Article

Application of Explainable AI (XAI) for Anomaly Detection and Prognostic of Gas Turbines with Uncertainty Quantification.

Ahmad Kamal Mohd Nor [1], Srinivasa Rao Pedapati [2] and Masdi Muhammad [2,*]

[1] Universiti Teknologi Petronas; ahmad_18002773@utp.edu.my
[2] Universiti Teknologi Petronas; Srinivasa.pedapati@utp.edu.my
* Correspondence: ahmad_18002773@utp.edu.my; Tel.: +6-012-500-6658

Abstract: XAI is presently in its early assimilation phase in Prognostic and Health Management (PHM) domain. However, the handful of PHM-XAI articles suffer from various deficiencies, amongst others, lack of uncertainty quantification and explanation evaluation metric. This paper proposes an anomaly detection and prognostic of gas turbines using Bayesian deep learning (DL) model with SHapley Additive exPlanations (SHAP). SHAP was not only applied to explain both tasks, but also to improve the prognostic performance, the latter trait being left undocumented in the previous PHM-XAI works. Uncertainty measure serves to broaden explanation scope and was also exploited as anomaly indicator. Real gas turbine data was tested for the anomaly detection task while NASA CMAPSS turbofan datasets were used for prognostic. The generated explanation was evaluated using two metrics: Local Accuracy and Consistency. All anomalies were successfully detected thanks to the uncertainty indicator. Meanwhile, the turbofan prognostic results show up to 9% improvement in RMSE and 43% enhancement in early prognostic due to SHAP, making it comparable to the best published methods in the problem. XAI and uncertainty quantification offer a comprehensive explanation package, assisting decision making. Additionally, SHAP ability in boosting PHM performance solidifies its worth in AI-based reliability research.

Keywords: XAI; SHAP; Uncertainty; PHM; Anomaly Detection; Prognostic.

1. Introduction

AI is a marvel of today's technological advancement. It marks the culmination of decades-long effort by the technical community in imitating biological reasoning. The expansion of data volume, the availability of open source development tools, the easing of collaboration between AI players and the 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 general public, but it reaches far into niche areas. Much progress has been felt especially in fields such as healthcare [1], defense [2], manufacturing [3], biology [4] robotics [5] and reliability [6] in the recent years.

Tech firms and external funders define the AI investment landscape at the moment, with machine learning startups being one of the most funded sectors since 2011 [7]. Approximately 30% augmentation in AI investment was registered from the 2010 to 2013 and 40% from the 2013 to 2016 [8]. To give an idea of the scale this represents, around $26 to $39 billion were invested in 2016.

Price Water Cooper (PwC) projects an equivalent of $15.7 trillion or 14% of added GDP value by the 2030 fueled by the growth in productivity and consumer demand due

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

to AI [9]. McKinsey, on the other hand, estimates an annual increase of 1.2% in global GDP, or $13 trillion by the 2030, driven by AI substitution of workforce and AI-driven industrial innovation [10].

1.1. Black Box Obstacle

However, the most commonly used and the most powerful AI methods are black box in nature. DL, for example, is opaque. In other word, the reason an output is produced by the model is unknown. Naturally, this presents an obstacle, a risk, in AI assimilation in elevated performance, high stake markets. Decision making in these areas depends much on supportive evidence, and not merely point-estimate prediction. Wrong forecast by AI models could prove disastrous in term of life, health, time, or financially.

Regulation bodies see red in this opacity and started introducing laws to protect user. The General Data Protection Regulation (GDPR) in the European Union (EU) went into effect in 2018 [11]. GDPR is a comprehensive set of regulations governing algorithmic responsibility, requiring openness, procedure, and supervision when computers are used to make major 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 [12]. It suggests some key requirements to make AI trustworthy.

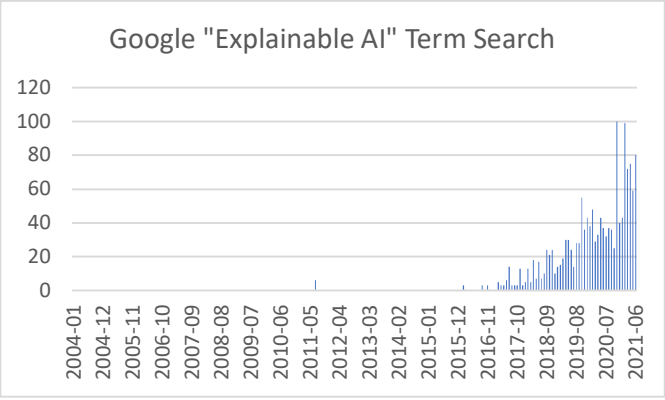
These laws and guidelines echo the same idea: AI transparency.

1.2. Explainable AI (XAI)

XAI is a discipline dedicated in making AI model discoverable and more transparent. While the term has existed early on, it recently picked up steam as a result of rising scrutiny in AI usage [13]. The accumulation of publications and the surge in interest expressed for the search term *Explainable AI* since 2016, shown here in Figure 1, reflect the growing interest in the field [14]. In 2017, DARPA launched the "Explainable AI (XAI) initiative", while the Chinese government published "The Development Plan for New Generation of Artificial Intelligence" in the same year, both aiming to proliferate XAI [13].

The need for XAI transcends regulations. XAI could prove to be rewarding than burdensome to AI community. Some of the incentives in incorporating XAI are as follows:

1. Justify decision, detect problem, and improve AI models.
2. Comply with the regulations, bias, ethics, reliability, accountability, safety, and security of AI use.
3. Enabling user to verify model's desirable properties, encouraging interactivity, gaining new insights on the model or the data and augment human intuition.
4. Allow user's task, effort, and resources to be more optimized and targeted.
5. Important when the cost of error is high or when the AI system is not yet proven to be reliable.
6. Foster the collaboration between experts, data scientists, users, and stakeholders.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

Figure 1. Interest shown for 'Explainable AI' term in Google search.

1.3. The State XAI in PHM

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. [15]. 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 asset's Remaining Useful Life (RUL) or leftover operating time before breakdown. Anomaly detection is the action of identifying unusual patterns going against the norm of operational indicators whereas diagnostic is the action of classifying failure and discovering the detailed root cause of failure. AI-based methods occupy a key position in PHM research as shown in [6]. XAI, on the other hand, is somewhat a novelty in PHM.

A systematic review conducted by the author in [16] summarizes the current state of XAI in PHM:

1. XAI assimilation in PHM is still in its early years. Nevertheless, it is gaining interest, with a spike in published works in 2020.
2. Interpretable model, rule & knowledge-based model as well as attention mechanism are the most commonly used XAI approach in PHM at the moment, as presented in **Figure 2.**
3. XAI is fast becoming vital to PHM, as it can be adapted as a tool to execute PHM tasks, as seen in the majority of diagnostic and anomaly detection works.
4. PHM performance is unaltered by XAI.
5. Identified gaps in PHM-XAI research comprises of lack in human participation, explainability metrics and uncertainty management.
6. Mostly real, industrial case studies were tested in previous works to demonstrate the effectiveness of XAI in PHM domain.

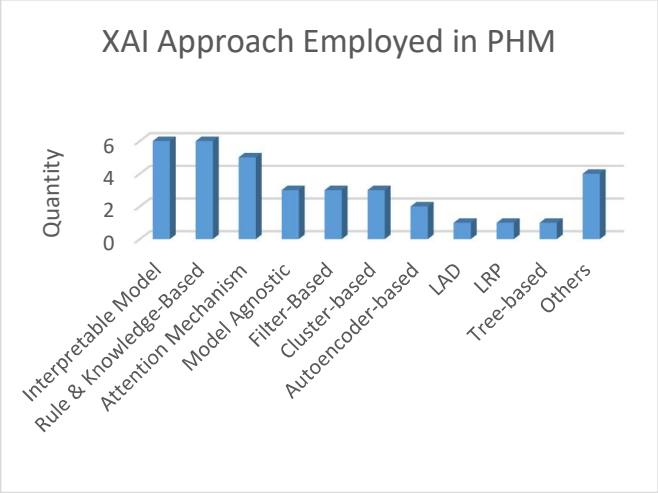


Figure 2. XAI approach in PHM works.

1.4. Related Works

This section elaborates some of the former PHM-XAI articles available. In presentation order: Interpretable model [17], tree-based [18], knowledge & rule-based [19], Logic Analysis of Data (LAD) [20], feature extraction-based [21], filter-based [22], cluster-based [23], attention-based [24], model-agnostic explainability [25] and Layer-Wise Relevance Propagation (LRP) [26].

 doi:10.20944/preprints202109.0034.v1

An interpretable logistic regression model with elastic net regularization is employed in high pressure plunger pump anomaly detection in [17]. Data is first equally divided, and statistic measures are calculated on each division. A rolling window operation is then applied on the extracted features where 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 associated with 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 normal/failure feature distributions and measuring their Kolmogorov–Smirnov distance.

A graphical diagnosis technique based on Convolutional Neural network (CNN) and extreme gradient boosting (XGBoost), applied on gas turbine failure problem is presented in [18]. It replaces portions of the CNN architecture with XG-boost, a machine learning 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 strong 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 will improve, allowing CNN to be more accurate. These signals may be sorted with XGBoost, improving feature order interpretability. To determine the accuracy, the original raw data obtained from the gas turbine is first fed into the CNN. The signal rankings from the initial raw data, as well as the accuracy gained by CNN, are then 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 [19] for interpretable aircraft structural damage diagnosis. This framework consists of a Residual Convolution Recurrent Neural Network (RCR-Net), a K-margin diagnostic method and a knowledge-directed interpretation approach. RCR-Net is a deep learning 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 diagnosis model is robust against noise. It focuses on the RCR-Net's most relevant segments automatically. Its health-condition detector uses 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 knowledge discovery in database (KDD) method and Failure Tree Analysis (FTA) is proposed in [20]. The KDD method, in specific LAD, extracts patterns from the process dataset and produces rule-based explanation describing the root cause of failure. This explanation is later translated into FTA logic reasoning. The ability of this method is demonstrated in an actuator system failure diagnosis.

The Spectrum Anomaly Detector with Interpretable Feature (SAIFE) is an Adversarial Autoencoders (AAE) based model applied on the problem of wireless spectrum anomaly detection [21]. LSTM 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 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, anomaly is detected based on the reconstruction error, classification error and the loss from the discriminator which is part of AAE generator-discriminator adversarial architecture, Anomaly localization is achieved by plotting the absolute reconstruction error.

TScatNet is proposed in [22] for bearing and drive train failure diagnosis. TScatNet collects domain-invariant features utilizing Morlet wavelet and uses these features for diagnosis purpose. TScatNet consists of a time-scattering (Scat) module of standard CNN having Morlet wavelets as convolutional filters and a Softmax module comprising of

 doi:10.20944/preprints202109.0034.v1

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 global averaging pooling (GAP) layer. The GAP layer 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.

Emission control system fault diagnosis method based on PCA clustering is presented in [23]. The sensor data is firstly treated with PCA for dimensionality reduction. This sensor data is mapped to relative air/fuel ratio target, which represents normal or degraded operation. The result of the PCA then undergo PCA-based clustering (Vectorized PCA-VPCA, Multilinear Principal Component Analysis-MPCA or Uncorrelated Multilinear PCA-UMPCA clustering). 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 diagnosis. This provides practitioners with an efficient and interpretable model of multichannel profile data in high-dimensional spaces to support the diagnosis 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 [24].

A feed forward neural network (FFNN) together with SHAP (global) and LIME (local) are employed to predict and explain the damage of prismatic cantilever steel beam in [25]. The frequencies and associated damage percentage ranging from 0% to 75% are used as input features and the distance, corresponding to 194 positions of damage are used as target of the FFNN.

