised fashion for learning interpretable features, while the reconstruction is fully unsupervised. The model learns the features during the semi-supervised training with partial data. During testing, an anomaly is detected based on the reconstruction error, classification error, and loss from the discriminator, which is part of the AAE generator-discriminator adversarial architecture. Anomaly localization is achieved by plotting the absolute reconstruction error.

The time-scattering convolutional network (TScatNet) is proposed in [24] for bearing and drive train failure diagnostic. The TScatNet collects domain-invariant features utilizes Morlet wavelet and uses these features for diagnostic purpose. The TScatNet consists of a time-scattering (Scat) module of a standard CNN having Morlet wavelets as convolutional filters and a softmax module comprising of global averaging pooling (GAP) and softmax layer. The Scat module transforms the input into scattering features maps. At testing phase, these maps are passed to the GAP layer, which aids in the simplification of testing processes and improves the stability of the derived scattering characteristic. The softmax layer maps each scattering feature into the probability value of fault categories.

An emission control system fault diagnostic based on principal component analysis PCA clustering [25] is presented in [26]. The sensor data is firstly treated with PCA for dimensionality reduction. These sensor data are mapped to relative air/fuel ratio target, which represents normal or degraded operation. Then, the result of the PCA undergo PCA-based clustering (vectorized PCA -VPCA-, multilinear PCA -MPCA-, or uncorrelated multilinear PCA -UMPCA-). The PCA-based clusters isolate fault events in a restricted number of clusters (scenarios), each one described by a reference pattern. Once the data have been partitioned into clusters (scenarios), practitioners analyze cluster patterns to get more insight for fault diagnostic. This framework provides practitioners with an efficient and interpretable model of multichannel profile data in high-dimensional spaces to support the diagnostic and finding root cause.

Classification of linear motion guide fault based on CNN applied to vibration signal and explainability with frequency domain-based grad-CAM (FG-CAM) are proposed to analyze frequencies that have a significant impact on fault conditions [27]. A feed-forward neural network (FFNN) together with global Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) are employed to predict and explain the damage of prismatic cantilever steel beam in [28]. The frequencies and associated damage percentage, ranging from 0% to 75%, are used as input features and the distances, corresponding to 194 positions of damage, are used as target of the FFNN.

Diagnostic of induction motor fault using CNN and LRP is proposed in [30]. The vibration time series data segments used as input are transformed into time-frequency image using continuous wavelet transform with Morlet wavelet, which is then processed by a CNN for classification. The LRP captures pixel-level representation of features contributing to the failure.

1.5. Research Objectives and Contributions

This work firstly elaborates how data uncertainty can be exploited as anomaly indicator in detection tasks. Then, it details a prognostic improvement method using SHAP global explanation, with both task predictions being explained by SHAP. Additionally, the uncertainty also serves to strengthen the explanation by broadening its scope. Local accuracy and consistency metrics are employed to assess the explanation. Real-world data from a gas turbine and NASA commercial modular aero-propulsion system simulation (CMAPPS) turbofan sensor datasets are respectively utilized for demonstrating the anomaly detection and prognostic capabilities.

The main contributions of the present work are four folds:

- (i) Firstly, the uncertainty, together with the XAI form a broader explanation scope, which bridge the gap identified in Section 1.4.
- (ii) Secondly, the SHAP ability to improve PHM task is proposed, which was absent from previous works as explained in [18].
- (iii) Thirdly, the application of explanation metrics is considered, which was nearly missing from former works as indicated in Section 1.4.
- (iv) Fourthly and lastly, this investigation reveals the practicality of DL uncertainty as an anomaly indicator using a real-world dataset.

The secondary contributions of this work are two folds:

- (i) We add knowledge to the AI-based PHM works that employ model agnostic approaches, which is insufficiently explored as testified in Figure 2.
- (ii) By verifying the local accuracy and consistency property of SHAP explanation, we confirm the efficiency, symmetry, dummy, and additivity properties of Shapley values.

Note that 100% of the anomalous data are successfully detected when using the uncertainty-based indicator. Moreover, the prognostic performance improved around 6% to 9% and an 43% improvement in early prognostic based on the SHAP global explanation.

2. Materials and Methods

2.1. Uncertainties in Deep Learning

Uncertainty in DL linked to the quality of input data is known as aleatoric uncertainty (AU). This uncertainty may happen due to noise, data acquisition error, or stochasticity captured in the input data, which is the usual situation encountered in the real world. This type of uncertainty cannot be reduced further by having more data if no improvement was done on the data acquisition technique. Uncertainty linked to the chosen parameters (weights) of a DL model is called epistemic uncertainty (EU) [31,32,33,34].

2.2. Multi-Outputs Bayesian LSTM

To enable the quantification of both uncertainties and generate explanation, a single input, multi outputs probabilistic LSTM is developed. The model consists of an input layer and then an LSTM layer, followed by a fully connected layer. The proceeding layers are the output layers. The first output layer is the AU layer, generating sequential outputs with data uncertainty. The second output layer is the EU one, also producing sequential outputs with parameter uncertainty. The last output layer produces the prediction to be explained. In this layer, the outputs from the LSTM are sliced to obtain only the first value of each sequence which are then grouped in a single explanation vector. The model was only trained to minimize the first and second outputs loss. For a simplified schematic of the whole model, see Structure 1 in Figure 3.

For anomaly detection, the model is fed with only healthy data while for prognostic, both healthy and failure inputs are involved. Next, we define the concepts of probabilistic layers and Bayesian hyperparameter optimization (BayesOpt).

Figure 3. Different LSTM structures used in this work.

