Through the Lens of Explainability: Enhancing Trust in Remaining Useful Life Prognosis Models

Kaouther Benguessoum[1], Raoni Lourenço[1] Vincent Bourel[2], and Sylvain Kubler[1]

[1] SnT, University of Luxembourg, 6 Rue Richard Coudenhove-Kalergi, L-1359 Luxembourg, Luxembourg
kaouther.benguessoum@uni.lu, raoni.lourenco@uni.lu, sylvain.kubler@uni.lu,
[2] CERATIZIT Luxembourg S.à r.l., L-8201 Mamer, Luxembourg

Abstract. Accurately estimating Remaining Useful Life (RUL) in industrial systems is crucial for optimizing maintenance strategies and extending the lifespan of assets. Data-driven RUL models leverage machine learning (ML) algorithms to extract patterns from operational data, excelling in capturing complex relationships. Despite advancements in RUL prognosis models, the black-box nature of machine learning algorithms poses challenges for industrial users, hindering trust and adoption. Explainable Artificial Intelligence (XAI) methods offer promising solutions by making complex models transparent and interpretable. This paper focuses on applying XAI methods to enhance trust in machine learning models for RUL prognosis. We emphasize a quantitative assessment of explanation mechanisms, including metrics such as consistency and robustness. Our study contributes to developing more trustworthy and reliable predictive maintenance strategies. We evaluate XAI methods explaining RUL models applied to a real-world scenario of industrial furnace data. Our findings aim to provide valuable insights for industrial practitioners, guiding them in selecting RUL prognosis techniques.

Keywords: remaining useful life, Industry 4.0, Explainable AI

1 Introduction

The accurate estimation of the Remaining Useful Life (RUL) of equipment in industrial systems is critical for ensuring optimal maintenance strategies, minimizing downtime, and maximizing the lifespan of assets. Over the years, researchers and practitioners have explored various approaches to RUL prognosis, leading to the development of physical and data-driven models. Physical models [7, 14], rooted in engineering principles, provide insights into the degradation processes of components. These models leverage knowledge of the system's design, materials, and operating conditions to estimate RUL. On the other hand, data-driven models [22, 13, 8, 24, 16, 20, 4] harness the power of machine learning algorithms to extract patterns and correlations from historical operational data. This paper

focuses on these models, which excel in capturing complex relationships that may be challenging for traditional physics-based approaches.

Despite the advancements in RUL prognosis models, industrial users often need help trusting the predictions generated by machine learning algorithms. The black-box nature of these models can hinder their adoption, mainly when critical decisions rely on their outputs. This lack of transparency raises concerns about the model's reliability, interpretability, and generalizability to varying operating conditions. Explainable artificial intelligence (XAI) methods are promising solutions for these trust issues. XAI methods aim to make the decision-making process of complex models more transparent and interpretable. By providing insights into the inner workings of machine learning models, XAI methods empower users to understand, validate, and ultimately trust the predictions made by these models [17]. XAI methods rely on different aspects: model specificity, scope, and the location of explanation [2]. For the model specificity, we have model-specific methods that can be only used on a specific type of model, such as Layer-wise Relevance Propagation (LRP) [3] that is designed to provide explanations for predictions made by deep neural networks. In contrast, model-agnostic methods, such as SHAP [15] and LIME [18], are designed to apply to different types of machine learning models as only the model's input and output are analyzed.

Recent works apply XAI methods to explain RUL machine learning models [17, 10, 21, 12]. In practice, decision-makers evaluate the generated explanations subjectively to select the best models. However, reviewing explanation methods to confirm their correctness objectively is equally essential. Although quantitative metrics exist to evaluate XAI outputs [23, 6], no previous study has evaluated explanations for data-driven RUL models.

In this work, we contribute to the field of RUL prognosis by investigating the application of XAI methods to enhance trust in machine learning models applied to a real industrial scenario. Our study emphasizes ensuring a quantitative assessment of the explanation mechanisms by metrics such as consistency and robustness, leveraging an XAI evaluation framework [9] specific for Deep Learning (DL) model explainers. By presenting a comprehensive analysis of these models, we aim to provide industrial users with valuable insights into the strengths and limitations of RUL prognosis techniques, ultimately contributing to developing more trustworthy and reliable predictive maintenance strategies.