Diagnosis of induction motor fault using CNN and LRP is proposed in [26]. The vibration time series data segments used as input are transformed into time-frequency image using Continuous Wavelet Transform (CWT) with Morlet wavelet which is then processed by CNN for classification. 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 anomaly detection task. Then it details a prognostic improvement method using SHAP global explanation. Both task's predictions were explained by SHAP. Additionally, the uncertainty also serves to strengthen the explanation by broadening its scope. Local Accuracy and Consistency metrics were used to assess the explanation. Real world data from a gas turbine and NASA CMAPPS turbofan datasets were respectively used for demonstrating the anomaly detection and prognostic capabilities.

The direct contributions of this work are four folds:

1. Firstly, the uncertainty, together with XAI form a broader explanation scope, which bridge the gap identified in 1.3.
2. Secondly, the SHAP ability to improve PHM task, which was absent from previous works as explained in [16].
3. Thirdly, the application of explanation metrics which was nearly missing from former works as indicated in 1.3.
4. Finally, this paper reveals the practicality of deep learning uncertainty as anomaly indicator using real world dataset.

The supplementary contributions of this work are two folds:

1. This work adds to AI-based PHM articles employing model agnostic approaches which is insufficiently explored as testified in **Figure 2**.
2. By verifying the local accuracy and consistency propriety of SHAP explanation, this work also verifies the *Efficiency*, *Symmetry*, *Dummy* and *Additivity* proprieties of Shapley values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

100% of the anomalous data were successfully detected thanks to the uncertainty-based indicator. Additionally, the prognostic performance improved around 6% to 9% as well as 43% improvement in early prognostic thanks to 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 parameter (weights) of DL model is called epistemic uncertainty (EU) [27,28,29,30].

2.2. Multi-Outputs Bayesian LSTM

To enable the quantification of both uncertainties and generate explanation, a single input, multi outputs probabilistic LSTM was developed. The model consists of an input layer, 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 layer, 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. For a simplified schematic of the whole model, refer to Structure 1 in **Figure 3**.

For anomaly detection, the model was fed with only healthy data while for prognostic, both healthy and failure inputs were involved.

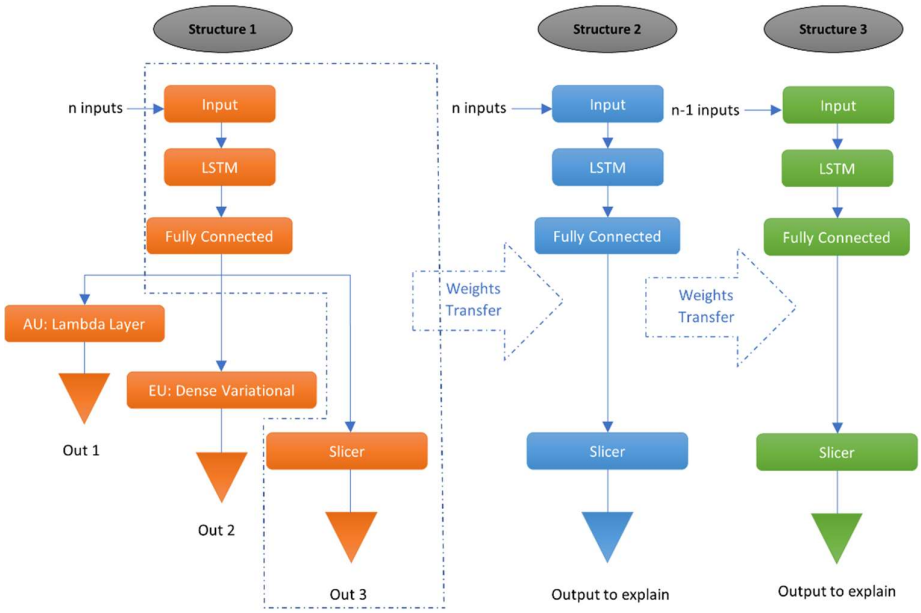


Figure 3. Different LSTM structures used in this work.

2.2.1. Probabilistic Layers

The AU layer is a probabilistic layer that learns and predicts the mean and variable standard deviation from the input coming from the LSTM layer to form the prediction range, translated into uncertainty distribution [31]. Thus, every point in an RUL sequence consists of a distribution of RUL prediction. In this work, normal distribution was used to model the uncertainty as it is easily understood.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

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 ELBO (Evidence Lower BOund) objective, $\mathcal{L}$ [32].

$$\mathcal{L}(q_\theta) = -\mathbb{E}_{q_\theta(w)}\left[-log\ P\ (y|X,w) - log\frac{P(w)}{q_\theta(w)}\right] \quad (1)$$

$$\mathcal{L}(q_\theta) = -\int dw q_\theta(w) log\ P\ (y|X,w) + \int dw q_\theta(w) log\frac{q_\theta(w)}{P(w)} \quad (2)$$

With $P(w)$ the prior, the approximation distribution $q_\theta(w)\ and\ P(y|X,w)$, the likelihood function relating all inputs X, all labels y and the weights w. The weights distribution can then be sampled to produce the output for a given input.

2.2.2. Bayesian Hyperparameter Optimization (BayesOpt)

The hyperparameters for the model were obtained via Bayesian hyperparameter optimization (BayesOpt) [33]. 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 & Uncetainty Visualization

Since noise could worsen the AU, data denoising was performed by applying Singular Value Decomposition (SVD) following the method shown in [34][35].

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

2.4. CUSUM Changepoint Detection for Anomaly Detection

The uncertainty mirrors the model's confidence in predicting. Since the model was 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. CUSUM changepoint detection was applied to identify the anomaly spikes with the appropriate control limit [36].

Given a sequence $x_1, x_2, x_3, \ldots, x_n$ with mean m_x and standard deviation σ_x, the upper U_i and lower L_i cumulative process sums are:

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

$$L_i = \begin{cases} 0, & i = 1 \\ min\left(0, L_{i-1} + x_i - m_x + \frac{1}{2}n\sigma_x\right) & i > 1 \end{cases} \quad (4)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

A process deviates at the sample x_j if it obeys $U_j > c\sigma_x$ or $L_j < -c\sigma_x$ with c the control limit.

The predetermined control limit, c is defined using healthy data prediction AU. Given σ_{AE} the standard deviation of the AU. σ_{AUmax} is the maximum and σ_{AUmean} is the mean of the standard deviations of the AU, σ_{AUstd} is the standard deviation of the standard deviations of the AU, c can be calculated as:

$$c = \frac{\sigma_{AUmax} - \sigma_{AUmean}}{\sigma_{AUstd}} \tag{5}$$

2.5. SHapley Additive exPlanations

SHAP is a game theoretic approach to explain the output of any machine learning model [37]. It evaluates the contribution of each feature to the prediction by using Shapley values. SHAP can be both global and local explainability approach. Shapley values determine the importance of a single feature by considering the outcome of each possible combination of features. In other word, the Shapley value is the average expected marginal contribution of a feature across all possible combination of features.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \tag{6}$$

Given g the explanation model. $z' \in \{0,1\}^M$ are the simplified features that describe the presence of interested feature in the feature's combination with $z' = 0$ means the interested feature are absent in the combination and $z' = 1$ signifying the feature are present. M is the maximum coalition size and $\phi_j \in R$ is the Shapley values for a feature j. The formula for Shapley value is:

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\dots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val(S \cup \{x_j\}) - val(S)\right) \tag{7}$$

S 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. $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S. $E_X(\hat{f}(X))$ is the average predicted value.

$$val_x(S) = \int \hat{f}(x_1,\dots,x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X)) \tag{8}$$

However, SHAP only accepts non probabilistic model. Thus, for generating explanation, another LSTM model was used, whose structure and weights resemble the input and the third output layers of the original model as depicted in Structure 2 in **Figure 3**.

SHAP force plot and waterfall plot, were used to explain the instance prediction while SHAP summary plot, a global visualization, explains by identifying the most contributing features in a sequence. In 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 summary plot, features are ordered according its absolute Shapley value. Those with important values occupy the top positions than less important features. The force plot was used to explain anomaly instances while the summary plot was exploited to explain and improve the prognostic performance. The waterfall plot, on the other hand,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

was employed to verify the consistency nature of the explanation as described later in Section 2.7.2.

2.6. Performance Evaluation

2.6.1. Model Predictive Performance

The average RMSE for 100 predictions was calculated between the predicted RUL (mean of RUL distribution) and the ground truth RUL [38,39].

$$RMSE = \left(\sqrt{\frac{1}{M}\sum_{i=1}^{M}(RUL_{tru}^{(i)} - Mean_{pred}^{(i)})^2}\right)\Big/_{100} \quad (9)$$

With $RUL_{trut}^{(i)}$ as the ground truth RUL for gas turbine i, $Mean_{pred}^{(i)}$ as the predicted RUL for gas turbine i and M as the total number of gas turbine.

2.6.2. Early Prediction Score

This metric was only applied in prognostic task. The scoring function, s, gives higher score for the same error in early prediction than late prediction. It penalizes late prediction than the early ones as the latter is more important that the former in any failure related forecasting problem [40,41]. The average score for 100 predictions was calculated.

$$s = \left(M\sum_{i=1}^{M} s_i\right)\Big/_{100} \quad (10)$$

$$s_i = \begin{cases} e^{\frac{-d_i}{13}} - 1, d_i < 0 \\ e^{\frac{di}{10}} - 1, d_i > 0 \end{cases} \quad (11)$$

$$di = \left(Mean_{pred}^{(i)} - RUL_{truth}\right) \quad (12)$$

2.7. Explanation Metrics

This subsection introduces the metrics for evaluating SHAP explanation [42].

2.7.1. Local Accuracy

This propriety states that the feature contributions must add up to the difference of prediction for x and the average. Starting from a normal SHAP notation:

$$f(x) = \Phi_0 + \sum_{j=1}^{M} \Phi_j x_j' \quad (13)$$

By posing $\Phi_0 = E_x(\hat{f}(X))$ and setting $x' = 1$, the Shapley Value *efficiency* propriety is found.

$$f(x) = \Phi_0 + \sum_{j=1}^{M} \Phi_j x_j' = E_x(\hat{f}(X)) + \sum_{j=1}^{M} \Phi_j \quad (14)$$

$$\sum_{j=1}^{M} \Phi_j = f(x) - E_x(\hat{f}(X)) \quad (15)$$

Where $f(x)$ is the prediction for x and $E_x(\hat{f}(X))$ is the average prediction.

2.7.2. Consistency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

This propriety 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. With $z'_{\setminus j} \Leftrightarrow z' = 0$. For any two models f and f', if:

$$f'_x(z') - f'_x(z'_{\setminus j}) \geq f_x(z') - f_x(z'_{\setminus j}) \tag{16}$$

for $z' \in \{0,1\}^M$, then:

$$\Phi_j(f', x) \geq \Phi_j(f, x) \tag{17}$$

$f_x(z')$ is the model with Structure 2 in **Figure 3** while $f'_x(z')$ is the same model but with different weights. $f_x(z'_{\setminus j})$ and $f'_x(z'_{\setminus j})$ are then 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.

To examine this propriety, the output of $f_x(z'), f_x(z'_{\setminus j}), f'_x(z'), f'_x(z'_{\setminus j})\ \Phi_j(f, x)$ and $\Phi_j(f', x)$ were extracted from the waterfall plot. **Eq. (16)** can then be calculated to verify **Eq. (17)**.

By validating this metric, the explanation also conforms to the *Symmetry, Dummy* and *Additivity* natures of Shapley values.

3. Results

3.1. Case Study 1: Anomaly Detection on Real 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 [43]. 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 [43], only several variables are useful for the DL model. The inputs-outputs are shown in **Table 2**. All the inputs were used to predict each of the output as depicted in **Figure 5** by four models denoted as $LSTM_{P2}, LSTM_{P4}, LSTM_{T4}$ and $LSTM_{N1}$

Figure 4 depicts a schematic diagram of the gas turbine under consideration.