- (i) Probabilistic layers: The AU layer is a probabilistic layer that learns and predicts the mean and standard deviation (SD) from the input coming from the LSTM layer to form the prediction range, translated into uncertainty distribution [35]. Thus, every point in an RUL sequence consists of a distribution of the RUL prediction. In this work, the normal or Gaussian distribution was used to model the uncertainty as it is easily understood. The EU layer, called the dense variational layer, learns, and predicts the weights distributions or the posterior distribution of the weights using variational inference by maximizing the evidence lower bound objective [35], $\mathcal{L}$ namely, defined as

$$\mathcal{L}(q_{\theta}) = -E_{q_{\theta}(w)} \left[-\log(P(y|X, w)) - \log \left(\frac{P(w)}{q_{\theta}(w)} \right) \right] = - \int dw q_{\theta}(w) \log(P(y|X, w)) + \int dw q_{\theta}(w) \log \left(\frac{q_{\theta}(w)}{P(w)} \right), \quad (1)$$

with $P(w)$ defined in (1) being the prior distribution, $q_{\theta}(w)$ being the approximation distribution, and $P(y|X, w)$ being the likelihood function relating all inputs X , all labels y , and the weights w . Then, the weights distribution can be sampled to produce the output for a given input.

- (ii) Bayesian hyperparameter optimization: The hyperparameters for the model are obtained via BayesOpt, executed in 250 evaluations [36]. Optimized hyperparameters help in reducing the EU. The explored hyperparameters and its search space are shown in Table 1.

Table 1. BayesOpt hyperparameters search space.

Parameters	Hidden Units	Fully Connected Layer Size	Mini Batch Size	Learning Rate
Space	10 to 1000	10 to 500	26 to 130	5e-4 to 1e-3

2.3. Data Denoising and Uncertainty Visualization

Since noise could worsen the AU, data denoising is performed by applying singular value decomposition following the method shown in [37,38].

The rolling SD of the prediction distributions characterizes the uncertainty. Increasing the trend in SD signifies a decreasing confidence in model's prediction and vice versa.

2.4. CUSUM Changepoint Detection for Anomaly Detection

The uncertainty mirrors the model confidence in predicting. Since the model is trained with only healthy data, the AU is expected to show a spike once anomalous input is tested, signaling that the distribution of data in question was not previously learned during the training phase. The changepoint detection is applied to identify the anomaly spikes with appropriate control limits such as the cumulative sum (CUSUM) control chart [1,39].

Given a sequence $\{x_1, \dots, x_n\}$ with mean m_x and SD σ_x , the lower (L_i) and upper (U_i) limits of a CUSUM control chart are stated as

$$L_i = \begin{cases} 0, & i = 1; \\ \min\left(0, L_{i-1} + x_i - m_x + \frac{1}{2}n\sigma_x\right), & i > 1; \end{cases} \quad (2)$$

$$U_i = \begin{cases} 0, & i = 1; \\ \max\left(0, U_{i-1} + x_i - m_x - \frac{1}{2}n\sigma_x\right), & i > 1. \end{cases} \quad (3)$$

A process deviates from control if x_j obeys $L_j < -c\sigma_x$ or $U_j > c\sigma_x$, with L_j and U_j defined in (2) and (3), where c is a predetermined constant defined using healthy data prediction AU and calculated as

$$c = \frac{\sigma_{AUmax} - \sigma_{AUmean}}{\sigma_{AU-SD}} \quad (4)$$

where σ_{AUmax} stated in (4) is the maximum of the SDs of the AU, σ_{AUmean} is the mean of the SDs of the AU, and σ_{AU-SD} is the SD of the SDs of the AU,

2.5. Shapley Additive Explanations

The SHAP is a game theoretic approach to explain the output of any ML model [40]. It evaluates the contribution of each feature to the prediction by using Shapley values. The SHAP can be both global and local explainability approaches. The Shapley values determine the importance of a single feature by considering the outcome of each possible combination of features. In other words, the Shapley value is the average marginal contribution of a feature across all possible combination of features formulated as

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j \quad (5)$$

where M established in (5) is the maximum coalition size, $\phi_j \in R$ is the Shapley value for a feature j , $z' \in \{0,1\}^M$ are the simplified features that describe the presence of interested feature in the feature combination with $z' = 0$ meaning the interested feature is absent in the combination and $z' = 1$ indicating the feature is present. The formula for Shapley value is given by

$$\phi_j(\text{val}) = \sum_{S \subseteq \{x_1, \dots, x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (\text{val}(S \cup \{x_j\}) - \text{val}(S)) \quad (6)$$

where S expressed in (6) is a subset of the features used in the model, x is the vector of feature values of the instance to be explained, and p is the number of features.

Note that the prediction for feature values in set S that are marginalized over features that are not included in set S is stated as

$$\text{val}_x(S) = \int \hat{f}(x_1, \dots, x_p) dP_{x \notin S} - E_x(\hat{f}(X)) \quad (7)$$

where $E_x(\hat{f}(X))$ is the average predicted value. However, the SHAP only accepts non-probabilistic model. Thus, for generating explanation, another LSTM model is used, whose structure and weights resemble the input and the third output layers of the original model as depicted in Structure 2 of Figure 3.

The SHAP force plot and waterfall plot are used to explain the instance prediction while the SHAP summary plot, a global visualization, explains by identifying the most contributing features in a sequence. In the force plot, each feature value is represented as a positive or negative force pushing or dragging the prediction while in waterfall plot, the features contribution, and its force, linking the instance prediction and the average prediction are depicted. In the summary plot, features are ordered according to its absolute Shapley value. Those with important values occupy top positions than less important features. The force plot is used to explain anomaly instances, while the summary plot explains and improves the prognostic performance. Furthermore, the waterfall plot is employed to verify the consistency nature of the explanation as described later in Section 2.7.

2.6. Performance Evaluation

Next, we present two performance evaluation metrics: the model predictive performance and the early prediction score.