In the remainder of this paper, we present the related work (Section 2), describe the methodology and evaluation metrics (Section 3), followed by the description and analysis of the experiments we conducted on real-world scenarios of industrial furnace data (Section 4), and conclude in Section 5.

2 Related Work

Machine Learning for RUL. In the early phases of data-driven models for Remaining Useful Life (RUL), traditional Machine Learning (ML) algorithms like Support Vector Machines (SVM) and Random Forests (RF) dominated. Wu et al. [22] illustrated the benefits of SVM in a *classification-regression* scheme,

showing enhanced performance in capturing intricate relationships. Multi-layer Perceptrons (MLPs) [13] and Extreme Learning Machines (ELM) [8] emerged as direct RUL prediction options, prioritizing simplicity for accuracy and reduced computational costs, particularly in specific contexts. The rise of deep learning, including Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM), became notable with their ability to handle complex patterns in massive IoT and Industry 4.0 datasets. LSTM, effective in handling sequentially correlated data [24, 16], and CNNs, showcasing remarkable accuracy [20], emphasize the significance of data-driven model choices [4].

XAI for RUL. Despite being nascent, using XAI for RUL prognosis is gaining acceptance [17]. Noteworthy contributions include a convolutional bidirectional LSTM for turbofan engine RUL prognosis [10], utilizing SHAP for validation. Another method [21] introduces RUL interpretation for deep learning, employing Shap to obtain interpretations at different hierarchies. Additionally, a proposed explainable regression framework for predicting machine RUL [12] utilizes SHAP and LIME for explanations.

XAI Evaluation Metrics. Various metrics gauge the effectiveness of XAI methods. Robustness, measured by average sensitivity through Monte Carlo sampling [23], and explanation complexity, assessed with the Gini Index [6], are relevant metrics. However, this work provides the first quantitative evaluation of XAI methods on data-driven RUL models using diverse metrics.

3 Methodology

XAI ensures the reliability of AI systems. However, when employing XAI for decision-making in high-stakes scenarios like RUL prediction, it is imperative to validate the accuracy of the generated explanations. This raises the following question: *How can we ensure the correctness of explanations to make informed decisions in RUL prognosis?*

We answer this question by providing a quantitative assessment of explanation mechanisms, which holds equal importance with subjective, inherently human-centric evaluations. Our approach, as illustrated in figure 1, uses state-of-the-art DL algorithms, specifically LSTM and CNN, for accurate RUL estimation. Subsequently, we delve into generating explanations using Layer-Wise Relevance Propagation (LRP) [3] and objectively evaluating the obtained explanations.

RUL Prognosis. To estimate the RUL, we propose to use CNN and LSTM, considering their demonstrated capabilities. CNN is chosen for its adept feature extraction capability. In this context, CNN handles sequential patterns by ingesting data divided into sliding windows. These sliding windows encapsulate features in one dimension and the corresponding time sequences of each feature

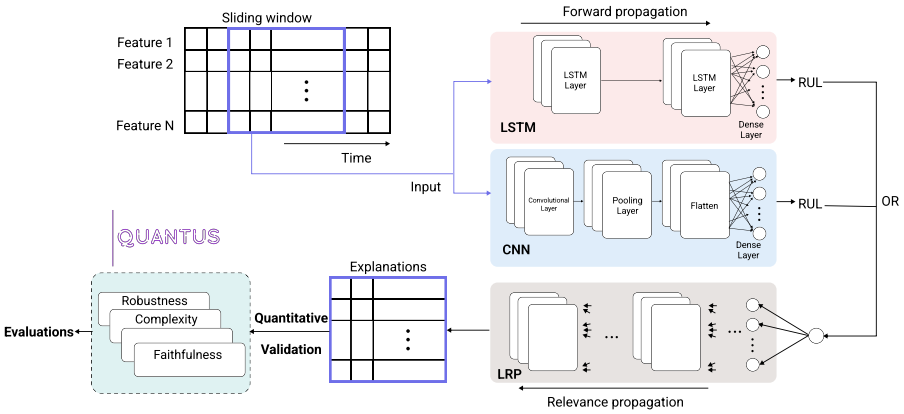

Fig. 1: Diagram of the proposed methodology for quantitative assessment.

