Enhancing Explainability in Predictive Maintenance : Investigating the Impact of Data Preprocessing Techniques on XAI Effectiveness

Mouhamadou Lamine NDAO [1,2], Genane Youness [1,2], Ndèye Niang [2], Gilbert Saporta [2]

[1] Laboratoire LINEACT CESI, IDFC, Nanterre, France
{mlndao, gyouness}@cesi.fr
[2] Laboratoire Cedric-MSDMA, Paris, France
{mouhamadou.ndao, genane.youness}@lecnam.net ; {n-deye.niang-keita, gilbert.saporta}@cnam.fr

Abstract

In predictive maintenance, the complexity of the data often requires the use of Deep Learning models. These models, called "black boxes", have proved their worth in predicting the Remaining Useful Life (RUL) of industrial machines . However, the inherent opacity of these models requires the incorporation of post-hoc explanation methods to enhance transparency. The quality of the explanations provided is then assessed using so-called evaluation metrics. Modeling is a whole process that includes an important data preprocessing phase, with parameter selection such as time window, smoothing parameter, or rectified RUL when dealing with multivariate time series dataset. We propose to analyze the impact of these preprocessing methods on the quality of explanations provided by the local post-hoc models LIME, KernelSHAP, and L2X, utilizing six evaluation metrics: stability, consistency, congruence, selectivity, completeness, and acumen. This analysis will be based on NASA's Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset with the LSTM model. Our findings reveal that the choice of specific pre-processing parameters can significantly improve predictive performance. Furthermore, the quality of explanations depends on the selection of explicability methods. In addition, a factorial analysis of the evaluation metrics reveals that they do not all point in the same direction. Indeed, understanding the nuanced relationships between evaluation metrics is essential for a comprehensive and accurate assessment of explainability methods.

Introduction and Background

Remaining useful life (RUL) is the key parameter for assessing tool degradation and ensuring the performance of industrial asset health management (PHM). Artificial intelligence (AI) techniques, and in particular the deep learning approach (Son and Oh 2022), have proven their worth in predicting remaining useful life (Ferreira and Gonçalves 2022). Despite their power, these methods are often regarded as "black boxes" due to their complex internal structures and lack of transparency. To boost confidence in AI adoption and address the interpretability challenge, eXplainable AI (XAI) has been presented as a solution to this problem, under the aegis of the US Defense Advanced Research Projects Agency (DARPA) (Gunning and Aha 2019). Various XAI techniques have been developed and classified into distinct categories (Molnar 2022) depending on the method used to generate explanations, the type of explanation, and the scope of the explanation technique. The explanation methods could be intrinsic or post-hoc, depending on whether the explanation mechanism operates during or after the learning phase; agnostic or specific, based on their applicability to AI models; global, cohort, or local, based on the scope of the explanation to be provided for a single prediction from the AI system, focus on a subspace in the decision space, or the entire set of predictions. Cyrus (Cyras et al. 2020) distinguishe three types of explanations : attributive (e.g. logical inference attribution, feature importance association), contrastive (e.g. counterfactuals) and actionable (e.g. guidelines towards a desired outcome, realizable actions).

Several publications support the idea that XAIs are essential for providing an accurate and comprehensible model for estimating RUL (Nor, Pedapait, and Muhammad 2021). However, one might question the reliability of these XAI methods. We might also ask how the quality of these explanations can be assessed, and whether it is possible to compare the results of two XAI methods. Various approaches have been proposed for evaluating the effectiveness of eXplainable AI (XAI) methods (Coroama and Groza 2022). These approaches include human-based qualitative methods as well as quantitative approaches aimed at quantitatively evaluating some of the properties that an explanation must satisfy. For quantitative evaluation, some methods focus on examining the relationship between data and explanations (Honegger 2018a), while others are based on the relation between data, predictions, and explanations (Solís-Martín, Galán-Páez, and Borrego-Díaz 2023).