- (i) Model predictive performance: The average root mean square error (RMSE) for 100 predictions is calculated between the predicted RUL (mean of RUL distribution) and the true RUL [41,42] given by

$$\text{RMSE} = \frac{\left(\sqrt{\frac{1}{M} \sum_{i=1}^M (\text{RUL}_{\text{true}}^{(i)} - \text{mean}_{\text{pred}}^{(i)})^2} \right)}{100} \quad (8)$$

where $\text{RUL}_{\text{true}}^{(i)}$ stated in (8) is the true RUL for gas turbine i , $\text{mean}_{\text{pred}}^{(i)}$ is the predicted RUL for gas turbine i , and M as the total number of gas turbines.

- (ii) Early prediction score: This metric is only applied in prognostic task. The scoring function, s namely, gives a higher score for the same error in early prediction than in late prediction. This metric penalizes the late prediction than the early prediction as the latter one is more important than the former in any failure-related forecasting problem [43,44]. The average score for 100 predictions is calculated as

$$s = \frac{(M \sum_{i=1}^M s_i)}{100} \quad (9)$$

where s_i defined in (9) is given by

$$s_i = \begin{cases} \frac{-d_i}{e^{13}} - 1, & d_i < 0; \\ \frac{d_i}{e^{10}} - 1, & d_i > 0; \end{cases} \quad d_i = (\text{mean}_{\text{pred}}^{(i)} - \text{RUL}_{\text{true}}^{(i)}).$$

2.7. Explanation Metrics

Now, we introduce the two metrics for evaluating the SHAP explanation [45]: local accuracy and consistency.

- (i) Local accuracy: This metric states that the feature contributions must add up to the difference of prediction for x and the average. Starting from a normal SHAP notation, we have that

$$f(x) = \Phi_0 + \sum_{j=1}^M \Phi_j x'_j, \quad (10)$$

where Φ_0 formulated in (10) is given by $\Phi_0 = E_x(\hat{f}(X))$ and setting $x'_j = 1$, the Shapley value efficiency property is found such that

$$f(x) = \Phi_0 + \sum_{j=1}^M \Phi_j x'_j = E_x(\hat{f}(X)) + \sum_{j=1}^M \Phi_j$$

and

$$\sum_{j=1}^M \Phi_j = f(x) - E_x(\hat{f}(X)),$$

where $f(x)$ is the prediction for x and $E_x(\hat{f}(X))$ is given in (7).

- (ii) Consistency: This metric states that, if a model changes so that the marginal contribution of a feature value increases or stays, the Shapley value also increases or stays the same, which is defined, for any two models f and f' , as

$$f'_x(z') - f'_x(z'_{\setminus j}) \geq f_x(z') - f_x(z'_{\setminus j}), \quad (11)$$

for $z' \in \{0, 1\}^M$ and $z'_{\setminus j} \Leftrightarrow z' = 0$, with $f_x(z')$ being the model with Structure 2 in Figure 3, while $f'_x(z')$ is the same model but with different weights. Note that $f_x(z'_{\setminus j})$ and $f'_x(z'_{\setminus j})$ are the models with Structure 3 in Figure 3, having the same weights as $f_x(z')$ and $f'_x(z')$ respectively, except for the input of interest. Then, we have that

$$\Phi_j(f', x) \geq \Phi_j(f, x), \quad (12)$$

To examine this metric, the outputs of $f_x(z')$, $f_x(z'_{\setminus j})$, $f'_x(z')$, $f'_x(z'_{\setminus j})$, $\Phi_j(f, x)$ and $\Phi_j(f', x)$ are extracted from the waterfall plot. Note that the expression stated in (11) can be calculated to verify the formula given in (12). By validating this metric, the explanation also conforms to the symmetry, dummy and additivity natures of the Shapley values.

3. Results

3.1. Case Study 1: Anomaly Detection on Real-World Gas Turbine Data

A one-year worth of data coming from a twin-shaft 18.8 MW industrial gas turbine was exploited. This equipment had been previously studied in [46]. The data consists predominantly of healthy data with some anomalies producing null (zero) and NaN sensor measurement. It comprises of 98 features ranging from temperature, pressure, speed, and position, totaling 8737 hours of recorded measurement. However, as stated in [46], only some variables are useful for the DL model. The inputs-outputs are shown in Table 2. All the inputs were employed to predict each of the outputs as depicted in Figure 5 by four models denoted as $LSTM_{p2}$, $LSTM_{p4}$, $LSTM_{T4}$ and $LSTM_{N1}$. Figure 4 shows a schematic diagram of the gas turbine under consideration.

Figure 4. 18.8 MW gas turbine scheme.

Table 2. Real gas turbine variables.

Ref	Input	Unit	Ref	Output	Unit
P_1	Compressor inlet pressure	Bar	P_2	Compressor outlet pressure	Bar
T_1	Compressor inlet temperature	K	P_4	Gas generator turbine outlet pressure	Bar
N_2	Power turbine rotational speed	RPM	T_4	Gas generator turbine outlet temperature	K
m_f	Fuel mass flow rate	kg/s	N_1	Gas generator rotational speed	RPM

Figure 5. Real gas turbine inputs and outputs modeling.

(i) Data preparation

Anomalous data in the order of 377 hours were firstly removed from the dataset. The rest of the data were divided into training and testing datasets. A sequence of data was set to 24 hours. Thus, the models were fed with 24 hours input and output the same length of prediction. Hourly data from 01 January 2018 to 26 November 2018 amounting to 7488 hours or 312 sequences were used for training and validation. The data from 26 November 2018 to 31 December 2018 amounting to 816 hours or 34 sequences were reserved for healthy state testing purpose. The anomalous hours were combined with the healthy data corresponding to the period before and after the anomaly to make up a sequence of 24 hours. The null anomaly, on the 8 April to 9 April at 11pm to 12am (sixth to seventh instances) were considered. The summary of the datasets is presented in Table 3.

Table 3. Real gas turbine datasets.

Dataset	Training & Validation	Testing	Null Anomaly
Date & Sequence of Interest	1 Jan to 26 Nov 2018	26 Nov to 31 Dec 2018	8 April to 9 April 2018 11pm to 12am 6th to 7th instances
Total Hours	7488	816	24
Total Sequence	312	34	1

(ii) Healthy data modeling performance

The average RMSE results for healthy testing data are presented in Table 4. However, in [42] did not specify any numerical results for the performance so that no comparison can be done.

