Explainable Predictive Maintenance is Not Enough: Quantifying Trust in Remaining Useful Life Estimation

Ripan Kumar Kundu and Khaza Anuarul Hoque
University of Missouri, Columbia, Missouri, USA
rkcgc@umsystem.edu,hoquek@umsystem.edu

ABSTRACT

Machine learning (ML)/deep learning (DL) has shown tremendous success in data-driven predictive maintenance (PdM). However, operators and technicians often require insights to understand what is happening, why it is happening, and how to react, which these black-box models cannot provide. This is a major obstacle in adopting PdM as it cannot support experts in making maintenance decisions based on the problems it detects. Motivated by this, several researchers have recently utilized various post-hoc explanation methods and tools, such as LIME, SHAP, etc., for explaining the predicted RUL from these black-box models. Unfortunately, such (post-hoc) explanation methods often suffer from the *disagreement problem*, which occurs when multiple explainable AI (XAI) tools differ in their feature ranking. Hence, explainable PdM models that rely on these methods are not trustworthy, as such unstable explanations may lead to catastrophic consequences in safety-critical PdM applications. This paper proposes a novel framework to address this problem. Specifically, first, we utilize three state-of-the-art explanation methods: LIME, SHAP, and Anchor, to explain the predicted RUL from three ML-based PdM models, namely extreme gradient boosting (XGB), random forest (RF), logistic regression (LR), and one feed-forward neural network (FFNN)-based PdM model using the C-MAPSS dataset. We show that the ranking of dominant features for RUL prediction differs for different explanation methods. Then, we propose a new metric *trust score* for selecting the proper explanation method. This is achieved by evaluating the XAI methods using four evaluation metrics: fidelity, stability, consistency, and identity, and then combining them into a single *trust score* metric through utilizing Kemeny and Borda rank aggregation methods. Our results show that the proposed method effectively selects the most appropriate explanation method from a set of explanation methods for estimated RULs. To the best of our knowledge, this is the first work that attempts to address and solve the disagreement problem in explainable PdM.

1. INTRODUCTION

Data-driven predictive maintenance (PdM) approaches based on black-box machine learning (ML)/deep learning (DL) models have achieved remarkable success in terms of predictive accuracy and capability of modelling complex systems (Cummins et al., 2021; Keleko, Kamsu-Foguem, Ngouna, & Tongne, 2022; Chen, Hong, & Zhou, 2022a; Jayasinghe, Samarasinghe, Yuenv, Low, & Ge, 2019). However, the complete repair plan and maintenance actions that must be performed based on the detected symptoms of damage and wear often require complex reasoning and planning processes involving many actors and balancing different priorities–which cannot be fully automated in many cases. Therefore, operators, technicians, and managers require insights to understand what is happening, why it is happening, and how to react. The decisions black-box PdM models make are often difficult for human experts to understand and act upon. Thus, adding explainability to these models can provide several benefits, such as (i) helping in improving the model's understanding and providing insight into why and how the model arrived at a specific decision and (ii) can help reliability engineers to develop more accurate PdM models by identifying the important features in a model.

Motivated by this, a new research direction has recently emerged, leveraging black-box ML/DL models with explainable artificial intelligence (XAI), which explains the predicted remaining useful life (RUL) (Vollert, Atzmueller, & Theissler, 2021; Khan, Ahmad, Khan, Khan, & Ahmad, 2022; Hong, Lee, Lee, Ko, Kim, & Hur, 2020; M. Baptista, Mishra, Henriques, & Prendinger, 2020; Hong, Lee, Lee, Ko, & Hur, 2020). But, unfortunately, the state-of-the-art XAI methods often suffer from the *disagreement problem* (Krishna et al., 2022), which occurs when two (or more) explanation methods do not agree on a model's feature ranking. For instance, let us consider explaining a feed-forward neural network (FFNN) PdM model trained using the C-MAPSS (FD001 for this example) dataset (Saxena, Goebel, Simon, &

Figure 1. For a single prediction, the local explanations are chosen when there is a disagreement between (a) SHAP and (b) LIME explanation method.

Eklund, 2008). Let us use two different state-of-the-art XAI tools, SHapley Additive Explanations (SHAP) (Lundberg & Lee, 2017), Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016), to explain this PdM model. Figure 1 shows the explanation, and we can clearly observe disagreements. Specically, in Figure 1a, SHAP identies features such as SensorMeasure9, SensorMeasure14, and SensorMeasure11 as most influential features ranked as 1,2, and 3, respectively, for RUL estimation. On the other hand, LIME identifies features such as SensorMeasure12, SensorMeasure4, and SensorMeasure9 as most influential features ranked as 1,2 and 3, respectively, for RUL estimation. It is worth mentioning that the effectiveness of PdM systems often depends less on the accuracy of the alarms the AI models raise than on the relevancy of the actions operators perform based on these alarms. Therefore, such disagreement can easily misguide the required insights by the operators and technicians to understand what is happening, why it is happening, and how to react, which can lead to catastrophic consequences in safety-critical applications. Indeed, this disagreement in explainable PdM raises a fundamental question: how to choose the correct explanation method for PdM models?

To answer the above question and resolve the disagreement problem in PdM, in this paper, we propose a novel framework for selecting the proper explanation method from a set of explanation methods given a batch of RUL estimations from a PdM model. Our contributions in this paper can be summarized as follows.

- First, we develop three ML-based and one DL-based PdM model. Specically, we use extreme gradient boosting (XGB), random forest (RF), logistic regression (LR), and one simple FFNN to predict RUL using the C-MAPSS dataset (Saxena et al., 2008). Then we employ three post-hoc explanation techniques, namely SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and Anchors (Ribeiro, Singh, & Guestrin, 2018) to explain the predicted RUL. Our results reveal that the ranking of dominant features for RUL prediction differs for different explanation methods. In addition to establishing and demonstrating the disagreement problem, this also indicates that no specic explanation method can perform optimally for a given RUL estimation.
- We propose a method to quantitatively measure PdM disagreement problems based on the quality of explanations. Specically, we apply four explainability evaluation metrics: delity, stability, consistency, and identity, to evaluate the disagreement among XAI methods.
- Finally, to resolve the PdM disagreement problem, we propose a novel *trust score* metric by combining the XAI evaluation metrics into a single metric to select the best RUL explanation method from a set of given explanation methods. Specically, we use Kemeny rank aggregation (Cachel, Rundensteiner, & Harrison, 2022) and Borda rank aggregation (Lestari, Adji, & Permanasari, 2018) methods for aggregating the rankings induced by these evaluation metrics. Our results show that the computed *trust score* can effectively select the accurate explanation method from a set of PdM explanation methods.

Our experimental results indicate that disagreement in explanation is more likely to occur due to their lack of a common foundational goal among these methods, and the models being explained are complex and highly non-linear. Moreover, we also observe that no explanation methods could consistently provide an accurate explanation. For example, for some samples, SHAP may perform better; for others, LIME or Anchors may perform better. For the case of the C-MAPSS dataset, we show that SHAP achieves a higher trust score in most cases for the predicted RUL explanation. To the best of our knowledge, our proposed method is the first work to pave the way for selecting the most trustworthy PdM explanation method.

2. RELATED WORKS

State-of-the-art ML/DL methods have shown great success in RUL prediction due to their ability to model highly non-linear, complex, and multi-dimensional systems with little prior prognostic experience. A brief overview of these works can be found in (Serradilla, Zugasti, Rodriguez, & Zurutuza, 2022; Wen, Rahman, Xu, & Tseng, 2022). However, as previously mentioned, these models are black-box without offering insights into their working mechanism and the reasons behind their decisions (M. Baptista et al., 2020). Hence, several researchers have attempted to apply XAI to explain PdM decisions in recent years (Vollert et al., 2021; Torcianti & Matzka, 2021; Ferraro, Galli, Moscato, & Sperl, 2022; Hong, Lee, Lee, Ko, & Hur, 2020; M. Baptista et al., 2020; Hong, Lee, Lee, Ko, Kim, & Hur, 2020; Khan et al., 2022; Cohen, Huan, & Ni, 2023; Arya, Saha, Hans, Rajasekharan, & Tang, 2023; M. L. Baptista, Goebel, & Henriques, 2022).