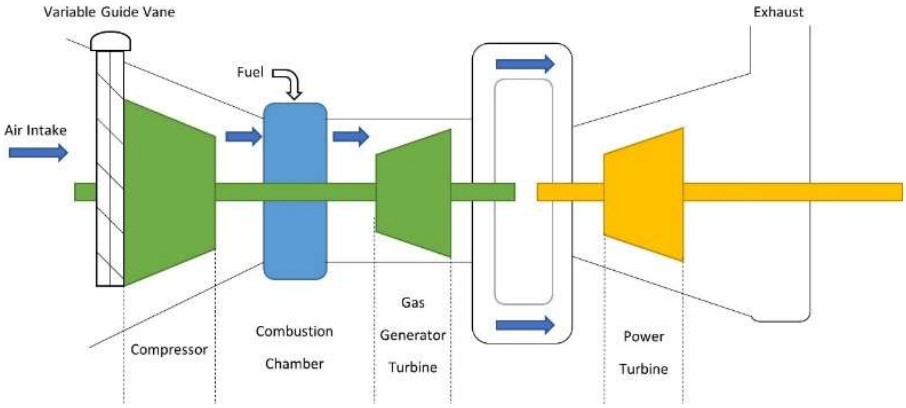


Figure 4. 18.8 MW gas turbine schematic.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

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

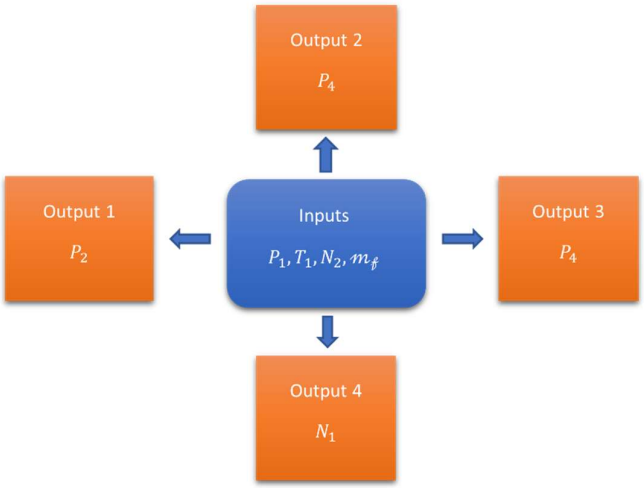


Figure 5. Real gas turbine inputs and outputs modelling.

3.1.1. Data Preparation

Anomalous data in the order of 377 hours was firstly removed from the dataset. The rest of the data was 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 01/01/18 to 26/11/18 amounting to 7488 hours or 312 sequences were used for training and validation. The data from 26/11/18 to 31/12/18 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 8th April to 9th April at 11pm to 12am (6th to 7th 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	1st Jan to 26th Nov 2018	26th Nov to 31st Dec 2018	8th April to 9th April 2018 11pm to 12am 6th to 7th instances
Total Hours	7488	816	24
Total Sequence	312	34	1

3.1.1. Healthy Data Modelling Performance

The average RMSE results for healthy testing data are presented in **Table 4.** [43] however do not specify any numeral results for the performance, thus no comparison can be done

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

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

3.1.2. Prediction with Null Anomaly

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

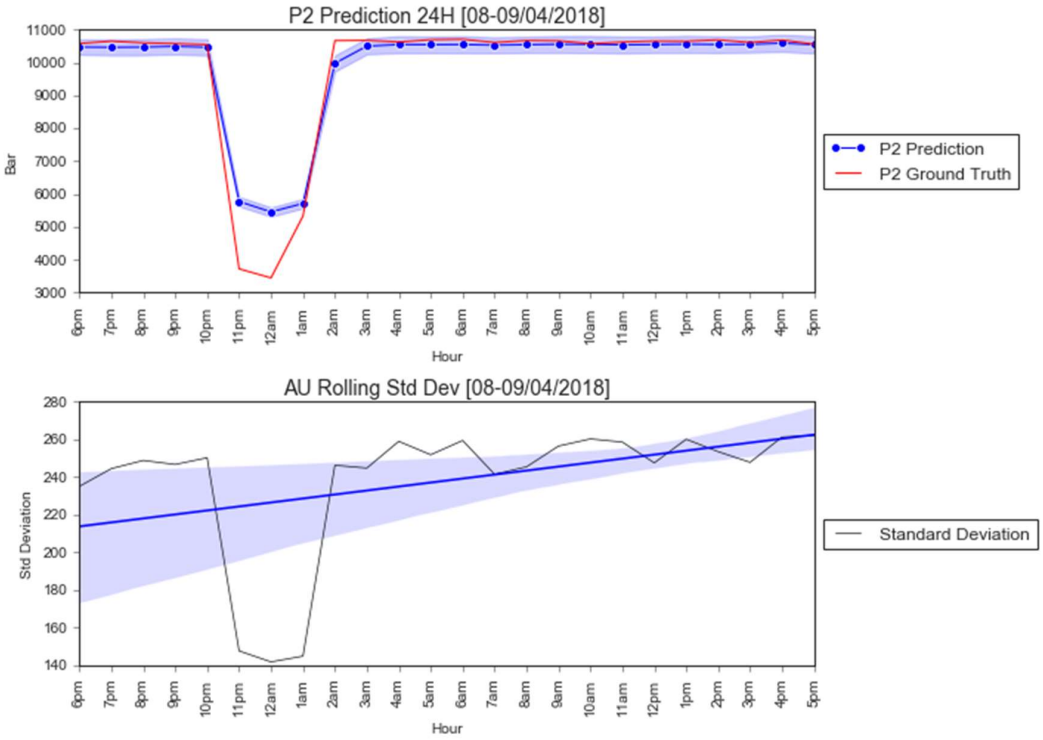


Figure 6. $LSTM_{P2}$ prediction containing null anomalous data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

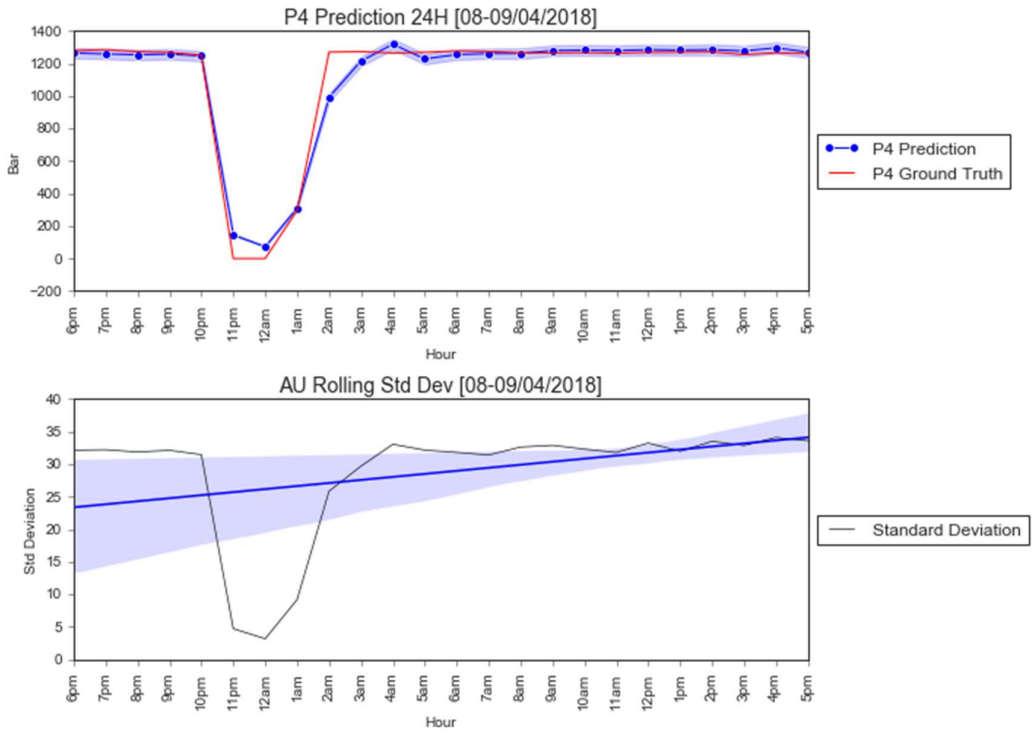


Figure 7. $LSTM_{P4}$ prediction containing null anomalous data.

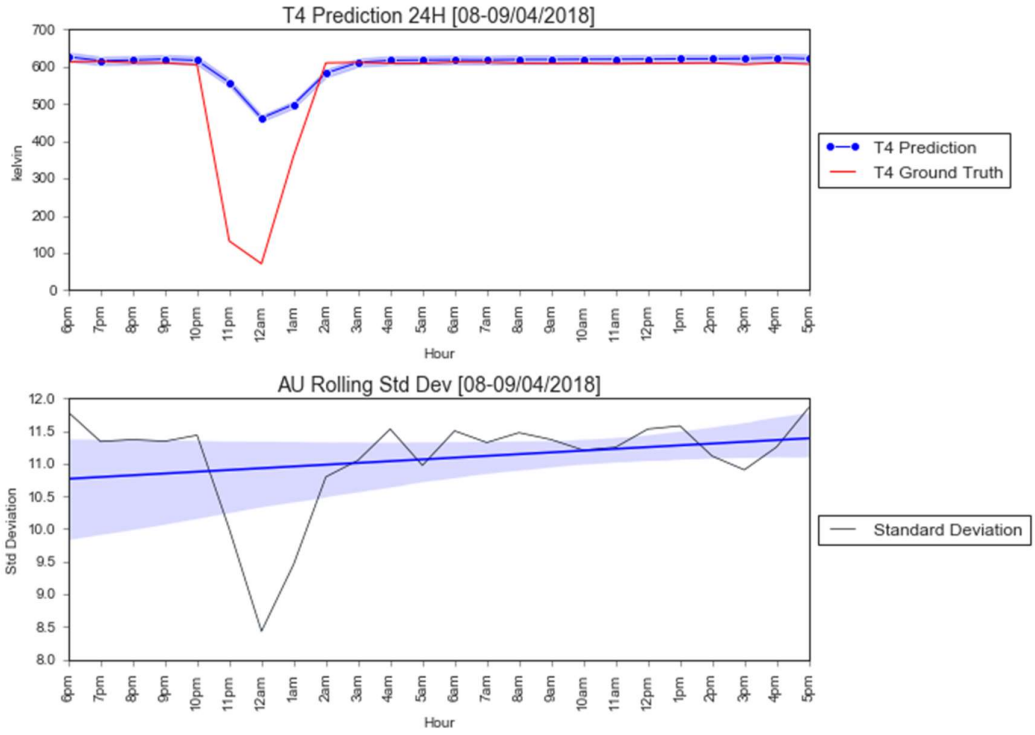


Figure 8. $LSTM_{T4}$ prediction containing null anomalous data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

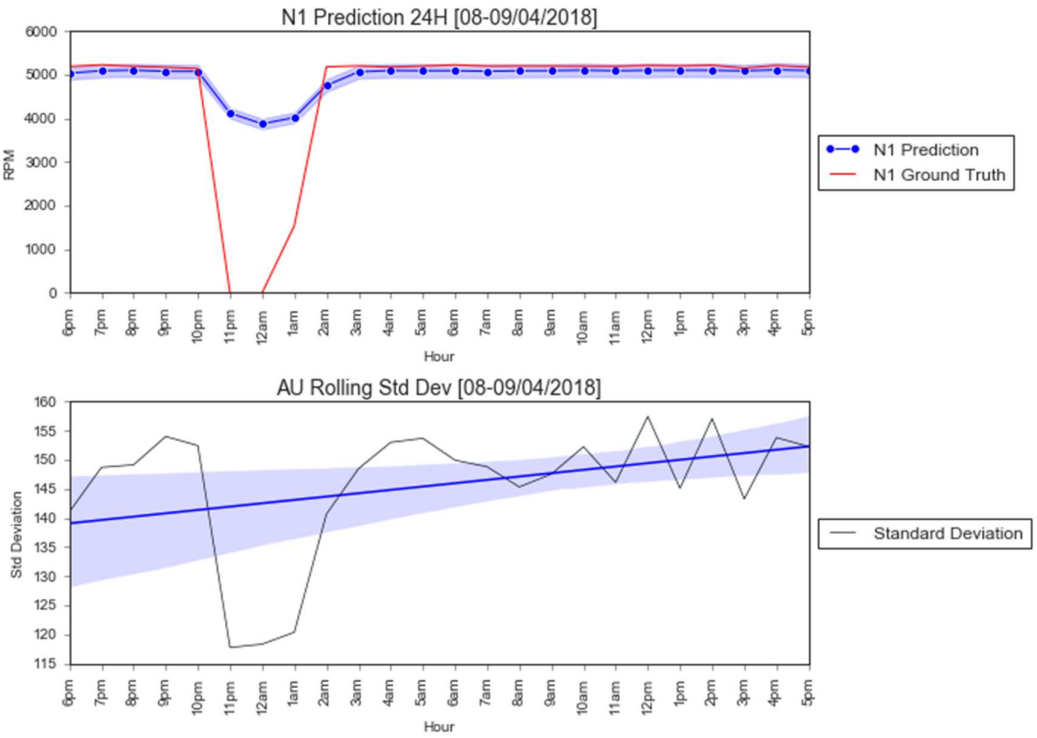


Figure 9. $LSTM_{N1}$ prediction containing null anomalous data.