Table 4. Average RMSE for 100 evaluations with AU and UE.

Model	RMSE with AU	RMSE with EU
$LSTM_{P2}$	387.73	387.90
$LSTM_{P4}$	22.98	36.54
$LSTM_{T4}$	12.79	42.23
$LSTM_{N1}$	128.14	34.74

(iii) Prediction with null anomaly

The predictions done for the sequence containing anomalous inputs for $LSTM_{P2}$, $LSTM_{P4}$, $LSTM_{T4}$, and $LSTM_{N1}$ with AU are, respectively, presented in Figures 6, 7, 8, and 9.

Figure 6. $LSTM_{P2}$ prediction containing null anomalous data.

Figure 7. LSTM_{P4} prediction containing null anomalous data.

Figure 8. LSTM_{T4} prediction containing null anomalous data.

Figure 9. LSTM_{N1} prediction containing null anomalous data.

(iv) Calculation of the control limit c

The variables and results for the control limit c are listed in Table 5.

Table 5. Control limit c variables.

Model	σ_{AUmax}	σ_{AUmean}	σ_{AUstd}	c
LSTM _{p2}	271.61	251.86	5.92	3.33
LSTM _{p4}	34.58	32.39	0.74	2.95
LSTM _{T4}	12.15	11.35	0.26	3.13
LSTM _{N1}	158.71	149.51	3.56	2.58

(v) Anomaly detection with CUSUM

The CUSUM chart for anomaly detection associated with the predictions and the control limit c are presented in Figure 10. The coordinates featured in the chart belong to the identified anomalies.

Figure 10. CUSUM chart for anomaly detection in 24h sequence corresponding to: (a) LSTM_{P2}, (b) LSTM_{P4}, (c) LSTM_{T4}, and (d) LSTM_{N1} predictions.

(vi) Anomaly sequence force plot visualization

The SHAP force plot for explaining the anomalies are shown in Figures 11 and 12. The marked areas corresponding to the sixth and seventh instances are the tested anomaly instances. For illustration purpose, only instances 1 to 9 are shown.

Figure 11. SHAP force plots for anomaly instances corresponding to: (a) $LSTM_{p2}$ and (b) $LSTM_{p4}$ predictions.

Figure 11. SHAP force plots for anomaly instances corresponding to: (c) LSTM_{T4} and (d) LSTM_{N1} predictions.

3.2. Case Study 2: Failure Prognostic on CMAPSS Turbofan Dataset

CMAPPS turbofan run failure datasets were published in 2008 by Nasa Prognostic Centre of Ames Research Centre consisting of four complete sets of training, testing, and true RUL for numerous turbofan engines. The synthetic data were obtained by simulating a variety of operational conditions and injecting faults of varying degradation degree.

The chosen FD001 training and testing datasets consisted each of 100 recorded turbofan degradations as summarized in Table 6. A single record corresponds to a turbofan whose health condition deteriorated after certain cycle, or failure start point, until breakdown [47]. Each turbofan fleet might be used in different operating conditions. As such the extent of degradation is different from one another. Each record is a time series comprising of time (in cycles), three operating conditions (OC) and 21 sensor signals as presented in Appendix A (Table 11). The RUL targets for the training dataset are not available, only the true RUL are given. The OC refers to different operating regimes combination of altitude (O-42K ft), throttle resolver angle (20-100), and match number (0-0.84) [47]. High level noise is incorporated, and the faults encountered are hidden by the effect of various operational conditions [48].

Table 6. Turbofan datasets

Dataset	Fault Modes	Operating Condition	Train Units	Test Units
#1	1	1	100	100

(i) Data preparation

Only strictly monotonic sensors are selected [46]. These sensors are useful as they best represent trending degradation contrary to irregular and unchanged signals. Thus, 14 sensor signals, corresponding to sensors 2, 3, 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 20 and 21 are used. Together with the three OCs, the total features used was 17.

To obtain the RUL labels for training, piece-wise linear degradation was assumed [49,50]. Each fleet health was considered stable in the beginning, followed by a linear deterioration after the failure start point until breakdown.

Originally, the RUL for a signal takes the value of the recorded signal last cycle, or the signal sequence length, and degraded linearly until zero as shown for Fleet 1 in Figure 12(a). The failure start point for each signal is identified using a CUSUM chart with the control limit c set to five standard deviations. Then, the mean of these failure start points is calculated, in this case, resulting to cycle 46. Combining the linear degradation obtained earlier and the mean failure start point, the transformed Fleet 1 RUL sequence is presented in Figure 12(b). To facilitate model generalization, all target RULs were capped to 50. The total signal sequence lengths and its respective RUL for training and testing datasets are presented in Appendix B (Figure 18).

Figure 12. RUL targets modelling: (a) initial, that is, the recorded signal sequence length is 192 and thus a linear degradation starting from RUL = 192 to RUL = 0 is modelled; (b) final: that is, a piece wise linear degradation, combining 12(a) with failure start point at cycle 46.

(ii) Prognostic performance

The RUL prediction for Fleet 1 and Fleet 18 using the 17 features are illustrated respectively in Figures 14 and 5. These fleets were chosen because the former fleet testing data length and true RUL follow the same trait as the training data while the latter fleet is not, as indicated in Appendix B. Thus, it is interesting to examine the uncertainty behavior between the two fleets.

The SHAP summary plot for the prediction of these fleets are depicted in Figure 13. Indeed, almost all summary plots for the 100 testing fleets show the same order of features as in Figure 13. Hence, we can choose the best set of features to improve the predictive performance. Accordingly, the model is also tested with the best 13 features or 75% and the best 9 features or 50% of the original 17 features. Table 7 lists the combination of features tested.