The discipline of XAI, which studies the development of explanation methods, has two main approaches: (i) designing inherently interpretable models such as rule lists, decision trees, etc., (ii) employing post-hoc explanation methods, i.e. SHAP, LIME, etc., to explain a black-box ML/DL model locally (for a specific sample) or globally (for the entire model space). For instance, Jakubowski et al. (Jakubowski, Stanisz, Bobek, & Nalepa, 2022) applied an inherently interpretable ML model, namely an Explainable Boosting Machine (EBM) for RUL prediction of a turbofan engine (C-MAPSS dataset) and provided explanations for their predictive RUL decision. However, this inherently interpretable model did not perform well regarding stability (Jakubowski et al., 2022), meaning that similar points showed different explanations. Moreover, inherently interpretable models typically depend on the different data properties and thus suffer from the dimensionality problem (Schmitt & Jula, 2007). Another limitation of their proposed EBM model is that it requires a higher training time (Nori, Jenkins, Koch, & Caruana, 2019), which is not feasible for real-world PdM applications. To tackle these problems and provide more meaningful explanations, researchers in (Hong, Lee, Lee, Ko, & Hur, 2020; Khan et al., 2022; Cohen et al., 2023) used different post-hoc explanation methods to explain the predicted RUL. For instance, authors in (Khan et al., 2022) applied different black-box models such as Gradient Boosting, MLP, and SVM for RUL prediction. Next, they employed post-hoc explanation methods, specifically SHAP and LIME, to explain the predicted RUL using the C-MAPSS datasets. On the other hand, Hong. et.al (Hong, Lee, Lee, Ko, & Hur, 2020) used SHAP for explaining the 1D-CNN, LSTM, and bidirectional LSTM-based PdM models predicted RUL. Then it was shown that SHAP's force plot and decision plot could help decision-makers understand which feature significantly affects increasing/decreasing the predicted RUL. A more detailed treatment of XAI in the PdM system is provided in this comprehensive survey article (Vollert et al., 2021). However, one of the problems with these explanation methods is that they often suffer from the disagreement problem (Krishna et al., 2022). From a conceptual standpoint, the misalignment of goals among explanation methods leads to an inconsistent view of explanations. For instance, the SHAP method is based on game theoretic concepts (Lundberg & Lee, 2017), whereas LIME is motivated by the function approximation method (Ribeiro et al., 2016). Such differences lead to conceptual and practical challenges to understanding and using explanation methods, thwarting progress in the explainable AI field and raising the question of which explanation method to use and when. Surprisingly, the disagreement problem in the explainable PdM domain is vastly under-explored.

Only a few works exist when it comes to evaluating the quality of the explanation of PdM models. For instance, Jakubowski et al. (Jakubowski et al., 2022) measured the stability and consistency of the generated explanation for the predicted RUL for the C-MAPSS dataset. Unfortunately, their proposed approach did not consider how to choose an accurate and trustworthy explanation for explaining the predictive RUL. This is important since unstable and inconsistent explanations may lead to an untrustworthy PdM model for the end-users. The authors in (Jakubowski et al., 2022) also showed that explanation methods such as LIME and SHAP might generate inconsistent and unstable explanations. However, their work does not explain to what extent this disagreement problem occurs in practice and how to resolve it. In another work (Arya et al., 2023), the authors proposed a two-stage global explanation method where explanations from the first explainer are explained by the second one to enable more explainability in RUL explanations. Though this work effectively explains the predicted RUL, it can not evaluate the trustworthiness of the generated explanation. Instead, it solves a binary classification problem by converting the importance weights of the features to binary values, which could lead to instability in the RUL explanation and information loss.

Indeed, it is essential to understand and quantify how often explanations produced by state-of-the-art methods disagree with each other and examine how reliability engineers can address such differences. Moreover, there is no single metric using which XAI methods can be evaluated comprehensively to select the most accurate explanation method for a given RUL. Thus, in this work, we propose an approach that integrates XAI evaluation metrics into a single metric, the trust score. More specifically, we apply a novel robust rank aggregation method for combining these multiple XAI evaluation metrics into a single explanation model selection criterion that provides a trustworthy explanation of the predicted RUL.

3. METHODOLOGY

This section presents our proposed methodology for a trustworthy PdM system as illustrated in Figure 2. First, the dataset from sensor obtained data is divided into training and testing datasets. Then we use the training dataset to train four ML/DL algorithms, specifically XGB, RF, LR, and a simple FFNN for RUL estimation. These four trained PdM models are then used to predict the RUL from the test dataset. The next phase in the framework uses post-hoc (local) explanation tools, specifically SHAP, LIME, and Anchors, to explain the predicted RUL by identifying dominating features for individual predicted RULs. Next, to evaluate the trustworthiness of the XAI tool, we calculate their respective trustworthiness metrics, also known as surrogate evaluation metrics: fidelity (FI), stability (SI), identity (ID), and consistency (CO). Consequently, we rank them using robust rank aggregation (RRA) methods for combining multiple surrogate explanation evaluation metrics into one metric. This enables us to choose the trustworthy XAI method on-the-fly for RUL estimation.

3.1. RUL Prediction Model Development

The initial step of our proposed methodology is to develop a PdM model for RUL estimation. Researchers have already published tons of work in this area, and a brief survey of them can be found in (Zhang, Si, Hu, & Lei, 2018; Lipu et al., 2018; Cummins et al., 2021; Chen, Hong, & Zhou, 2022b). It is worth mentioning that though this is the first step of our methodology, developing a PdM model does not capture this paper's main contribution. Instead, this work is more focused on evaluating the trustworthiness of these models so that a trustworthy PdM model can be devised. To build accurate PdM models, we use three ML models, namely XGB, RF, LR, and one simple FFNN. We chose these models as they are common in RUL estimations (Jafari & Byun, 2022; Jiao et al., 2023; Sharma & Bora, 2022; Ni, Ji, & Feng, 2022; Tong, Miao, Mao, Wang, & Lu, 2022; Rauf, Khalid, & Arshad, 2022; Wu, Zhang, & Chen, 2016). We train them using the C-MAPSS dataset (Saxena et al., 2008). We use the grid search method (Shekar & Dagnew, 2019) for the hyperparameter tuning of these models. For training the FFNN model, we use the learning rate of 0.001. Furthermore, to prevent the FFNN model from overfitting, we use an early-stopping strategy with a patience value of 30 while training. Furthermore, we use a 10-fold cross-validation technique to validate the performance of these ML and DL-based PdM models.

In addition to RUL estimation, we also use the same PdM models for classification by converting the RUL values into a classification problem similar to the work (Remadna, Terrissa, Al Masry, & Zerhouni, 2022), where the class labels are: good condition, moderate condition, and warning condition. We will need the RUL classification results to calculate the explanation evaluation metrics, which require the corresponding label of the predicted RUL. To assign the labels, we define the engine's condition with the life ratio (LRO), the ratio between the current and end cycles. $LRO = 0$ indicates that the component has just started its degradation, whereas $LRO = 1$ means it has completely degraded. We labeled the good condition as 0 if $LRO <= 0.6$, the moderate condition as 1 if $0.6 < LRO \quad 0.8$, and the warning condition as 2 if $LRO > 0.8$.