3.1.3. Control limit c calculation

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

Table 5. Control limit c variables.

Model	σ_{AEmax}	σ_{AEmean}	σ_{AEstd}	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

3.1.4. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

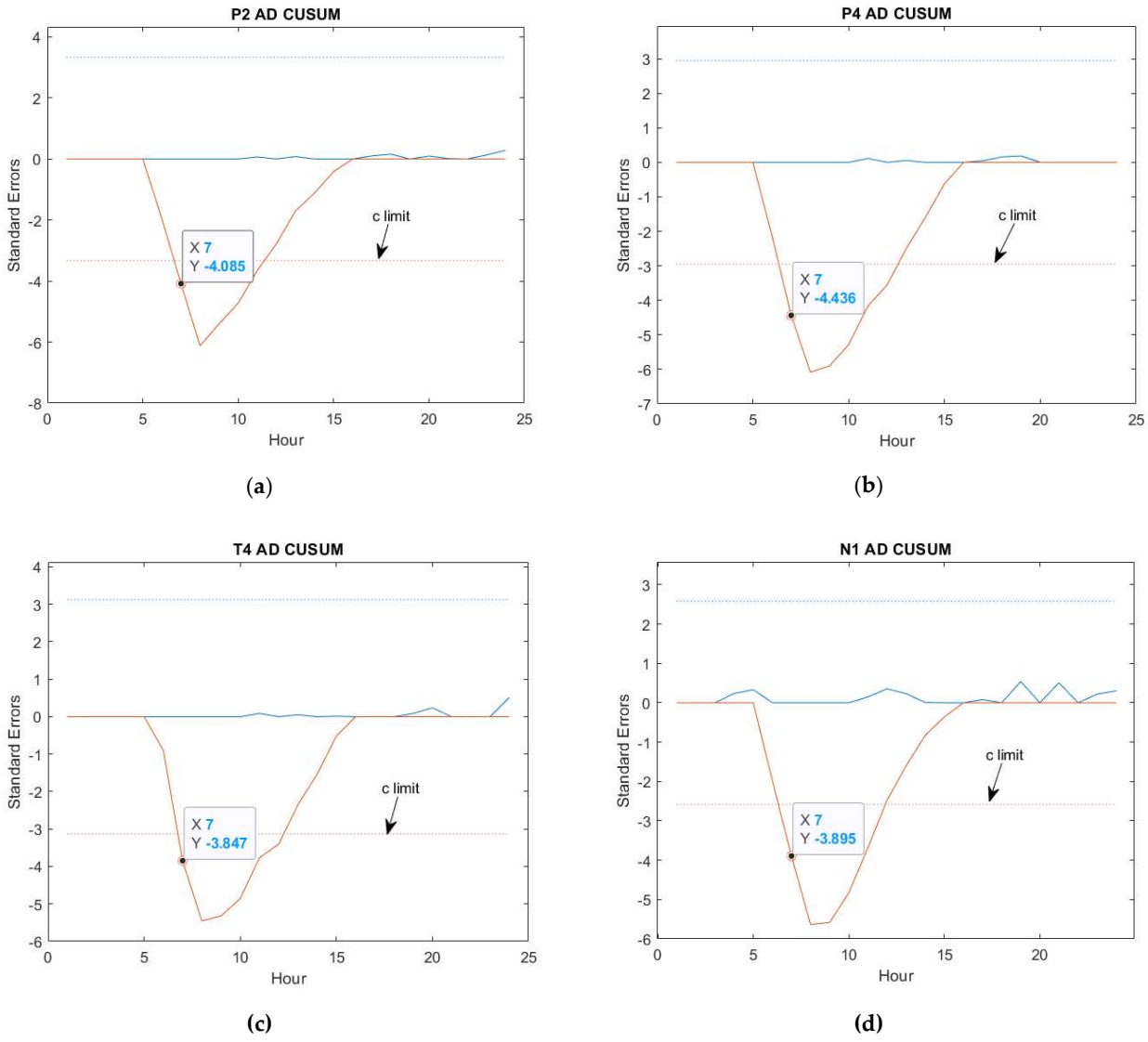
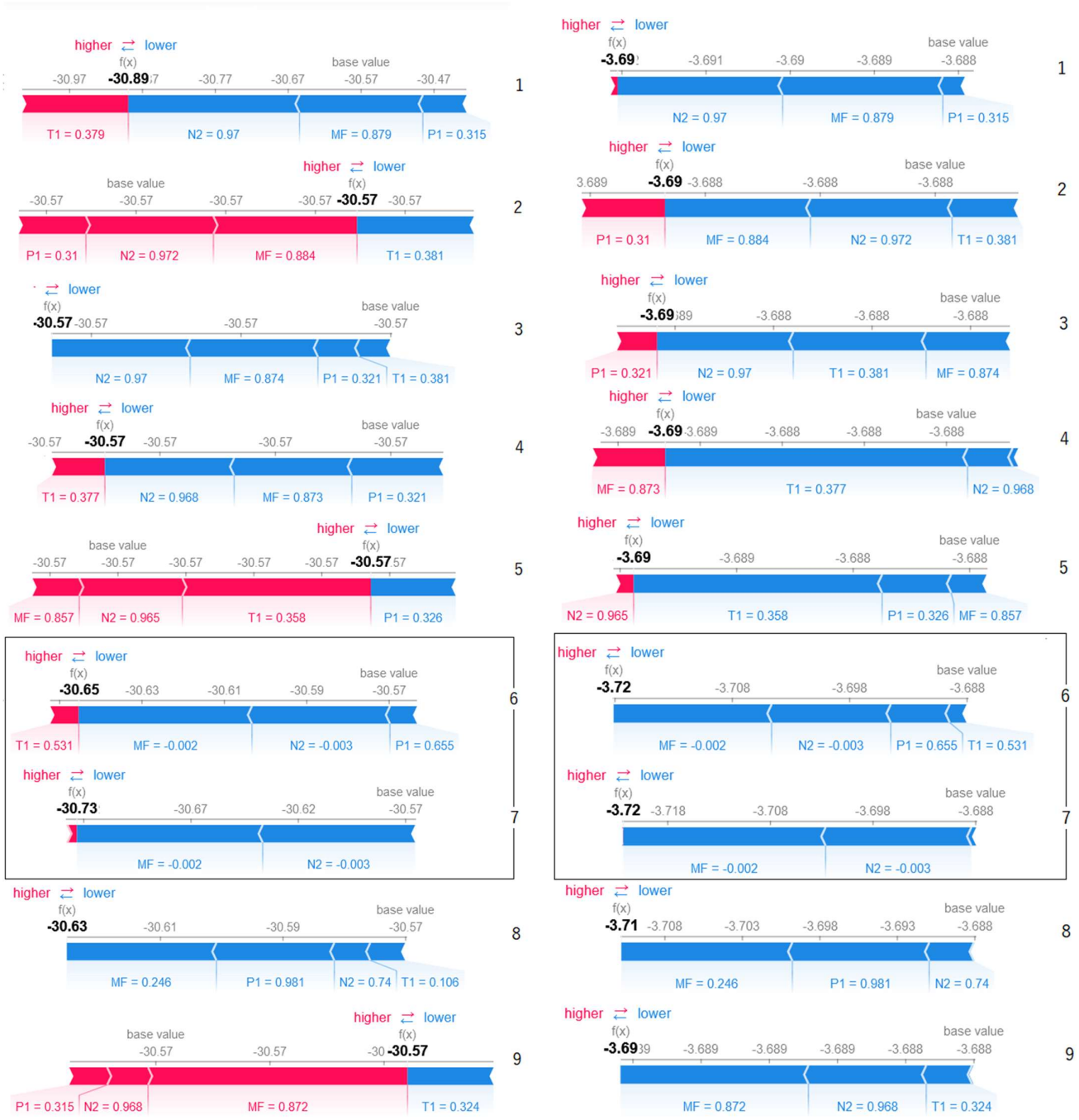


Figure 10. CUSUM chart for anomaly detection in 24h sequence (**a**) Anomaly detected in $LSTM_{P2}$ prediction; (**b**) Anomaly detected in $LSTM_{P4}$ prediction; (**c**) Anomaly detected in $LSTM_{T4}$ prediction; (**d**) Anomaly not detected in $LSTM_{N1}$ prediction.

3.1.5. Anomaly Sequence Force Plot Visualization

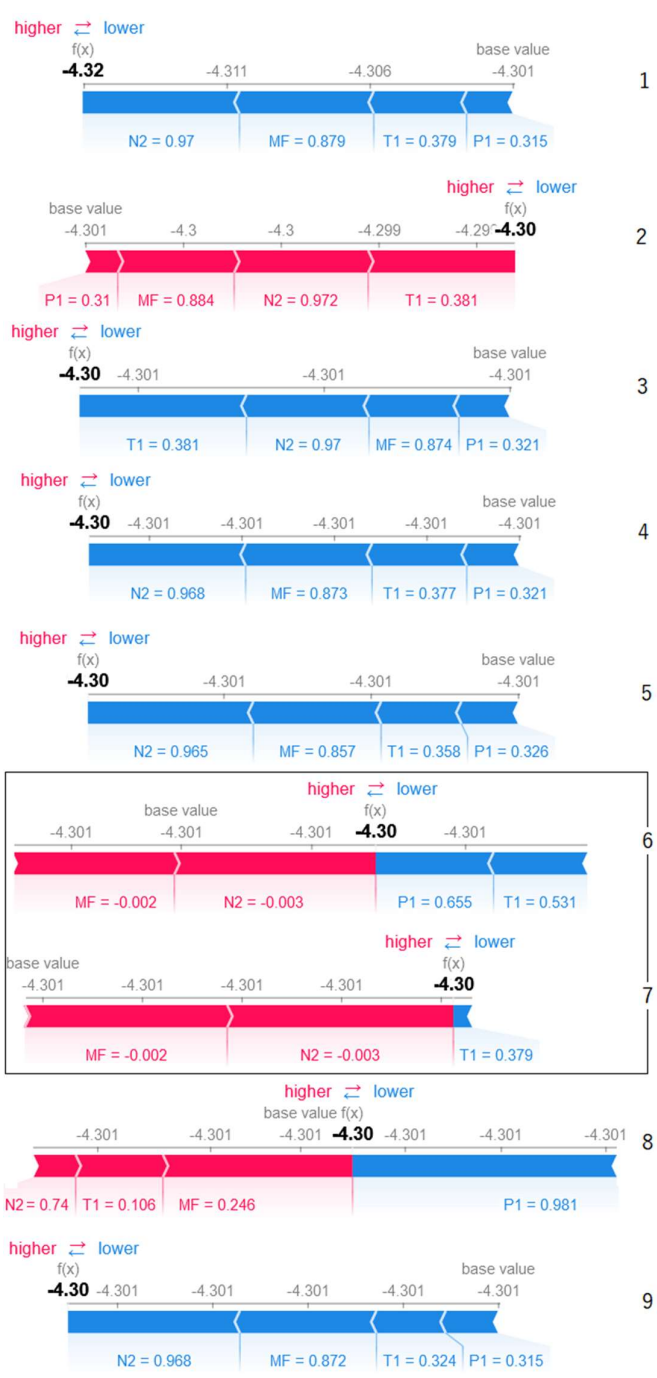
The SHAP force plot for explaining the anomalies are shown in **Figure 11.** The marked areas corresponding to the 6th and 7th instances are the tested anomaly instances. For illustration purpose, only instance 1 to 9 are shown.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1