Table 7. Combination of features tested.

Combination	Features According to Contribution Order
17 Features	S8, S11, S4, S13, S15, OC1, S3, OC3, S12, S2, S21, S14, S20, S17, OC2, S7 and S9
13 Features	S8, S11, S4, S13, S15, OC1, S3, OC3, S12, S2, S21, S14 and S20
9 Features	S8, S11, S4, S13, S15, OC1, S3, OC3 and S12

Figure 13. SHAP summary plot for (a) Fleet 1 and (b) Fleet 18 sequence predictions.

The average prognostic RMSE and early scoring results for 100 predictions using the 17, 13 and 9 features are presented in Table 8. As shown in this table, the model performed better in mostly all metrics with 13 features. It shows around 9% improvement in RMSE with AU and 6% in RMSE with EU as well as 43% improvement in early scoring with AU and EU compared to 17 features.

Table 8. Average prognostic performance with 17, 13 and 9 features.

Results	17 Features	13 Features	9 Features
RMSE with AU	16.20	14.68	14.75
RMSE with EU	17.09	16.04	15.56
Score with AU	724.13	409.10	412.37
Score with EU	897.38	507.63	518.52

Figure 14. Fleet 1 RUL prediction with AU and EU.

Figure 15. Fleet 18 RUL prediction with AU and EU.

(iii) Performance comparison with existing methods

The results using 13 features with AU compared with published methods are respectively presented in Tables 9 and 10.

Table 9. RMSE comparison with existing methods.

Results	Proposed Method	DBN [49]	ELM [49]	RNN [50]	DCNN [50]	BiLSTM [51]
RMSE with AU	14.68	15.21	17.27	13.44	12.61	13.65

where BiLSTM: bidirectional LSTM; DBN: deep belief network; DCNN: deep convolutional neural network; ELM: extreme learning machine; RNN: recurrent neural networks

Table 10. Early score comparison with existing methods

Results	Proposed Method	DBN [49]	ELM [49]	RNN [50]	DCNN [50]	BiLSTM [51]
Score with AU	4.09×10^2	4.18×10^2	5.23×10^2	3.39×10^2	2.74×10^2	2.95×10^2

3.3. Explanation Evaluation

(i) Local accuracy verification

The waterfall plot of the first instance on the first sequence of $LSTM_{P2}$ prediction is shown in Figure 16. From the illustration, note that the sum of Shapley values or contributions is equal to $f(x) - E_x(\hat{f}(X))$, with $f(x) = -31.15$, and $E_x(\hat{f}(X)) = -30.573$. The values are more accurately shown on top of the plot with the order of: $P1$, $T1$, $N2$ and m_f .

Figure 16. Waterfall plot of the first instance on the first sequence of $LSTM_{P2}$.

(ii) Consistency verification

The contribution of variable m_f on the first test data instance is investigated as an example. For each output, the difference and contribution of m_f , Φ_{MF} , is calculated as

$$P2_{tot'_1} = LSTM_{P2.2}(z'_1) - LSTM_{P2.2}(z'_{1 \setminus MF}) = -3.273; \Phi_{MF_{P2.1}}(f', x) = -0.050;$$

$$P2_{tot_1} = LSTM_{P2}(z'_1) - LSTM_{P2}(z'_{1 \setminus MF}) = -14.893; \Phi_{MF_{P2.1}}(f', x) = -0.198;$$

$$P4_{tot'_1} = LSTM_{P4.2}(z'_1) - LSTM_{P4.2}(z'_{1 \setminus MF}) = -0.962; \Phi_{MF_{P4.1}}(f', x) = -0.002;$$

$$P4_{tot_1} = LSTM_{P4}(z'_1) - LSTM_{P4}(z'_{1 \setminus MF}) = -1.182; \Phi_{MF_{P4.1}}(f', x) = -0.048;$$

$$T4_{tot'_1} = LSTM_{T4.2}(z'_1) - LSTM_{T4.2}(z'_{1 \setminus MF}) = 0.219; \Phi_{MF_{T4.1}}(f', x) = 0.002;$$

$$T4_{tot_1} = LSTM_{T4}(z'_1) - LSTM_{T4}(z'_{1 \setminus MF}) = -1.004; \Phi_{MF_{T4.1}}(f', x) = -0.006;$$

$$N1_{tot'_1} = LSTM_{N1.2}(z'_1) - LSTM_{N1.2}(z'_{1 \setminus MF}) = 9.017; \Phi_{MF_{N1.1}}(f', x) = 0.749;$$

$$N1_{tot_1} = LSTM_{N1.2}(z'_1) - LSTM_{N1.2}(z'_{1 \setminus MF}) = -3.030; \Phi_{MF_{N1.1}}(f', x) = -0.075;$$

with $P2_{tot'_1} > P2_{tot_1}$, $P4_{tot'_1} > P4_{tot_1}$, $T4_{tot'_1} > T4_{tot_1}$, and $N1_{tot'_1} > N1_{tot_1}$. Thus, $\Phi_{MF_{P2.1}} > \Phi_{MF_{P2.1}}$; $\Phi_{MF_{P4.1}} > \Phi_{MF_{P4.1}}$; $\Phi_{MF_{T4.1}} > \Phi_{MF_{T4.1}}$; and $\Phi_{MF_{N1.1}} > \Phi_{MF_{N1.1}}$ as seen above. These results are illustrated in the waterfall plots in Figure 17.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 17. Waterfall plot for local feature contribution of (a) $\Phi_{MF_P2-1}(f', x)$, (b) $\Phi_{MF_P2-1}(f', x)$, (c) $\Phi_{MF_P4-1}(f', x)$, (d) $\Phi_{MF_P4-1}(f', x)$, (e) $\Phi_{MF_T4-1}(f', x)$, (f) $\Phi_{MF_T4-1}(f', x)$, (g) $\Phi_{MF_N1-1}(f', x)$, and (h) $\Phi_{MF_N1-1}(f', x)$.