3.2. RUL Explanation

This section discusses how the explanation block provides predicted RUL explanations with feature importance scores for individual samples during predictions. Explanations in XAI can be generally categorized as global and local explanations. A global explanation aims to identify features crucial for the overall prediction, whereas a local explanation identifies features dominating an individual sample's prediction. However, our work focuses on the local explanation method in which samples are randomly chosen from the test dataset containing all the features. To calculate the feature importance score, we use three post-hoc explanation tools, namely, LIME (Ribeiro et al., 2016) and (SHAP) (Lundberg & Lee, 2017)), Anchor (Ribeiro et al., 2018). SHAP is a feature importance explanation approach that assigns a feature significance value to each prediction. As previously mentioned, it is based on the mathematical foundation of Shapley values from cooperative game theory (Lundberg & Lee, 2017). For a given set of input samples and ML/DL models, the goal of SHAP is to explain the prediction of input samples by calculating the contribution of each feature to the prediction. On the other hand, the LIME-based explanation method generates explanations by approximating the underlying model with an interpretable one to show what feature contributed to the output from that single sample. More specifically, LIME trains an interpretable model on a newly generated dataset consisting of perturbed samples around the original data point and the corresponding predictions. Then, LIME weights the proximity of sampled data points to the original data and generates an explanation. Consequently, Anchor explains the prediction in the form of rules which accurately capture the important factors driving a given predic-

Figure 2. An overview of a trustworthy RUL explanation from a set of explanation methods of explainable predictive maintenance framework.

tion. The assumption is that the prediction is always the same for the given instances on which the anchor holds. However, as previously mentioned, one of the problems with these explanation methods is they often suffer from the disagreement problem(Krishna et al., 2022). Thus, it is crucial to investigate whether different explanation methods can produce the same or different explanations that are inconsistent and inaccurate. To find the answer to this question, we evaluate the performance of these explanations using a diverse set of explanation metrics in the next section.

3.3. RUL Explanation Evaluation Metrics

The researchers have proposed several metrics to measure the explainability of AI models (Elkhawaga, Elzeki, Abuelkheir, & Reichert, 2023; Nauta et al., 2023). A comprehensive survey of these metrics for evaluating explanation methods is available (Zhou, Gandomi, Chen, & Holzinger, 2021). Their survey highlights two characteristics of a high-quality explanation: how well it approximates the model and how human-understandable it is. Our work focuses on four explanation evaluation metrics: fidelity, stability, identity, and consistency. We choose these four metrics because they offer a comprehensive evaluation framework for XAI methods and are widely adopted by researchers and practitioners to assess the strengths and limitations of different explanation methods (Elkhawaga et al., 2023; Zhou et al., 2021). Thus, these metrics help us measure the accuracy and fairness of the generated explanation for a predicted RUL and how easily a user understands the explanation. Evaluating these metrics helps to measure the XAI disagreement problem in a principled manner.

Fidelity: When evaluating an explanation method, the most common question is "To what extent does it accurately represent the underlying decision-making process?" In other words, do the important features highlighted in the explanation represent the most important features of the model? Explanations that precisely identify the most dominating features of the underlying models for RUL prediction have high fidelity. More specifically, fidelity is the concordance of the predictions between the applied XAI methods and the complex black-box ML/DL models, which can be defined as:

$$F_{(x;f;\ )} = \frac{|top(k;W) \setminus top(k;w)|}{k} \tag{1}$$

In Equation (1), w represents the ground truth weights of a black-box model f for a given input x and explanation method $\ $. The fidelity metric F is defined as the percent of the top k features from explanation W, which are also in the top k features (where k is a function that returns the indices of the k largest elements of a given input x) from w. Note, W is the given local explanation for the predicted RUL and is denoted as $W = \ (x; f\ ))$. The main idea behind this fidelity metric is that slight modifications to unimportant features k should not significantly impact the black-box model f prediction. If the model prediction changes significantly, the explanation has low fidelity and fails to capture important features that are crucial to the model prediction. Note, an explanation with a low fidelity score can be useless (Carvalho, Pereira, & Cardoso, 2019).

Identity: The second metric used to evaluate the XAI explanation is identity. It assumes that if there are two identical instances, such as the actual and predicted RUL classes, they must have identical explanations (Parimbelli et al., 2023). If this is not the case, then either the explanation model generates an explanation that is not identical or the PdM model predicted the wrong RUL class.

Stability: The idea that similar observations should receive similar explanations means that small changes in the observations will lead to low changes in the explanations. This property is known as the explanation model's stability or robustness. We use the Lipschitz indicator proposed by Alvarez-Melis and Jaakkola (Alvarez-Melis & Jaakkola, 2018) to measure an explanation method's stability or robustness. This

Lipschitz indicator provides the robustness of the explanation methods to endure slight perturbations to the input that do not affect the model's prediction represented as:

$$L_X(x_i) = \max_{x_j \in N_\epsilon(x_i)} \frac{||f_i - f_j||}{||x_i - x_j||} \quad (2)$$

In Equation (2), $L_X(x_i)$ represents the stability of the data point x_i from test set X where $x_i \in X$, f_i and f_j are the feature importance score of the instances x_i and x_j. $N_\epsilon(x_i)$ is the neighborhood or ball of radius ϵ centered at x_i, which is defined as all data points that have L2 norm distance to data point x_i is smaller than ϵ. The general idea behind measuring an explanation's stability at a point x_i is that we add some noise to x_i to generate similar points and then find the neighborhood of the point x_i and the maximum dissimilarity. Then take the average Euclidean distance between x_i the data point explanation and those of similar data points. A lower stability value indicates that the model performs better in explanation.

Consistency: The consistency metric quantifies the similarity between the explanations generated by various explanation methods for predictions of different black-box models. The main intuition behind this metric is that if an explanation for a single observation is measured multiple times, each of the measured explanations should be similar. If this is not the case, then either the black-box model is not making a good prediction or the explanation method is not providing a proper explanation. To measure the consistency of a given instance, we compute numerous explanations for that instance and then measure the average L1 distance between the original explanation and each of the new explanations similar to the work (Bobek, Bałaga, & Nalepa, 2021). The consistency metric can be expressed as follows:

$$C(G_{i_1}; G_{i_2}) = \frac{1}{||G_{i_1} - G_{i_2}||_2 + 1} \quad (3)$$

where $C(G_{i_1}; G_{i_2})$ is the measured consistency of i^{th} observation and G_{i_1} and G_{i_2} are the feature importance of the i^{th} observation for the given two explanation models.

3.4. Robust Rank Aggregation

In previous sections, we discussed how explanation evaluation metrics provide more insight into the generated explanations from the explanation methods. However, these explanation methods often generate disagreeing explanations in practice and lack a principled approach for reliability engineers/managers to select suitable explanations. Thus, there is a need to derive a method that can be used to select the most accurate explanation method for a given observation (predicted RUL) from a set of explanation methods. For this, we take advantage of extensive research that has been conducted on the topic of rank aggregation in ML and their applications in different fields (Klementiev, Roth, & Small, 2008; Dwork, Kumar, Naor, & Sivakumar, 2001; Waad, Brahim, & Limam, 2013). This section formalizes the robust rank aggregation framework for selecting suitable explanation methods by utilizing Kemeny rank aggregation (Cachel et al., 2022) and Borda rank aggregation (Lestari et al., 2018) methods. We use these two methods to combine multiple XAI evaluation metrics into a single metric trust score as an accurate explanation model selection criterion.

