doi:10.20944/preprints202109.0034.v3

Article

Abnormality Detection and Failure Prediction Using Explainable Bayesian Deep Learning: Methodology and Case Study of Real-World Gas Turbine Anomalies

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 (V.L.) and ahmad_18002773@utp.edu.my (A.K.K.N)

Abstract: Mistrust, amplified by numerous artificial intelligence (AI) related incidents, has caused the energy and industrial sectors to be amongst the slowest adopter of AI methods. Central to this issue is the black-box problem of AI, which impedes investments and fast becoming a legal hazard for users. Explainable AI (XAI) is a recent paradigm to tackle this challenge. Being the backbone of the industry, the prognostic and health management (PHM) domain has recently been introduced to XAI. However, many deficiencies, particularly lack of explanation assessment methods and uncertainty quantification, plague this young field. In this paper, we elaborate a framework on explainable anomaly detection and failure prognostic employing a Bayesian deep learning model to generate local and global explanations from the PHM tasks. An uncertainty measure of the Bayesian model is utilized as marker for anomalies expanding the prognostic explanation scope to include model's confidence. Also, the global explanation is used to improve prognostic performance, an aspect neglected from the handful of PHM-XAI publications. The quality of the explanation is finally examined employing local accuracy and consistency properties. The method is tested on real-world gas turbine anomalies and synthetic turbofan data failure prediction. Seven out of eight of the tested anomalies were successfully identified. Additionally, the prognostic outcome showed 19% improvement in statistical terms and achieved the highest prognostic score amongst best published results on the topic.

Keywords: Anomaly detection; Bayesian methods; black-box models, CUSUM method; explainable artificial intelligence; prognostic and health management; singular value decomposition.

1. Introduction

1.1 Artificial Intelligence

Artificial intelligence (AI) is officially the hype of the century, unravelling possibilities that once reside only in our imagination. AI is currently serving numerous fields and constantly breaking fresh boundaries. Its capability is consumed by the mass public and reaching far into specialized domains. Intensive race between world powers to harness its power stimulates consistent stream of fund to support AI based projects in all parts of the globe. With AI technology presently within reach by literally everyone, the age of AI has just begun.

How does one define AI? According to a survey conducted by Artificial General Intelligence Sentinel Initiative (AGISI) in 2018, the most agreeable definition for AI voted by experts is that stated in [1]. This describes AI as having the faculty of adaptation and improvisation, despite establishing limited knowledge and resources. The description further implies the autonomicity and learning capacity of the system.

The European Commission's portrayal of AI is somewhat like the former definition, without the concept of restriction albeit carefully specifying the system's partial degree of autonomy [2]. These depictions paint us the picture of a system, capable of reasoning and operating with partial or no supervision at all, thus potentially beneficial, or dangerous, to the human being. The type and task of AI methods are commonly classified into seven categories as follows:

1. Machine learning (ML): in addition, based on deep learning (DL) and predictive analytics.
2. Natural language processing: Translation, classification, information extraction.
3. Speech: This is visualized as speech to text and text to speech.
4. Expert systems: inference engine and knowledge base.
5. Planning, schedule, optimization: Reduction and classical probabilistic as well as temporal.
6. Robotic: Reactive machine, limited memory, theory of mind, and self-aware.
7. Vision: Image recognition and computer/machine vision.

Such a vast catalogue of ability naturally finds its worth in many applications. Globally, the impact of AI is more anticipated in key economic and social pillars such as manufacturing, transportation, healthcare, business analytics, finance, and retails [3,4]. Likewise, research on AI stretches over other niche domains like entertainment [5], law enforcement [6], security [7], safety [8], defense [9], construction [10], investment [11], and mining operation [12]. The list goes on with the endless possibility, with new fronts being opened by researchers on daily basis.

ML and DL have emerged as the most popular and powerful tools in solving technical challenges. Their nonlinearity power, ever-increasing data volume, availability of open-source development tools within reach by everyone, together with enhanced and affordable computing power, push DL to the forefront of AI tools. Some of the notable DL achievements throughout the decade are mentioned here. In speech recognition field, DL outperformed the Gaussian mixture modeling-based systems in automatic speech recognition with record accuracy [13]. Alpha Go, an AI game system, beat world champions, Lee Sedol and Ke Jie in Go game match in 2016 and 2017 respectively [14,15]. In robotic, the OpenAI five robot system beat the world champion team in 2019 in Dota game tournament [14]. In 2021, CoAtNet-7 achieved 90.88% accuracy in ImageNet image classification dataset [16].

When charting the AI investment landscape, we can mention the following. Price Water Cooper (PwC) estimated that AI could uplift global GDP by 14% or 15.7$ trillion by 2030, with China and the United States as the biggest beneficiaries of this impact [17]. In 2019, the United States (US) possessed the most investment under the form of private AI companies representing around 64% of global share followed by China. The rest of the world trails the US and China, contracting around 400% in investment value from 2015 to 2019. During the said epoque, transportation, customer relation, and business analytics received the biggest specific investments in the US while transportation, security and arts attracted more investment in China. Globally, transportation and business analytic sectors constitute important investment grounds [18]. Soon, AI will fully replace capital and labor as the new factors of production, being the main driver of productivity [17]. The labor market will experience profound change where less workforce generating higher value will be required. To thrive or merely survive the competition, increasing AI assimilation to replace low skilled works is expected to be the future agenda in the industries [19].

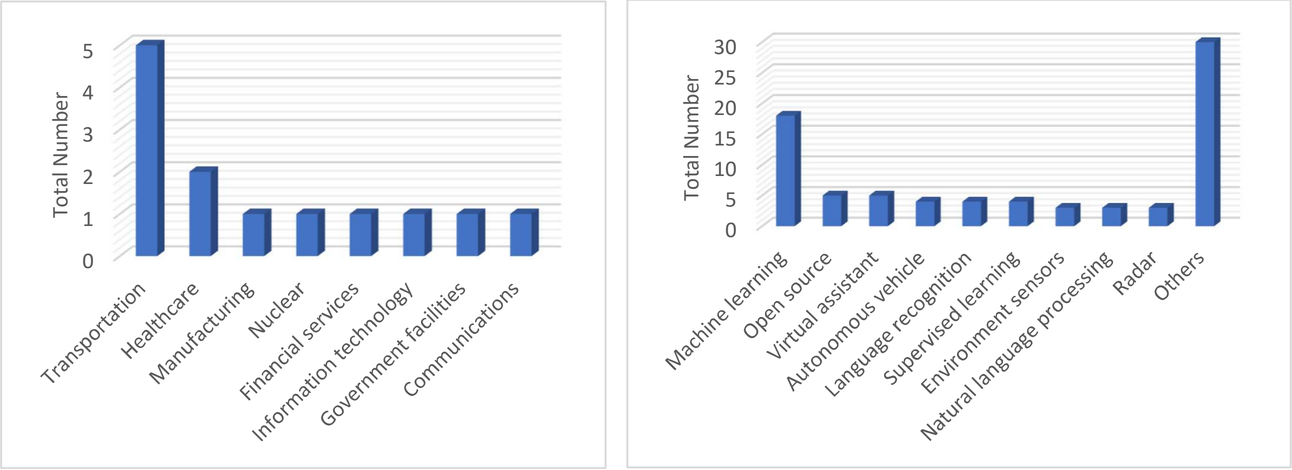
According to the World Intellectual Property Organization (WIPO), the number new of AI patents registered tripled from 2013 to 2017, mirroring the intensive efforts led by the technical community in exploiting AI potential to overcome challenges [20]. Geographically, the top world economies, majorly in Asia, occupy the biggest share in AI patent registration headed by Japan (43%), followed by US (20%), Europe Union(EU)-28 (10%), China (10%), South Korea (10%) and Germany (3%). The primary sectors where patent and trademark registrations are concentrated correspond to computers and electronics, machinery, information technology services, and transportation [21].

Surprisingly, the industrial, manufacturing and energy sectors are amongst the slowest to adopt AI in their day-to-day operations [17]. Considering the continuous improvement in these areas, this slowness seems to be improbable. However, one can understand that there is a confidence issue from the industrial actors in blindly accepting AI decisions.

Trust is thus the primary obstacle in AI implementation. In the mentioned domains, this mistrust is more related to performance issue. In other fields, some types of problems might arise. The Center for Security and Emerging Technology (C-NET) defines the category of AI malfunction as follows [22]:

1. Failures of robustness: The system is subjected to unusual or unforeseen inputs, causing failure.
2. Failures of specification: The system is attempting to do something that is subtly different from what the developer of user anticipated, which might result in surprising behavior or consequences.
3. Failures of assurance: In operational mode, the system cannot be fully supervised or regulated.

The AI incident database documents the growing AI incidents since 2019 [23]. This repertoire exhibits several information worth noting. As per today, the top domains where incidents are reported are transportation, healthcare, manufacturing, and nuclear as presented in Figure 1(a). Most of the incidents are caused by ML issues as shown in Figure 1(b). These facts strengthen the belief about why the industrial and energy sectors are hesitant in using AI. In a more serious note, 8% of the incidents resulted in loss of lives.



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

Figure 1. AI incidents overview for (**a**) incidents' domain and (**b**) incidents' causes.

Secondly, AI regulation. The heightened incidents and its consequences risk to stall investments and prompted the call for regulations. The laws intend not to punish but to foster a responsible AI culture. A summary on global AI regulations can be found in [24]. Included is the General Data Protection Regulation (GDPR). The GDPR is the strictest regulation to date issued by the EU. This regulation would affect developers around the world whose system's output are related to the EU. It classifies AI systems into three categories as follows:

1. Limited and minimal-risk, high-risk, and unacceptable-risk.
2. The unacceptable-risk system will not be authorized anymore while the high-risk system will be conditioned to strict requirements.
3. The minimal-risk system will also be subject to a few conditions.

Under GDPR, offences could incur fine of up to €30 million or 6 % of global revenue with the use of illegal systems and the breach of the data-governance requirements by employing hazardous systems could result in the heftiest penalties.

In brief, there are six system's qualities demanded towards a responsible AI ecosystem stated as:

1. Transparency: An AI system mechanism should be understood.
2. Reliability and safety: An AI system should work as intended and safe to use.
3. Privacy and security: AI systems should respect confidentiality and protected.
4. Fairness: AI systems should behave equally toward all human being
5. Inclusiveness: AI systems should inspire and promote human participation.
6. Accountability: Responsibility measures must be available when AI system malfunction.

One can note that most of the provisions in the law focus on the issue of transparency, fairness, privacy, and data security related to AI algorithms. The transparency refers to the mechanism of AI methods in obtaining their output. In fact, transparency is the main key in minimizing AI malfunctions and achieving the AI quality goals mentioned before. This is due to the black-box characteristic of some AI techniques.

DL, being the most powerful AI method now, is a black-box model so that it is opaque. Though very effective, its mechanism in generating forecast is unknown. Naturally, this opacity thwarts AI dissemination in high stakes areas, such as the industry and the energy sectors, where incomprehensible outcome could lead to incorrect prediction. In turn, this can provoke disastrous effects in term of lives, safety and financially. Obviously, the experts of the domain demand more than mere point estimate prediction to convince them in taking the correct course of action. Thus, the ball lies in the research community hands to diminish this mistrust. Then, third, explainable AI (XAI) enters.

Note that XAI is a field dedicated in making AI model transparent to human through various approaches. Though this notion is known for decades, global attention garnered in XAI shows a notable rise more recently, reflected by the increasing initiatives by various parties including the Defense Advanced Research Projects Agency (DARPA) since 2016 [25]. This sudden spike in interest on XAI is partly due to emerging laws as mentioned previously. The steady accumulation in general and specialized review articles on XAI translates the growing interest in XAI from the research community [26-30].

The advantages of XAI, however, far outweighs the need of regulations based on:

1. Justify model's decision, detecting its problem, especially during the trial period of AI model, strengthening reliability and safety.
2. Comply with the regulations, transparency that leads to accountability, enhanced security, and data privacy.
3. Help to understand AI reasoning and decrease problems related to fairness in AI use.
4. Assist practitioners in verifying the required proprieties of AI system from developer.
5. Promote interactivity and expand human creativity by discovering new perspective on the model or the data.
6. Allow resources to be more optimized, avoiding wastage.
7. Foster collaboration between experts, data scientists, users, and stakeholders.

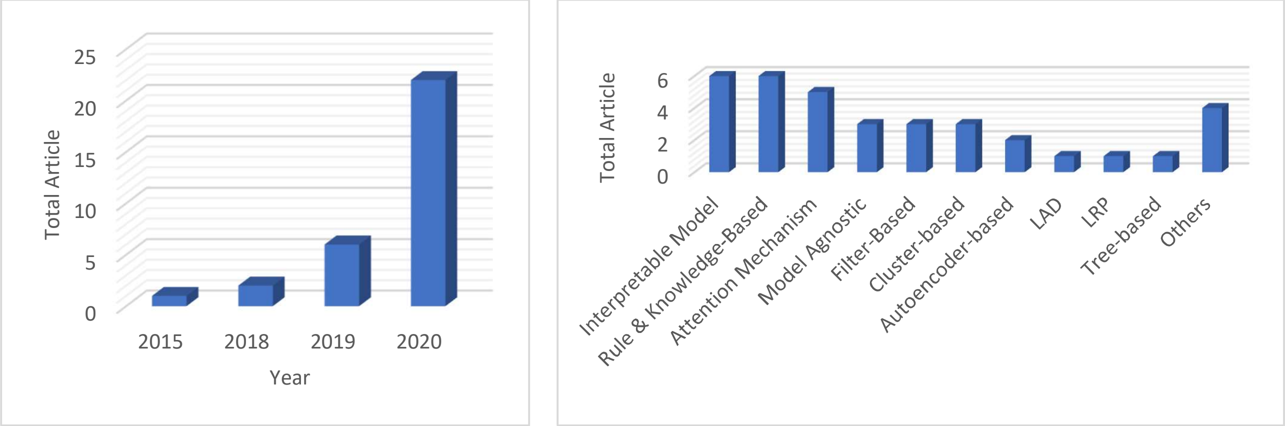
Several published articles have organized XAI approaches into distinct taxonomies [31-33]. This paper briefly describes the categorization according to [31] which falls into two general classes. Firstly, transparent models, which are directly interpretable due to their simple structure or comprehensible visualization such as linear or logistic regression, decision tree and rule-based methods. Secondly, post-hoc explainability, where explanation is generated after the model to be explained is trained. Included in this category is model agnostic approach, an external method that can be used with any AI model. In addition, post-hoc explainability is applied for shallow ML models (tree ensembles, random forests and multiple classifier systems, support vector machine). Hence, approaches related to DL such as neural networks (model simplification, feature relevance), techniques appropriate only for certain DL models as convolutional neural network (CNN) and recurrent neural network (RNN), layer wise propagation (LRP), class activation mapping (CAM), gradient weighted class activation mapping and for hybrid-transparent-opaque models (knowledge-based and case-based reasonings).

As the backbone of the industry, prognostic, and health management (PHM) is a set of frameworks exploiting sensor signals to safeguard the health state of industrial assets by identifying and examining, tracking degradation, and to estimate failure evolution [34]. To achieve this goal, three main activities comprising of anomaly detection, failure prognostic, and diagnostic are employed, that is:

1. The first consists of identifying outliers in the system's output data [35].
2. The second task englobes the determination of remaining useful life (RUL); and
3. Lastly, the classification and identification of root cause of failure [36,37].

In recent years, AI has become a predominant tool in reliability-based research [38].