4. Discussion

4.1. Explainable Anomaly Detection

Note that 100% of the tested null anomalies were successfully detected with the help of the AU indicator and CUSUM changepoint detection as illustrated in Figure 10. The AU spiked, representing the unsurety of the model when it was fed with anomalous data, surpassing the healthy threshold c limit at the instances of anomaly for all outputs.

The force plots local explanation, linked to the anomaly instances shown in Figure 11 highlight that m_f , fuel mass flow rate and N2, power turbine rotational speed, were the responsible features causing the anomaly. During the initial instances before the anomalies, all features contributed to the prediction. When the consecutive anomalies occurred, the force of both features were amplified. In the seventh instance, all other feature forces were eclipsed, showing mostly m_f and N2. However, on the eighth instance, the distribution of contributing forces became normal, with all the features taking part in the prediction. The red colored bar in the plot pushed the prediction positively while the blue colored bar dragged the prediction negatively. The width of the bar represents its contributing force magnitude while the values on these bars are the normalized test data values. The base value is the average output of the model during training phase. To improve the anomaly detection, one could reduce the c limit value, resulting to a faster detection. Nevertheless, by doing so, the risk of false alarm increases. Considering that the tested anomalies are merely stochastic disturbance rather than a continuous one, the present c limit definition is deemed acceptable.

4.2. Explainable Prognostic

Figure 15 depicts the prognostic result of Fleet 18. From this figure, note that the AU shows a rising trend, signaling that the model is increasingly uncertain of its prediction, reflecting the predicted RUL sequence which is far from the true RUL. Nonetheless, the AU for Fleet 1 prediction indicates a decreasing trend as presented in Figure 14, mirroring the good prediction that the model made. Thus, this model becoming more and more confident of its sequential estimation. Meanwhile, the EU measure manifests very small change in nearly the same scale for both fleets, which is expected for the weight uncertainty. This uncertainty should not be influenced much by the change in input data.

The summary plot global explanation ordered the features according to their contribution power in the sequence prediction as shown in Table 7. The top 5 variables influencing the prediction are physical fan speed, static pressure at high pressure compressor (HPC) outlet, total temperature at low pressure turbine (LPT) outlet, corrected fan speed, and bypass ratio. The model predictive performance increased around 6% to 9%, while its early prediction showed 43% improvement with only 13 of the most influencing features. The model performed a little worse with only 9 features compared to 13, though it was still better than using all 17 inputs. The predictive power decreased by 0.5% and increased by 3% with AU and EU, respectively, while the early prediction ability decreased by 0.8% and 2% with AU and EU, respectively, compared to 13 features. However, considering that only 9 features were used instead of 13, this small performance drop is perfectly tolerable. Weighing all factors, one could even justify that the model with 9 features is better than the one with 13 features. The enhanced result is comparable to the best methods outcome in the CMAPPS FD001 dataset. It is true that some works fare better than the proposed framework. This is firstly due to a more complex structure adoption. The DCNN and RNN, studied in [52] for example, has five convolutional layers and five recurrent layers to learn the data while the BiLSTM presented in [53] possesses two BiLSTM layers and two fully connected layers. Secondly, the mentioned methods only produce point estimates results, without any quantification of uncertainty. Obviously, model generalization is easier in this case. Consequently, without uncertainty measure, these works can only be experimental and cannot be applied in real-world applications.

4.3. Explanation Evaluation

This work demonstrated that the SHAP explanation satisfies the local accuracy and consistency criteria. By fulfilling these properties, the explanation also conforms to the efficiency, symmetry, dummy and additivity natures of the Shapley values. Efficiency affirms that the sum of the feature contributions is equal to the difference between the instance prediction and the average prediction of all instances. Symmetry implies that two feature value contributions should be identical if they contribute equally to all feasible coalitions. Dummy states that a feature that does not affect the predicted value should have a Shapley value of zero regardless of the coalition it is part of. Moreover, the additivity denotes that, for an ensemble prediction, for a specific feature, we can calculate the Shapley value of the feature in each individual ensemble, average them, and get the Shapley value for the feature for the whole ensemble.

5. Conclusions

In this article, we have applied the Shapley additive explanations agnostic approach to the outputs of a Bayesian long short-term memory in anomaly detection and prognostic tasks of gas turbines using real-world and simulated datasets. The forecast uncertainty, generated by the Bayesian model, broaden the explanation scope to include model confidence, strengthening the explanation. It was also exploited as anomaly indicator. The Shapley additive explanations were used to enhance prognostic performance by identifying the most contributing features in the prediction. All the anomalous instances were detected owing to the uncertainty indicator. Moreover, the model root mean square error increased around 6% to 9%, while its early prediction ability showed 43% improvement using the Shapley additive explanations. These results are comparable to the best existing methods in the problem. Therefore, the generated explanation has verified the local accuracy and consistency properties, validating the efficiency, symmetry, dummy and additivity natures of the Shapley values. In summary, this investigation showed how the Shapley additive explanations and deep learning uncertainty form a broader explanation scope while simultaneously demonstrating the Shapley additive explanation ability in enhancing the prognostic and health management performance, highlighting its potential as an easy to use, flexible and powerful explainable artificial intelligence technique.

Author Contributions: Conceptualization A.K.M.N.; methodology A.K.M.N., V.L.; software, A.K.M.N.; validation, A.K.M.N., V.L.; formal analysis, A.K.M.N.; investigation, A.K.M.N.; data curation, A.K.M.N.; writing—original draft preparation, A.K.M.N.; writing—review and editing, A.K.M.N., V.L.; visualization, A.K.M.N.; supervision, S.R.P., M.M., V.L.

Funding: This research was partially funded by Universiti Teknologi Petronas Foundation -YUTP- (A.K.M.N., S.R.P., M.M) and by FONDECYT, project grant number 1200525 (V.L.) from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation

Data Availability Statement: The data presented in this study are openly available in github.com/Kamalnor?tab=repositories (accessed on 24 August 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table 11. Turbofan sensor description.