We use Kemeny rank aggregation to find a barycentric or a median ranking by picking a distance on the set of rankings, known as the Kemeny-Young problem (Waad et al., 2013). However, despite having many desirable qualities, Kemeny rank aggregation may suffer from NP-hard problems (Baumeister & Rothe, 2016). Thus, as an alternative, we use the Borda rank aggregation methods to find an efficient approximate solution in which each explanation model receives awards from each evaluation metric (e.g., fidelity, stability, consistency). For instance, let us assume a set of candidate explanation methods $E = \{e_1, e_2, ..., e_N\}$ where $N > 0$ represent the total number of explanation methods. Let us also assume a set of XAI evaluation metrics $M = \{m_1, m_2, ..., m_J\}$, where $J > 0$ is the total number of evaluation metrics and each $m_i \in M$ represents an N dimensional vector where each $v_k \in m_i$ contains the quantitative evaluation of metric m_i. To choose the best explanation method, we calculate the trust score based on given E and M using robust rank aggregation function F which can be defined as: $F(.) : M \rightarrow R$, where each element $r_i \in R$ (R is set with N elements) represents the aggregated evaluation metric value associated with each explanation method. We calculate the trust score for each rank aggregation method (Kemeny and Borda) by quantifying the agreement between the aggregated rankings and the ground truth or a reference ranking. This trust score (TS) provides a fair ranking on the performance of aggregated rank and selects the best explanation method for a given predicted RUL, which can be defined as follows:

$$TS = \frac{1}{J} \sum_{p=1}^{N} \sum_{q=1}^{N} Rank\_agr_{score}(p, q) \quad (4)$$

In Equation (4), $Rank\_agr_{score}(p, q)$ represents the pairwise agreement score between explanation methods p and q in the aggregated rankings and the reference ranking. More specifically, it measures how consistently two rankings agree with each other. Moreover, it combines or merges multiple individual rankings into a single aggregated ranking. We calculate the $Rank\_agr_{score}(p, q)$ between two rankings using Kendall's tau (τ) distance. Kendall's tau (τ) distance measures the number of pairwise disagreements between the two rankings. A higher trust score indicates better agreement be-

tween the aggregated and reference rankings, suggesting a more reliable and trustworthy rank aggregation method.

4. DATASET & EXPERIMENTAL SETUP

This section explains the experimental setup and data we used to validate our proposed method. We used Scikit-Learn (Pedregosa et al., 2011) to train the ML models and TensorFlow-2.4 (Sergeev & Del Balso, 2018) to train and evaluate our FFNN model. We use 50, 100, and 200 nodes in each consecutive layer, ReLU activation, categorical cross-entropy and mean squared error loss function, Adam optimizer, and 300 training epochs. For explaining the ML/DL-based PdM models, we used the SHAP (Lundberg & Lee, 2017), the Anchor (Ribeiro et al., 2018), and the InterpretML (for LIME) (Nori et al., 2019) library. The ML/DL-based PdM models were trained on an Intel Core i9 Processor and 32GB RAM option with NVIDIA GeForce RTX 3080 Ti GPU.

4.1. Dataset

To validate the effectiveness of the proposed approach, we use the Commercial Modular AeroPropulsion System Simulation (C-MAPSS) dataset (Saxena et al., 2008). This dataset comprises 22 different features from the sources, such as pressure, fan speed, fuel, coolant flow, temperature, etc., and 3 operational parameters (settings). The dataset consists of four fleets of engines FD001, FD002, FD003, and FD004, in which each fleet has an approximately equal number of train and test instances, as shown in Table 1. The training data capture run-to-failure trajectories, whereas the testing data preserves sensor readings of engines up to a given point in time with a known RUL. The RUL at a given point for each turbofan engine can be determined based on the total number of completed cycles. We used 70% samples from this dataset for training the ML/DL-based PdM models and their remaining 30% samples for testing.

4.2. Evaluation Metrics

We use the most widely used error performance metrics to evaluate the RUL estimation accuracy: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Similarly, to evaluate the performance of classification models, we use standard metrics such as balanced accuracy and F1-score (Brodersen, Ong, Stephan, & Buhmann, 2010). We choose balanced accuracy instead of accuracy as balanced accuracy is known to perform better with imbalanced data, and converting regression problems to classification problems often suffers from class imbalance problem (Brodersen et al., 2010). Therefore, we used balanced accuracy to avoid such situations.

Table 1. Number of train and test engine units in each fleet of the C-MAPSS dataset

	FD001	FD002	FD003	FD004
Train	100	260	100	249
Test	100	259	100	248
Op. cond./fault modes	1/1	6/1	1/2	6/2

5. RESULTS

This section presents the results obtained from evaluating the performance of PdM models, RUL local explanation, XAI methods performance, and robust rank aggregation.

Table 2. Performance of 10-Fold Cross Validation on C-MAPSS dataset in RUL prediction

Model	MAE				RMSE			
	FD001	FD002	FD003	FD004	FD001	FD002	FD003	FD004
XGB	13.75	15.72	14.43	18.45	14.05	16.32	14.67	17.95
RF	13.34	15.91	14.87	19.64	13.84	22.15	15.31	21.05
LR	17.55	18.71	16.23	25.87	17.76	23.03	18.32	26.92
NN	9.98	11.73	10.54	12.89	12.11	14.81	13.13	14.64

Table 3. Performance of 10-Fold Cross Validation on C-MAPSS dataset in Classification Task

Model	Balanced Accuracy%				F1-Score			
	FD001	FD002	FD003	FD004	FD001	FD002	FD003	FD004
XGB	91.5	90.3	89.7	89.3	92.6	91.4	91.2	92.5
RF	89.5	88.7	88.1	87.5	91.8	90.8	91	92.1
LR	87.2	86.8	84.5	85.1	90.3	89.2	88.9	89.5
NN	92.7	91.5	90.4	91.5	93.4	93.5	92.3	93.1

5.1. Performance of PdM Models

Table 2 shows the performance of the developed PdM models. The FFNN model performs better than other ML models for regression (RUL estimation) and classification. For instance, the FFNN model's RMSE score is 9.98, 11.73, 10.54, and 12.89 for FD001–FD004 datasets. In contrast, the XGB model's RMSE score is 13.75, 15.72, 14.43, and 18.45. The MAE scores for these models also show the same trend. For classification, the performance of the FFNN and XGB models are quite close to each other. For instance, the balanced accuracy score of the FFNN model is 92.70%, 91.50%, 90.40%, and 91.50% for FD001–FD004 datasets. In contrast, the balanced accuracy score of the XGB model is 91.50%, 90.30%, 89.70%, and 89.30%. Overall, the NN model performs better for regression and classification tasks than other ML models due to their high predictive performance capabilities.

5.2. Explanation of RUL

The results of the RUL (local) explanation utilizing SHAP for the FFNN and XGB models and FD001 dataset are shown in Figure 3. In Figure 3, the yellow and green bars denote predicted RUL probabilities and the mean absolute score (MAS) for that individual outcome. MAS is actually calculated as logits or log odds. To convert these logits into a probability, we sum them up and pass them through the logistic link function (Zou, Hu, Tian, & Shen, 2019). This logistic link

(a) (b)

Figure 3. For a single prediction in the FD001 dataset, the local explanations provided by SHAP in which the actual value of RUL of the component is 114 while the predicted value is 111.87 by (a) FFNN, (b) and XGB model.

(a) (b)

Figure 4. For a single prediction in the FD001 dataset, the local explanations provided by LIME in which the actual value of RUL of the component is 114 while the predicted value is 111.87 by (a) FFNN, (b) and XGB model.

Table 4. The delity metric of SHAP, LIME, and Anchor methods computed for 10 randomly selected test samples from the FD001–FD004 datasets for LR, XGB, RF, and NN models.