In the realm of RUL prediction, Long Short-Term Memory (LSTM) models are widely adopted as the analysis model. Proposed by (Hochreiter and Schmidhuber 1997), LSTM has showcased its effectiveness in providing valuable short- and long-term information (Vollert and Theissler 2021). It stands out as a powerful model for processing temporal data with complex structures, effectively addressing the challenge of optimizing backpropagation gradients to

adjust network weights. However, implementing such models involves various phases, among which the preprocessing phase is of great importance. In multivariate time series, this phase includes several critical steps, such as windowing, smoothing parameter tuning, and data truncation.

This study aims to verify whether the quality of explanations provided by XAI models can be influenced by this preprocessing phase adapted to multivariate time series. First, we fit a one-layer LSTM model, then we apply various XAI methods, and we compare its performance based on six selected evaluation metrics to assess the quality of the explanations. Then, we study how variations in preprocessing parameters (smoothing α, time window TW, rectified RUL RUL_{early}) influence explanation quality and identify optimal parameter combinations that result in high-quality explanations. Finally, we use a factorial approach to explore the relationships between the six evaluation metrics and ascertain their concordance in evaluating explanations.

Materials and Methods

Notations We will use the following notations:

- N : number of observations (number of Engines)
- $X = (x_i^t)_{(i \in N, t \in T)}$ the set of observations with i the Engine and t the time
- Y_t : the RUL observed at time t
- f : the prediction function of the analysis model
- $\epsilon = \{\epsilon_i\}_{(i \in N)}$: feature importance, meaning explanation for Engine i
- ρ : Spearman's rank correlation coefficient

eXplainable Artificial Intelligence (XAI)

In the context of RUL prediction, we are particularly interested in those parts of the Engine responsible for the degradation of the service life of a given Engine. Thus, in our analysis, for comparison purposes, we will focus on three approaches: LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016), KernelSHAP (Lundberg and Lee 2017) and L2X (Chen et al. 2018), which are considered perturbation-based local explanation methods. All three approaches are post-hoc, local, and "surrogate" models. Indeed, it is based on a transparent model to explain the prediction of an observation by a "black box" model. We denote (x, y) an observation in (X, y), and g the learning function of the surrogate model (e.g. linear regression).

For LIME approach, the main idea is to create a set of observations from x (denoted X_h) by the distribution h, train the linear model g on this sample with a sparsity constraint, and then use the regression coefficients phi_i as the effect of the different variables involved in the prediction. To explain a result, they rely on three phases: sampling, learning and explanation extraction. KernelSHAP uses game theory to assign a SHAP value, to each feature, describing its contribution to the final prediction. For simplicity, we will write SHAP referring to KernelSHAP. L2X attempts to find the subset of features most informative regarding corresponding prediction for this instance. The subset is determined by a feature selector, through variational approximation, uniquely optimized to maximize the mutual information MI (Latham and Roudi 2009) between features and the corresponding prediction.

The explanation is generated by perturbing the features. In the context of this study, given that we are dealing with time series, we adopt the perturbation approach proposed by (Solís-Martín, Galán-Páez, and Borrego-Díaz 2023), which is more appropriate with time series.

The quality of the explanations generated by these post-hoc methods is assessed by XAI evaluation metrics that verify certain properties that these explanations must respect, such as robustness, stability, fidelity, representativeness, etc. These metrics are briefly presented below.

XAI evaluation metrics

Doshi and al. (Doshi-Velez and Kim 2017) have outlined three categories of evaluation approaches for XAI models:

- **Human-grounded Evaluation:** encompassing methods based on general human assessment;
- **Application-grounded Evaluation:** involves approaches relying on human assessment specific to a particular application, with a predominant emphasis on expert opinions within the relevant domain;
- **Functionally-grounded Evaluation:** pertains to approaches utilizing mathematical functions to evaluate the quality of post-hoc models quantitatively.