in the other dimension. This approach enables CNN to capture the temporal dependencies effectively. LSTM is enlisted for its ability to manage sequentially correlated data. Its architecture allows precise control over the remembering and forgetting processes of both previous and current states. This makes it suitable for tasks where understanding the sequential evolution of data is needed, as is the case in RUL prediction.

Explanations generation. Both chosen DL algorithms function as black boxes, offering decisions without revealing any information about the factors influencing them. To address this opacity, we employ LRP as our XAI method. LRP facilitates the identification of each input feature's contribution to the model's output by operating in reverse, starting from the output and assigning relevance scores to each feature, indicating their importance in shaping the final prediction.

Example 1. Let's consider a machine with features like Temperature, Pressure, and Vibration, and we want to predict its Remaining Useful Life (RUL). After RUL prediction and LRP application, we obtain relevance values: 0.2 for Temperature, 0.3 for Pressure, and 0.5 for Vibration. These values represent each feature's contribution to the model's output. The obtained results can be visualized via a heatmap, providing a clear and intuitive understanding of the relative importance of each feature, as illustrated in the figure 2.

The utilization of LRP is motivated by two key factors: i) LRP offers a model-specific approach that seamlessly aligns with the internal architecture of DNN. ii) Given the sequential nature of the data, LRP provides a granular understanding of the contributions of each feature within each time sequence.

Explanations Evaluation. Following the generation of LRP explanations for both the CNN and LSTM, ensuring the reliability of these explanations is

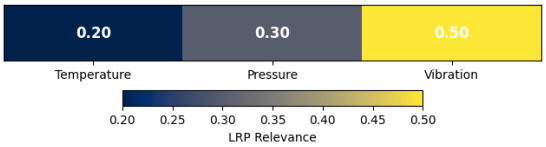

Fig. 2: Heatmap illustrating the relevance values obtained through LRP for features in RUL of a machine.

paramount for the crucial task of selecting the most suitable model (either LSTM or CNN). Recognizing the limitations of relying solely on subjective evaluations, as human assessments may vary, our methodology emphasizes objective evaluations of the obtained explanations. In our work, we propose to use Quantus [9], an XAI quantification toolkit, to attain a holistic evaluation of the generated explanation from multiple perspectives: Robustness, Complexity, Faithfulness, and more.

4 Experiments

4.1 Setup

We evaluated our work in a real-world dataset comprising furnace runs, chosen for their significance in manufacturing processes. Despite their vital role in the steel industry, no prior work has estimated the RUL of furnaces. The tabular dataset includes information about runs from seven distinct furnaces, features from different sensors, and the occurrence of failures recorded since 2022. Due to confidentiality reasons, we cannot make the dataset publicly available.To frame the dataset as an RUL problem, we estimate the remaining number of runs a furnace will undergo before a failure occurs. In line with this objective, we use a linear degradation function.

Beside CNN and LSTM models used to predict the RUL, we've also employed a state-of-the-art (SOTA) method for predicting RUL, known as DA-LSTM [19], that shows superiority in prediction accuracy. This model, proposed in the paper, introduces a lightweight approach combining dual attention and Long Short-Term Memory (LSTM). The attention mechanism (AM) captures complex degradation features, while the LSTM is used to compensate for the limitations of AM in modeling sequential information. After obtaining LRP explanations, we employed Quantus to assess the robustness and complexity of the generated explanations using two metrics for each:

– Avg-Sensitivity [23] assesses the robustness of the explanations by evaluating their stability and sensitivity in the face of input perturbations. It quantifies how much an explanation changes in response to insignificant perturbations in the input. Essentially, if the input being explained undergoes a slight modification without significantly altering the model's prediction, we expect

minimal changes in the corresponding explanation as well. Additionally, explanations exhibiting high sensitivity may be more vulnerable to adversarial attacks, as noted by [1]