PHM-XAI is still a very young discipline. As testified by the recent systematic review on PHM-XAI presented in [39] and shown in Figure 2(a), several peer reviewed journal articles treating the subject is still small but steadily rising. Several explainability approaches have been explored by the PHM-XAI researchers. To forge trust in AI and facilitate its legal use in the industry, it is urgent to disseminate XAI know-how to PHM players in both the research and industrial domains.



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

Figure 2. Overview of PHM-XAI domain for (**a**) PHM-XAI publications over the years and (**b**) PHM-XAI published approaches.

1.2 Research Gaps and Opportunities

The review presented in [39] further lists several deficiencies plaguing the research in PHM-XAI that need to be remedied promptly, considering:

1. Lack of human involvement: Human engagement is crucial for assessing the generated explanation as the latter is meant for them. Furthermore, human-AI cooperation could contribute to the integration of human related sciences, augmenting the PHM-XAI field. In addition, human participation is urged for the development of interactive AI, where experts and AI system work hand in hand, providing more assurance in AI system's output.
2. The use of explanation evaluation metrics is practically absent: These measures are important for researchers and developers when evaluating the quality of an explanation.
3. Insufficiency in uncertainty management: Uncertainty quantification safeguards the system against adversarial examples where false explanation could be generated from unseen, new data. Moreover, it provides users with supplementary confidence

in trusting AI methods prediction compared to point estimation models. It is therefore unconceivable for a working AI system to be devoid of this feature.

Therefore, the review article summarized some research opportunities in PHM-XAI stated as:

1. As shown in Figure 2(b), model agnostic explainability, LRP and logic analysis of data (LAD) are less explored, but they possess great potential as they could be used with any black-box models without altering its performance. LAD can be combined with fault tree analysis for complex risk management.
2. While SHapley Additive exPlanations (SHAP) is an established model agnostic method and employed in PHM-XAI works, note that it was not exploited to improve PHM task's performance.

Addressing the weakness and seizing the opportunity, this article demonstrates the application of the SHAP model agnostic approach in explaining and improving anomaly detection and failure prognosis tasks taking a case study related to gas turbine systems. Abrupt disturbances in a real-world gas turbine modeling are tested for detection. Then, the root cause of degradation in a turbofan prognostic problem using simulated data is deciphered. SHAP local and global explanations are utilized to improve the prognostic performance. Prediction uncertainty, specifically aleatoric uncertainty, issued from a DL model to be explained, served a dual purpose: (i) as anomaly indicator, monitored using cumulative sum (CUSUM) changepoint detection; and (ii) to bolster explanation in terms of the confidence of the model in its output. Additionally, these uncertainties were minimized based on denoising and hyperparameters optimization operations, a crucial aspect seldom ignored in probabilistic DL articles. Decreased uncertainties amplified anomaly detection ability and increased the accuracy of prognosis. Then, the explanation produced is evaluated utilizing local accuracy and consistency metrics.

The main contributions of this work are as follow:

1. We combine SHAP and DL uncertainty to constitute a wider explanation scope, where the first one explains the decision of the model, while the latter one describes its output confidence.
2. We demonstrate the SHAP global explanation's ability to improve prognostic task's performance, which was absent from previous works.
3. We apply explanation evaluation metrics, which is clearly deficient from previous PHM-XAI literature.
4. We show the potential of DL uncertainty as anomaly indicator for a real-world industrial dataset, which validates its capability.
5. We minimize DL uncertainties for enhancing prognostic accuracy. Additionally, the small aleatoric uncertainty enables a more visible aspect spiking effect caused by anomalous data.

The secondary contributions are the following:

6. We add model agnostic explainability to the collection of PHM-XAI articles, which is still lacking in the moment.
7. We prove the local accuracy trait of the explanation validates the efficiency property of Shapley values, while confirming the consistency characteristic justifies the additivity and symmetry proprieties of these values.

1.3 Related Works

PHM-XAI works associated with anomaly detection and failure prognostic are summarized. In the order of presentation: (i) interpretable model [40,41]; (ii) extraction-based approach [42]; (iii) decision rules and knowledge-based explanation [43]; (iv) attention mechanism [44]; (v) model agnostic [45]; and (vi) visual explanation technique [46].

The dynamic structure–adaptive symbolic approach (DSASA), a cross-domain life prediction model, is elaborated in [40] for slewing bearings RUL prediction. The DSASA presents internal model structures visibly, takes historical run-to-failure data into account, and dynamically adapts real-time deterioration. In a nutshell, multi-signal-based health indicators are fed into three genetic programming algorithms for symbolic life modeling. This modeling visually displays the life process in the manner of legible mapping relationships and obtains ideal RUL prediction results. Then, the DSASA reconstructs original life expressions from the initial symbolic life model and uses dynamic coupling terms and its exponents to track the real-time asset deterioration. The recorded performance is better than the previously employed method for the case study and contributed by XAI ability.

An interpretable structured-effect neural network (SENN) stated as

$$\mathrm{SENN}_{\Theta}(t;\, X_{t,} \ldots X_1) = \lambda(t) + \beta^{\mathrm{T}} X_t + \mathrm{RNN}_{\Theta}\,(X_{t,} \ldots X_1, t), \tag{1}$$

consists of a non-parametric baseline, a linear component of the current condition and a recurrent component as proposed in [41] for turbofan prognostic application, with the model being represented in (1). Here, the first component, $\lambda(t)$ namely, is the non-parametric part consisting of lifetime probabilistic model. The second component is a linear form that can be employed with raw sensor readings, X_t say, where the importance of features may be evaluated based on the linear coefficients. The third component, RNN_{Θ}, refers to recurrent neural network with weights Θ. Thus, the recurrent component needs to explain less variance of the data compared to pure neural network structure. The performance of the model surpasses other traditional ML methods except the LSTM. However, XAI does not contribute to this performance.

An autoencoder with explanation discriminator is employed in [42] for continuous batch washing equipment anomaly detection. The autoencoder's reconstruction error, which is the anomaly indicator, is utilized by the discriminator to measure the precision and accuracy measurement of the anomaly detection task. The discriminator rescales the reconstruction error using a sigmoidal function giving value 0 as normal, 1 as anomaly and between 0 and 1 as warning. The performance of the proposed method is comparable to the best technique, isolation forest, previously employed for the problem, assisted by XAI approach.

The Fused-AI interpretabLe Anomaly Generation System (FLAGS), which combines both knowledge-driven (KD) and data-driven (DD) abilities, is presented in [43] for anomaly detection, failure recognition and root cause analysis of train. The FLAGS consists of three stages as follows:

1. In the first phase, both KD and DD fault recognition (FR) and root cause analysis (RCA) using data from failure mode/effect analysis (FMEA) and fault tree analysis (FTA), are employed simultaneously. The data streams and case-specific context data are used as inputs. Faults from the KD or outliers from the DD are produced with interpretation of the detected anomalies and stored inside a knowledge graph (KG).
2. In the second phase, the detected anomalies are shown in a dynamic dashboard complete with the raw data and interpretation, where the user modification is authorized. This is also stored in the KG.
3. Then, in the third phase, the information in the KG, which are anomalies, the feedback, and all contextual meta-information, is used to improve the AD, FR and RCA techniques of both KD and DD. The reported accuracy is good for anomaly detection, being it better than other standalone DD methods, partly because of the XAI approach.

The self-monitoring, analysis, and reporting technology (SMART), depicted in [44], is utilized to detect and predict failure in hard-drives through SMART statistics in the Attention-augMENted DEep aRchitecture (AMENDER) model. The SMART statistics daily record is incorporated into vectors through the feature integration layer. Then, these vectors are fed into the temporal dependency extraction layer consisting of gated recurrent unit (GRU), whose output can be considered as a compact representation of the SMART temporal sequence of the observed days. The attention distribution is calculated from the healthy context vector and the SMART compact representation. The healthy context vector is the high-level feature representation of healthy hard-drives. The resultant distribution, together with the GRU hidden state, produce attentional hidden state of the corresponding days. This attention mechanism enables the model to focus on failure advancement. Then, the attentional hidden state may be used to determine the health of the hard-drive for the associated day. The model's performance is better than other tested methods in both hard-drive health status classification and prognostic. The attention mechanism contributed to this performance, besides being the mechanism for diagnostic.

A fouling prediction in crossflow heat exchanger, using feed-forward neural network architecture with LIME model agnostic explainability, is described in [45]. The model is fed with operational data, such as inlet fluid temperatures, ratio of fouled fluid flow rates to flow rates under clean conditions, and output fluid temperatures from the heat exchanger and predicts fouling resistances of the equipment. Note that the predictive accuracy is very good.

A comprehensive visual explanation tool applied to turbofan engine prognostic is suggested in [46]. This online diagnostic, prognostic, and situation awareness system works with streaming data and is divided into the following sections: (i) ML–based classifier; (ii) visualization dashboard for health state monitoring; (iii) cybersecurity command centre, and (iv) high-performance local servers. The visualization dashboard displays real-time predictive analytics to reveal potential flaws, risks, and harmful attacks. In the form of heat maps, users may view the input and output. One heat-map for each sensor input and related engine at each time step. The network weights of each layer may be examined by practitioners to see how each feature contributes to the output of the following layer. The network weights are represented by the line thickness. As the weight values increase, the thicker the lines increase as well. Practitioners may also customize model hyper-parameters like the number of layers, hidden units, weights in each layer, regularizer types, and regularizer parameters, to integrate their expertise into the learning process.

This article is organized as follow. The methodology is described in Section 2. The case study, results and discussion are presented in Sections 3 and 4, respectively. Finally, the concluding remarks are in given in the last Section 5.

2. Methodology

2.1 Multi Output Bayesian LSTM and Uncertainty Quantification Layers

A single input and multi outputs LSTM model is employed for anomaly detection and RUL estimation tasks. The model, denoted by f_x, comprises an input layer, where input data are fed, a single LSTM layer, a fully connected or dense layer, and two output layers, such as presented in Figure 3. The LSTM layer produces sequential prediction by employing a gating mechanism to retain important memory or forget negligible ones. This structure enables the accumulation of important information, a crucial ability in anomaly monitoring and degradation tracking tasks. The input data's matrix multiplication and addition with the weights and bias factors of the model happen in the dense layer. Then the forecast, altogether with uncertainty, are enabled by the probabilistic nature of the output layers.

Two types of uncertainties are defined in DL models. The first type is the aleatoric uncertainty, linked to noise, acquisition error and randomness in the dataset. Thus, the first output layer has aleatoric uncertainty, which learns and predicts using the sequential output of the LSTM layer as input, based on the mean and standard deviation of the output distributions as depicted in layer "dense2" of f_x in Figure 3. The prediction range reflects the uncertainty.

The second type is the epistemic uncertainty, corresponding to the uncertainty of the weights of the DL model. Hence, the second output layer has epistemic uncertainty, also known as the dense variational layer. This layer learns and predicts the posterior distribution of the weights using variational inference by maximizing the evidence lower bound (ELBO) objective function stated as

$$\mathcal{L}(q_\theta) = -\mathbb{E}_{q_\theta(w)}\left[-\log(P\,(y|x,w)) - \log\left(\frac{q_\theta(w)}{P(w)}\right)\right] \tag{2}$$

$$\mathcal{L}(q_\theta) = -\int dw q_\theta(w)\log(P\,(y|x,w)) + \int dw q_\theta(w)\log\left(\frac{q_\theta(w)}{P(w)}\right), \tag{3}$$

where $P(y|x,w)$ stated in (2)-(3) is the corresponding likelihood function that links the input x, the output y and the weights w. Note that $P(w)$ is the prior or the initial distribution of the weights, whereas $q_\theta(w)$ is the approximated distribution once the training of the DL model is completed. The sampled values $q_\theta(w)$ is the prediction output.

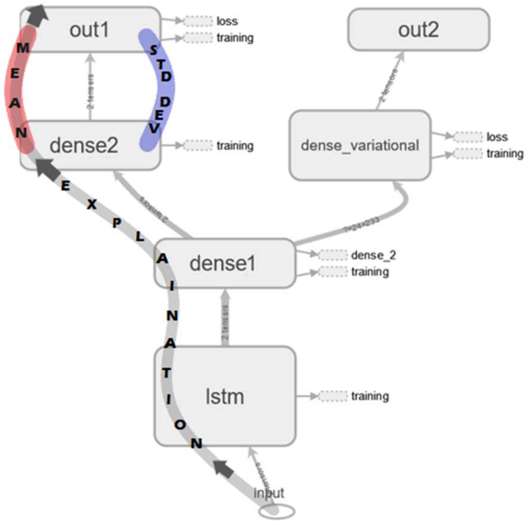


Figure 3. Architecture of f_x.

2.2 Minimization of Uncertainties, Anomaly Detection, and RUL Estimation

The Gaussian or normal distribution, a well understood and commonly used probability model, was utilized to describe both types of uncertainties. The aleatoric and epistemic uncertainties are represented by the rolling standard deviation of the predicted distribution's sequence. The only possible way to reduce the aleatoric uncertainty of the recorded data is removing its noise. Thus, the data are firstly denoised using the singular value decomposition algorithm following the methodology stated in [48,49]. The denoised data are later utilized to optimize the DL hyperparameters with Bayesian hyperparameter optimization (BayesOpt), whose limits are shown in Table A1 in the Appendix [50]. The BayesOpt optimized the model and decreased the epistemic uncertainty.

For anomaly detection, the model is trained with healthy data as it is expected that the aleatoric uncertainty shows a spike when the model encounters an abrupt anomalous observation. This spike, or changepoint, is detected using a CUSUM algorithm with a specified control limit C as stated in [51]. Note that C is determined employing the prediction's aleatoric uncertainty of the healthy data. Given $\mathrm{AU}_{\mathrm{stdmax}}$, $\mathrm{AU}_{\mathrm{stdmean}}$ and $\mathrm{AU}_{\mathrm{stdstd}}$ corresponding to the maximum, mean, and standard deviation of the standard deviations of the aleatoric uncertainties, respectively, the specified control limit defined as $C = (\mathrm{AU}_{\mathrm{stdmax}} - \mathrm{AU}_{\mathrm{stdmean}})/\mathrm{AU}_{\mathrm{stdstd}}$.

Given a sequence $y_i, \dots, y_n$ of process measurements with mean μ_x and standard deviation σ_x, the lower and upper cumulative process sums are defined as

$$u_i = \begin{cases} 0, & i = 1, \\ \max\left(0, u_{i-1} + y_i - \mu_x - \frac{1}{2} n\sigma_x\right), & i > 1, \end{cases} \tag{4}$$

$$l_i = \begin{cases} 0, & i = 1, \\ \min\left(0, l_{i-1} + y_i - \mu_x + \frac{1}{2} n\sigma_x\right), & i > 1, \end{cases} \tag{5}$$

where u_i and l_i stated in (4) and (5) are the lower and upper cumulative process sums.

Deviation is detected at point y_j if $u_j > C\sigma_x$ or $l_j < -C\sigma_x$. For prognostic purpose, the DL model is trained with both healthy and degradation data. The trend of the aleatoric uncertainty reflects the confidence of the model in its prediction. The rising aleatoric uncertainty trend mirrors a growing uncertainty, while the contrary represents increasing confidence of the DL model.

2.3 Model Performance Assesment and SHAP Explainability

The root mean squared error (RMSE) and early prognostic metric are employed. The first metric is applied to evaluate both the anomaly detection and prognostic tasks while the second one is only used for prognostic.