(a) (b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1



(c) **(d)**

Figure 11. SHAP force plots for anomaly instances (**a**) Force Plot null anomaly instances in $LSTM_{P2}$ prediction; (**b**) Force Plot null anomaly instances in $LSTM_{P4}$ prediction; (**c**) Force Plot null anomaly instances in $LSTM_{T4}$ prediction; (**d**) Force Plot null anomaly instances in $LSTM_{N1}$ prediction.

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

CMAPPS (Commercial Modular Aero Propulsion System Simulation) Turbofan run to-failure datasets were published in 2008 by Nasa Prognostic Centre (PCoE) of Ames Research Centre consisting of 4 complete sets of training, testing, and ground truth RUL for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

numerous turbofan engines. The simulated data was obtained by simulating a variety of operational conditions and injecting faults of varying degradation degree.

The chosen FD001 training and testing datasets consist 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 [44]. 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 (Cycle), 3 Operating Conditions (OC) and 21 sensor signals as presented in **Appendix A**. The RUL targets for the training dataset are not available, only the ground truth RUL are given. The OC refers to different operating regimes combination of Altitude (O-42K ft.), Throttle Resolver Angle (20-100), and Mach Number (0-0.84) [44]. High level noise is incorporated, and the faults encountered are hidden by the effect of various operational conditions [45].

Table 6. Turbofan datasets

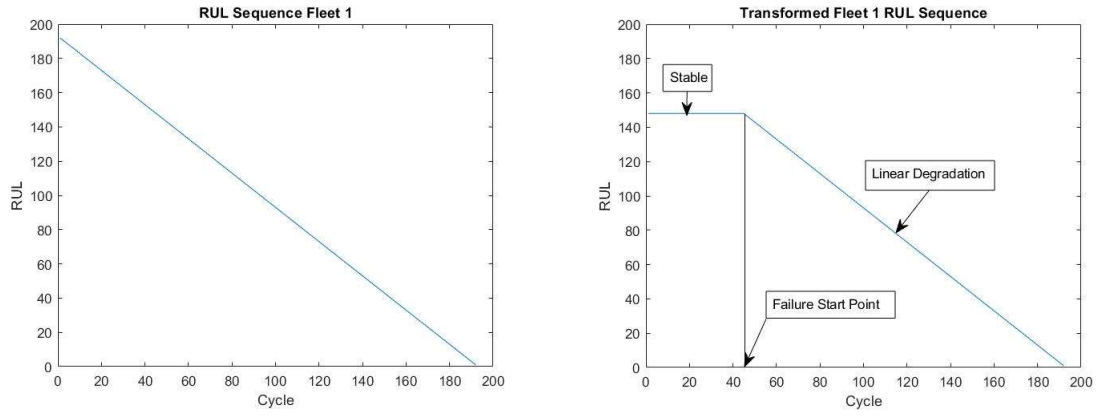
Dataset	Fault Modes	Operating Condition	Train Units	Test Units
#1	1	1	100	100

3.2.1. Data Preparation

Only strictly monotonic sensors were selected [45]. These sensors are useful as they best represent trending degradation contrary to irregular and unchanged signals. 14 sensor signals, corresponding to sensors 2,3,4,7,8,9,11,12,13,14,15,17,20 and 21 were used. Together with the three OC's, the total features used was 17.

To obtain the RUL labels for training, piece-wise linear degradation was assumed [46,47]. Each fleet's 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 took the value of the recorded signal's 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 was identified using CUSUM with the control limit c set to 5 standard deviations. Then, the mean of these failure start points was 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's 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**.



(**a**) (**b**)

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

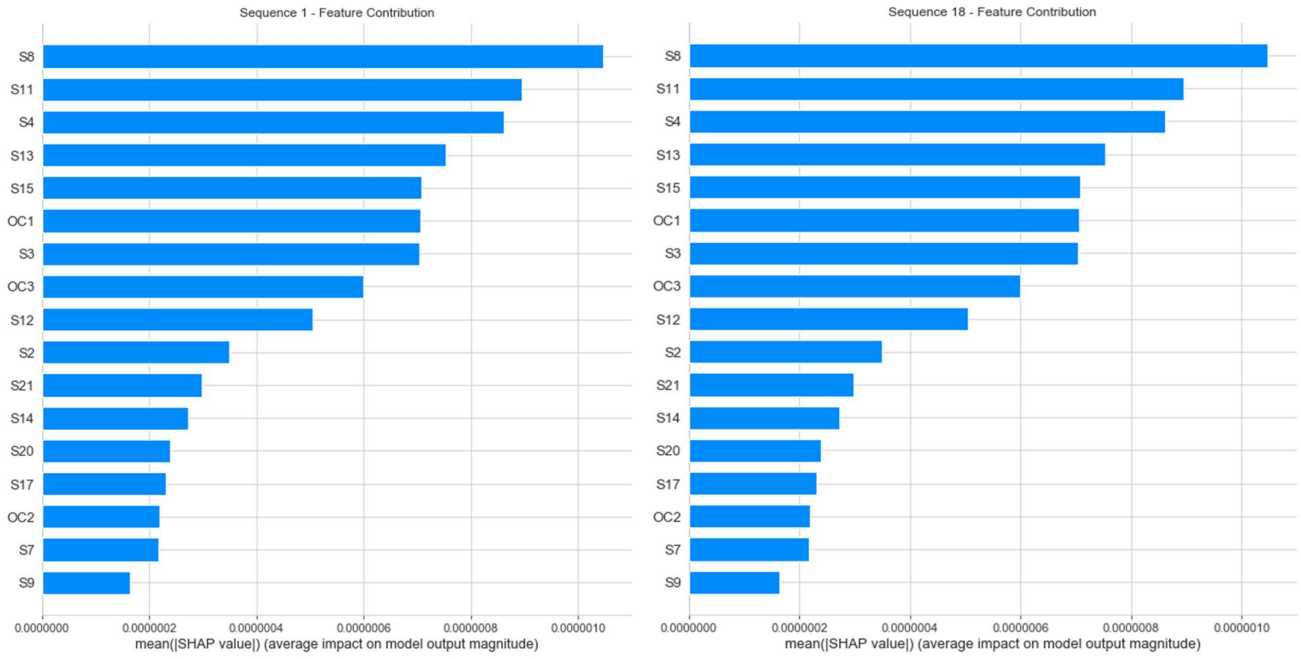
3.2.2. Prognostic Performance

The RUL prediction for Fleet 1 and Fleet 18 using the 17 features are illustrated respectively in **Figure 14** and **Figure 15.** These fleets were chosen because the former fleet's testing data length and ground truth RUL follow the same trait as the training data while the latter fleet is not, as indicated in **Appendix B.** It is thus interesting to examine the uncertainty behavior between the two.

The SHAP summary plot for the prediction of these fleets are depicted in **Figure 13**. As a matter of fact, almost all summary plots for the 100 testing fleets show the same order of features as **Figure 13.** One can thus choose the best set of features to improve the predictive performance. Accordingly, the model was 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



(**a**) (**b**)

Figure 13. SHAP summary plot (**a**) Summary plot for Fleet 1 sequence prediction; (**b**) Summary plot for Fleet 18 sequence prediction.

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, the model performed better in mostly all metrics with 13 features. It shows around 9% improvement in RMSE with AU

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

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

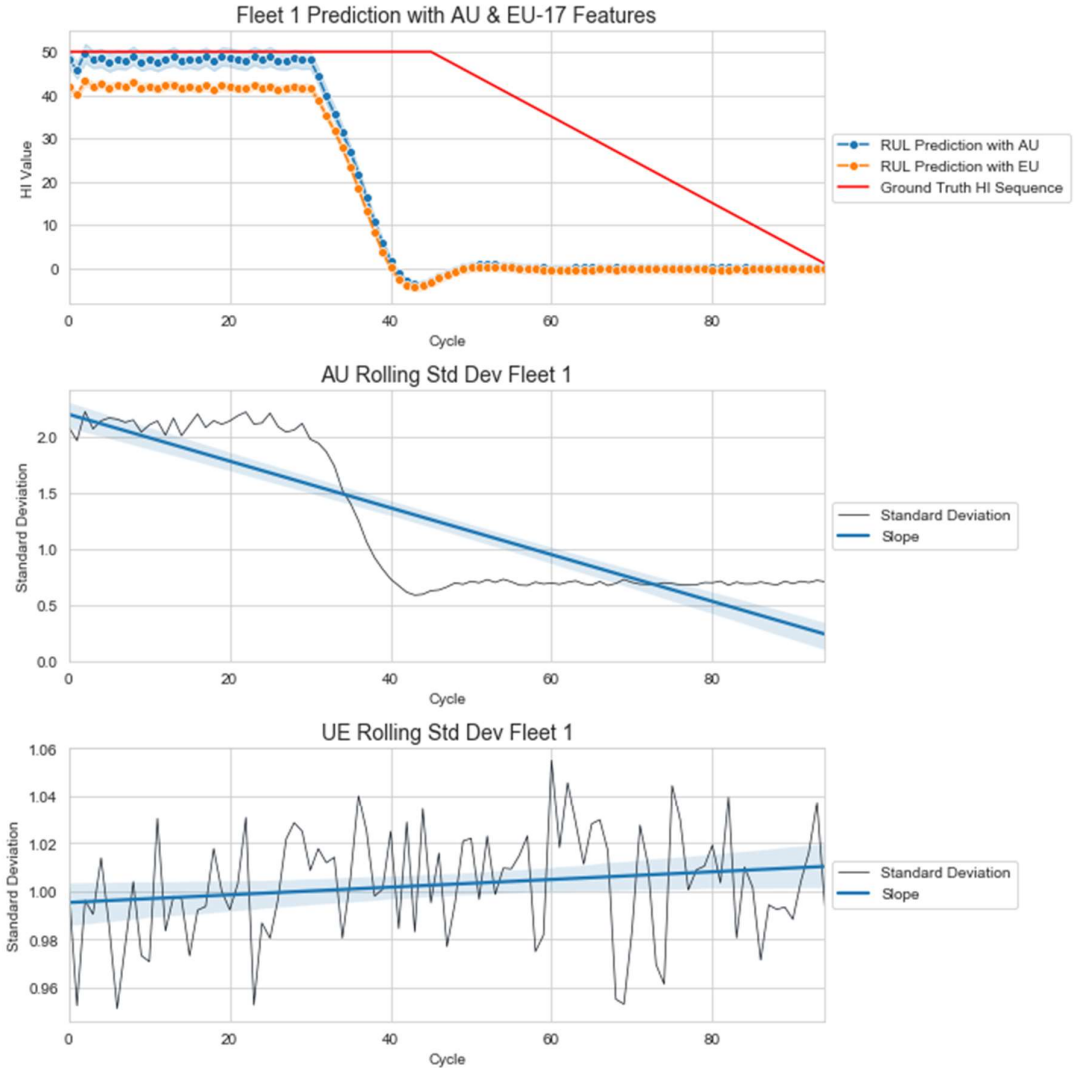


Figure 14. Fleet 1 RUL prediction with AU and EU.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