Sensor	References	Description	Unit
S1	T_2	Total temperature fan inlet	$^{\circ}\text{R}$
S2	T_{24}	Total temperature at LPC outlet	$^{\circ}\text{R}$
S3	T_{30}	Total temperature at HPC outlet	$^{\circ}\text{R}$
S4	T_{50}	Total temperature at LPT outlet	$^{\circ}\text{R}$
S5	P_2	Pressure at fan inlet	Psia
S6	P_{15}	Total pressure in bypass-duct	Psia
S7	P_{30}	Total pressure at HPC outlet	Psia
S8	N_f	Physical fan speed	RPM
S9	N_c	Physical core speed	RPM
S10	E_{pr}	Engine pressure ratio (P_{50}/P_2)	N/A
S11	P_{s30}	Static pressure at HPC outlet	psia
S12	Φ	Ratio of fuel flow to P_{s30}	Pps/psi
S13	NR_f	Corrected fan speed	RPM
S14	NR_c	Corrected core speed	RPM
S15	BPR	Bypass ratio	N/A
S16	farB	Burner fuel-air ratio	N/A
S17	htBleed	Bleed enthalpy	N/A
S18	N_{f_dmd}	Demanded fan speed	RPM
S19	$PCNfR_dmd$	Demanded corrected fan speed	RPM
S20	W_{31}	HPT coolant bleed	lbm/s
S21	W_{32}	LPT coolant bleed	lbm/s

Appendix B

Figure 18. Data sequence length and associated RUL for training data related to (a) sequence length and (b) RUL; and for testing data associated with (c) sequence length and (d) true RUL.

References

1. Aykroyd, R.G., Leiva, V., Ruggeri, F. Recent developments of control charts, identification of big data sources and future trends of current research. *Technological Forecasting and Social Change* 2019, 144, 221-232.
2. Davenport, T., Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthcare Journal* 2019, 6, 94-98.
3. Bistrón, M., Piotrowski, Z. Artificial intelligence applications in military systems and their influence on sense of security of citizens. *Electronics* 2021, 10, 871.
4. Zhang, J., Arinez, J., Chang, Q., Gao, R., Xu, C. Artificial intelligence in advanced manufacturing: current status and future outlook. *Journal of Manufacturing Science and Engineering* 2020, 142, 110804.
5. Mahmud, M., Kaiser, M.S., McGinnity, T.M., Hussain, A. Deep learning in mining biological data. *Cognitive Computation* 2021, 13, 1-33.
6. Raj, M., Seamans, R. Primer on artificial intelligence and robotics. *Journal of Organization Design* 2019, 8, 11.
7. Nor, A.K.M., Pedapati, S.R., Muhammad, M. Reliability engineering applications in electronic, software, nuclear and aerospace industries: A 20-year review (2000–2020). *Ain Shams Engineering Journal*, 2021, in press doi: 10.1016/j.asej.2021.02.015
8. Mou, X. Artificial intelligence: investment trends and selected industry uses. *The World Bank* 2019, 143357, pp. 1-8.
9. Bughin, J., Hazan, S., Ramaswamy, Chui, M., Allas, T., Dahlstrom, P., Henke, N., Trench, M. Artificial intelligence: the next digital frontier? McKinsey and Company 2017, pp. 1-80, apo.org.au/node/210501 (accessed 24 August 2021).
10. Gillham, J. The economic impact of artificial intelligence on the global economy. Price Waterhouse Coopers, London, UK, 2017 doi:10.13140/RG.2.2.21506.38083.
11. Bughin, J., Seong, J., Manyika, J., Chui, M., Joshi, R. Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey and Company 2018, retrieved from: www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy (accessed on 24 August 2021).
12. Figueroa Barraza, J., López Droguett, E., Ramos Martins, M.. Towards interpretable deep learning: A feature selection framework for prognostics and health management using deep neural networks. *Sensors* 2021, 21, 5888.
13. Kaminski, M.E. The right to explanation, explained. *34 Berkeley Technology Law Journal* 2019, 189-218.
14. European Commission High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. Retrieved from: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 24 August 2021).
15. Xu F., Uszkoreit H., Du Y., Fan W., Zhao D., Zhu J. Explainable AI: A brief survey on history, research areas, approaches and challenges. In Tang J., Kan MY., Zhao D., Li S., Zan H. (eds.) *Natural Language Processing and Chinese Computing*. NLPCC 2019. vol 11839. Springer, Cham.
16. Adadi, A., Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018, 6, 52138-52160.
17. Sheppard, J.W., Kaufman, M.A., Wilmer, T.J. IEEE standards for prognostics and health management. *IEEE Aerospace and Electronic Systems Magazine*, 24, 34-41.
18. Nor, A.K.M., Rao, S.P., Muhammad, M. Explainable AI (XAI) for PHM of industrial asset: A state-of-the-art, prisma-compliant systematic review. *arXiv e-prints*, 2021.
19. Langone, R., Cuzzocrea, A. Skantzos, N. Interpretable anomaly prediction: Predicting anomalous behavior in Industry 4.0 settings via regularized logistic regression tools. *Data and Knowledge Engineering* 2010, 130, 101850.
20. Zhou, D., Yao, Q., Wu, H., Ma, S., Zhang, H. Fault diagnostic of gas turbine based on partly interpretable convolutional neural networks. *Energy* 2020, 200, 117467.
21. Zhou, Y., Hong, S., Shang, J., Wu, M., Wang, Q., Li, H., Xie, J. Addressing noise and skewness in interpretable health-condition assessment by learning model confidence. *Sensors*. 2020, 20, 7307.
22. Waghen, K., Ouali, M.S. (2019). Interpretable logic tree analysis: A data-driven fault tree methodology for causality analysis. *Expert Systems with Applications*, 136, 376-391.
23. S. Rajendran, W. Meert, V. Lenders and S. Pollin. Unsupervised wireless spectrum anomaly detection with interpretable features. *IEEE Transactions on Cognitive Communications and Networking* 2019, 5, 637-647.
24. Liu, C., Qin, C., Shi, X., Wang, Z., Zhang, G., Han, Y. TScatNet: An interpretable cross-domain intelligent diagnostic model with antinoise and few-shot learning capability. *IEEE Transactions on Instrumentation and Measurement* 2021, 70, 3506110.
25. Martín-Barreiro, C., Ramírez-Figueroa, J.A., Cabezas, X., Leiva, V., Galindo-Villardón, M.P. Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data. *Sensors* 2021, 21,4094.
26. Pacella, M. Unsupervised classification of multichannel profile data using PCA: An application to an emission control system. *Computers and Industrial Engineering* 2018, 122, 161-169.
27. Kim, M. S., Yun, J. P., Park, P. An explainable convolutional neural network for fault diagnosis in linear motion guide. *IEEE Transactions on Industrial Informatics* 2020, 17, 4036-4045.
28. Onchis, D., Gillich, G.R. Stable and explainable deep learning damage prediction for prismatic cantilever steel beam. *Computers in Industry* 2021, 125, 103359.
29. Palacios, C.A., Reyes-Suarez, J.A., Bearzotti, L.A., Leiva, V., Marchant, C. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy* 2021, 23, 485.
30. Grezmak, J., Zhang, J., Wang, P., Loparo, K.A., Gao, R.X. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnostic. *IEEE Sensors Journal* 2020, 20, 3172-3181.