The RMSE is utilized to examine the model's predictive performance with aleatoric and epistemic uncertainties [52]. In order to obtain a meaningful measure, the mean performance for 100 predictions is calculated. The RMSE measures how spread the errors are between the predictions and the true RULs being is defined as

$$\mathrm{RMSE}_{\mathrm{mean}} = \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{RUL}_{\mathrm{true}}^{(i)} - \mathrm{RUL}_{\mathrm{mean}}^{(i)}\right)^2}\right)\Big/_{100} \tag{6}$$

where $\mathrm{RUL}_{\mathrm{true}}^{(i)}$ stated in (6) is the true RUL for asset i, $\mathrm{RUL}_{\mathrm{mean}}^{(i)}$ is the predicted RUL (the mean of predicted RUL distribution) for asset i and N as the total number of assets. In addition, we define the early prognostic metric as

$$S = {}^{\left(N\sum_{i=1}^{N} s_i\right)}\Big/_{100} \tag{7}$$

where

$$s_i = \left\{\begin{matrix} e^{\frac{-d_i}{13}} - 1, d_i < 0 \\ e^{\frac{di}{10}} - 1, d_i > 0 \end{matrix}\right\}, \; d_i = \left(\mathrm{Mean}_{\mathrm{pred}}^{(i)} - \mathrm{RUL}_{\mathrm{truth}}\right)$$

The metric S stated in (7) gives higher score for errors of similar amplitude in early prediction than late prediction as the former is more important in failure estimation. Note that s_i is the individual asset's prognostic score, while d_i is the individual asset's prognostic error. Here also, the mean of 100 prediction scores, with aleatoric and epistemic uncertainties, are calculated.

The SHAP is a technique to explaining any machine learning model's output mechanism based on game theory [53]. It uses Shapley values to assess the contribution of each feature to the prediction. The formula for Shapley value is given by

$$\phi_j(\mathrm{val}) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(\mathrm{val}(S \cup \{x_j\}) - \mathrm{val}(S)\right) \tag{8}$$

The Shapley value of feature j, ϕ_j namely, defined in (8), is the average marginal contribution of feature j's value over all probable combinations of features values regarding the prediction. Note that S is a subset of the total p features and x is the instance's vector to be explained. The prediction for feature values in set S that are marginalized over those excluded from set S is $\mathrm{val}_x(S)$ defined as

$$\mathrm{val}_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - \mathrm{E}_X(\hat{f}(X)), \tag{9}$$

where $\mathrm{E}_X(\hat{f}(X)$ stated in (9) is the expected value of all predictions. The description of the SHAP is provided next. Given the model of explanation, e namely, the coalition vector, and $z' \in \{0,1\}^N$, with $z' = 1$ indicating that the feature is present in the coalition, while $z' = 0$ points to the contrary, N is the maximum coalition size, we have that

$$e(z') = \phi_0 + \sum_{j=1}^{N} \phi_j z'_j \tag{10}$$

where, as mentioned, ϕ_j expressed in (10) is the Shapley value of feature j.

The SHAP can explain both global and local outputs. However, it is not compatible with probabilistic DL and only accepts a single output vector for explanation. Thus, a workaround, in the form of a non-probabilistic model labelled as f'_x, is developed as shown in Figure 4(a). Note that f'_x has the same layers and weights as those figured along the explanation path in f_x, except the weights in dense2 of f_x. Here, only the weights corresponding to the mean are used and transferred to f'_xwhile the weights associated with the standard deviation are ignored. The output layer out3 in f'_x slices only the first value of each sequence vector and arranges them in a single vector for the SHAP explanation.

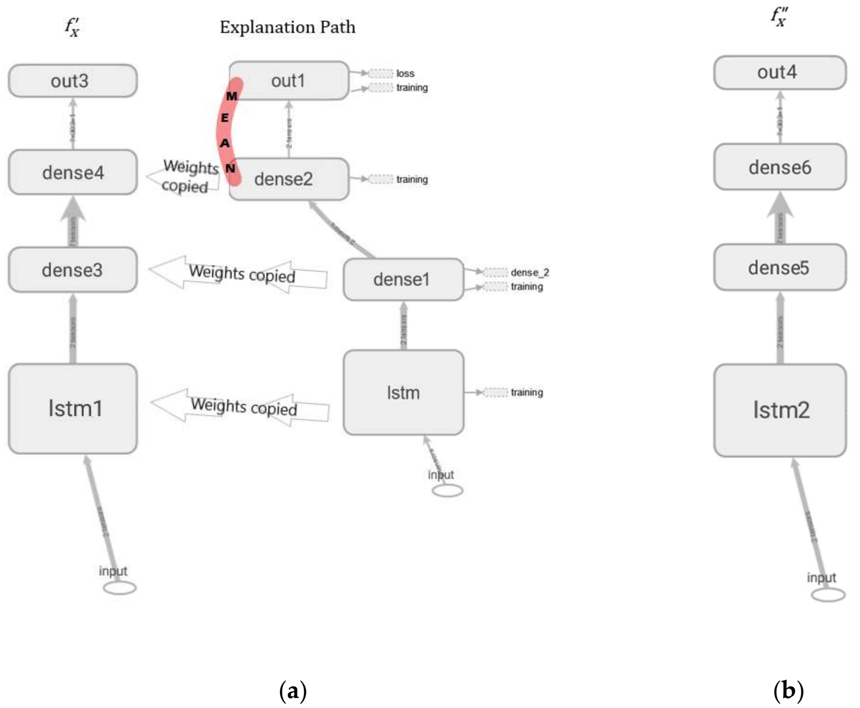


Figure 4. Explanation models architectures for (**a**)f'_x and (**b**) f''_x.

2.4 Explanation Visualization

Three means of visualization are used for illustrating the local and global explanations as follows:

1. Local: Force plot and Waterfall plot, which highlights the positive or negative forces of features influencing an instance's output. The former, as it only shows features inputs values and forces directions, is utilized for explaining anomalous instances. The latter one, as it shows features contributions and forces directions, is used to verify the local accuracy and consistency properties of the explanation elaborated in the next subsection.
2. Global: Summary plot, which highlights the most contributing features in a sequence. The plot arranges the features according to its contributing power and its forces directions. Here, the explanation is exploited to enhance the prognostic accuracy by employing only the most contributing features. The model is initially tested with all the features followed by only using 75% of the best features. Therefore, the performances of the different settings are analyzed and compared with published results.

The first explanation property to be verified is local accuracy of SHAP as stated in [54]. It establishes that the sum of the feature contributions, Φ namely, is equal to prediction of x or $f(x)$, minus the average prediction, $\mathrm{E}_x(\hat{f}(X\,))$ say.

From the definition of the SHAP given in (10), posing $\hat{f}(x') = e(z')$ and $\phi_0 = \mathrm{E}_x(\hat{f}(X))$, we get that

$$\hat{f}(x') = \mathrm{E}_x(\hat{f}(X)) + \sum_{j=1}^{N} \Phi_j x'_j \qquad (11)$$

where

$$\textstyle\sum_{j=1}^{N} \Phi_j = f(x) - \mathrm{E}_x(\hat{f}(X\,)). \qquad (12)$$

By setting $x'= 1$ in (11), the efficiency property of the Shapley values defined in (12) is retrieved. This property is examined using waterfall plot.

The second property is consistency, which states that if a model is modified, resulting to either the unchanged or increased marginal contribution of a feature, the Shapley value also follows the marginal contribution's trend as defined in [54].

Let v' be the complete set of features, $v'_{\setminus j} \Leftrightarrow v' = 0$, and the absence of feature j from the set of features v', for models f' and f''. Thus, if

$$f'_x(v') - f'_x(v'_{\setminus j}) \geq f''_x(v') - f''_x(v'_{\setminus j}) \qquad (13)$$

for $v' \in \{0, 1\}^N$, then

$$\Phi_j(f', x) \geq \Phi_j(f'', x) \qquad (14)$$

where $f'_x(v')$ is calculated from f'_x and $f''_x(v')$ from the model f''_x shown in Figure 4(b), having the same layers as f'_x but with different weights. Observe that $f'_x(v'_{\setminus j})$ and $f''_x(v'_{\setminus j})$ are obtained by removing the weights of the feature j from f'_x and f''_x respectively. In order to calculate the expression presented in (13), a waterfall plot is used to obtain the values of $f_x(v'), f_x(v'_{\setminus j}), f'_x(v'), f'_x(v'_{\setminus j})$ and to confirm $\Phi_j(f, x)$ and $\Phi_j(f', x)$ in the inequality formulated in (14).

3. Results

3.1 Case Study 1: Real Gas Turbine Anomaly Detection

Data from an 18.8 MW, twin-shaft industrial gas turbine from Petronas Angsi Oil Platform in Terengganu, Malaysia, recorded over one year period, or 8737 hours, are used in this study. Note that 98 sensor signals, comprising of various pressure, temperature, velocity, and positional readings make up the largely healthy data. While the features number is overwhelming, only several were used in modeling the gas turbine as indicted in [55]. The inputs and outputs utilized are shown in Tables 1 and 2 respectively. Four DL networks using f_x architecture labelled as $\text{Bayes\_LSTM}_{\text{N1}}$,$\text{Bayes\_LSTM}_{\text{P2}}$, $\text{Bayes\_LSTM}_{\text{P4}}$ and $\text{Bayes\_LSTM}_{\text{T4}}$ are fed with all the inputs to predict each output.

Table 1. List of inputs.

Ref	Input	Unit
N_2	Power turbine rotational speed	RPM
P_1	Compressor inlet pressure	Bar
m_f	Fuel mass flow rate	kg/s
T_1	Compressor inlet temperature	K

Table 2. List of outputs.

Ref	Output	Unit
N_1	Gas generator rotational speed	RPM
P_2	Compressor outlet pressure	Bar
P_4	Gas generator turbine outlet pressure	Bar
T_4	Gas generator turbine outlet temperature	K

First, we preprocess the data. The anomaly part is separated from the dataset and the healthy part is split into training. Validation and testing datasets as shown in Table 3.

Sequence of input and output are set to 24 hours. The only abrupt, null sensor's reading instances, consider as anomalies from 12 am to 1 am on 20/03/18 to 21/03/18 and 11 pm to 12 am on 08/04/18 to 09/04/18, which are chosen from the anomaly data collection and joined with the neighboring healthy data to put together a sequence of 24 hours. Both anomalies are set to be on the 12th to 13th instances of the sequences.

Table 3. Gas turbine dataset's summary.

Dataset	Date	Quantity (hour)
Training	01/01/18 – 23/10/18	6672
Testing	26/11/18 – 30/12/18	816
Validation	23/10/18 – 26/11/18	816
Anomaly 1	20/03/18 – 21/03/18	24
Anomaly 2	08/04/18 – 09/04/18	24
Unused Data		385
Total		8737

The RMSE results of $\text{Bayes\_LSTM}_{\text{N1}}$, $\text{Bayes\_LSTM}_{\text{P2}}$, $\text{Bayes\_LSTM}_{\text{P4}}$ and $\text{Bayes\_LSTM}_{\text{T4}}$ predictions with both aleatoric and epistemic uncertainties are shown in Table 4. The best result comes from $\text{Bayes\_LSTM}_{\text{N1}}$, where both RMSEs are low while the worst corresponds to the P2 model. The high difference between the training and testing data sets might be the cause of $\text{Bayes\_LSTM}_{\text{P2}}$ poor performance. Note that $\text{Bayes\_LSTM}_{\text{P4}}$ produces an interesting outcome, where the RMSE between aleatoric and epistemic uncertainties are not in the same order. This result could be improved by extending the BayesOpt evaluations to minimize the epistemic uncertainty. Nevertheless, no performance comparison may be done due to the inexistence of a benchmark result.

Table 4. RMSE results with AU and EU.

Model	RMSE aleatoric uncertainty	RMSE epistemic uncertainty
$\text{Bayes\_LSTM}_{\text{N1}}$	20.40	27.11
$\text{Bayes\_LSTM}_{\text{P2}}$	702.49	787.87
$\text{Bayes\_LSTM}_{\text{P4}}$	11.10	92.15
$\text{Bayes\_LSTM}_{\text{T4}}$	32.68	49.74

For illustration purpose, the anomalies modelled with $\text{Bayes\_LSTM}_{\text{N1}}$ and aleatoric uncertainty are shown in Figures 5 and 6. The aleatoric uncertainty anomaly spike on the 12th and 13th instances can be noted in these figures.

Predictions for $\text{Bayes\_LSTM}_{\text{P2}}$, $\text{Bayes\_LSTM}_{\text{P4}}$ and $\text{Bayes\_LSTM}_{\text{T4}}$ are in Figures A1, A2 and A3 for 20/03/18-21/03/18 and Figures A6, A7 and A8 for 08/04/18-09/04/18 respectively in Appendix 2. The parameters $\text{AU}_{\text{stdmax}}$, $\text{AU}_{\text{stdmean}}$, $\text{AU}_{\text{stdstd}}$ and C calculated from each model are listed in Table 5.

The CUSUM charts for anomalies predicted from $\text{Bayes\_LSTM}_{\text{N1}}$ obtained from the parameters in Table 5 are shown in Figure 7. As illustrated, both aleatoric uncertainty spikes are detected by the CUSUM method with the formulated control limit C as it is the case for all the other models except for Figure A4(a) in Appendix 2. The CUSUM charts of the anomalies predicted from $\text{Bayes\_LSTM}_{\text{P2}}$, $\text{Bayes\_LSTM}_{\text{P4}}$ and $\text{Bayes\_LSTM}_{\text{T4}}$ are in Figure A4 for 20/03/18-21/03/18 and Figure A9 for 08/04/18-09/04/18 in Appendix 2.

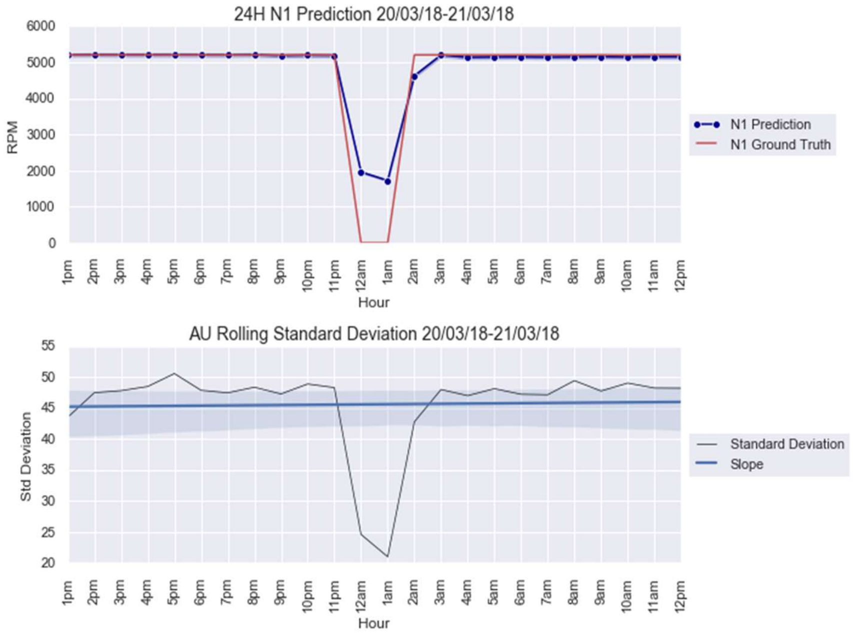


Figure 5. Bayes_LSTM$_{N1}$anomaly modeling 20/03/18-21/03/18.