XAI methods	Models	FD001	FD002	FD003	FD004
SHAP	LR	0.875	0.843	0.795	0.892
	XGB	0.975	0.953	0.925	0.898
	RF	0.912	0.905	0.883	0.934
	NN	0.998	0.956	0.986	0.971
LIME	LR	0.910	0.905	0.918	0.886
	XGB	0.904	0.953	0.925	0.898
	RF	0.943	0.937	0.856	0.892
	NN	0.912	0.889	0.898	0.893
Anchor	LR	0.863	0.843	0.795	0.892
	XGB	0.890	0.878	0.892	0.879
	RF	0.881	0.907	0.887	0.865
	NN	0.924	0.905	0.894	0.934

Table 5. The identity metric of SHAP, LIME, and Anchor methods computed for 10 randomly selected test samples from the FD001–FD004 datasets for LR, XGB, RF, and NN models.

XAI methods	Models	FD001	FD002	FD003	FD004
SHAP	LR	0.032	0.0054	0.0019	0.00056
	XGB	0.242	0.437	0.295	0.159
	RF	0.465	0.513	0.503	0.485
	NN	0.798	0.752	0.787	0.734
LIME	LR	0.0	0.0	0.0	0.0
	XGB	0.0242	0.0193	0.0157	0.172
	RF	0.0805	0.081	0.061	0.074
	NN	0.08	0.053	0.079	0.071
Anchor	LR	0.0	0.0	0.0	0.0
	XGB	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0
	NN	0.018	0.014	0.009	0.012

(a) (b)

Figure 5. For a single prediction in the FD001 dataset, the local explanations provided by Anchor in which the actual value of RUL of the component is 114 while the predicted value is 111.87 for (a) FFNN, (b) and XGB model.

Table 6. The stability metric of SHAP, LIME, and Anchor methods computed for 10 randomly selected test samples from the FD001–FD004 datasets for LR, XGB, RF, and NN models.

XAI methods	Models	FD001	FD002	FD003	FD004
SHAP	LR	0.416	0.429	0.443	0.427
	XGB	0.339	0.353	0.331	0.319
	RF	0.302	0.325	0.336	0.317
	NN	0.273	0.295	0.301	0.289
LIME	LR	0.507	0.537	0.525	0.519
	XGB	0.473	0.493	0.498	0.465
	RF	0.406	0.443	0.418	0.425
	NN	0.387	0.415	0.395	0.408
Anchor	LR	0.786	0.797	0.811	0.792
	XGB	0.687	0.703	0.719	0.749
	RF	0.745	0.762	0.716	0.704
	NN	0.642	0.669	0.638	0.655

Table 7. The consistency metric of SHAP, LIME, and Anchor methods computed for 10 randomly selected test samples from the FD001–FD004 datasets for LR, XGB, RF, and NN models.

XAI methods	Models	FD001	FD002	FD003	FD004
SHAP	LR	0.0014	0.0009	0.0008	0.001
	XGB	0.189	0.176	0.183	0.165
	RF	0.332	0.315	0.216	0.197
	NN	0.063	0.095	0.031	0.089
LIME	LR	0.143	0.106	0.125	0.113
	XGB	0.103	0.89	0.98	0.95
	RF	0.166	0.153	0.147	0.175
	NN	0.0087	0.059	0.0755	0.0418
Anchor	LR	0.0001	0.0001	0.0001	0.0001
	XGB	0.0032	0.0034	0.0064	0.0009
	RF	0.0143	0.0117	0.0122	0.0091
	NN	0.00	0.00	0.00	0.00

Table 8. Performance evaluation of the XAI methods for the top-1 selected model (NN-based RUL prediction) in the FD001 dataset for 10 randomly selected test samples.

	SHAP	LIME	Anchor
Fidelity	0.953	0.923	0.913
Stability	0.351	0.328	0.531
Consistency	0.09	0.082	0.0002
Identity	0.753	0.612	0.094

function is calculated for the feature ranking, where logits represent the sigmoid's midpoint for a sample. The features with large MAS values are classed as important as they have a higher average impact on the model output. The x-axis represents the model's output MAS (the probabilities of feature importance in RUL prediction), and the y-axis lists the model's features. In other words, the features in the green color mean that the condition of the component is good (the reason behind the high RUL value), while the yellow indicates that the conditions of the component are degraded (contributed to the reduction of the predicted RUL value). It is worth mentioning that MAS values in the SHAP explanation are also known as absolute Shapley values. From Figure 3a, it is observed that features such as SensorMeasure12, SensorMeasure15, etc., are the most influential features for RUL prediction, which has positive probabilities of 0.3 indicating that the condition of the component is good. On the other hand, features such as SensorMeasure9 and SensorMeasure14 have negative probabilities that indicate that when the RUL degrades, these features have a higher impact on the RUL prediction. Likewise, the local explanation of the predicted RUL using XGB model is shown in Figure 3b. From Figure 3b, it is observed that features such as SensorMeasure9, SensorMeasure4, SensorMeasure9 etc., are the most dominating features for RUL prediction, which has positive probabilities of 0.5 and features such as SensorMeasure12, SensorMeasure20 etc., have negative probabilities for RUL value degradation.

The results of the local explanation utilizing LIME for the same samples from the FD001 dataset for the FFNN and XGB models are shown in Figure 4. Like SHAP, LIME explanations are also composed of several values, including the predicted RUL value (i.e., 111.87). In Figure 4a, the features such as SensorMeasure11, SensorMeasure15, SensorMeasure4 etc., on the right side in orange color are the ones that contribute to increasing the prediction RUL values, while the ones in blue color are the features that have a negative effect or decrease the predicted RUL value. For example, features such as SensorMeasure15, SensorMeasure13, SensorMeasure6, SensorMeasure7 on the right side indicate that the machine component is in good condition and its RUL is supposed to be high. However, the features such as SensorMeasure11, SensorMeasure4, SensorMeasure8, Sen-