This work will focus on "Functionally-grounded Evaluation" and elaborate on six available evaluation metrics. Each of these metrics evaluates a specific property (Nauta et al. 2023). Thus, they could complement each other in evaluating an XAI method. We consider three types of metrics :

1. **Without perturbation:** which includes no disturbance to the dataset in its calculation process.
   - **Stability :** It evaluates the robustness of an XAI method. According to this metric (Honegger 2018b), if two observations are similar regarding X, they should be similar regarding the explanations ϵ : so we should have
   $$\rho_i = \rho(XX_i, E_i) \;\; \rho_i > 0 \;\; \forall\, i \in N \qquad (1)$$
   where: $XX_i = \{d(x_i, x_0), \ldots, d(x_i, x_n)\}$ and $E_i = \{d(\epsilon_i, \epsilon_0), \ldots, d(\epsilon_i, \epsilon_n)\}$. The greater the stability of the XAI model, the more intuitive the interpretation.
2. **Perturbation of irrelevant variables:** which consists in perturbing the most important variables in the calculation process.
   - **Coherence:** evaluates the coherence property of an XAI method by calculating the difference between the prediction error p_e^i with real data and the prediction error after perturbation of unimportant features according to the XAI method e_e^i. It is given by $\alpha_i = \left|p_e^i - e_e^i\right|$. When the model indexes unimportant variables as unimportant, a perturbation of unimportant

variables should not significantly affect prediction errors. As a result, there should be a minimal difference between prediction errors. Therefore, the smaller α, the better the consistency of the explainability approach.

- **Congruence:** This metric corresponds to the standard deviation of coherence given by :

$$\delta = \sqrt{\frac{\sum (\alpha_i - \bar{\alpha})^2}{N}} \tag{2}$$

where $\bar{\alpha} = \frac{1}{N} \sum \alpha_i$. It assesses the variability of coherence; hence, it should be minimized.

- **Completeness:** assesses the representativeness. Given by $\gamma_i = \frac{e_e^i}{p_e^i}$, it evaluates the ratio between the initial prediction error and the error following a perturbation in the initial data. The closer it is to 1, the better the quality of the explainer.

3. **Perturbation of the most important variables:** in which the least important variables are perturbed.

- **Acumen:** (Solís-Martín, Galán-Páez, and Borrego-Díaz 2023) evaluates the robustness of an XAI method. It is used to check whether the importance of the variable according to the XAI model does not depend on its arrangement (rank) in the data. Indeed, according to the author, if a variable is important according to an XAI model, it must remain so after perturbation. It is calculated as follows :

$$\omega = 1 - \frac{\sum_{f_i \in \mathcal{I}} \frac{p_a(f_i)}{N}}{M} \tag{3}$$

Where $\mathcal{I}$ is a subset of the M variables before perturbation; $p_a(f_i)$ returns the variable rank f_i in terms of importance after perturbation.

- **Selectivity:** measures the selectivity property of an XAI method. Unlike the previous metrics, which are based on the disturbance of the unimportant variables in the explainability model, this metric is based on the disturbance of the most important variables. It is calculated as follows:

(a) Order the variables according to their importance given by an explainability method;

(b) Introduce a perturbation into the data by substituting random variables for the most important variables, then calculate the prediction error for each perturbation within the regression framework.

Experimentation

Data Overview

In the area of predictive maintenance, the main problem is the lack of real data, especially when it comes to RUL prediction. Thus, the leading works (Solís-Martín, Galán-Páez, and Borrego-Díaz 2023), (Baptista, Goebel, and Henriques 2022) carried out in this field are based on NASA's C-MAPSS dataset (Saxena et al. 2008). These simulated data provide a comparative framework for the results obtained by the various studies. Our experiment uses the *FD004* subset of the C-MAPSS dataset. This simulated dataset consists of an aircraft turbine Engine life cycle. It contains 21 sensors, 6 operational conditions characterized by Mach number, altitude, and Engine ambient temperature, and 2 failure modes. In the training set, the failure amplifies until the system fails; in the test set, the time series ends just before the system fails. The aim is to predict the number of operational cycles remaining before failure in the test set. We retained 14 of the 21 sensors, excluding those that showed no variation.

Preprocessing Approach

In our previous work (Youness and Aalah 2023), we studied in detail the roles and functionalities of features during the preprocessing stages. We used feature clustering as part of the pre-processing pipeline and conducted an in-depth analysis to assess its various aspects and impact on model performance. In this study time-series preprocessing is conducted in three phases: exponential smoothing, time window, and rectified RUL, which can affect the interpretability of XAI models.

First, we normalized the sensor readings for each operational condition using the min-max normalization technique, which assigns values between [-1,1].

Subsequently, an exponential smoothing process is applied to produce an accurate RUL estimate despite noise in the data. Exponential smoothing assigns different weights to historical observations based on their recency. The choice of the smoothing parameter α in exponential smoothing determines the level of emphasis on recent observations. The fitted values used the smoothing parameter α according to the following equation:

$$\hat{y}_{t+1|t} = \alpha y_t + (1 - \alpha)\hat{y}_{t|t-1} \tag{4}$$

Setting α close to 1 indicates fast learning, meaning that forecasts are based on the most recent values, while a value close to 0 indicates slow learning.

After reducing noise by exponential smoothing, a fixed-length sliding time window TW is applied to convert the multivariate time series dataset. Longer time windows can indeed have a significant impact on the performance of a model. Still, they may also introduce challenges such as increased computational complexity and potential delays in adapting to rapidly changing patterns. This parameter needs to be adjusted.

In an addition step, a rectified RUL value, denoted RUL_{early} , is set. This value sets a threshold for the RUL, defining the system as "healthy" until it reaches this predefined point. This will allow the model to focus on learning from the RUL_{early} cycle, regardless of the Engine's previous lifetime.

Our study analyzes the influence of smoothing parameter α, time window TW, and rectified remaining RUL RUL_{early} on model performance and the quality of explanations provided by XAI methods.

Experimental setting

We have employed the following experimental design to evaluate the impact of these three preprocessing parameter

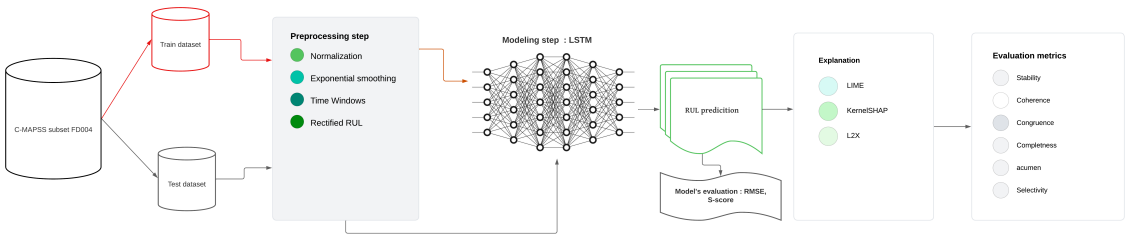

Figure 1: Flowchart for a triplet of preprocessing parameters (α, TW, RUL_{early})

choices. Based on the values used in the literature (Vollert and Theissler 2021), we have selected the following possible values for each parameter.

$$\alpha = [0.1, 0.2, 0.3, 0.5]$$

$$TW = [20, 25, 30, 35, 40, 60]$$

$$RUL_{early} = [100, 120, 130, 140, 150]$$

In total, we have 120 trials. In each trial, our approach consists of training LSTM model with a single layer considering $(\alpha, TW, RUL_{early})$ in the preprocessing step. Then we :

1. Evaluate model performance using RMSE, and S-score;
2. Generate an explanation of predictions using a local approach: LIME, KernelSHAP and L2X ;
3. Evaluate the quality of the results of these XAI methods via 6 evaluation metrics: stability, coherence, completeness, congruence, selectivity and acumen.

The flowchart illustrates the steps of our proposed work (Figure 1). The hyperparameters for the LSTM model include one hidden layer with 64 nodes, a dropout rate of 0.2, a batch size of 120, training for 20 epochs with a learning rate of 0.001, and utilizing the Adam optimizer.

Results and discussion

Best preprocessing parameters

The results indicate that the parameters leading to the best model correspond to α = 0.5, TW = 40, and RUL_{early}= 100, as illustrated in Table 1.

α	TW	RUL_{early}	$RMSE$	R^2	*S-score*
0.10	60	100	**9.84**	0.88	731.43
0.50	40	100	10.14	**0.90**	**506.47**
0.10	25	120	10.38	0.87	1195.34
0.50	30	120	10.40	0.89	688.00
0.20	35	150	10.48	0.90	579.74

Table 1: Top 5 Performance Models by RMSE, R^2 and S-score according to preprocessing parameters

Figure 2 shows that the weakest performance occurs when TW is set at 60. A similar examination is carried out for the smoothing parameter α. The results show that the most favorable models in terms of RMSE and S-score are obtained

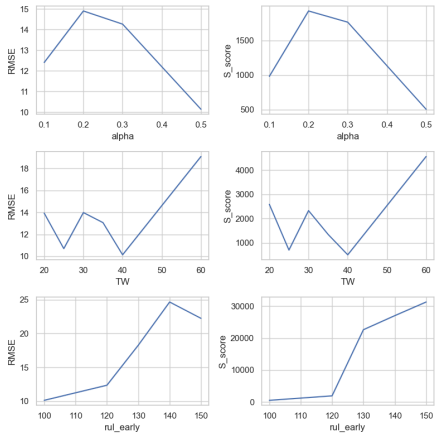

Figure 2: Performance Models by RMSE and S-score when varying one Preprocessing Parameter (e.g. the time window TW) and fixing the two others (e.g. α and rectified RUL: $\alpha = 0.5$, $RUL_{early} = 100$)

when α is set to 0.5, while the least favorable model is associated with $\alpha = 0.2$. In rectified RUL, the highest performance is observed when RUL_{early} is set to 100. The analysis indicates the model achieves optimal performance with an RMSE equal to 10.14 and S-Score equal to 506.47, using the Optimal Preprocessing Parameters (OPP): $\alpha = 0.5$, $TW = 40$, and $RUL_{early} = 100$. This performance stands out favorably compared to literature benchmarks in the context of RUL prediction using the C-MAPSS *FD004* dataset, as illustrated in Table 2.

Authors	**Approach**	**RMSE (std)**	**S-score(std)**
(Wang et al. 2022)	B-LSTM	16.24	5220
(Qin et al. 2022)	SD-TemCapsNet	16.49	804
(Youness and Aalah 2023)	FC+LSTM	16.14(0.96)	1299(255)
(Cao 2023)	DCNN-BiLSTM	13.77	–
(Arunan et al. 2024)	ChangePoint-LSTM	18.69	–
(Wahid et al. 2024)	TCRSCANet	16.23	1107
(Proposed method)	OPP+LSTM	**10.14**	**506**

Table 2: Comparison of different approaches using *FD004* (OPP = Optimal Preprocessing Parameters)

Given the Optimal Preprocessing Parameters associated with the best model, we proceed to analyze the explanation

of its predictions using LIME, KernelSHAP, and L2X.

Prediction explanation for Engine 20 with LIME, KernelSHAP, and L2X

To study whether an XAI method can lead to different explanations. We explain the prediction related to Engine 20 with the three XAI approaches. According to LIME, the sensors with the most significant influence on RUL prediction for Engine 20 are T50, Ps30, W31, and T24. KernelSHAP, on the other hand, identifies BPR, htBleed, W31, and W32 as the critical sensors determining the RUL prediction for Engine 20. Lastly, L2X points out that the key sensors contributing to the RUL prediction for Engine 20 are T30, T50, and NRc (Figure 3). We, therefore, obtain an explanation that differs according to the XAI method used. This underlines the importance of evaluating explanations' quality and helps determine which approach offers the most insightful explanation.

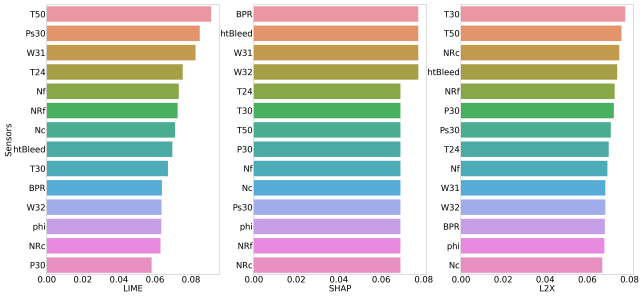

Figure 3: Features importance in the prediction of the RUL of Engine 20 according to LIME, KernelSHAP and L2X

Such evaluations assist in selecting the most suitable approach for interpreting a model's predictions. In this research, we opt for 6 evaluation metrics that will enable us to compare the quality of explanations obtained using the three approaches: LIME, KernelSHAP, and L2X.

Evaluation metric analysis for LIME, KernelSHAP, and L2X

Stability remains consistent across the three XAI models when we evaluate the quality of explanations based on the six metrics presented in Table 3. In most instances, LIME and KernelSHAP provide the same quality explanations on completeness, coherence, and congruence. However, KernelSHAP remains the best, providing superior explanations, especially on additional metrics such as selectivity and acumen. Subsequently, we delve into the dynamics of these metrics, considering preprocessing parameter choices while keeping one parameter constant.

Quality of the explanation based on preprocessing parameters

The analysis of the quality of explanations based on the preprocessing parameters shows that for some metrics, there

	XAI methods		
Metrics	**LIME**	**KernelSHAP**	**L2X**
Stability ↑	**1.00**	**1.00**	**1.00**
Coherence ↓	**0.17**	**0.17**	0.41
Completeness ↑	**0.83**	**0.83**	0.58
Congruence ↓	**0.20**	0.21	0.37
Selectivity ↑	0.65	**0.68**	0.57
Acumen ↑	0.08	**0.41**	0.03

Table 3: Quality of Explanations Provided by LIME, KernelSHAP, and L2X According to Six Metrics with Optimal Preprocessing Parameters

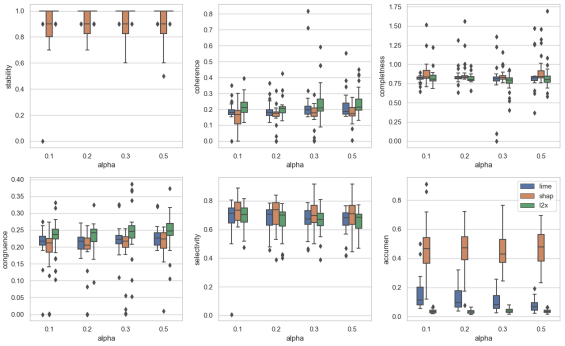

Figure 4: Quality of explanations for the three approaches (LIME, KernelSHAP, and L2X) depending on α

is minimal variation when the smoothing parameters α and rectified RUL (RUL_{early}) change. This is particularly noticeable in the case of stability. Examination of Figures 4, 5, and 6 reveals consistent values for LIME and L2X, while for KernelSHAP, the shape of the boxplot remains identical across different α values and for 4 out of 5 scenarios of rectified RUL.

KernelSHAP exhibits the greatest metric variation of the three XAI methods, characterized by a larger variance around the mean. Regardless of the preprocessing parameter value, LIME and kernelSHAP consistently deliver better average explanation quality. This distinction is particularly evident when assessing the acumen metric. Conversely, for the completeness metric, the variation in quality is less significant across all three approaches, with completeness values often concentrated around the mean. This analysis clearly shows that the three pre-processing parameters exert varying effects on the quality of the explanations provided by the three XAI methods considered. Moreover, the perception of this variation in explanation quality varies according to the different evaluation measures. It is, therefore, essential to examine the relationships between the different evaluation measures.

Relation between evaluation metrics

Once we have defined the metrics, we can classify them according to their interpretation. In the case of acumen, stability, completeness, and selectivity, a higher value means a

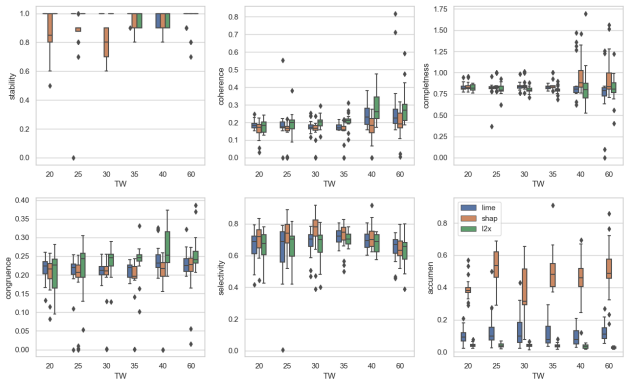

Figure 5: Explanation's quality of LIME, KernelSHAP and L2X depending on TW

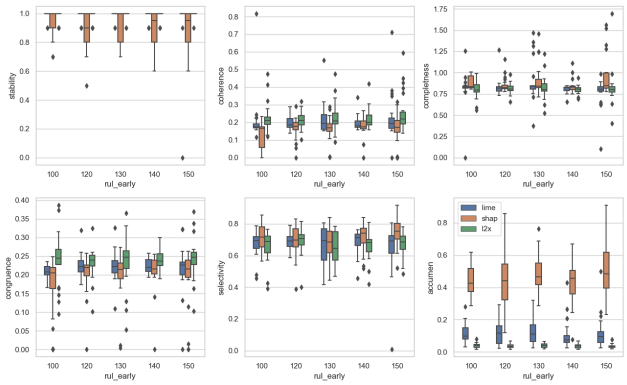

Figure 6: Explanation's quality of LIME, KernelSHAP and L2X depending on RUL_{early}

better quality explanation. Conversely, values close to 0 are considered preferable for coherence and congruence.

However, empirical analysis of the relationship between metrics using the Principal Component Analysis (PCA) (Smith 2002) approach reveals a distinct grouping structure for these six metrics (Figure 7). A significant positive correlation is observed between coherence and congruence in all three XAI methods. In essence, models with high coherence values also tend to have high congruence values, which is consistent with the definition of these measures. Notably, coherence and congruence stand in direct contrast to completeness and acumen. Moreover, selectivity does not exhibit a consistent relationship across the three approaches. This indicates that selectivity is not intrinsically correlated with the other parameters. In addition, stability, which should be aligned with comprehensiveness and acumen, sometimes clashes with them. This is unexpected since, by definition, stability is supposed to reach high values when the quality of the explanation is comparable to acumen and completeness.

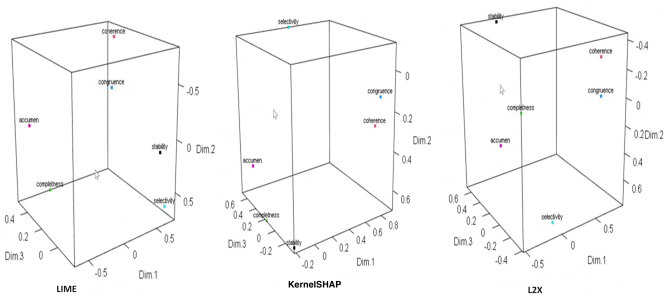

Figure 7: Relationships between evaluation metrics by XAI method in 3D view

Conclusion and future works

The present study highlights the significance of preprocessing parameter choices in improving predictive performance and explanation quality. We have conducted an in-depth analysis of the impact of certain time series preprocessing procedures, such as the time window, the smoothing parameter, and the rectified RUL, on the explanatory effectiveness of three XAI methods, namely LIME, KernelSHAP, and L2X. This analysis brought attention to the nuanced relationships between different evaluation metrics, shedding light on the complexity of assessing the quality of explanations. It revealed that explanation quality is not sensitive to all preprocessing parameters. For some parameters, explanation quality can vary. This variation is not captured by all the evaluation metrics used. So, even though there is a wide range of metrics, each evaluation metric has its specificity and enables a specific property to be evaluated.

As an empirical result, analyzing the relationships between metrics reveals a notable contradiction. Contrary to the expected definitions, parameters that are presumed to have a positive relationship display negative relationships instead. This contradiction raises the question of these parameters' real significance and importance in the evaluation process.

Therefore, future efforts will be directed toward a comprehensive analysis of these parameters, incorporating additional evaluation measures such as fidelity, identity, separability, etc. Another approach could be to merge these quantitative measures into a standardized synthetic indicator capable of evaluating explanations with greater reliability. It should be noted that this study was based on a single type of model architecture (LSTM) and a single C-MAPSS dataset (*FD004*). It is, therefore, necessary to explore a wider range of datasets and model architectures to validate the applicability of the results and improve generalization. In addition, a human evaluation of the explanations would be an effective way of assessing the quality of the explanations.

References

Arunan, A.; Qin, Y.; Li, X.; and Yuen, C. 2024. A change point detection integrated remaining useful life estimation model under variable operating conditions. *Control Engineering Practice* 144:105840.

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

Cao, G. 2023. Remaining useful life prediction of aircraft engines using dcnn-bilstm with k-means feature selection. In *International Symposium on Artificial Intelligence and Robotics*, 354–365. Springer.

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, 883–892. PMLR.

Coroama, L., and Groza, A. 2022. Evaluation metrics in explainable artificial intelligence (xai). In Guarda, T.; Portela, F.; and Augusto, M. F., eds., *Advanced Research in Technologies, Information, Innovation and Sustainability*, 401–413. Cham: Springer Nature Switzerland.

Cyras, K.; Badrinath, R.; Mohalik, S. K.; Mujumdar, A.; Nikou, A.; Previti, A.; Sundararajan, V.; and Feljan, A. V. 2020. Machine reasoning explainability.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ferreira, C., and Gonçalves, G. 2022. Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems* 63:550–562.

Gunning, D., and Aha, D. 2019. Darpa's explainable artificial intelligence (xai) program. *AI magazine* 40(2):44–58.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Honegger, M. 2018a. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*.

Honegger, M. 2018b. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *CoRR* abs/1808.05054.

Latham, P. E., and Roudi, Y. 2009. Mutual information. *Scholarpedia* 4(1):1658.

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

Molnar, C. 2022. *Interpretable Machine Learning*. NA, 2 edition.

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

Nor, A. K. B. M.; Pedapait, S. R.; and Muhammad, M. 2021. Explainable ai (xai) for phm of industrial asset: A state-of-the-art, prisma-compliant systematic review.

Qin, Y.; Yuen, C.; Shao, Y.; Qin, B.; and Li, X. 2022. Slow-varying dynamics-assisted temporal capsule network for machinery remaining useful life estimation. *IEEE Transactions on Cybernetics* 53(1):592–606.

Ribeiro, M. T.; Singh, S.; and 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*, 1135–1144.

Saxena, A.; Goebel, K.; Simon, D.; and Eklund, N. 2008. Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management (PHM)*, 1–9.

Smith, L. I. 2002. A tutorial on principal component's analysis.

Solís-Martín, D.; Galán-Páez, J.; and Borrego-Díaz, J. 2023. On the soundness of xai in prognostics and health management (phm). *Information* 14(5):256.

Son, S., and Oh, K.-Y. 2022. Integrated framework for estimating remaining useful lifetime through a deep neural network. *Applied Soft Computing* 122:108879.

Vollert, S., and Theissler, A. 2021. Challenges of machine learning-based rul prognosis: A review on nasa's c-mapss data set. In *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 1–8. IEEE.

Wahid, A.; Breslin, J. G.; Intizar, Muhammad Ali, a.; and al. 2024. Tcrscanet: Harnessing temporal convolutions and recurrent skip component for enhanced rul estimation in mechanical systems. *Human-Centric Intelligent Systems* 1–24.

Wang, X.; Huang, T.; Zhu, K.; and Zhao, X. 2022. Lstm-based broad learning system for remaining useful life prediction. *Mathematics* 10(12):2066.

Youness, G., and Aalah, A. 2023. An explainable artificial intelligence approach for remaining useful life prediction. *Aerospace* 10(5).