31. Peng, W., Ye, Z., Chen, N. Bayesian deep-learning-based health prognostics toward prognostics uncertainty. *IEEE Transactions on Industrial Electronics* 2020, 67, 2283-2293.
32. Li, G., Yang, L., Lee, C. G., Wang, X., Rong, M. A Bayesian deep learning RUL framework integrating epistemic and aleatoric uncertainties. *IEEE Transactions on Industrial Electronics*. 10.1109/TIE.2020.3009593.
33. Qin, Y., Liu, Z., Liu, C., Li, Y., Zeng, X., Ye, C. Super-Resolved q-Space deep learning with uncertainty quantification *Medical Image Analysis*, 2021, 67, 101885.
34. Chai, Y., Bian, Y., Liu, H., Li, J., Xu, J. Glaucoma diagnosis in the Chinese context: An uncertainty information-centric Bayesian deep learning model. *Information Processing and Management* 2021, 58, 102454.
35. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J et al. Tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation 2016, USENIX Association, USA, 265–283.
36. Wu, J., Chen, X.Y, Zhang, H., Xiong, L.D, Lei, H., Deng, S.H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology* 2019, 17, 26-40.
37. Epps, B.P., Krivitzky, E.M. Singular value decomposition of noisy data: Noise filtering. *Experiments in Fluids* 2019, 60, 126.
38. Epps, B.P., Krivitzky, E.M. Singular value decomposition of noisy data: Mode corruption. *Experiments in Fluids* 2019, 60, 121.
39. CUSUM. The MathWorks Inc., Natick, Massachusetts, USA. Retrieved from: www.mathworks.com/help/signal/ref/cusum (accessed on 24 August 2021).
40. Lundberg, S.M., Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 2017 pp. 4768-4777.
41. Wang, Q., Zheng, S., Farahat, A., Serita, S., Gupta, C. Remaining useful life estimation using functional data analysis. In *Proceedings of the IEEE International Conference on Prognostics and Health Management* 2019, pp. 1-8, San Francisco, CA, USA.
42. Ge, Y., Sun, L., Ma, J. An improved PF remaining useful life prediction method based on quantum genetics and LSTM. *IEEE Access* 2019, 7, 160241-160247.
43. Ge, Y., Wu, J., Jiang, X. A prediction method using Bayesian theory for remaining useful life. *International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)* 2019, Zhangjiajie, China, pp. 856-862.
44. Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., et al. A light gradient boosting machine for remaining useful life estimation of aircraft engines. In *21st International Conference on Intelligent Transportation Systems (ITSC)* 2018, Maui, Hawaii, USA, pp. 3562-3567.
45. Molnar, C. Interpretable machine learning. A guide for making black box models explainable. Online Book 2019, christophm.github.io/interpretable-ml-book (accessed on 24 August 2021)
46. Tahan, M., Muhammad, M., Karim, Z.A. A multi-nets ANN model for real-time performance-based automatic fault diagnosis of industrial gas turbine engines. *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 2017, 39, 2865-2876.
47. Ramasso, E., Saxena, A. Performance benchmarking and analysis of prognostic methods for CMAPSS datasets. *International Journal of Prognostics and Health Management* 2014, 5, 1-15.
48. Saxena, A., Goebel, K., Simon, D., Eklund, N. Damage propagation modeling for aircraft engine run-to-failure simulation. In *IEEE International Conference on Prognostics and Health Management* 2008, pp. 1-9.
49. Li, J. Li, X., He, D. Domain adaptation remaining useful life prediction method based on ADABN-DCNN. In *IEEE Prognostics and System Health Management Conference*, Qingdao, China, 2019, pp. 1-6,
50. Heimes, F.O. Recurrent neural networks for remaining useful life estimation. In *IEEE International Conference on Prognostics and Health Management* 2008, pp. 1-6.
51. Zhang, C., Lim, P., Qin, A.K., Tan, K.C. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics. *IEEE Transactions on Neural Networks and Learning Systems* 2016, 28, 2306-2318.
52. Li, X., Ding, Q., Sun, J.Q. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering and System Safety* 2018, 172, 1-11. 2018
53. Wang, J., Wen, G., Yang, S., Liu, Y. Remaining useful life estimation in prognostics using deep bidirectional LSTM neural network. In *IEEE Prognostics and System Health Management Conference* 2018, pp. 1037-1042.