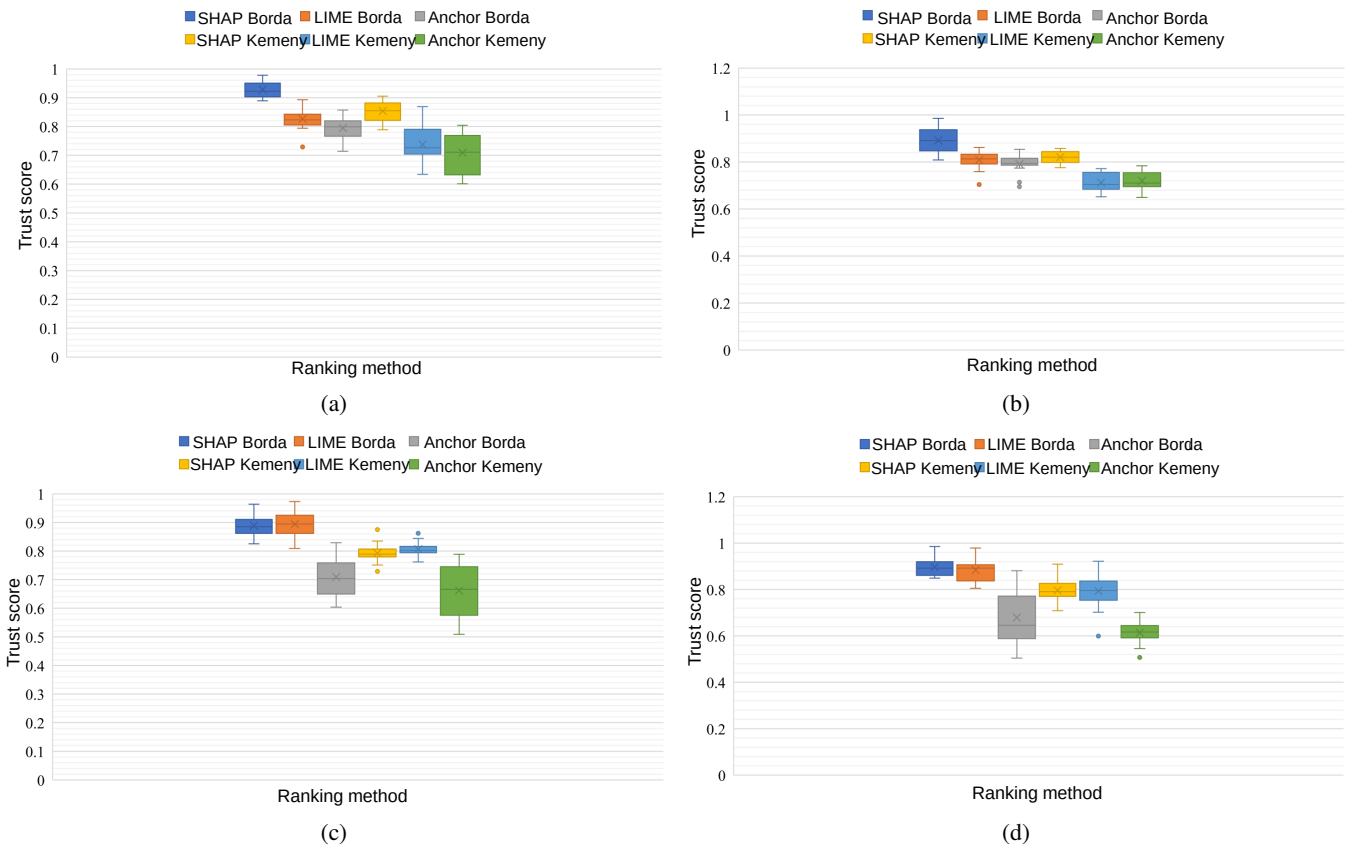

Figure 6. Performance of the top-1 selected model (NN-based RUL prediction). Box plots of the measured *trust score* of the explanation method selected by XAI evaluation metric sets (**a**) FD001, (**b**) FD002, (**c**) FD003, and (**d**) FD004 dataset.

sorMeasure9 on the left side in blue color indicate that the condition of the component is degraded and thus reducing the predicted RUL value. On the other hand, in Figure 4b, features such as *SensorMeasure2*, *SensorMeasure4*, *SensorMeasure9*,etc., appear to be the most predictive features for increasing the predicted RUL value and features such as *SensorMeasure15*, *OpSet1*, *SensorMeasure6*, etc., are the most influential features for reducing the predicted RUL value for the XGB model.Interestingly, the SHAP and LIME explanation methods provide different feature rankings for the same sample in the predicted RUL explanation, which leads to disagreement problems in the explanation. A disagreement problem occurs when two explanation methods do not agree on the feature ranking (Krishna et al., 2022). For instance, in Figure 3a, *SensorMeasure9* is the most influential feature for decreasing the predicted RUL value in the SHAP explanation, while *SensorMeasure11* is the most influential feature for decreasing the predicted RUL value in the LIME explanation as shown in Figure 4a.

Next, we analyze the relation between top features and the FFNN and XGB models output utilizing an Anchor-based local explanation to provide a deeper insight into the predicted RUL explanation. Figure 5 presents the results of the top feature for the FFNN and XGB models and FD001 dataset. We observe that the *conditions* necessary for the RUL prediction are very specific. For instance, as shown in Figure 5a, if features such as *Operational setting_2*, *SensorMeasure12*, *SensorMeasure14*, *SensorMeasure7*, etc., contribute with given measurement values and conditions, then the FFNN model predicts the RUL of 111.87 with a precision of 0:832 and coverage of 0:232. On the contrary, if features such as *SensorMeasure20*, *SensorMeasure14*, *SensorMeasure9*, etc., contribute with given measurement values and conditions, then the XGB model predicts the RUL of 111.87 with a precision of 0:752 and coverage of 0:192 (Figure 5b). A reliability engineer can use these conditions to observe the features' contribution and corresponding values for the specific RUL prediction. Like LIME and SHAP explanation methods, Anchor explanation methods also provide different features with corresponding values for predicted RUL explanations through a set of conditions. One key point we observe from these three explanation methods is that they may disagree with each other for the same sample. However, different explanation methods have different goals, leading to an inconsistent view of explanation (Krishna et al., 2022). Thus, only looking at the feature ranking does not provide sufficient disagreement with the predicted RUL explanation. Therefore, we use four

XAI evaluation metrics to capture the intuitions behind the predicted RUL explanation disagreement. These metrics capture specific aspects of the explanation disagreement concerning their corresponding feature importance score. In the next section, we will evaluate the XAI methods against these metrics.

5.3. Performance of RUL Explanation

In this section, we evaluate the performance of SHAP, LIME, and Anchor for generating RUL explanations fidelity, stability, identity, and consistency metrics. To perform the evaluation and show the disagreement problem, we randomly select 10 samples from the test dataset. Table 4 summarizes the fidelity scores of predicted RUL explanations using SHAP, LIME, and Anchor FD001–FD004 datasets. We observe that SHAP performs better regarding the FFNN model across the sub-dataset. For instance, the RUL explanation using the SHAP method for the FFNN model and FD003 dataset exhibits a 0:986 fidelity score, almost $1{:}09x$, and $1{:}12x$ higher than that of the LIME and Anchor methods. This indicates that minor explanation disagreement occurred with LIME and Anchor explanation methods. However, we notice that for the LR model, SHAP does not perform well compared to LIME for the RUL explanation across the sub-datasets. This is because LR is built on linear models, and LIME is also a local function approximation method; thus, if the LR model predicted inaccurate RUL, as a consequence, the local surrogate model also predicted the inaccurate RUL(Lundberg & Lee, 2017). However, this is not true for other classes of ML/DL models.

Furthermore, Table 5 summarizes the identity metric of benchmark methods in the RUL explanation. From Table 5, we observe that SHAP performs relatively consistently and has a higher identity score across the sub-datasets. This indicates that two identical instances always have the same explanation generated by SHAP. This is not the case for either LIME or Anchor, highlighting that these methods generate potentially unstable explanations. For instance, SHAP performs with an identity score of 0:798 for the FFNN model and FD001 dataset, which is approximately $9{:}98x$ and $44{:}5x$ higher than LIME and Anchor, highlighting an identical explanation.

Therefore, Table 6 summarizes stability metrics scores for the explanation methods. We observe that the SHAP method achieves higher stability in the model explanation. The FFNN model for all of the sub-datasets with the SHAP explanation achieves the lowest stability value. For instance, the RUL explanation using the SHAP method results in a stability score of 0:273, 0:302, 0:339, and 0:416 for the FFNN, RF, XGB, and LR models, respectively, for the FD001 dataset. This is indeed $1{:}41x$, $1{:}34x$, $1{:}39x$, and $1{:}21x$ lower than the LIME and $2{:}35x$, $2{:}47x$, $2{:}02x$, and $1{:}58x$ lower than the Anchor methods for the FFNN, RF, XGB, and LR models, respectively for the FD001 dataset. This indicates that SHAP performs better in predicted RUL explanations, highlighting that this method generates potentially stable explanations. Furthermore, Table 7 presents the mean consistency between ML/DL models. We observe that the consistency scores are relatively far from the ideal value of 1:0, which indicates the explanation methods are inconsistent, raising the disagreement problem. The highest consistency value is observed between the three pairs of models: RF and XGB with SHAP, RF with LIME, and LR with LIME. Overall, the explanations of the same model with different explanation methods give lower consistency, which **strongly indicates** that *no single explanation method provides a faithful explanation for a given predicted RUL.*

To elucidate further, we evaluated the fidelity, stability, identity, and consistency score of the XAI methods for the FNNN model using 10 more samples from the FD001 dataset. The results are shown in Table 8. We observe that the SHAP achieves higher fidelity, consistency, and identity values, while LIME achieves higher stability values for the same. Unfortunately, Anchor does not perform well in this case because of their set of if-then rules to explain the predicted RUL. These results show that different methods may generate unstable explanations for different batches of samples, leading to disagreement among explanation methods.