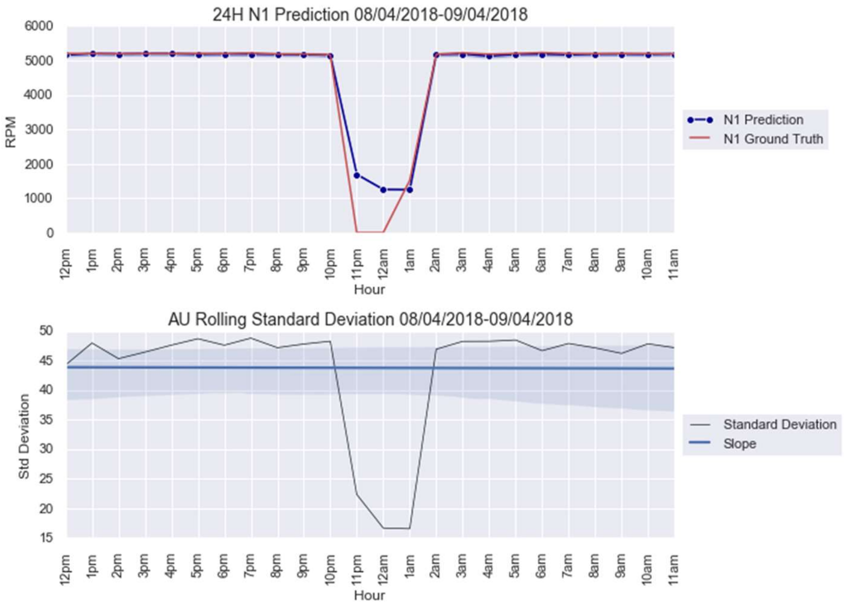
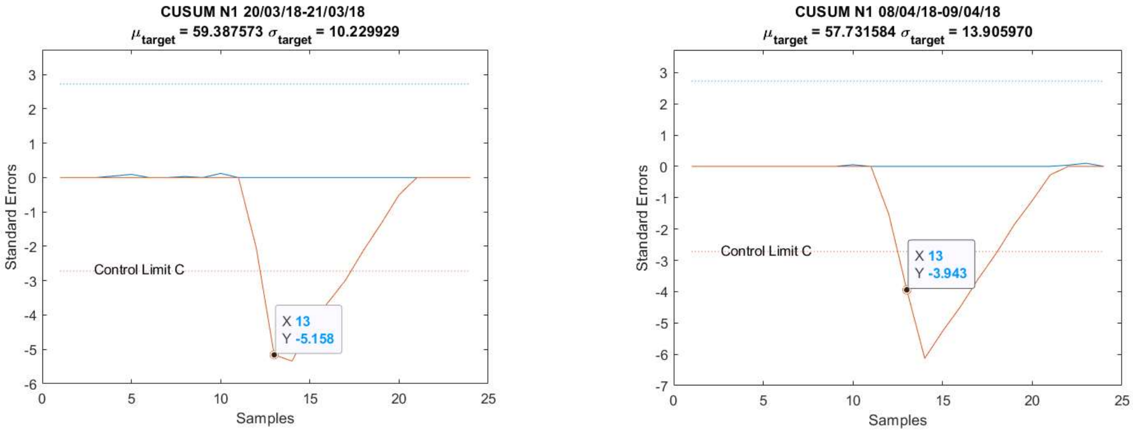


Figure 6. Bayes_LSTM$_{N1}$anomaly modeling 18/04/18-19/04/18.

Table 5. CUSUM chart's parameters obtained from DL models.

Model	$\mathbf{AU_{stdmax}}$	$\mathbf{AU_{stdmean}}$	$\mathbf{AU_{stdstd}}$	$\boldsymbol{C}$
Bayes_LSTM$_{N1}$	52.10	47.92	1.53	2.72
Bayes_LSTM$_{P2}$	87.35	82.70	1.87	2.49
Bayes_LSTM$_{P4}$	22.90	21.46	0.47	3.07
Bayes_LSTM$_{T4}$	14.20	13.27	0.30	3.05



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

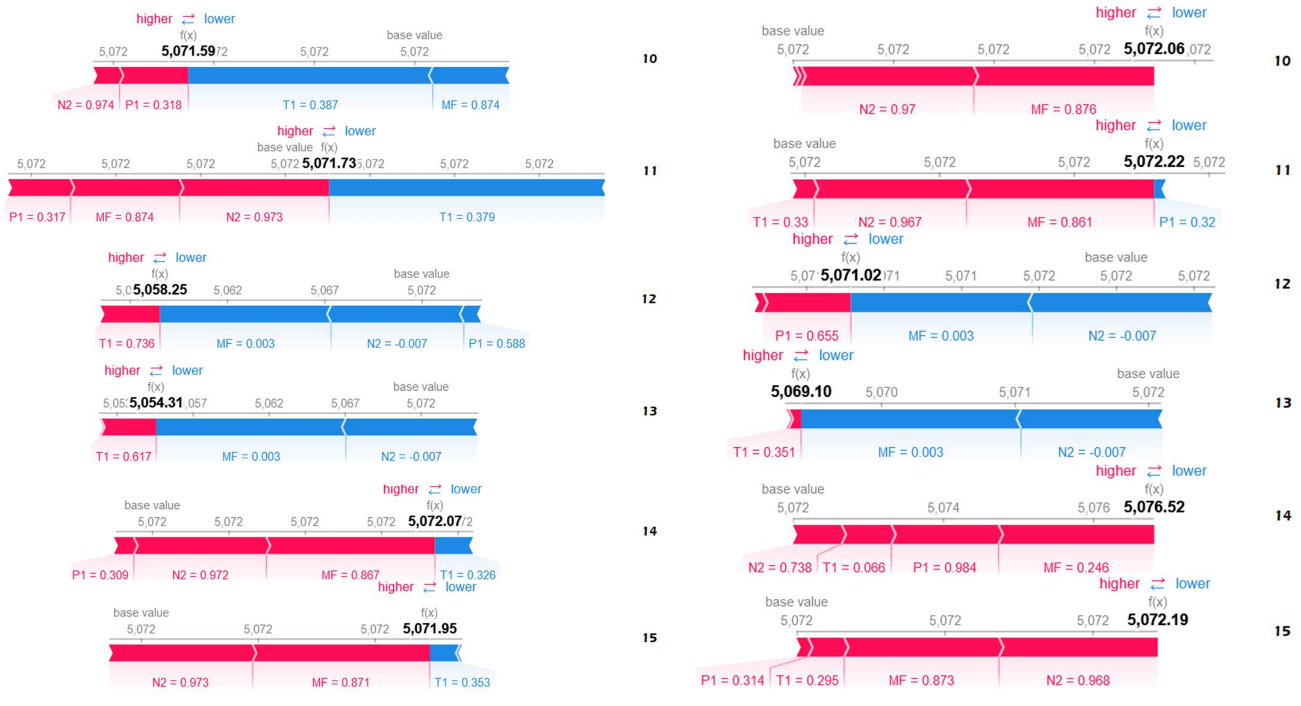
Figure 7. CUSUM chart $\text{Bayes\_LSTM}_{\text{N1}}$ predictions for (**a**) anomaly 20/03/18-21/03/18 and (**b**) anomaly 18/09/18-19/09/18.

The force plots for the anomalies predicted from $\text{Bayes\_LSTM}_{\text{N1}}$ are shown in Figure 8. Only instances 10 to 15 are displayed for illustration purpose. The red color signifies that the feature in question is pushing the prediction positively to increase the output value, $f(x)$, while the blue color indicates the contrary. In Figure 8, $f(x)$ relates to N_1's RPM. The length of the colored bar represents its force amplitude or impact on the prediction and the values associated with the features are the normalized value of the features. The base value is the mean of training data outputs. As depicted in Figure 8 as well as Figures A5 and A10 in Appendix 2, N_2 is the anomalous feature due to its negative normalized value.

From the figures mentioned, note that N_1 and N_2 influences initially positive in instance 11, becoming negative in the 12th and 13th instances, except for T_4 prediction. In addition, P_1 and T_1 positive influence grows on the 12th and 13th instances compared to previous instances except for N_1 prediction. In Figure 8, N_1 and N_2 forces become dominant on the 12th and 13th instances, making the predictions to be less than the base value. In Figures A5 and A10, P_1 and T_1 are generally the major forces, causing the outputs to be greater than the base value. From the illustrations, observe that most features assert positive impact, pushing the output value higher.

3.2 Case Study 2: Turbofan Engines Failure Prognostic

The NASA turbofan run to failure FD001 dataset, produced by the Nasa Prognostic Centre (from Ames Research Centre), was exploited for the prognostic study [56]. This synthetic time series data were generated by modeling a variety of operational scenarios and inserting defects with diverse degrees of deterioration. The original data comprises of training, testing, and true RUL for 100 turbofan engines as summarized in Table 6. Thus, there are 100 turbofan records, referring to turbofans health that declined until breakdown after a given cycle, or failure start point (FSP). Note that 21 sensor signals, described in Table A2 in Appendix 1, working per cycle and three operating conditions (OC) form the recorded data. The OC corresponds to diverse operating regimes, combination of altitude, throttle resolver angle, and mach number that conceal the extent of degradation of each turbofan. On top of this, high-level noise is blended to the dataset [44].



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

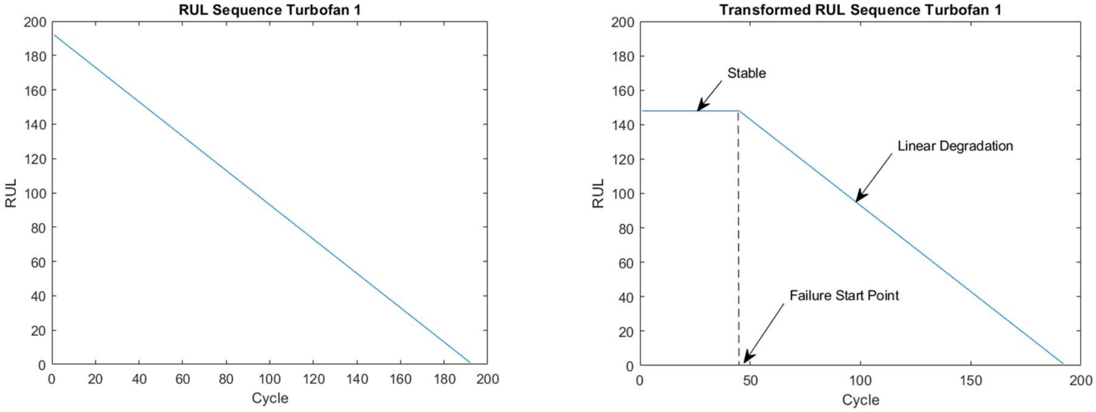
Figure 8. Force plots for (a) anomaly 20/03/18-21/03/18 and (b) anomaly 18/04/18-19/04/18.

Table 6. FD001 dataset summary.

Dataset	Fault Mode	OC	Training Data	Testing Data
#1	1	2	100	100

As in case study 1, we first preprocess the data. Out of the 21 signals, only 14 sensors whose signals trend are strictly monotonic are selected as they best represent degradation contrary to irregular and unchanged signals. The total inputs, including three OC's, are 17.

A piece-wise linear degradation assumption is adopted, where the RUL is assumed stable before the FSP and decreased linearly thereof until failure. In the initial phase, the RUL is equal to the value of the recorded signal's last cycle and decreases linearly as illustrated for Turbofan 1 in Figure 9(a) without the FSP. Then, the CUSUM method with C equals to 5 standard deviations are used to calculate the FSPs of the signals of the concerned turbofan. The mean of these FSPs is set as the FSP of the turbofan. The combination of linear degradation obtained earlier and the FSP results in the final RUL sequence as shown in Figure 9(b). The obtained RULs are limited at 50 to ease model's generalization.



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

Figure 9. RUL targets for Turbofan 1 for (**a**) Linear degradation without FSP and (**b**) Transformed and final RUL targets.

Note that some testing data with long sequence lengths are associated with very small true RULs that differed from the characteristic of the training data. Hence, it is anticipated that the model to perform more poorly on these 'abnormal' data. The prognostic results of turbofan 1 and turbofan 18 are examined as the former data's characteristic bore similarity to the training data's input-output nature while the latter resembles the abnormal data's trait.

Next, we provide the results with 100% features. The RMSE and score results with aleatoric and epistemic uncertainties are presented in Table 7.

Table 7. RMSE results with AU and EU

RMSE with	RMSE with	Score with	Score with
Aleatoric Uncertainty	Epistemic Uncertainty	Aleatoric Uncertainty	Epistemic Uncertainty
17.94	18.41	1025.31	1231.10

The 3D representation of turbofan 1 prognostic with aleatoric uncertainty is shown in Figure 10 to provide the full picture of the modeling. As noted in this illustration, the range of prediction or uncertainty decreased along the cycle, signaling growing model's confidence in its prediction. For the rest of the work, only the 2D presentations are shown.

The 2D depictions of turbofan 1 and turbofan 18 with aleatoric and epistemic uncertainties are presented in Figure 11. Looking at the aleatoric uncertainty rolling standard deviation slope of each prediction, one can observe decreasing trend for turbofan 1 and the contrary for turbofan 18. Hence, the model expresses increasing confidence in the former and decreasing confidence in the latter one. The different aleatoric uncertainty outcomes are translated by the model's prognostic outputs that show better performance for turbofan 1 than for turbofan 18. In Figure 11, observe that the RUL prediction with aleatoric uncertainty agrees with the true RUL in the early cycle before showing degradation and failure earlier than the true RUL curve, which is a demanded quality for prognostic modeling. The prediction oscillates at the end of the degradation phase before stabilizing at the failure stage. Meanwhile, a small gap separates the RUL prediction with aleatoric and epistemic uncertainties during the early cycle before both seemingly coincide during the degradation phase onward until failure. This is not the case for turbofan 18, where both prognostics are way off from the true RUL.

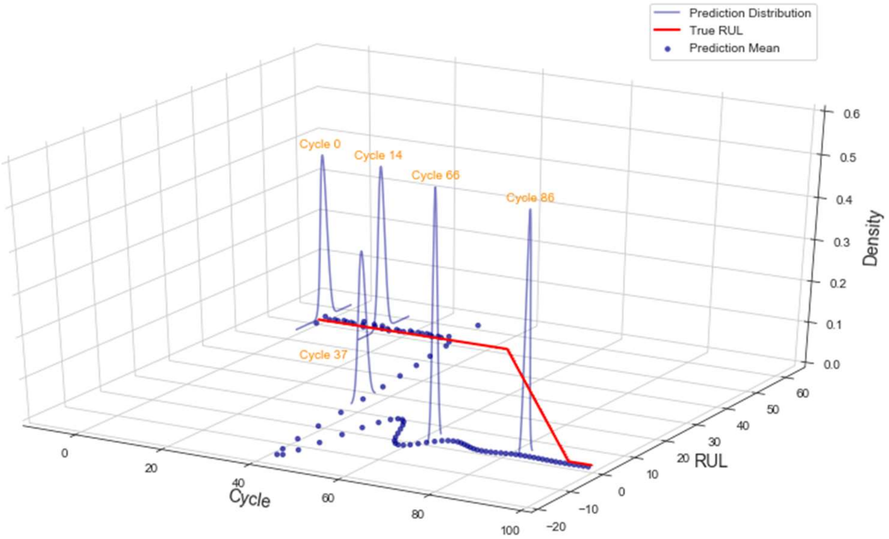
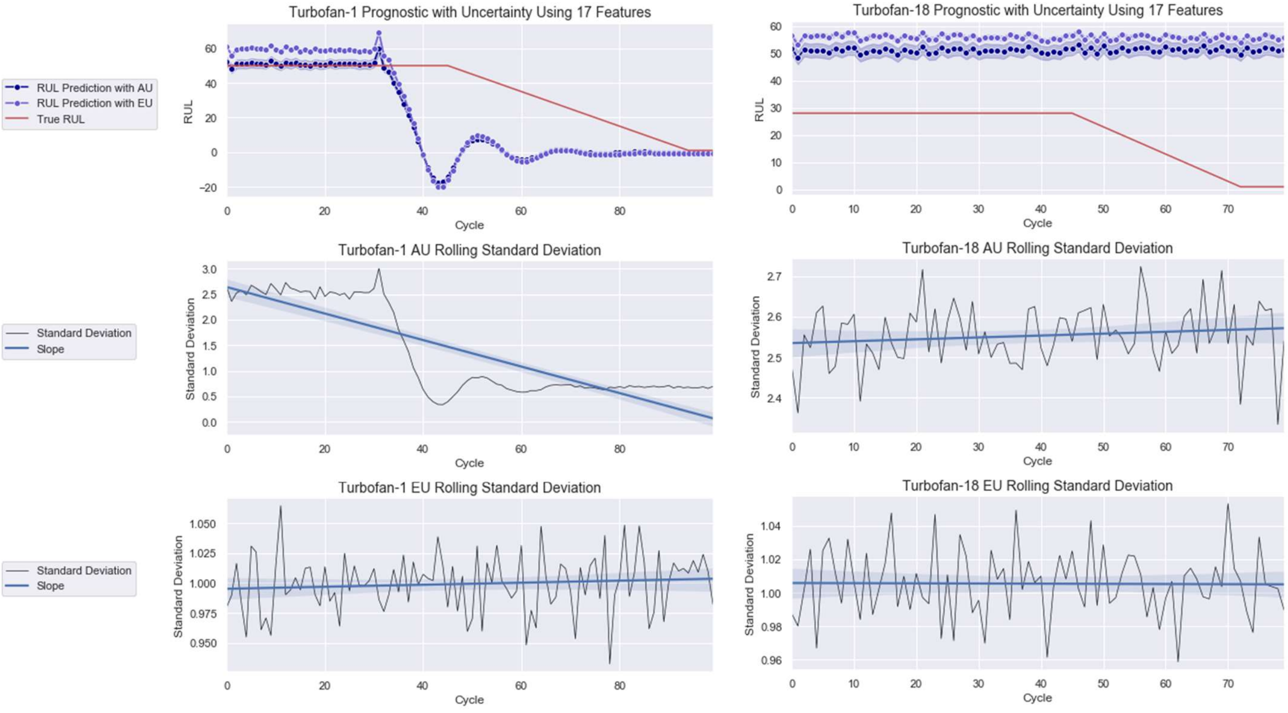


Figure 10. 3D rendering of turbofan 1 prognostic.

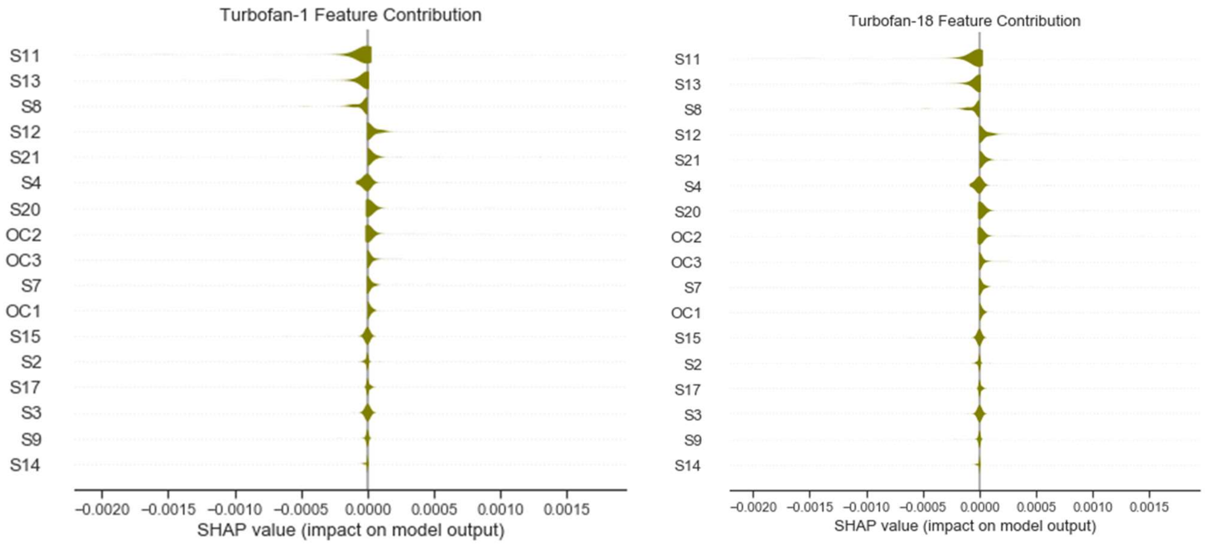


(a) **(b)**

Figure 11. Prognostic modeling with 17 features for (**a**) Turbofan 1 and (**b**) Turbofan 18.

The global explanation for 100% features is provided next. The feature contributions and their directions, issued from f_x' explanation model, are presented in the summary plots in Figure 12. Both plots seem similar, but if they differed, one should prioritize choosing the more confident prediction, in this case, turbofan 1.

doi:10.20944/preprints202109.0034.v3



(a) **(b)**

Figure 12. Summary plots 17 features for prognostic for (**a**) Turbofan 1 and (**b**) Turbofan 18.

Though the top contributing features influenced the predictions more negatively, most of the features had positive impact on the estimates. The features, according to their contributing power, are ordered in Table 8. Note that 75% of the original features, or 13 features that are selected to improve the prognostic modeling, are shown in Italic characters.

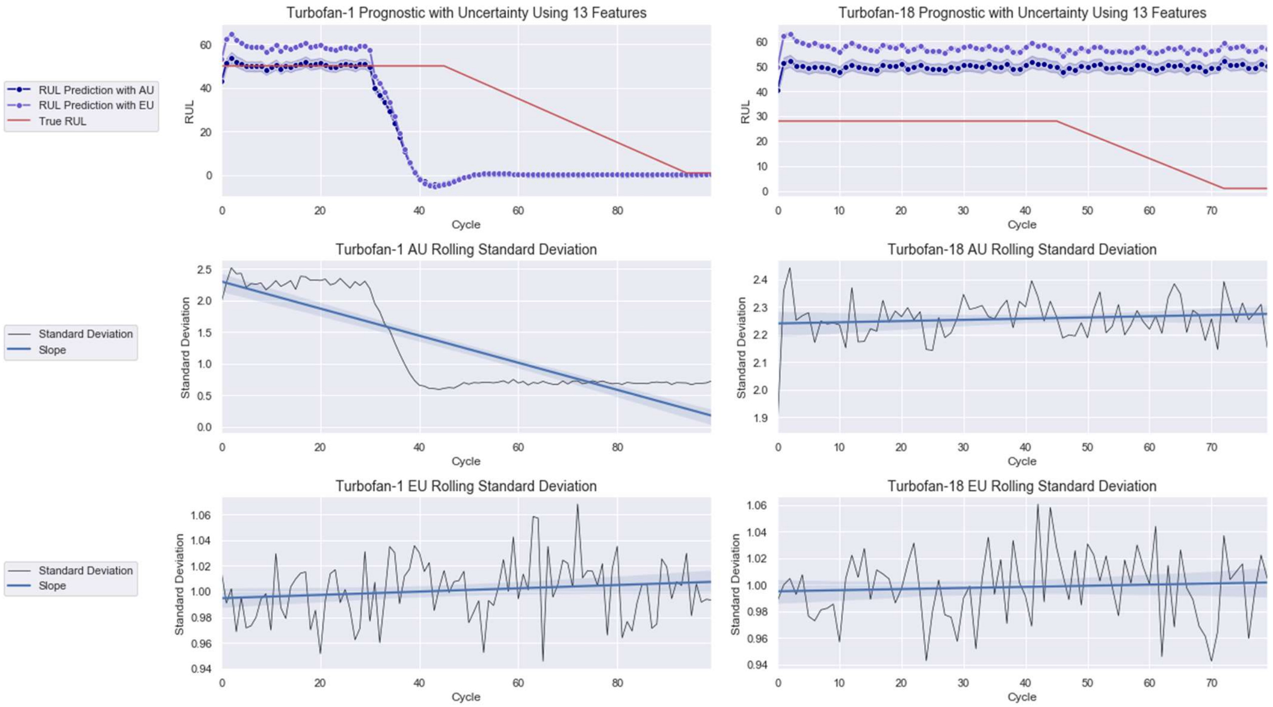
Table 8. 17 Features contributions according to contribution order.

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

Now, the performance and prognostic results with 75% features are reported. The RMSE and score outcomes with the selected features are presented in Table 9. As observed, the RMSE results with aleatoric and epistemic uncertainties show drastic improvement from the previous results, with the score, however, being worse. Outcomes for turbofan 1 and turbofan 18 are depicted in Figure 13(a) and Figure 13(b) respectively. The same manifestation of aleatoric uncertainty slopes trends as previous results is observed, matching the prognostic outcomes. Turbofan 1 modeling shows improvement as the oscillation at the end of the degradation phase decreases before stabilizing in the failure phase. The aleatoric uncertainty level for turbofan 18 improves in general from previous result. In the global explanation for 75% features, the features contributions, and their directions, which are mostly having positive impacts in the predictions, are presented in summary plots in Figure 14.

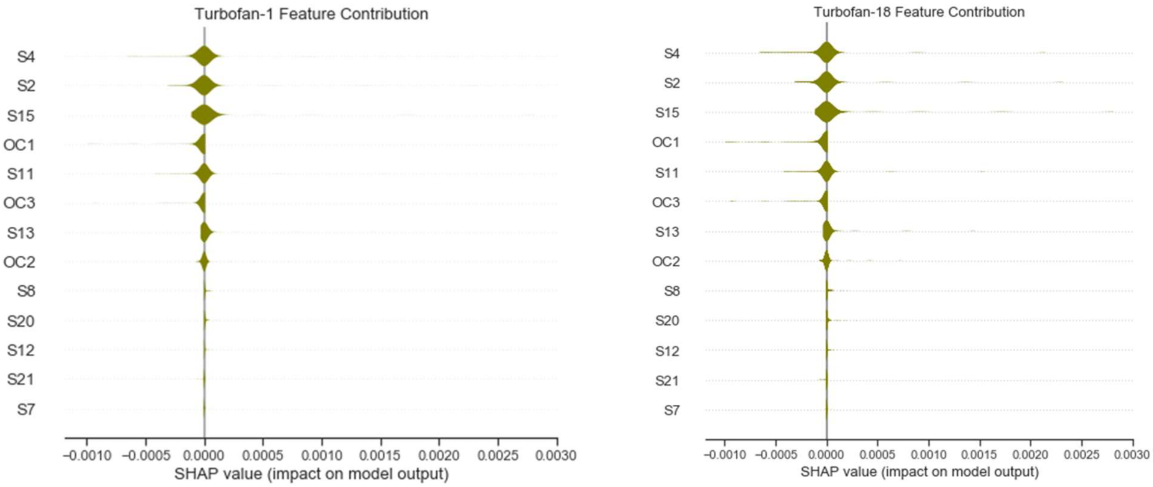
Table 9. RMSE results with aleatoric and epistemic uncertainties

RMSE with Aleatoric Uncertainty	**RMSE with Epistemic Uncertainty**	**Score with Aleatoric Uncertainty**	**Score with Epistemic Uncertainty**
14.59	15.87	431.99	594.88



(a) **(b)**

Figure 13. Prognostic modeling with 13 features for (**a**) Turbofan 1 and (**b**) Turbofan 18.



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

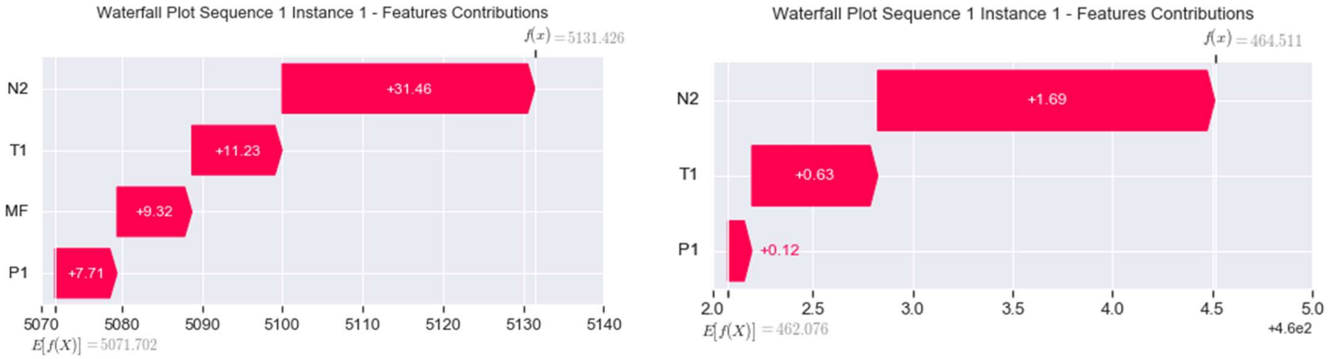
Figure 14. Summary plots 13 features for prognostic for (**a**) Turbofan 1 and (**b**) Turbofan 18.

When conducting the performance comparison, only the best RMSE and score with the aleatoric uncertainty obtained previously are compared with the best published works according to the year of publication. As presented in Table 10, the results are on par with these methods, with the prognostic score occupying the top position amongst all the techniques. Predictions issued from f'_x and f''_x for T_4 are presented in Figure 16.

Table 10. RMSE comparison with published methods

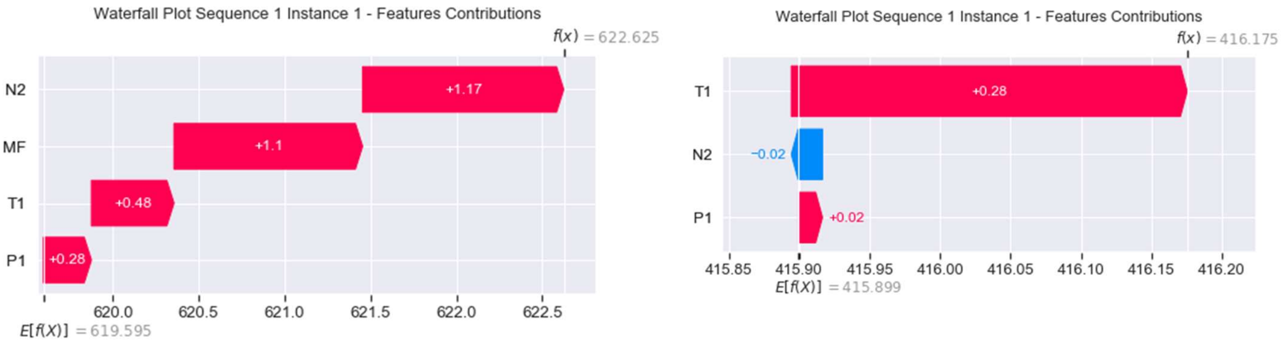
Year	Methods	RMSE	Score
2017	VAE+RNN	14.80	419
2018	CNN+FNN	12.61	274
2019	CNN+LSTM-FNN	12.56	231
2021	Proposed method	14.59	431

For the illustration purpose, we carry out an explanation evaluation using the Waterfall plots of the real-world gas turbine's T_4 prediction. Note that predictions issued from f'_x and f''_x are presented in Figures 15 and 16 respectively. The feature removed is N_1.



(a) **(b)**

Figure 15. Waterfall plots f'_x predictions for (**a**) $f'_x(v')$ and (**b**) $f'_x(v'_{\setminus N_1})$.