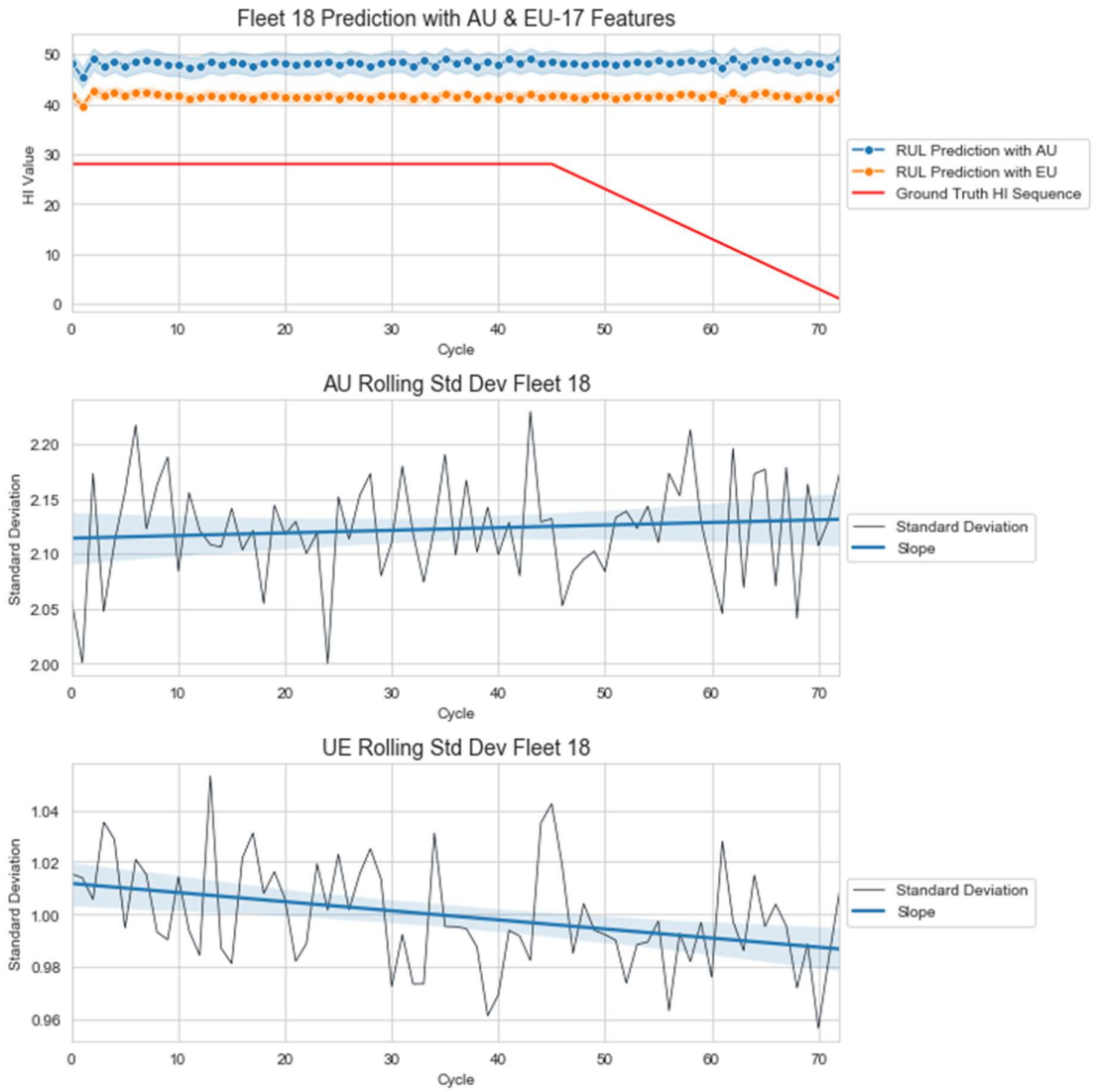


Figure 15. Fleet 18 RUL prediction with AU and EU.

3.2.2. Performance Comparison with Published Methods

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

Table 9. RMSE comparison with published methods

Results	Proposed Method	DBN [48]	ELM [48]	RNN [49]	DCNN [49]	BiLSTM [50]
RMSE with AU	14.68	15.21	17.27	13.44	12.61	13.65

Table 10. Early score comparison with published methods

Results	Proposed Method	DBN [48]	ELM [48]	RNN [49]	DCNN [49]	BiLSTM [50]
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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

3.3. Explanation Evaluation

3.3.1. 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, it can be verified 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 $m_{\oint}$.

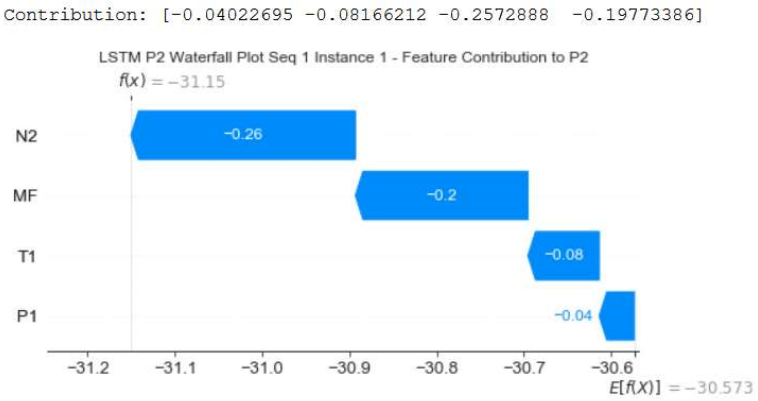


Figure 16. Waterfall plot of the first instance on the first sequence of $LSTM_{P2}$.

3.3.2. Consistency Verification

The contribution of variable $m_{\oint}$ on the first test data instance was investigated as an example. For each output, the difference and contribution of $m_{\oint}$, Φ_{MF}, was calculated:

$$P2\_tot'_1 = LSTM_{P2-2}(z'_1) - LSTM_{P2-2}(z'_{1_{\setminus MF}}) = -3.273;\ \Phi_{MF\_P2\_1}(f', x) = -0.050$$

$$P2\_tot_1 = LSTM_{P2}(z'_1) - LSTM_{P2}(z'_{1_{\setminus MF}}) = -14.893;\ \Phi_{MF\_P2\_1}(f', x) = -0.198$$

$$P4\_tot'_1 = LSTM_{P4-2}(z'_1) - LSTM_{P4-2}(z'_{1_{\setminus MF}}) = -0.962;\ \Phi_{MF\_P4\_1}(f', x) = -0.002$$

$$P4\_tot_1 = LSTM_{P4}(z'_1) - LSTM_{P4}(z'_{1_{\setminus MF}}) = -1.182;\ \Phi_{MF\_P4\_1}(f', x) = -0.048$$

$$T4\_tot'_1 = LSTM_{T4-2}(z'_1) - LSTM_{T4-2}(z'_{1_{\setminus MF}}) = 0.219;\ \Phi_{MF\_T4\_1}(f', x) = 0.002$$

$$T4\_tot_1 = LSTM_{T4}(z'_1) - LSTM_{T4}(z'_{1_{\setminus MF}}) = -1.004;\ \Phi_{MF\_T4\_1}(f', x) = -0.006$$

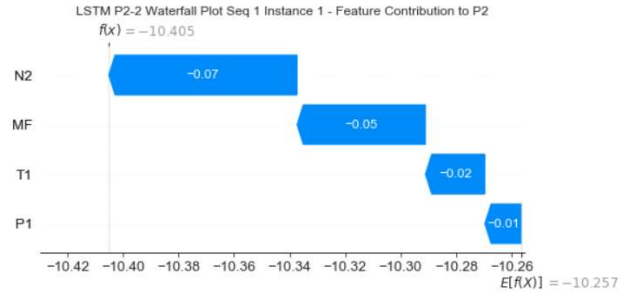
$$N1\_tot'_1 = LSTM_{N1-2}(z'_1) - LSTM_{N1-2}(z'_{1_{\setminus MF}}) = 9.017;\ \Phi_{MF\_N1\_1}(f', x) = 0.749$$

$$N1\_tot_1 = LSTM_{N1-2}(z'_1) - LSTM_{N1-2}(z'_{1_{\setminus MF}}) = -3.030;\ \Phi_{MF\_N1\_1}(f', x) = -0.075$$

These equal: $P2\_tot'_1 > P2\_tot_1$; $P4\_tot'_1 > P4\_tot_1$; $T4\_tot'_1 > T4\_tot_1$ and $N1\_tot'_1 > N1\_tot_1$, thus $\Phi_{\text{MF\_P2\_1}'} > \Phi_{\text{MF\_P2\_1}}$; $\Phi_{\text{MF\_P4\_1}'} > \Phi_{\text{MF\_P4\_1}}$; $\Phi_{\text{MF\_T4\_1}'} > \Phi_{\text{MF\_T4\_1}}$ and $\Phi_{\text{MF\_N1\_1}'} > \Phi_{\text{MF\_N1\_1}}$ as seen above. These results are illustrated in the waterfall plots in **Figure 17.**

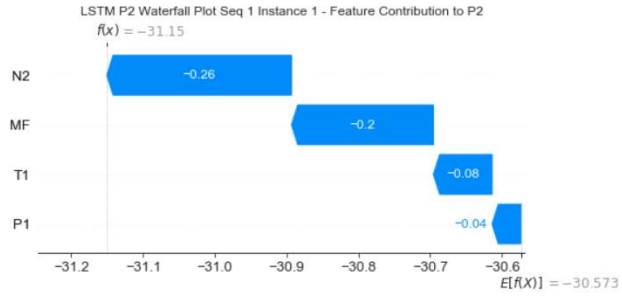
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

Contribution: [-0.01300738 -0.02142382 -0.06788444 -0.04628381]



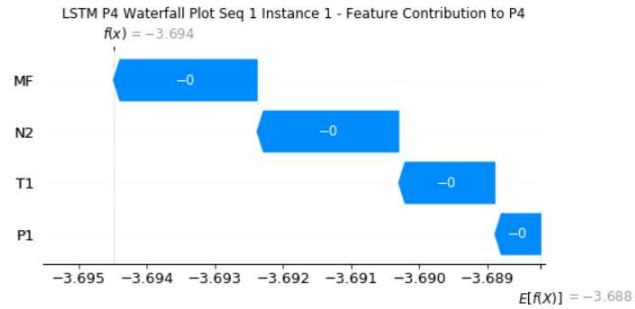
(a)

Contribution: [-0.04022695 -0.08166212 -0.2572888 -0.19773386]



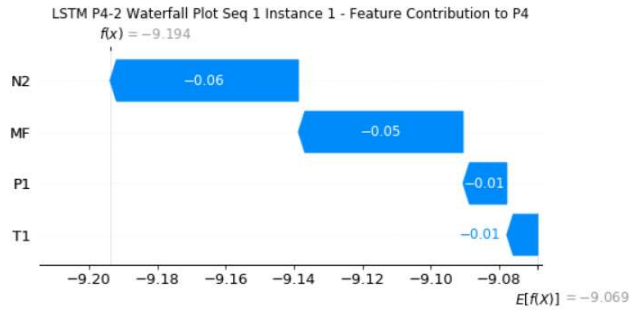
(b)