5.4. Calculating Trust Scores for Identifying the Best Suitable Explanation

In the previous section, we observed that no explanation method could consistently provide an accurate explanation. For example, for some samples, SHAP may perform better, and for others, LIME may perform better. Moreover, there is no single metric using which XAI methods can be evaluated in a comprehensive manner. Thus, combining the XAI method evaluation metrics into a single metric *trust score* is important to select the best explanation method (from a set of explanation methods) for a given set of samples. We obtain the *trust score* by applying the robust rank aggregation methods, namely Kemeny and Borda rank aggregation method. Finally, it is worth mentioning that we used the FFNN model for this analysis because it showed the best performance in predicting the RUL.

Figure 6 shows the trust scores for all sub-datasets. In each box plot, we compare the distribution of the average trust scores for Borda and Kemeny's aggregation method with respect to each explanation method for the FFNN model. Our results show that SHAP performs relatively well in generating a stable and consistent explanation compared to the LIME method in all cases. In contrast, Anchor often generates inconsistent explanations compared to SHAP and LIME methods. For instance, as shown in Figure 6 a and b, the RUL

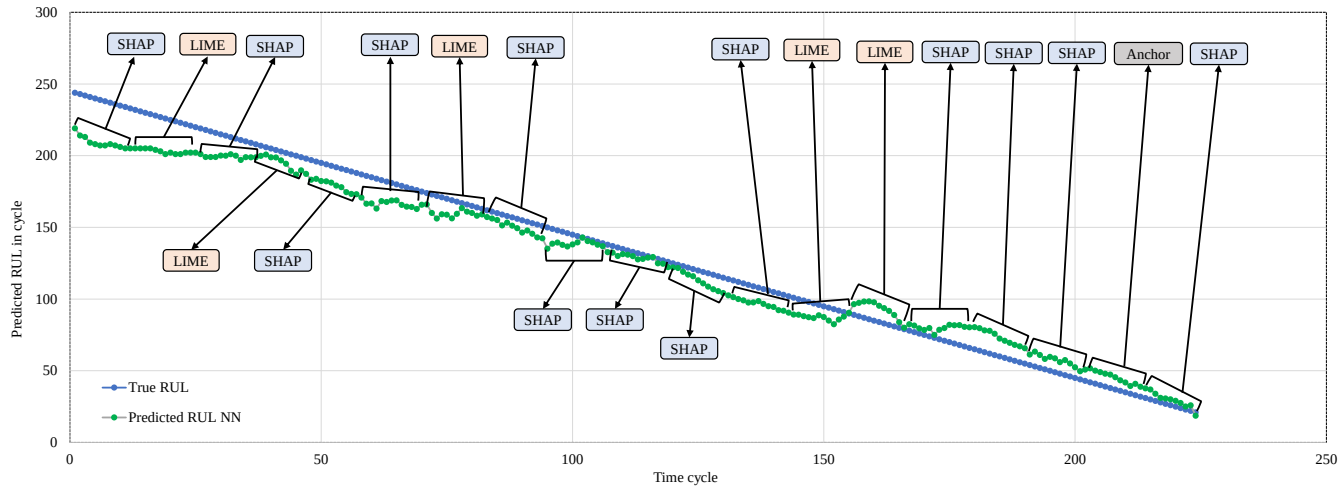

Figure 7. An overview of a trustworthy RUL explanation from a set of explanation methods of explainable predictive maintenance framework.

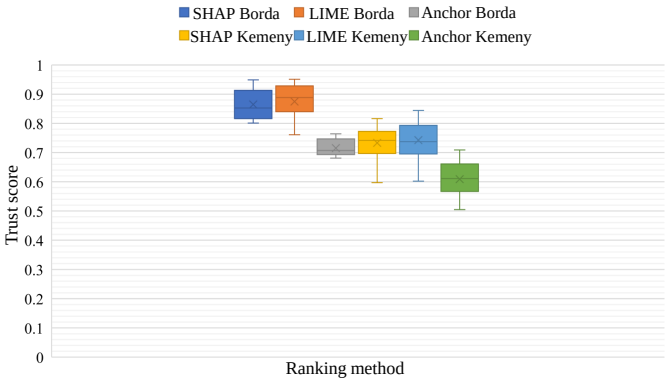

Figure 8. Performance of the top-1 selected model (NN-based RUL prediction). Box plots of the measured *trust score* of the explanation method selected by the XAI evaluation metric set for the next 10 selected test samples for the FD001 dataset.

explanation based on SHAP performs better with Borda and Kemeny methods, while LIME and Anchor do not. In addition, we observe that the distribution of trust score for the SHAP explanation method has the highest median score for both rank aggregation methods, whereas LIME and Anchor have the lowest median. On the other hand, from Figure 6 c, we observe that LIME performs comparatively better than SHAP methods and has the highest median score for both Borda and Kemeny rank aggregation methods. Interestingly, Anchor also achieved the lowest median for both Borda and Kemeny rank aggregation methods.

To evaluate our proposed method further, we randomly choose 10 more samples from the FD001 dataset and provide their average trust scores using Borda and Kemeny's aggregation method, as shown in Figure 8. We observe that the LIME explanation method performs better than the SHAP and Anchor explanation methods. Thus, the proposed trust score provides insight into choosing the best suitable explanation method for a specific RUL prediction. Indeed, the proposed trust score works better than any randomly chosen individual explanation evaluation metric. One interesting fact that we observe is that when performing robust rank aggregation for trust score calculation using all evaluation metrics for a batch size of 10 (10 samples from the dataset), the Borda rank aggregation method performs better than the Kemeny rank aggregation method while aggregating the evaluation metrics and quantifying the trust score. The reason behind that is the fact that some metrics have ranking performance worse, such as identity and consistency, which contain lots of 0 value (see Tables 5 and 7), leading to incomplete or biased rankings and affecting the overall quality of the Kemeny rank aggregation (Cachel et al., 2022).

Finally, to demonstrate how the proposed trust score can help in RUL estimation, we plot the RUL estimations using the FFNN model and FD001 dataset in Figure 7. For each batch of RUL values (10 RUL values/batch), we calculate the trust score of SHAP, LIME, and Anchor and then select the best one to label the batch in Figure 7. We observe that in most cases, SHAP is selected more often than the LIME and Anchor explanation methods. Unsurprisingly, Anchor is selected only once in the whole dataset due to their poor performance in the RUL explanation, as discussed in Section 5.3. *For reliability engineers, such trust scores can guide them in choosing the most suitable and accurate explanation methods to enable trustworthy and explainable predictive maintenance.*

6. LIMITATIONS OF THE PROPOSED METHOD

Although our proposed framework sheds light on a unique problem that poses a critical challenge in adopting explain-

able predictive maintenance, our approach also has a few limitations. For instance, we validated our proposed approach only on three ML and one simple DL model. However, with promising PdM use cases, we have yet to apply our proposed method on more complex DL architectures such as LSTM, GRU, Transformer Networks, etc. Furthermore, we applied three post-hoc explanation methods and four XAI evaluation metrics in our work. These metrics do not provide information on the source of bias in explanations. Thus, in future work, it would be interesting to develop novel evaluation metrics that can help reliability engineers readily discern a reliable explanation from an unreliable one when there is a disagreement. Moreover, we only applied feature importance-based explanation methods in this work. Therefore, in the future, we plan to conduct further research with other explanation methods such as example-based explanation, counterfactual explanation, visual explanation, etc.