(a) **(b)**

Figure 16. Waterfall plots f''_x predictions for (**a**) $f''_x(v')$ and (**b**) $f''_x(v'_{\setminus N_1})$.

Next, we calculate the local accuracy. Applying the expression given in (12) on Figure 15(a), we obtain:

- $\sum_{j=1}^{N} \Phi_j$ = 31.46 + 11.23 + 9.32 + 7.71 = 59.72.
- $f(x) - \mathrm{E}_x(\hat{f}(X))$ = 5131.426 – 5071.702 = 59.724 ≃ 59.72.

Now, applying the expression given in (12) on Figure 15(b), we get:

- $\sum_{j=1}^{N} \Phi_j$ = 1.69 + 0.63 + 0.12 = 2.44.
- $f(x) - \mathrm{E}_x(\hat{f}(X))$ = 464.511 – 462.076 = 2.435 ≃ 2.44.

Then, applying the expression given in (12) on Figure 16(a), we have:

- $\sum_{j=1}^{N} \Phi_j$ = 1.17 + 1.1 + 0.48 + 0.28 = 3.03.
- $f(x) - \mathrm{E}_x(\hat{f}(X))$ = 622.625 – 619.595 = 3.03.

Hence, applying the expression given in (12) on Figure 16(b), we reach:

- $\sum_{j=1}^{N} \Phi_j = 0.28 - 0.02 + 0.02 = 0.28$.
- $f(x) - \mathrm{E}_x(\hat{f}(X)) = 416.175 - 415.899 = 0.276 \simeq 0.28$.

Observe that the calculations confirm the local accuracy property of the explanation. Next, we evaluate the consistency property similarly, that is, applying the expression given in (13) on Figures 15 and 16, we obtain that:

- $f_x'(v') = 5131.426$; $f_x'(v'_{\setminus N_1}) = 464.511$; $f_x'(v') - f_x'(v'_{\setminus N_1}) = 4{,}666.915$, and
- $f_x''(v') = 622.625$; $f_x''(v'_{\setminus N_1}) = 416.175$; $f_x''(v') - f_x''(v'_{\setminus N_1}) = 206.45$.
- Thus, $f_x'(v') - f_x'(v'_{\setminus N_1}) > f_x''(v') - f_x''(v'_{\setminus N_1})$.

Applying the expression given in (14) on Figures 15 and 16, we get that:_

- $\Phi_j(f', x) = 9.32$; $\Phi_j(f'', x) = 1.1$, thus $\Phi_j(f', x) > \Phi_j(f'', x)$

Therefore, once again, the calculations confirm now the consistency property of the explanation.

4. Discussion

The insights gained from the study as well as its limitation and future opportunities are elaborated in this section.

4.1 Anomaly Detection

This paper firstly proposed an anomaly detection framework based on deep learning aleatoric uncertainty and CUSUM changepoint detection. Bayesian deep learning models, capable of generating uncertainties, were trained using only healthy data. Thus, it is expected that the aleatoric uncertainty, which is influenced by the input data quality, is stable for healthy data and shows abnormality when encountering abrupt anomalies. As demonstrated in Figures 7, A4 and A9, the strategy yielded 87.5% success or 7 out of 8 anomalies detected in real-world gas turbine dataset. The achievement was partly due to the minimization of aleatoric uncertainty by the mean of singular value decomposition denoising. As observed, the aleatoric uncertainty around healthy data prediction is so small because of denoising except for Figure A4(a), where aleatoric uncertainty variation was too big to be minimized by singular value decomposition denoising. Without this operation beforehand, the anomaly spikes risk is invisible from the rest of the prediction's aleatoric uncertainty, hindering effective anomaly detection.

The force plot for local explanation uncovers the dynamic caused by N_2 anomaly to the predictions. Note that N_1 seems to follow N_2 behavior, changing force direction from positive to negative and dragging the prediction lower. The two features influence seems amplified in instance 13 due to the consecutive N_2 anomaly. Also, observe that P_1 and T_1 positive influences rose, increasing the prediction. It is also learnt that most features are exerting positive impact that pushed the output value higher. Nonetheless, whether N_2 influenced N_1, P_1 and T_1 is not certain and could be investigated by other means such as partial dependence plot in the future.

Since the investigation only focused on abrupt anomalies, it is recommended to apply the technique on long consecutive anomalies and examine the generated explanation.

Additionally, this work defined the calculation of control limit C using aleatoric uncertainty level calculations. However, one can see from Figures 7, A4 and A9 in Appendix 2 that the anomalies were only identified on the 13th or higher instance, even when the disturbances had already started from the 12th instance. Faster detection could be possible with a proper definition of control limit C. One could lower the limit but a risk of having more false alarms exists, especially when the range of aleatoric uncertainty is important such as in Figure A4(a). As can be seen in this figure, using 1/3 of C as control limit to identify anomaly on the 12th instance leaded to many erroneous detections.

4.2 Failure Prognostic

Secondly, the deep learning model was employed for failure prognostic purpose. This time, it was fed with both the healthy and failure recorded data. The aleatoric uncertainty in this task served as confidence indicator, expressing the uncertainty of the model in its output. Based on the graphical results in Figures 11, 13 and 15, the aleatoric uncertainty indicator matched all the prognostic modelings, where it increases when the prediction was bad and decreases when the prognostic was good. This feature is vital in failure prediction especially in the absence of true RUL. Then, practitioners could judge the quality of the prediction for important decision making.

The global explanation in the form of summary plot helped to improve the performance of deep learning model. By only using the best contributing features, the RMSE result obtained was on par with the best published techniques in this problem. Interestingly, all the OCs played important role in the prediction and made it to the final selection. While the results coming from frequentist models may seems a bit better, this is mainly due to their more complex structures as their designations suggest. The Bayesian deep learning model employed in this work only consisted of a single LSTM and dense layer that limits its nonlinearity modeling power compared to the other methods. Furthermore, the frequentist models could never be utilized in real-life applications and its usage scope is limited to experimental purpose as they are devoid of uncertainty quantification. Hence, one could incorporate more complex network to the existing Bayesian deep learning model in the future to enhance its performance. Moreover, feature selection could be done in another angle where features are chosen according to their influence direction rather than their contributing power to investigate the effect on the performance.

The aleatoric uncertainty indicator provided another dimension in explanation, where it indicated which prediction was reliable before explanation when the XAI approach takes place. This feature enabled the differentiation between explanation of reliable outputs and unreliable ones, helping users and developers to obtain a deeper insight into the artificial intelligence decision. This distinction facilitates user to prioritize on explaining either one of the output types for fast decision making. Time is always a natural constraint in this situation. The prioritization in turn will lead to resources optimization for the task at hand. Furthermore, the distinction aided in selecting which global explanation to use for improving the model. Obviously, it would be wiser to choose explanation from a more confident prediction than a lesser one.

One can notice that the epistemic uncertainty level in all the plots hovered around the same range. This is normal as the uncertainty for the weights is fixed once the training was done.

4.3 Safeguarding Security and Explanation Evaluation

Uncertainty quantification excels in minimizing adversarial example risk. This issue arises when new and unseen data, either unintentionally generated or engineered by attackers is fed to the network. Adversarial example could fool deep learning models. Obviously, frequentist models are unable to detect this abnormality. Bayesian model, however, can signal its presence in the form of rising uncertainty. While this work focused on mechanical failure assumption, it is equally important to investigate failure due to adversarial example as well.

The explanation generated conforms to the local accuracy and consistency properties. The former one also equals to efficiency nature of the Shapley values. Certifying the latter one, also it means justifying the symmetry and additivity qualities of the Shapley values. The first characteristic asserts that the Shapley values of two features should be equal if their contributions to all probable coalitions are even. The final attribute denotes that for an ensemble prediction, for a specific feature, one can calculate the Shapley value of the feature in each individual ensemble, averaging them, and getting the Shapley value for the feature for the whole ensemble.

While the explanation fulfils several demanded general qualities, the need to evaluate explanation based of PHM criteria such as security and safety, cost, and time are still present. This aspect is also echoed in [39]. Therefore, it is crucial for PHM-XAI researchers to develop explanation metrics satisfying PHM needs.

5. Conclusions

Opacity of artificial intelligence models constitutes an operational and legal risks that could potentially derail investments of intelligence artificial in the energy and industrial sectors. To promulgate the assimilation of artificial intelligence in real world prognostic and health management applications, this article tackles the challenges afflicting PHM-XAI domain, in specific, lack of explanation assessment and uncertainty quantification. Prognostic and health management tasks relating to anomaly detection and failure prognostic of gas turbine engine were investigated. The Shapley additive explanations model agnostic approach was employed to generate local and global explanations from a Bayesian deep learning model. The former one for the anomaly explanation while the latter one for the failure prediction. The global explanation was also exploited to improve the prognosis performance. The deep learning model was able to predict with uncertainty whose trend served as anomaly marker that changes intensely with abnormal data. The anomaly detection strategy succeeded in identifying seven out of eight available abnormalities, while the best selected features from the global explanation enhanced prognostic performance to be on par with the best results in the problem. The Shapley additive explanations were finally validated with the local accuracy and consistency characteristics of explanation.

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: Not applicable.

Data Availability Statement: The data and code presented in this study are openly available in https://github.com/AhmadNor (accessed on 9 January 2022).

Acknowledgments: The authors would like to thank Petronas for the data used in this work.

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

Appendix 1: Tables

Table A1. BayesOpt hyperparameters ranges.

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

Table A2. Turbofan datasets sensors description.

Sensor	Description	Unit
S1	Total temperature fan inlet	^{0}R
S2	Total temperature at low pressure compressor (LPC) outlet	^{0}R
S3	Total temperature at high pressure compressor (HPC) outlet	^{0}R
S4	Total temperature at low pressure turbine (LPT) outlet	^{0}R
S5	Pressure at fan inlet	psia
S6	Total pressure in bypass-duct	psia
S7	Total pressure at HPC outlet	psia
S8	Physical fan speed	rpm
S9	Physical core speed	rpm
S10	Engine pressure ratio (P50/P2)	N/A
S11	Static pressure at HPC outlet	psia
S12	Ratio of fuel flow to Ps30	Pps/psi
S13	Corrected fan speed	rpm
S14	Corrected core speed	rpm
S15	Bypass ratio	N/A
S16	Burner fuel-air ratio	N/A
S17	Bleed enthalpy	N/A
S18	Demanded fan speed	rpm
S19	Demanded corrected fan speed	rpm
S20	HPT coolant bleed	lbm/s
S21	LPT coolant bleed	lbm/s

Appendix 2: Figures

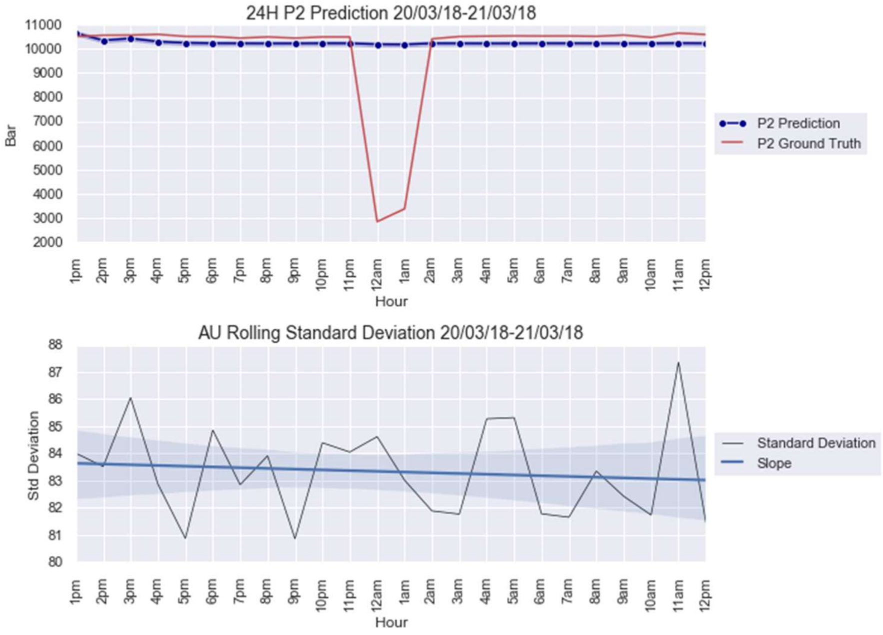


Figure A1. Bayes_LSTM$_{P2}$ anomaly modeling 20/03/18-21/03/18.

doi:10.20944/preprints202109.0034.v3

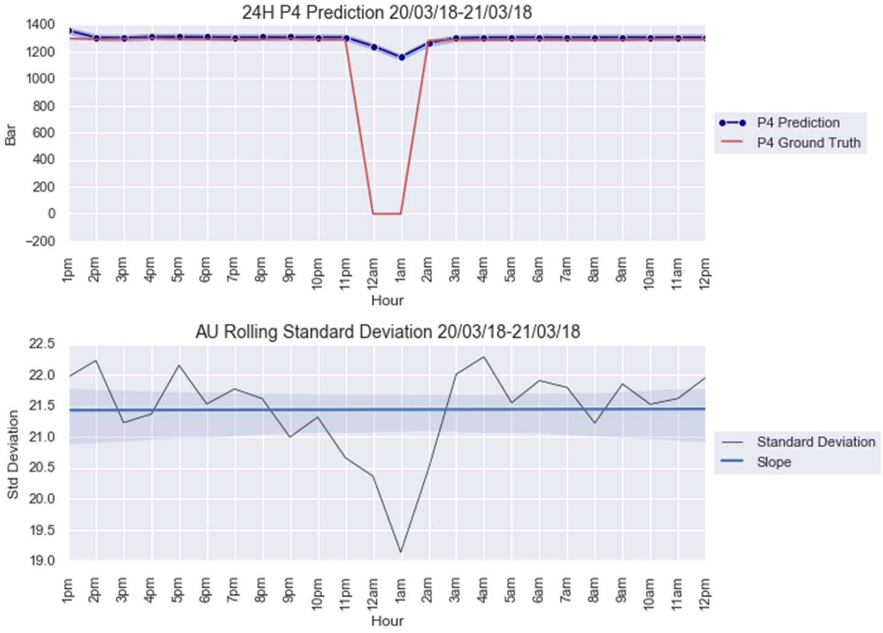


Figure A2. Bayes_LSTM$_{P4}$ anomaly modeling 20/03/18-21/03/18.

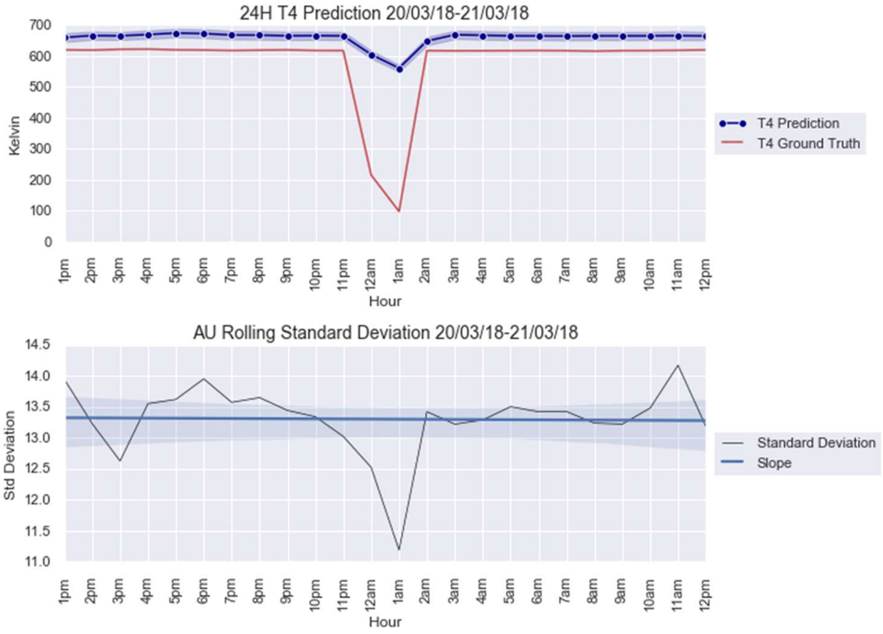
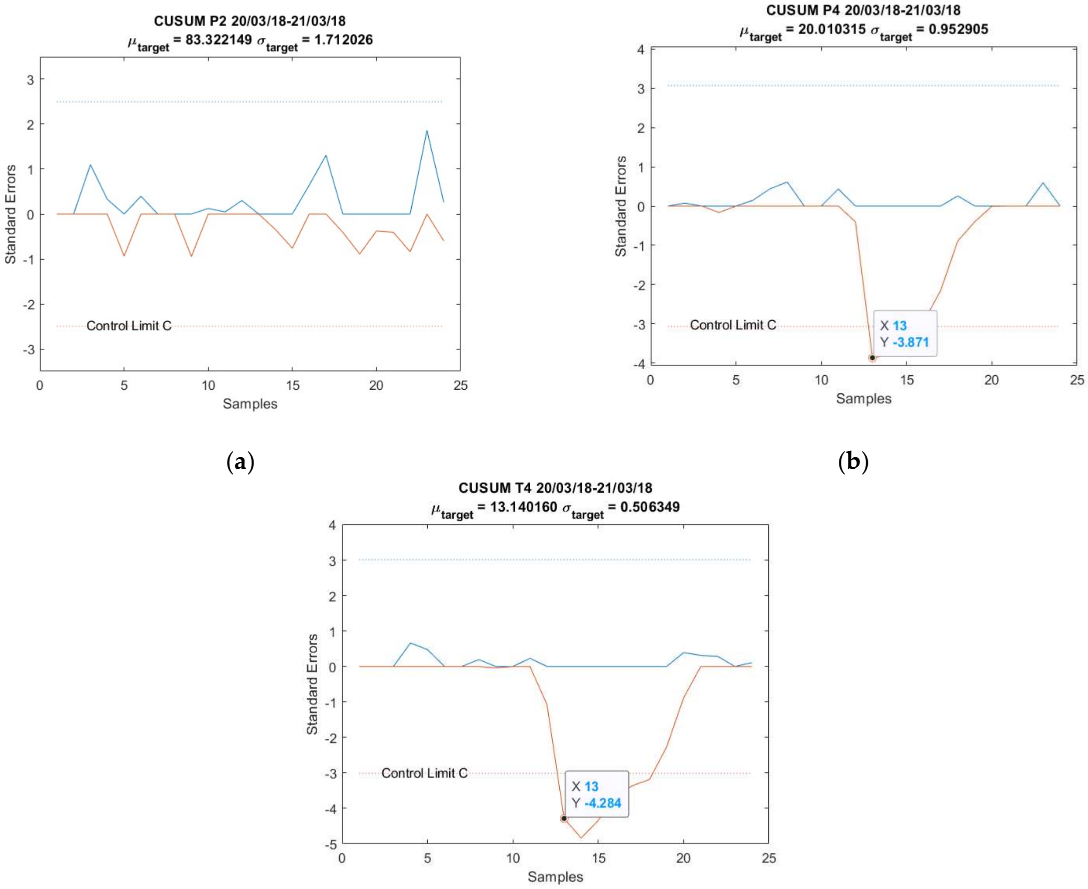


Figure A3. Bayes_LSTM$_{T4}$ anomaly modeling 20/03/18-21/03/18.



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

(**c**)

Figure A4. CUSUM charts anomaly 20/03/18-21/03/18 for (a) anomaly from $\text{Bayes\_LSTM}_{\text{P2}}$; (b) anomaly from $\text{Bayes\_LSTM}_{\text{P4}}$; and (c) anomaly from $\text{Bayes\_LSTM}_{\text{T4}}$

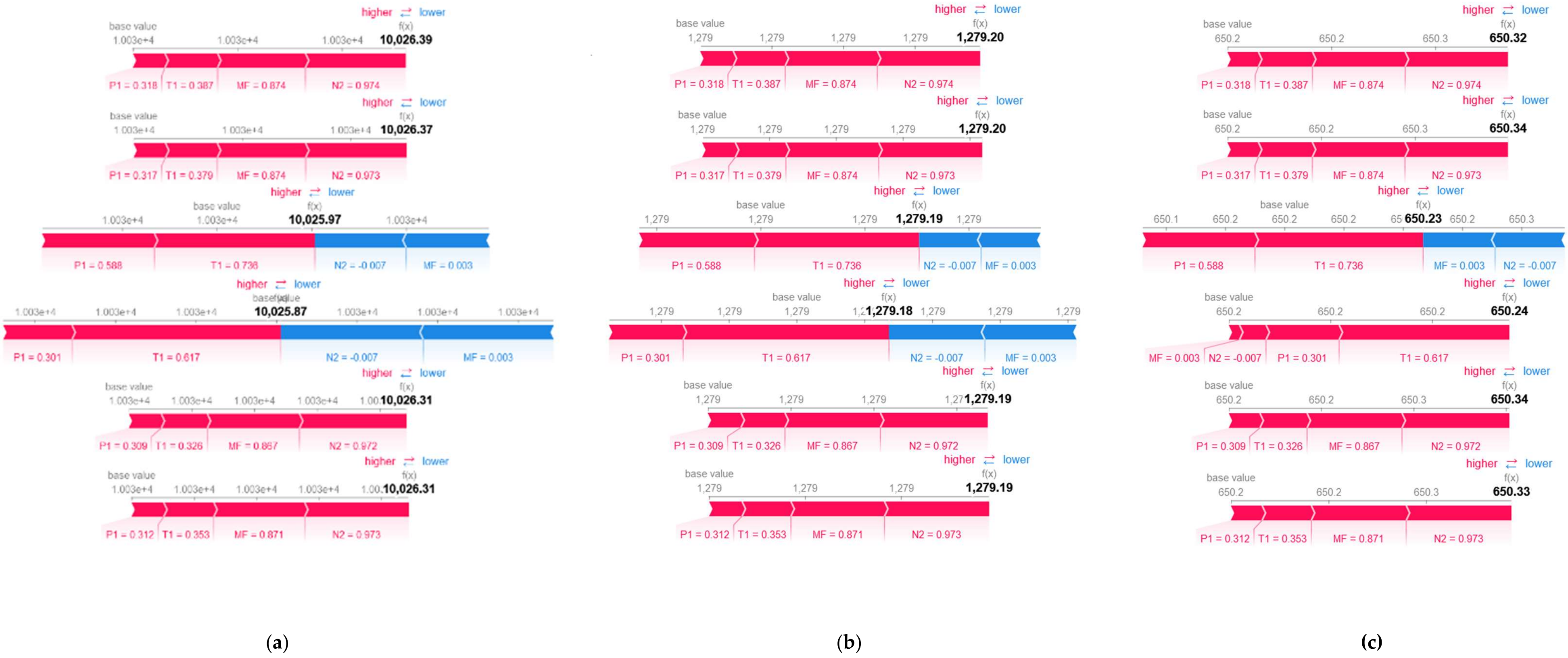


Figure A5. Force plots anomaly 20/03/18-21/03/18 for (a) anomaly from Bayes_LSTM$_{P2}$; (b) anomaly from Bayes_LSTM$_{P4}$; and (c) anomaly from Bayes_LSTM$_{T4}$.

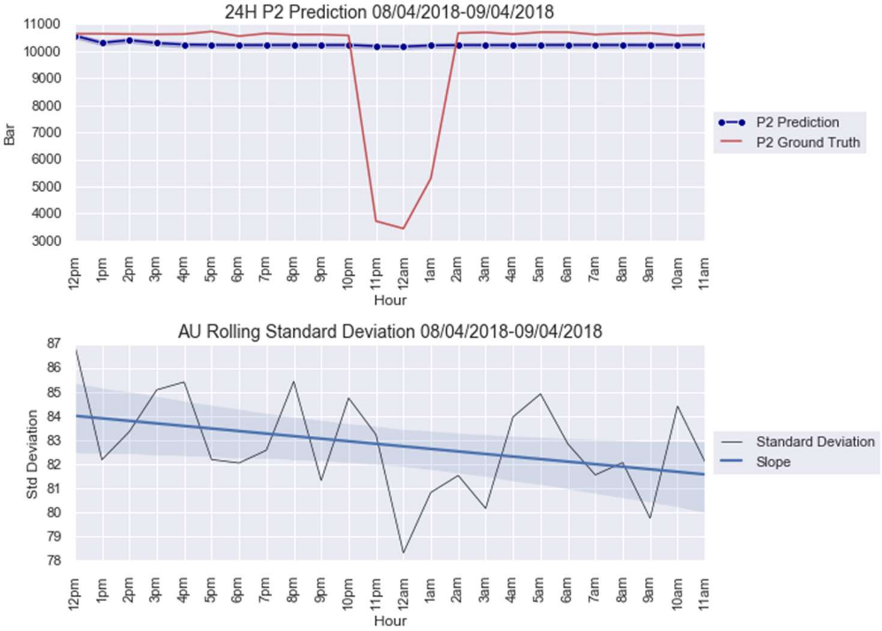


Figure A6. Bayes_LSTM$_{P2}$ anomaly modeling 18/04/18-19/04/18.

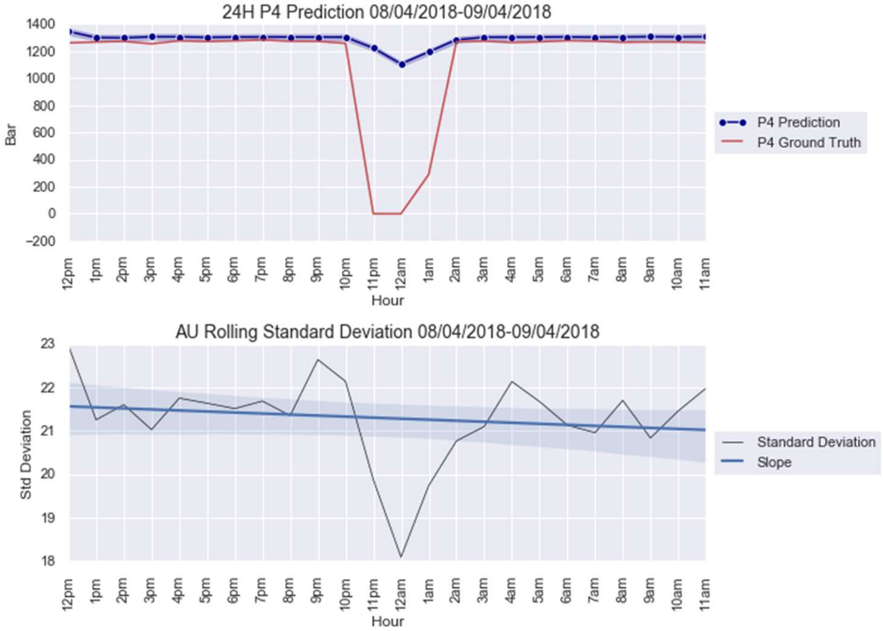


Figure A7. Bayes_LSTM$_{P4}$ anomaly modeling 18/04/18-19/04/18.

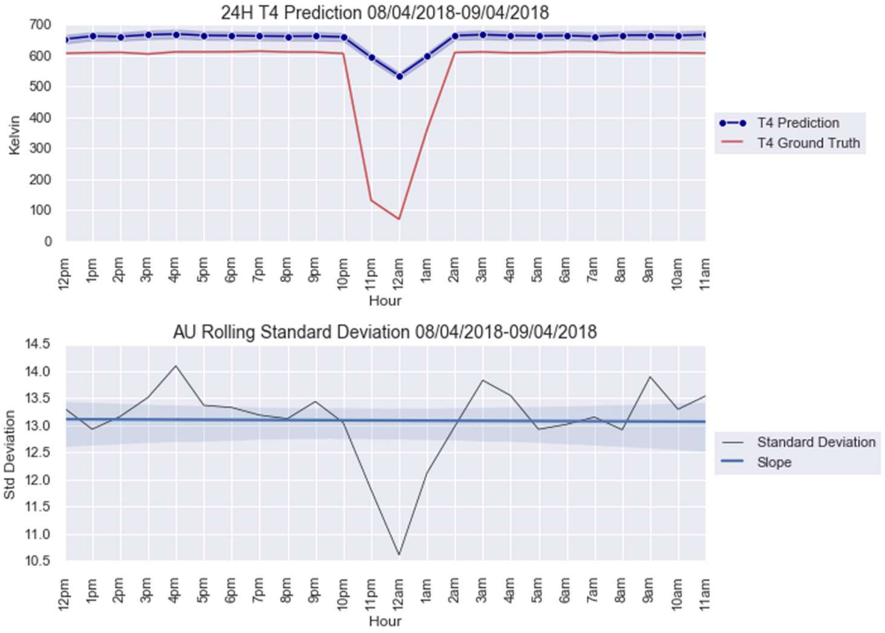
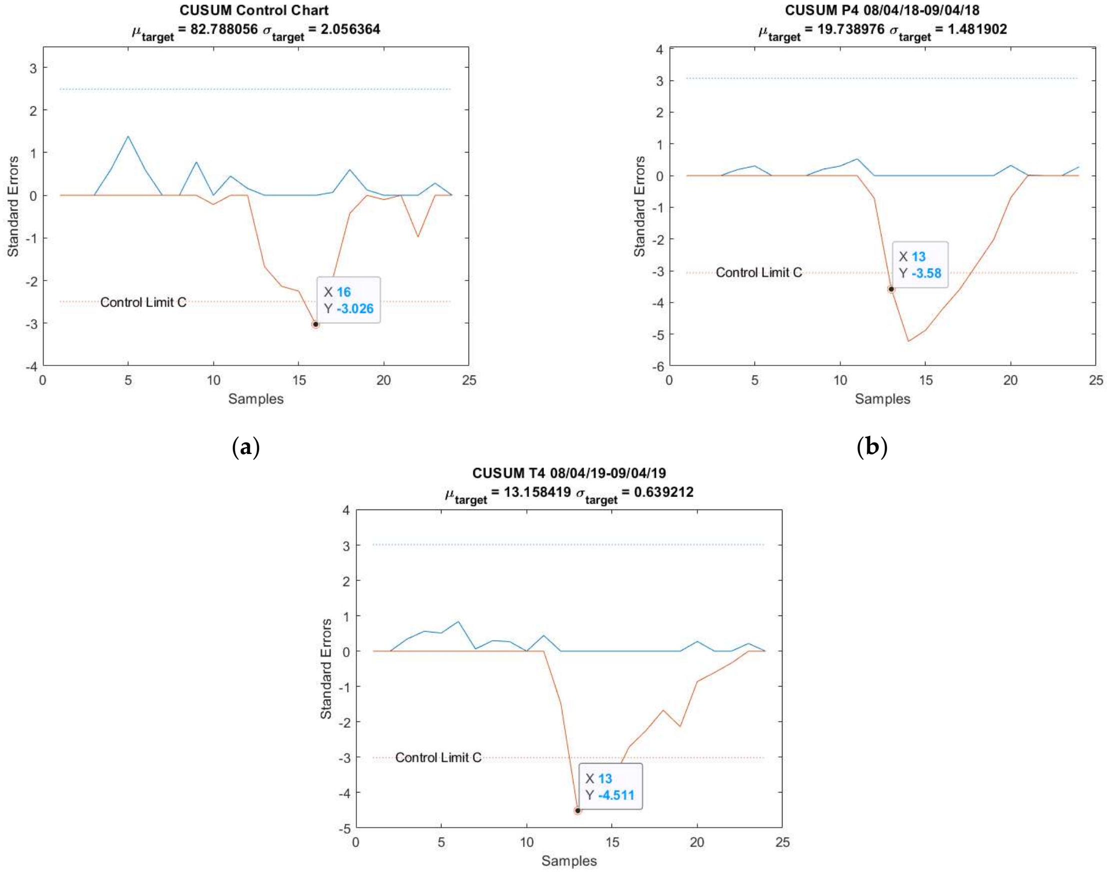


Figure A8. Bayes_LSTM$_{T4}$ anomaly modeling 18/04/18-19/04/18.