Contribution: [-0.00067266 -0.00141481 -0.00207995 -0.00211023]



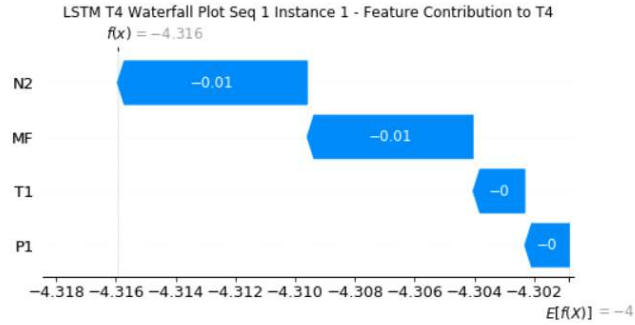
(c)

Contribution: [-0.01281367 -0.00904393 -0.05510879 -0.04815559]



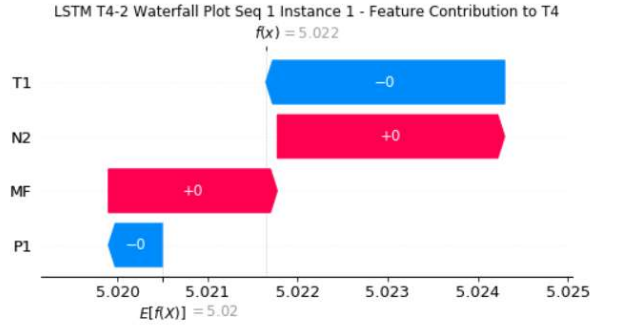
(d)

Contribution: [-0.00148626 -0.00173069 -0.00635258 -0.00555439]



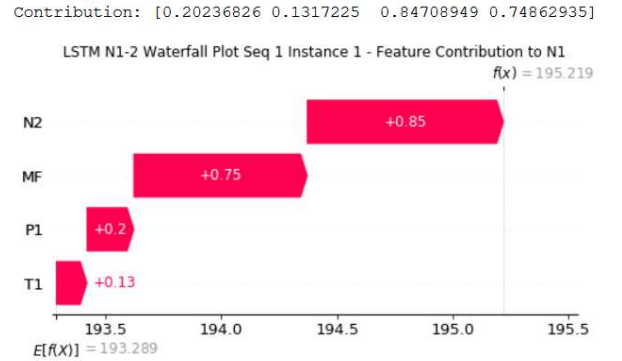
(e)

Contribution: [-0.00059942 -0.00265166 0.00252425 0.0018752]



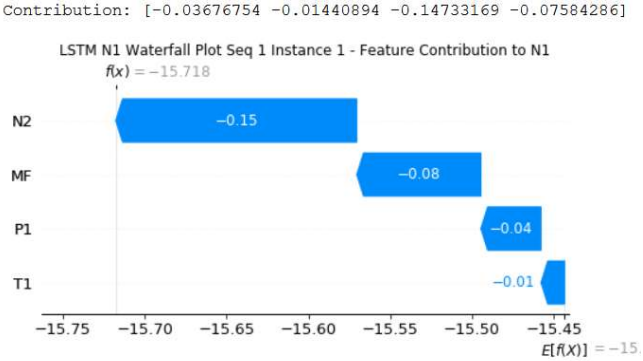
(f)

Contribution: [0.20236826 0.1317225 0.84708949 0.74862935]



(g)

Contribution: [-0.03676754 -0.01440894 -0.14733169 -0.07584286]



(h)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

Figure 17. Waterfall plot for local feature contibution: (**a**) Waterfall plot $\Phi_{MF\_P2\_1}(f', x)$; (**b**) Waterfall plot $\Phi_{MF\_P2\_1}(f', x)$; (**c**) Waterfall plot $\Phi_{MF\_P4\_1}(f', x)$; (**d**) Waterfall plot $\Phi_{MF\_P4\_1}(f', x)$; (**e**) Waterfall plot $\Phi_{MF\_T4\_1}(f', x)$; (**f**) Waterfall plot $\Phi_{MF\_T4\_1}(f', x)$; (**g**) Waterfall plot $\Phi_{MF\_N1\_1}(f', x)$; (**h**) Waterfall plot $\Phi_{MF\_N1\_1}(f', x)$.

4. Discussion

4.1. Explainable Anomaly Detection

100% of the tested null anomalies were successfully detected with the help of 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 as 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 7th instance, all other feature forces were eclipsed, showing mostly m_f and N2. However, on the 8th 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 lower the c limit value, resulting to a faster detection. However, 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. As can be seen, 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 ground truth RUL. The AU for Fleet 1 prediction, however, indicates a decreasing trend as presented in **Figure 14**, mirroring the good prediction the model had made. The model thus becoming more and more confident of its sequential estimation. Meanwhile, the EU measure, manifest very small change in nearly the same scale for both fleets which is expected for the weight's uncertainty. This uncertainty should not be influenced much by the change in input data.

The summary plot global explanation ordered the features according to its 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 HPC outlet, total temperature at LPT outlet, corrected fan speed and bypass ratio. The model's 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 than 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 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 in [49] for example, has respectively five convolutional and five recurrent layers to learn the data while the BiLSTM in [50] 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's 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 SHAP explanation satisfies the Local Accuracy and Consistency criteria. By fulfilling these proprieties, the explanation also conforms to the *Efficiency*, *Symmetry*, *Dummy* and *Additivity* natures of 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 values' 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. Finally, the *Additivity* denotes that for an ensemble prediction, for a specific feature, one 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

This article elaborates the application of SHAP model agnostic approach in explaining the outputs of a Bayesian LSTM in anomaly detection and prognostic tasks of gas turbines using real and simulated datasets. The forecast uncertainty, generated by the Bayesian model, broaden the explanation scope to include model's confidence, strengthening the explanation. It was also exploited as anomaly indicator. SHAP global explanation was 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's RMSE increased around 6% to 9% while its early prediction ability showed 43% improvement thanks to SHAP. These results are comparable to the best published methods in the problem. Finally, the generated explanation verifies the Local Accuracy and Consistency proprieties, and by doing so validates the *Efficiency*, *Symmetry*, *Dummy* and *Additivity* natures of Shapley values. This paper shows how SHAP and deep learning uncertainty form a broader explanation scope while simultaneously demonstrating SHAP ability in enhancing PHM performance, highlighting its potential as an easy to use, flexible and powerful XAI technique.

Supplementary Materials: The following are available online at https://github.com/Kamalnor/Real-Gas-Turbine, Python codes and prediction visualization for $LSTM_{P2}$, $LSTM_{P4}$, $LSTM_{T4}$, & $LSTM_{N1}$ $LSTM_{P2-2}$, $LSTM_{P4-2}$, $LSTM_{T4-2}$, & $LSTM_{N1-2}$ for Consistency metric, as well as MATLAB files for optimized hyperparameters and anomaly detection. The following are available online at https://github.com/Kamalnor/Turbofan-FD001, Python codes and prediction visualization for turbofan prognostic with 17, 13 and 9 features as well as MATLAB files for optimized hyperparameters.

Author Contributions: Conceptualization, Ahmad Kamal Mohd Nor; methodology Ahmad Kamal Mohd Nor; software, Ahmad Kamal Mohd Nor; validation, Ahmad Kamal Mohd Nor; formal analysis, Ahmad Kamal Mohd Nor; investigation, Ahmad Kamal Mohd Nor; resources, Ahmad Kamal Mohd Nor; data curation, Ahmad Kamal Mohd Nor; writing—original draft preparation, Ahmad Kamal Mohd Nor; writing—review and editing, Ahmad Kamal Mohd Nor; visualization, Ahmad Kamal Mohd Nor; supervision, Srinivasa Rao Pedapati & Masdi Muhammad.

Funding: This research was funded by Universiti Teknologi Petronas Foundation (YUTP) and the APC was partially funded by the Centre of Graduates Studies, Universiti Teknologi Petronas.

Data Availability Statement: The data presented in this study are openly available in https://github.com/Kamalnor?tab=repositories.

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

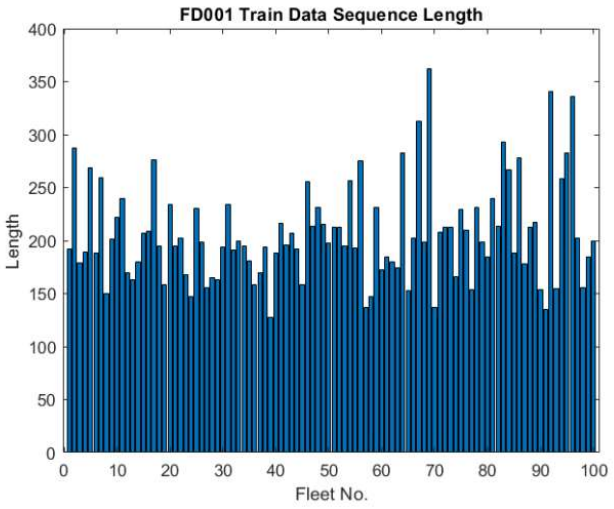
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 September 2021 doi:10.20944/preprints202109.0034.v1

Appendix A

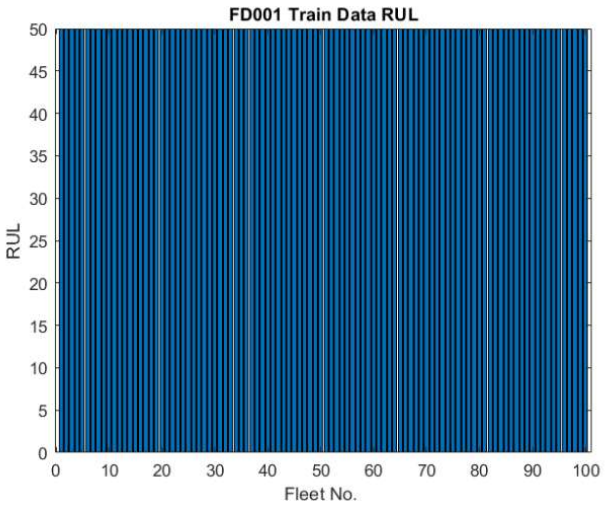
Table 11. Turbofan Sensors' Description.

Sensor	References	Description	Unit
S1	T_2	Total temperature fan inlet	^{0}R
S2	T_{24}	Total temperature at LPC outlet	^{0}R
S3	T_{30}	Total temperature at HPC outlet	^{0}R
S4	T_{50}	Total temperature at LPT outlet	^{0}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	Nf	Physical fan speed	RPM
S9	Nc	Physical core speed	RPM
S10	Epr	Engine pressure ratio (P50/P2)	N/A
S11	Ps_{30}	Static pressure at HPC outlet	psia
S12	Phi	Ratio of fuel flow to Ps30	Pps/psi
S13	NRf	Corrected fan speed	RPM
S14	NRc	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	Nf_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

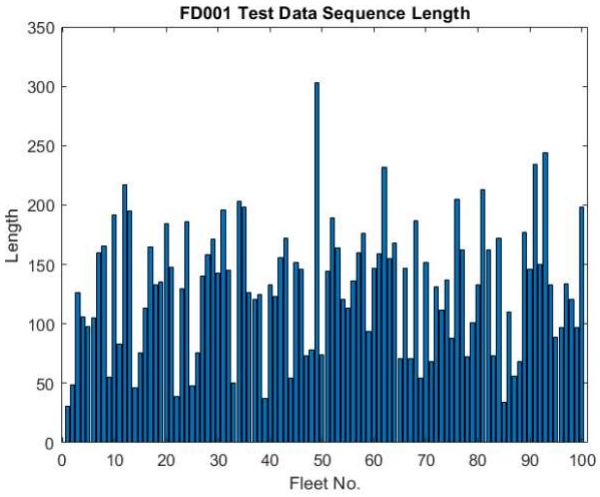
Appendix B



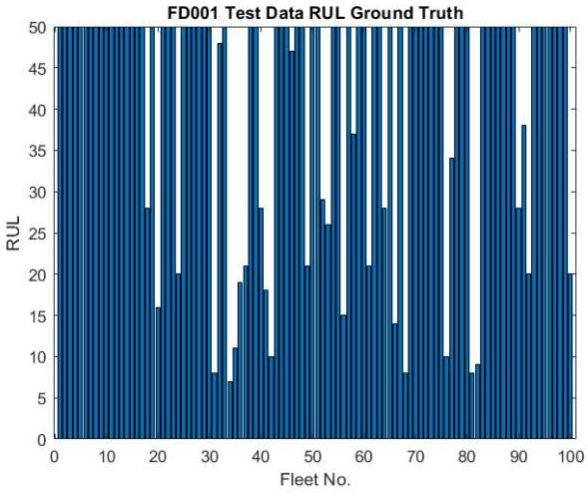
(**a**)



(**b**)



(c)



(d)

Figure 18. Data sequence length & associated RUL (**a**) Training data sequence length; (**b**) Training data RUL; (**c**) Testing data sequence length; (**d**) Ground truth RUL.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94.
2. Bistron, M.; Piotrowski, Z. Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. Electronics 2021, 10, 871. https://doi.org/10.3390/electronics10070871
3. Zhang, Jianjing & Arinez, Jorge & Chang, Qing & Gao, Robert & Xu, Chengying. (2020). Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook. Journal of Manufacturing Science and Engineering. 1-53. 10.1115/1.4047855.
4. Mahmud, M., Kaiser, M.S., McGinnity, T.M. et al. Deep Learning in Mining Biological Data. Cogn Comput 13, 1–33 (2021). https://doi.org/10.1007/s12559-020-09773-x
5. Raj, M., Seamans, R. Primer on artificial intelligence and robotics. *J Org Design* **8,** 11 (2019). https://doi.org/10.1186/s41469-019-0050-0
6. Ahmad Kamal M. Nor, Srinivasa R. Pedapati, Masdi Muhammad, Reliability engineering applications in electronic, software, nuclear and aerospace industries: A 20 year review (2000–2020), Ain Shams Engineering Journal, 2021, ISSN 2090-4479, https://doi.org/10.1016/j.asej.2021.02.015.
7. Mou, Xiaomin. 2019. Artificial Intelligence: Investment Trends and Selected Industry Uses. EMCompass, no. 71; International Finance Corporation, Washington, DC. © International Finance Corporation. https://openknowledge.worldbank.org/handle/10986/32652 License: CC BY 3.0 IGO.
8. Bughin, J., Eric Hazan, S. Ramaswamy, Michael Chui, Tera Allas, Peter Dahlstrom, Nicolaus Henke and Monica Trench. "Artificial intelligence: the next digital frontier?" (2017)
9. Gillham, Jonathan. (2017). The Economic Impact of Artificial Intelligence on the Global Economy.
10. *Jacques Bughin* et al., "*Notes from the AI Frontier*: *Modeling the Impact of AI on the World Economy*," McKinsey Global Institute, September 2018. Retrieved from: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy on 24th August 2021.
11. Margot E. Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Tech. L.J. 189 (2019), *available at* https://scholar.law.colorado.edu/articles/1227. Doi: 10.15779/Z38TD9N83H.
12. 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 on 24th August 2021.
13. Xu, Feiyu & Uszkoreit, Hans & Du, Yangzhou & Fan, Wei & Zhao, Dongyan & Zhu, Jun. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. 10.1007/978-3-030-32236-6_51.
14. Adadi, Amina & Berrada, Mohammed. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2870052.
15. J. W. Sheppard, M. A. Kaufman and T. J. Wilmer, "IEEE Standards for Prognostics and Health Management," in IEEE Aerospace and Electronic Systems Magazine, vol. 24, no. 9, pp. 34-41, Sept. 2009, doi: 10.1109/MAES.2009.5282287.
16. NOR, A. K. B. M., Rao PEDAPATI, S., and MUHAMMAD, M., "Explainable AI (XAI) for PHM of Industrial Asset: A State-of-The-Art, PRISMA-Compliant Systematic Review", arXiv e-prints, 2021.
17. Langone, Rocco & Cuzzocrea, Alfredo & Skantzos, Nikolaos. (2020). Interpretable Anomaly Prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. Data & Knowledge Engineering. 130. 101850. 10.1016/j.datak.2020.101850.

18. Dengji Zhou, Qinbo Yao, Hang Wu, Shixi Ma, Huisheng Zhang, Fault diagnosis of gas turbine based on partly interpretable convolutional neural networks, Energy, Volume 200, 2020, 117467, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2020.117467.
19. 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(24):7307. https://doi.org/10.3390/s20247307
20. Waghen, Kerelous & Ouali, Mohamed-Salah. (2019). Interpretable Logic Tree Analysis: A Data-Driven Fault Tree Methodology for Causality Analysis. Expert Systems with Applications. 136. 10.1016/j.eswa.2019.06.042.
21. S. Rajendran, W. Meert, V. Lenders and S. Pollin, "Unsupervised Wireless Spectrum Anomaly Detection With Interpretable Features," in IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 3, pp. 637-647, Sept. 2019, doi: 10.1109/TCCN.2019.2911524.
22. C. Liu, C. Qin, X. Shi, Z. Wang, G. Zhang and Y. Han, "TScatNet: An Interpretable Cross-Domain Intelligent Diagnosis Model with Antinoise and Few-Shot Learning Capability," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-10, 2021, Art no. 3506110, doi: 10.1109/TIM.2020.3041905.
23. Pacella, Massimo. (2018). Unsupervised Classification of Multichannel Profile Data using PCA: an application to an Emission Control System. Computers & Industrial Engineering. 122. 10.1016/j.cie.2018.05.029.
24. M. S. Kim, J. P. Yun and P. Park, "An Explainable Convolutional Neural Network for Fault Diagnosis in Linear Motion Guide," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2020.3012989.
25. Onchis, Darian & Gillich, Gilbert-Rainer. (2021). Stable and explainable deep learning damage prediction for prismatic cantilever steel beam. Computers in Industry. 125. 103359. 10.1016/j.compind.2020.103359.
26. J. Grezmak, J. Zhang, P. Wang, K. A. Loparo and R. X. Gao, "Interpretable Convolutional Neural Network Through Layer-wise Relevance Propagation for Machine Fault Diagnosis," in IEEE Sensors Journal, vol. 20, no. 6, pp. 3172-3181, 15 March15, 2020, doi: 10.1109/JSEN.2019.2958787.
27. W. Peng, Z. Ye and N. Chen, "Bayesian Deep-Learning-Based Health Prognostics Toward Prognostics Uncertainty," in IEEE Transactions on Industrial Electronics, vol. 67, no. 3, pp. 2283-2293, March 2020, doi: 10.1109/TIE.2019.2907440.
28. G. Li, L. Yang, C. Lee, X. Wang and M. Rong, "A Bayesian Deep Learning RUL Framework Integrating Epistemic and Aleatoric Uncertainties," in IEEE Transactions on Industrial Electronics, doi: 10.1109/TIE.2020.3009593.
29. Yu Qin, Zhiwen Liu, Chenghao Liu, Yuxing Li, Xiangzhu Zeng, Chuyang Ye, Super-Resolved q-Space deep learning with uncertainty quantification, Medical Image Analysis, Volume 67, 2021, 101885, ISSN 1361-8415, https://doi.org/10.1016/j.media.2020.101885.
30. Yidong Chai, Yiyang Bian, Hongyan Liu, Jiaxing Li, Jie Xu, Glaucoma diagnosis in the Chinese context: An uncertainty information-centric Bayesian deep learning model, Information Processing & Management, Volume 58, Issue 2, 2021, 102454, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2020.102454.
31. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). USENIX Association, USA, 265–283. Retrieved from: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Lambda on 24th August 2021.
32. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). USENIX Association, USA, 265–283. Retrieved from: https://www.tensorflow.org/probability/api_docs/python/tfp/layers/DenseVariational on 24th August 2021.
33. Wu, J. & Chen, X.-Y & Zhang, H. & Xiong, L.-D & Lei, H. & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology. 17. 26-40. 10.11989/JEST.1674-862X.80904120.
34. Epps, Brenden & Krivitzky, Eric. (2019). Singular value decomposition of noisy data: noise filtering. Experiments in Fluids. 60. 10.1007/s00348-019-2768-4.
35. Epps, Brenden & Krivitzky, Eric. (2019). Singular value decomposition of noisy data: mode corruption. Experiments in Fluids. 60. 10.1007/s00348-019-2761-y.
36. cusum, Copyright 2015-2018 The MathWorks, Inc. Retrieved from: https://www.mathworks.com/help/signal/ref/cusum.html on 24th August 2021.
37. Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.
38. Q. Wang, S. Zheng, A. Farahat, S. Serita and C. Gupta, "Remaining Useful Life Estimation Using Functional Data Analysis," 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, 2019, pp. 1-8, doi: 10.1109/ICPHM.2019.8819420.
39. Y. Ge, L. Sun and J. Ma, "An Improved PF Remaining Useful Life Prediction Method Based on Quantum Genetics and LSTM," in IEEE Access, vol. 7, pp. 160241-160247, 2019, doi: 10.1109/ACCESS.2019.2951197.
40. Y. Ge, J. Wu and X. Jiang, "A Prediction Method Using Bayesian Theory for Remaining Useful Life," 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), Zhangjiajie, China, 2019, pp. 856-862, doi: 10.1109/QR2MSE46217.2019.9021252.

41. F. Li et al., "A Light Gradient Boosting Machine for Remaining Useful Life Estimation of Aircraft Engines," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018, pp. 3562-3567, doi: 10.1109/ITSC.2018.8569801.
42. Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
43. T.B., Mohammadreza & Muhammad, Masdi & Abdul Karim, Zainal Ambri. (2017). 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. 39. 1-12. 10.1007/s40430-017-0742-8.
44. Ramasso, Emmanuel & Saxena, Abhinav. (2014). Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets. International Journal of Prognostics and Health Management. 5. 1-15.
45. Saxena, Abhinav & Goebel, Kai & Simon, Don & Eklund, Neil. (2008). Damage propagation modelling for aircraft engine run-to-failure simulation. International Conference on Prognostics and Health Management. 10.1109/PHM.2008.4711414.M
46. J. Li, X. Li and D. He, "Domain Adaptation Remaining Useful Life Prediction Method Based on AdaBN-DCNN," 2019 Prognostics and System Health Management Conference (PHM-Qingdao), Qingdao, China, 2019, pp. 1-6, doi: 10.1109/PHM-Qingdao46334.2019.8942857.
47. Heimes, F. (2008). Recurrent neural networks for remaining useful life estimation, In IEEE int. conf. on prognostics and health management.
48. C. Zhang, P. Lim, A. K. Qin, and K. C. Tan, ''Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics,'' IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2306–2318, Oct. 2017. doi: 10.1109/TNNLS.2016.2582798.
49. X. Li, Q. Ding, and J.-Q. Sun, ''Remaining useful life estimation in prognostics using deep convolution neural networks,'' Rel. Eng. Syst. Saf., vol. 172, pp. 1–11, Apr. 2018. doi: 10.1016/j.ress.2017.11.021 .
50. J. Wang, G. Wen, S. Yang, and Y. Liu, ''Remaining useful life estimation in prognostics using deep bidirectional LSTM neural network,'' in Proc. Prognostics Syst. Health Manage. Conf. (PHM-Chongqing), Chongqing, China, 2018, pp. 1037–1042. doi: 10.1109/PHM-Chongqing.2018.00184.