- Local Lipschitz Estimate [1] is another measure that evaluates the consistency and stability of generated explanations. It measures changes in the output caused by variations in the input. We specifically consider local stability, focusing on neighboring inputs. This metric states that similar inputs should lead to relatively consistent explanations. It relies on the local Lipschitz continuity, which assesses the smoothness of the function within a neighborhood of each point
- Sparseness [6] measures the complexity of explanations. Its primary objective is to determine if a minimal set of features is enough to explain the model. This entails assessing whether we can effectively describe the model's workings using only a concise and essential set of features (i.e. only the features with significant contributions are included). If this criterion is met, we categorize the explanation as simple and human friendly. The sparseness in our case is quantified using the Gini Index [11].
- Complexity [5] also enables us to assess the features utilized in the explanation. Given humans' limited capacity to process vast amounts of information at once, explanations with fewer features are preferable. In this case, complexity is measured by the fractional contribution of each feature to the total magnitude of the attribution. If every feature received equal attribution, the explanation would be complex. Conversely, the simplest explanation would highlight only one feature.

4.2 Results

Table 1 displays experimental results in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), revealing that both LSTM and DA-LSTM outperform the CNN model in terms of accuracy. This superiority is due to LSTM's ability to capture sequential and temporal dependencies in the data more effectively than CNN. While CNN excels in capturing local patterns, LSTM is designed to capture long-range dependencies, making it better suited for the RUL prediction a task that might require understanding events occurring far back in time.

Table 2 presents a comparison of the evaluation metrics for explanations generated by CNN, LSTM, and DA-LSTM models. Notably, CNN-generated explanations are more straightforward and robust. This superiority can be attributed to the architectural characteristics of CNN, which tend to reduce the dimensionality of the input to emphasize only essential features. This results in generating simpler explanations focused on important features. Furthermore, its hierarchical feature extraction method organizes information in layers, simplifying the process of assigning importance to different input parts. This allows for a stable and reliable way of assigning importance or relevance to different parts of the input data. In contrast, explanations derived from LSTM and DA-LSTM

models may yield more variability due to the sequential nature of their processing, leading to inconsistencies based on the order of events. This inconsistency impacts both the robustness of the generated explanations as well as the features used for explanation.

Table 1: Experimental results of CNN and two LSTM-based models in terms of RMSE and MAE.

Method	RMSE	MAE
CNN	36.59	17.38
LSTM	16.89	7.88
DA-LSTM	9.59	5.83

Table 2: Quantitative evaluation of LRP explanations in terms of robustness and complexity.

LRP of	Robustness		Complexity	
	Avg-Sens [23]	Lipschitz [1]	Spars [6]	Complex [5]
CNN	0.35	9.97	0.26	4.85
LSTM	0.78	15.49	0.66	4.32
DA-LSTM	0.80	17.54	0.72	4.50

While LSTM models demonstrate better performance in predicting RUL, CNN provides more robust and simpler explanations. Hence, a trade-off between performance and explainability is essential, as it underscores the need for careful consideration in model selection, particularly in real-world applications where decision-makers rely on model predictions to inform actions. It is important to recognize that high performance alone does not guarantee the quality of explanations based on which decisions are made. Additionally, these results may not universally apply to all CNN and LSTM models as data characteristics play an equally significant role. Nevertheless, emphasizing the importance of quantitative evaluation metrics such as robustness, simplicity, and consistency is essential for assessing the reliability and trustworthiness of model explanations.

5 Conclusion

In this study, we introduced an approach for validating RUL prediction explanations through quantitative evaluation. Using CNN and LSTM algorithms

for RUL prediction and employing LRP for explanation generation, we evaluated the explanations generated for both models. Our evaluation focused on assessing the robustness and complexity of the models, revealing the critical importance of quantitative evaluation in enhancing decision-making based on explanations. Our findings indicate that while LSTM-based models outperform in RUL prediction accuracy, CNN offers more robust and simpler explanations. This underscores the necessity of finding a trade-off between performance and explainability. Looking ahead, our future endeavors involve expanding the scope by exploring additional evaluation metrics and addressing diverse use cases.

Acknowledgment. This research was funded in whole or in part by the Luxembourg National Research Fund (FNR), grant reference 18435508.

Bibliography

[1] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Tarek Berghout and Mohamed Benbouzid. A systematic guide for predicting remaining useful life with machine learning. *Electronics*, 11(7):1125, 2022.

[5] Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

[6] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020.

[7] Yiwei Cheng, Kui Hu, Jun Wu, Haiping Zhu, and Xinyu Shao. A convolutional neural network based degradation indicator construction and health prognosis using bidirectional long short-term memory network for rolling bearings. *Advanced Engineering Informatics*, 48:101247, 2021.

[8] Zehai Gao, Cunbao Ma, Jianfeng Zhang, and Weijun Xu. Enhanced online sequential parallel extreme learning machine and its application in remaining useful life prediction of integrated modular avionics. *IEEE Access*, 7:183479–183488, 2019.

[9] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[10] Chang Woo Hong, Changmin Lee, Kwangsuk Lee, Min-Seung Ko, Dae Eun Kim, and Kyeon Hur. Remaining useful life prognosis for turbofan engine using explainable deep neural networks with dimensionality reduction. *Sensors*, 20(22):6626, 2020.

[11] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[12] Talhat Khan, Kashif Ahmad, Jebran Khan, Imran Khan, and Nasir Ahmad. An explainable regression framework for predicting remaining useful life

of machines. In *2022 27th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE, 2022.

[13] Meghdad Khazaee, Ahmad Banakar, Barat Ghobadian, Mostafa Agha Mirsalim, and Saeid Minaei. Remaining useful life (rul) prediction of internal combustion engine timing belt based on vibration signals and artificial neural network. *Neural Computing and Applications*, 33:7785–7801, 2021.

[14] Jialin Li, Xueyi Li, and David He. A directed acyclic graph network combined with cnn and lstm for remaining useful life prediction. *IEEE Access*, 7:75464–75475, 2019.

[15] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[16] Meng Ma and Zhu Mao. Deep-convolution-based lstm network for remaining useful life prediction. *IEEE Transactions on Industrial Informatics*, 17(3):1658–1667, 2020.

[17] Ahmad Kamal Bin Mohd Nor, Srinivasa Rao Pedapait, and Masdi Muhammad. Explainable ai (xai) for phm of industrial asset: A state-of-the-art, prisma-compliant systematic review. *arXiv preprint arXiv:2107.03869*, 2021.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[19] Jiayu Shi, Jingshu Zhong, Yuxuan Zhang, Bin Xiao, Lei Xiao, and Yu Zheng. A dual attention lstm lightweight model based on exponential smoothing for remaining useful life prediction. *Reliability Engineering & System Safety*, 243:109821, 2024.

[20] Biao Wang, Yaguo Lei, Tao Yan, Naipeng Li, and Liang Guo. Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery. *Neurocomputing*, 379:117–129, 2020.

[21] Yilin Wang, Yuanxiang Li, Yuxuan Zhang, Yongsheng Yang, and Lei Liu. Rushap: A unified approach to interpret deep learning model for remaining useful life estimation. In *2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)*, pages 1–6. IEEE, 2021.

[22] Ji-Yan Wu, Min Wu, Zhenghua Chen, Xiaoli Li, and Ruqiang Yan. A joint classification-regression method for multi-stage remaining useful life prediction. *Journal of Manufacturing Systems*, 58:109–119, 2021.

[23] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Kaisheng Zhao, Jing Zhang, Shaowei Chen, Pengfei Wen, Wang Ping, and Shuai Zhao. Remaining useful life prediction method based on convolutional neural network and long short-term memory neural network. In *2023 Prognostics and Health Management Conference (PHM)*, pages 336–343. IEEE, 2023.