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

(**c**)

Figure A9. CUSUM charts anomaly 08/04/18-09/04/18 for (a) snomaly from Bayes_LSTM$_{P2}$; (b) anomaly from Bayes_LSTM$_{P4}$; and (c) anomaly from Bayes_LSTM$_{T4}$.

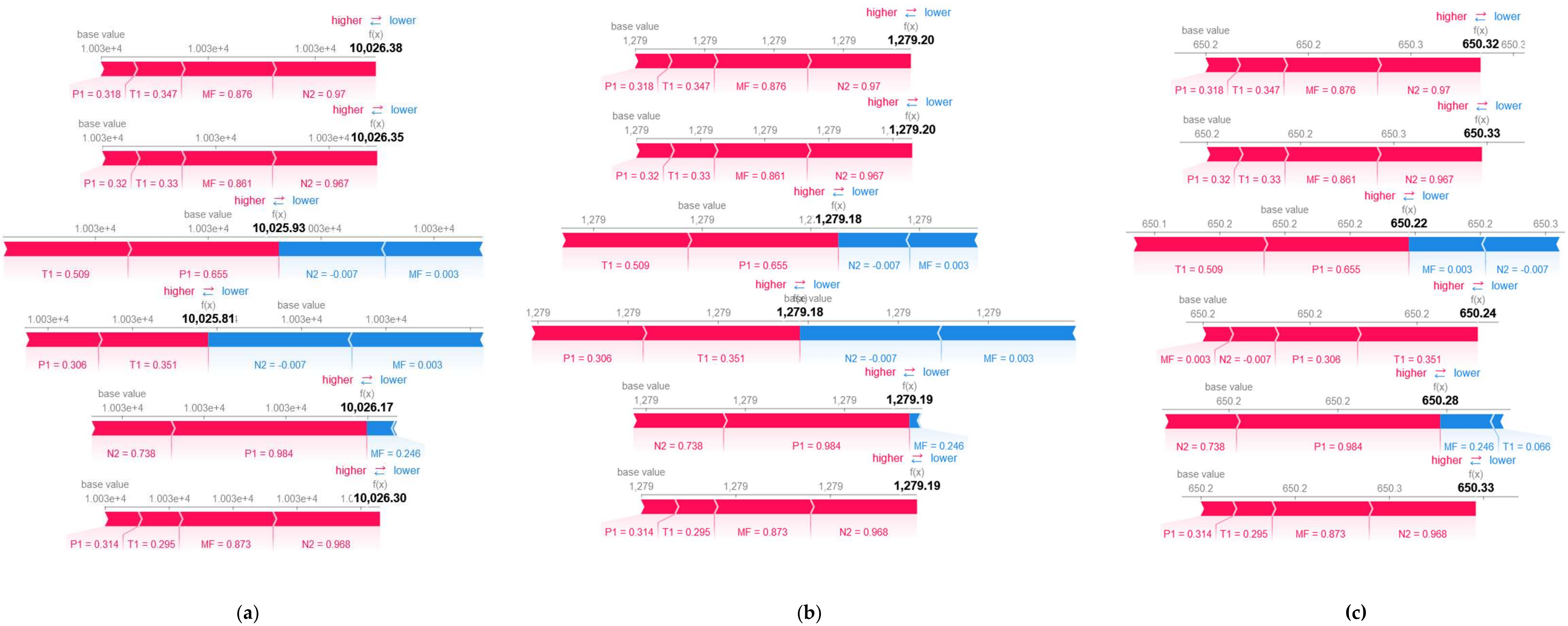


Figure A10. Force plots anomaly 08/04/18-09/04/18 for (a) anomaly from Bayes_LSTM$_{P2}$; (b) anomaly from Bayes_LSTM$_{P4}$; and (c) anomaly from Bayes_LSTM$_{T4}$.

References

1. Monett, D., Lewis, C. (2018). Getting Clarity by Defining Artificial Intelligence—A Survey. 10.1007/978-3-319-96448-5_21.
2. EU Commission's High-Level Expert Group on AI. (2018). A Definition of AI: Main Capabilities and Disciplines. (Retrieved on 22 Dec 2021 from: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf).
3. EU Commission. (2020). Critical industrial applications: report on current policy measures and policy opportunities. Retrieved on 22 Dec 2021 from: https://op.europa.eu/en/publication-detail/-/publication/fe5a340a-93fb-11ea-aac4-01aa75ed71a1/language-en.
4. Deloitte (2019). Scenarios and potentials of AI's Commercial Application in China. China AI Industry Overview. (Retrieved on 22 Dec 2021 from: https://www2.deloitte.com/cn/en/pages/innovation/articles/china-ai-industry-whitepaper-intelligence-driven-by-innovation.html).
5. Amato, G., Behrmann, M., Bimbot, F., Caramiaux, B., Falchi, F., Garcia, A., Geurts, J., Gibert, J., Gravier, G., Holken, H.,et al. (2019). AI in the media and creative industries.
6. Petit, N. (2018). Artificial Intelligence and Automated Law Enforcement: A Review Paper http://dx.doi.org/10.2139/ssrn.3145133
7. Raimundo, R, and Albérico, R. (2021). The Impact of Artificial Intelligence on Data System Security: A Literature Review, Sensors 21, no. 21: 7029. https://doi.org/10.3390/s21217029
8. Bates, D.W., Levine, D., Syrowatka, A. et al. (2021). The potential of artificial intelligence to improve patient safety: a scoping review. npj Digit. Med. 4, 54. https://doi.org/10.1038/s41746-021-00423-6
9. Qiu, Shilin, Qihe Liu, Shijie Zhou, and Chunjiang Wu. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies, Applied Sciences 9, no. 5: 909. https://doi.org/10.3390/app9050909
10. Momade, M.H., Durdyev, S., Estrella, D. and Ismail, S. (2021), "Systematic review of application of artificial intelligence tools in architectural, engineering and construction", Frontiers in Engineering and Built Environment, Vol. 1 No. 2, pp. 203-216. https://doi.org/10.1108/FEBE-07-2021-0036
11. Buczynski, W., Cuzzolin, F., Sahakian, B. (2021). A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life. Int J Data Sci Anal 11, 221–242 (2021). https://doi.org/10.1007/s41060-021-00245-5
12. Jung D, Choi Y. (2021). Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. Minerals. 2021; 11(2):148. https://doi.org/10.3390/min11020148
13. Deng (2015). Advancements in Deep Learning Theory and Applications: Perspective in 2020 and beyond.
14. Saadat, Md, Shuaib, Muhammad. (2020). Advancements in Deep Learning Theory and Applications: Perspective in 2020 and beyond. 10.5772/intechopen.92271.
15. Sejnowski, Terrence. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences. 117. 201907373. 10.1073/pnas.1907373117.
16. Dai, Zihang, Hanxiao Liu, Quoc V. Le and Mingxing Tan. "CoAtNet: Marrying Convolution and Attention for All Data Sizes." ArXiv abs/2106.04803 (2021): n. pag.
17. Rao, Anand S. and Verweji, Gerard. (2017). Sizing the prize: What's the real value of AI for your business and how can you capitalise? Retrieved on 22 Dec 2021 from: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html.
18. Arnold, Z., I. Rahkovsky and T. Huang. (2020). Tracking AI Investment: Initial Findings from the Private Markets, https://doi.org/10.51593/20190011
19. PwC (2017). Leveraging the upcoming disruptions from AI and IoT. How Artificial Intelligence will enable the full promise of the Internet-of-Thing. Retrieved on 22 Dec 2021 from: https://www.pwc.com/gx/en/industries/tmt/publications/ai-and-iot.html
20. WIPO. (2019). WIPO Technology Trends 2019: Artificial Intelligence. Geneva: World Intellectual Property. Organization. Retrieved on 22 Dec 2021 from: https://www.wipo.int/publications/en/details.jsp?id=4386.
21. Dernis H., Gkotsis P., Grassano N., Nakazato S., Squicciarini M., van Beuzekom B.,Vezzani A. (2019). World Corporate Top R&D investors: Shaping the Future of Technologies and of AI. A joint JRC and OECD report. EUR 29831 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-09670-2 , doi:10.2760/16575, JRC117068
22. Zachary Arnold and Helen Toner. (2021). AI Accidents: An Emerging Threat. https://doi.org/10.51593/20200072
23. McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. AAAI.
24. Karni Chagal-Feferkorn. (2020). AI Regulation in the World: A Quarterly Update October-December 2020, AI + Society Initiative, available at: https://techlaw.uottawa.ca/sites/techlaw.uottawa.ca/files/ai-regulation-in-the-world_2020_q4_final.pdf
25. Gunning, D., Vorm, E., Wang, J.Y. and Turek, M. (2022), DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters e61. https://doi.org/10.1002/ail2.61
26. Streich, J., Romero, J., Gazolla, J., Kainer, D., Cliff, A., Prates, E. et al. Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? Current Opinion in Biotechnology, 2020, 61, 217-225.
27. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J. Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 2020, 3, 26.

28. Tjoa, E., Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 2021 doi:10.1109/tnnls.2020.3027314.
29. Chen, K. et al. Neurorobots as a means toward neuroethology and explainable AI. Frontiers in Neurorobotics 2020, 14, 570308.
30. Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., He, Z. Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. Journal of the American Medical Informatics Association, 2020, 27, 1173-1185.
31. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. (2018). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Volume 58, Pages 82-115, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.12.012.
32. 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.
33. Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy 23, no. 1: 18. https://doi.org/10.3390/e23010018
34. Sheppard, John, Kaufman, Mark, Wilmering, Timothy. (2008). IEEE Standards for Prognostics and Health Management. IEEE Aerospace and Electronic Systems Magazine. 24. 97 - 103. 10.1109/AUTEST.2008.4662592.
35. Thudumu, Srikanth, Branch, Philip, Jin, Jiong, Singh, Jugdutt. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data. 7. 10.1186/s40537-020-00320-x.
36. Guo, Jian, Li, Zhaojun, Li, Meiyan. (2019). A Review on Prognostics Methods for Engineering Systems. IEEE Transactions on Reliability. PP. 1-20. 10.1109/TR.2019.2957965.
37. Gao, Zhiwei, and Xiaoxu Liu. (2021). An Overview on Fault Diagnosis, Prognosis and Resilient Control for Wind Turbine Systems, Processes 9, no. 2: 300. https://doi.org/10.3390/pr9020300
38. Nor, A.K.M., Pedapati, S.R., Masdi, M., (2021). Reliability engineering applications in electronic, software, nuclear and aerospace industries: A 20 year review (2000-2020). Ain Shams Engineering Journal. 12. 10.1016/j.asej.2021.02.015.
39. Nor, A.K.M., Pedapati, S.R., Masdi, M., Leiva, V. (2021). Overview of Explainable Artificial Intelligence for Prognostic and Health Management of Industrial Assets Based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Sensors 21, no. 23: 8020. https://doi.org/10.3390/s21238020
40. Ding, Peng, Jia, Minping, Wang, Hua. (2020). A dynamic structure-adaptive symbolic approach for slewing bearings' life prediction under variable working conditions. Structural Health Monitoring. 20. 10.1177/1475921720929939.
41. Kraus, M., Feuerriegel, S. (2019). Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. Decision Support Systems. 125. 113100. 10.1016/j.dss.2019.113100.
42. Antonio Luca, Alfeo, Cimino, Mario Giovanni C.A., Manco, Giuseppe, Ritacco, Ettore, Vaglini, Gigliola. (2020). Using an autoencoder in the design of an anomaly detector for smart manufacturing. Pattern Recognition Letters. 136. 10.1016/j.patrec.2020.06.008.
43. Steenwinckel, B., Paepe, D.D., Hautte, S.V., Heyvaert, P., Bentefrit, M., Moens, P., Dimou, A., Bossche, B.V., Turck, F.D., Hoecke, S.V., Ongenae, F. (2021). FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning. Future Gener. Comput. Syst., 116, 30-48.
44. Wang, J., Bao, W., Zheng, L., Zhu, X., Yu, P.S. (2019). An Attention-augmented Deep Architecture for Hard Drive Status Monitoring in Large-scale Storage Systems. ACM Transactions on Storage (TOS), 15, 1 - 26. https://doi.org/10.1145/3340290
45. Sundar, Sreenath, C. Rajagopal, Manjunath, Zhao, Hanyang, Kuntumalla, Gowtham, Meng, Yuquan, Chang, Ho, Shao, Chenhui, Ferreira, Placid, Miljkovic, Nenad, Sinha, Sanjiv, Salapaka, Srinivasa. (2020). Fouling modeling and prediction approach for heat exchangers using deep learning. International Journal of Heat and Mass Transfer. 159. 120112. 10.1016/j.ijheatmasstransfer.2020.120112.
46. Le, Dy, Vung, Pham, Nguyen, Huyen, Dang, Tommy. (2019). Visualization and Explainable Machine Learning for Efficient Manufacturing and System Operations. 3. 20190029. 10.1520/SSMS20190029.
47. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.
48. Epps, Brenden, Krivitzky, Eric. (2019). Singular value decomposition of noisy data: noise filtering. Experiments in Fluids. 60. 10.1007/s00348-019-2768-4.
49. Epps, Brenden, Krivitzky, Eric. (2019). Singular value decomposition of noisy data: mode corruption. Experiments in Fluids. 60. 10.1007/s00348-019-2761-y.
50. Martinez-Cantin, Ruben. (2014). BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits. Journal of Machine Learning Research. 15. 3735-3739.
51. MATLAB CUSUM, Detect Small Changes in Mean Using Cumulative Sum. Copyright 2015-2018 The MathWorks, Inc. (Retrieved on 22 Dec 2021 from: https://www.mathworks.com/help/signal/ref/cusum.html).
52. Song, Yan, Gao, Shengyao, Li, Yibin, Jia, Lei, Li, Qiqiang, Pang, Fuzhen. (2020). Distributed Attention-Based Temporal Convolutional Network for Remaining Useful Life Prediction. IEEE Internet of Things Journal. PP. 1-1. 10.1109/JIOT.2020.3004452.

53. Lundberg, S.M., Lee, S. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.
54. Molnar, Christoph. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. (Retrieved on 22 Dec 2021 from: https://christophm.github.io/interpretable-ml-book/).
55. 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.
56. Saxena, Abhinav, Goebel, Kai, Simon, Don, Eklund, Neil. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. International Conference on Prognostics and Health Management. 10.1109/PHM.2008.4711414.
57. A. L. Ellefsen, S. Ushakov, V. Æsøy and H. Zhang. Validation of Data-Driven Labeling Approaches Using a Novel Deep Network Structure for Remaining Useful Life Predictions. IEEE Access, 7, 71563-71575, 2019.