7. CONCLUSION

In this work, we proposed a trustworthy RUL explanation framework by demonstrating and solving the disagreement problem among the state-of-the-art XAI tools. Specifically, we first developed three ML- and one DL-based PdM models. Then we applied three post-hoc explanation methods: SHAP, LIME, and Anchor, to explain the predicted RUL to demonstrate and evaluate their disagreement using four evaluation metrics, i.e., fidelity, consistency, stability, and identity. Finally, to solve this disagreement, we proposed a novel *trust score* by combining their rankings using a robust rank aggregation approach from different explanation evaluation metrics for selecting the best explanation method for a given batch of RUL samples. We illustrated the effectiveness of our proposed method using NASA's turbofan engine C-MAPSS dataset. Our results showed that the SHAP explanation method performed relatively well compared to the LIME method. Our results also showed that the Borda rank aggregation method performed better than the Kemeny method in selecting a suitable explanation method, with the highest trust score. We believe our proposed method can help identify the proper explanation method to guide reliability engineers in making correct decisions in safety-critical PdM applications.

REFERENCES

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Arya, V., Saha, D., Hans, S., Rajasekharan, A., & Tang, T. (2023). Global explanations for multivariate time series models. In *Proceedings of the 6th joint international conference on data science & management of data (10th acm ikdd cods and 28th comad)* (pp. 149–157).

Baptista, M., Mishra, M., Henriques, E., & Prendinger, H. (2020). Using explainable artificial intelligence to interpret remaininguseful life estimation with gated recurrent unit.

Baptista, M. L., Goebel, K., & Henriques, E. M. (2022). Relation between prognostics predictor evaluation metrics and local interpretability shap values. *Artificial Intelligence*, *306*, 103667.

Baumeister, D., & Rothe, J. (2016). Preference aggregation by voting. *Economics and computation: An introduction to algorithmic game theory, computational social choice, and fair division*, 197–325.

Bobek, S., Bałaga, P., & Nalepa, G. J. (2021). Towards model-agnostic ensemble explanations. In *Computational science–iccs 2021: 21st international conference, krakow, poland, june 16–18, 2021, proceedings, part iv* (pp. 39–51).

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124).

Cachel, K., Rundensteiner, E., & Harrison, L. (2022). Manirank: Multiple attribute and intersectional group fairness for consensus ranking. In *2022 ieee 38th international conference on data engineering (icde)* (pp. 1124–1137).

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chen, D., Hong, W., & Zhou, X. (2022a). Transformer network for remaining useful life prediction of lithium-ion batteries. *IEEE Access*, *10*, 19621-19628. doi: 10.1109/ACCESS.2022.3151975

Chen, D., Hong, W., & Zhou, X. (2022b). Transformer network for remaining useful life prediction of lithium-ion batteries. *Ieee Access*, *10*, 19621–19628.

Cohen, J., Huan, X., & Ni, J. (2023). Shapley-based explainable ai for clustering applications in fault diagnosis and prognosis. *arXiv preprint arXiv:2303.14581*.

Cummins, L., Killen, B., Thomas, K., Barrett, P., Rahimi, S., & Seale, M. (2021). Deep learning approaches to remaining useful life prediction: a survey. In *2021 ieee symposium series on computational intelligence (ssci)* (pp. 1–9).

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). *Rank aggregation revisited.* Citeseer.

Elkhawaga, G., Elzeki, O., Abuelkheir, M., & Reichert, M. (2023). Evaluating explainable artificial intelligence methods based on feature elimination: A functionality-grounded approach. *Electronics*, *12*(7), 1670.

Ferraro, A., Galli, A., Moscato, V., & Sperlì, G. (2022). Evaluating explainable artificial intelligence tools for hard disk drive predictive maintenance. *Artificial Intelligence Review*, 1–36.

Hong, C. W., Lee, C., Lee, K., Ko, M.-S., & Hur, K. (2020). Explainable artificial intelligence for the remaining useful life prognosis of the turbofan engines. In *2020 3rd ieee international conference on knowledge innovation and invention (ickii)* (pp. 144–147).

Hong, C. W., Lee, C., Lee, K., Ko, M.-S., Kim, D. E., & Hur, K. (2020). Remaining useful life prognosis for turbofan engine using explainable deep neural networks with dimensionality reduction. *Sensors*, *20*(22), 6626.

Jafari, S., & Byun, Y.-C. (2022). Xgboost-based remaining useful life estimation model with extended kalman particle filter for lithium-ion batteries. *Sensors*, *22*(23), 9522.

Jakubowski, J., Stanisz, P., Bobek, S., & Nalepa, G. J. (2022). Performance of explainable ai methods in asset failure prediction. In *Computational science–iccs 2022: 22nd international conference, london, uk, june 21–23, 2022, proceedings, part iv* (pp. 472–485).

Jayasinghe, L., Samarasinghe, T., Yuenv, C., Low, J. C. N., & Ge, S. S. (2019). Temporal convolutional memory networks for remaining useful life estimation of industrial machinery. In *2019 ieee international conference on industrial technology (icit)* (pp. 915–920).

Jiao, Z., Wang, H., Xing, J., Yang, Q., Yang, M., Zhou, Y., & Zhao, J. (2023). A lightgbm based framework for lithium-ion battery remaining useful life prediction under driving conditions. *IEEE Transactions on Industrial Informatics*.

Keleko, A. T., Kamsu-Foguem, B., Ngouna, R. H., & Tongne, A. (2022). Artificial intelligence and real-time predictive maintenance in industry 4.0: a bibliometric analysis. *AI and Ethics*, *2*(4), 553–577.

Khan, T., Ahmad, K., Khan, J., Khan, I., & Ahmad, N. (2022). An explainable regression framework for predicting remaining useful life of machines. In *2022 27th international conference on automation and computing (icac)* (pp. 1–6).

Klementiev, A., Roth, D., & Small, K. (2008). Unsupervised rank aggregation with distance-based models. In *Proceedings of the 25th international conference on machine learning* (pp. 472–479).

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Lestari, S., Adji, T. B., & Permanasari, A. E. (2018). Performance comparison of rank aggregation using borda and copeland in recommender system. In *2018 international workshop on big data and information security (iwbis)* (pp. 69–74).

Lipu, M. H., Hannan, M., Hussain, A., Hoque, M., Ker, P. J., Saad, M. M., & Ayob, A. (2018). A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations. *Journal of cleaner production*, *205*, 115–133.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1–42.

Ni, Q., Ji, J., & Feng, K. (2022). Data-driven prognostic scheme for bearings based on a novel health indicator and gated recurrent unit network. *IEEE Transactions on Industrial Informatics*, *19*(2), 1301–1311.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Parimbelli, E., Buonocore, T. M., Nicora, G., Michalowski, W., Wilk, S., & Bellazzi, R. (2023). Why did ai get this one wrong?—tree-based explanations of machine learning model predictions. *Artificial Intelligence in Medicine*, *135*, 102471.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Rauf, H., Khalid, M., & Arshad, N. (2022). Machine learning in state of health and remaining useful life estimation: Theoretical and technological development in battery degradation modelling. *Renewable and Sustainable Energy Reviews*, *156*, 111903.

Remadna, I., Terrissa, L. S., Al Masry, Z., & Zerhouni, N. (2022). Rul prediction using a fusion of attention-based convolutional variational autoencoder and ensemble learning classifier. *IEEE Transactions on Reliability*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).

Schmitt, E. J., & Jula, H. (2007). On the limitations of linear models in predicting travel times. In *2007 ieee intelligent transportation systems conference* (pp. 830–835).

Sergeev, A., & Del Balso, M. (2018). Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint*