A Comparative Assessment of eXplainable AI Tools in Predicting Hard Disk Drive Health*

(extended abstract)

Flora Amato[1], Antonino Ferraro[1], Antonio Galli[1,*], Valerio La Gatta[1], Francesco Moscato[2], Vincenzo Moscato[1], Marco Postiglione[1], Carlo Sansone[1] and Giancarlo Sperlì[1]

[1]Department of Electrical Engineering and Information Technology, University of Naples Federico II, Italy
[2]Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Italy

Abstract

In addressing the challenge of optimizing maintenance operations in Industry 4.0, recent efforts have focused on predictive maintenance frameworks. However, the effectiveness of these frameworks, largely relying on complex deep learning models, is hindered by their lack of explainability. To address this, we employ eXplainable Artificial Intelligence (XAI) methodologies to make the decision-making process more understandable for humans. Our study, based on a previous work, specifically explores explanations for predictions made by a recurrent neural network-based model designed for a three-dimensional dataset, used to estimate the Remaining Useful Life (RUL) of Hard Disk Drives (HDDs). We compare the explanations provided by different XAI tools, emphasizing the utility of global and local explanations in supporting predictive maintenance tasks. Using the Backblaze Dataset and a Long Short-Term Memory (LSTM) prediction model, our developed explanation framework evaluates Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) tools. Results show that SHAP outperforms LIME across various metrics, establishing itself as a suitable and effective solution for HDD predictive maintenance applications.

Keywords

eXplainable Artificial Intelligence, Predictive Maintenance, LSTM-based model, Deep Learning

1. Introduction

Over the past decade, numerous companies have increasingly turned their attention to Artificial Intelligence (AI) and Machine Learning (ML) techniques. This shift is driven by the potential of these technologies to design models that support practitioners across various tasks, leveraging abundant data. Notable applications include predictive maintenance [1], product recommendation [2], and labor market analysis [3].

A paradigm shift is evident towards the adoption of more sophisticated models based on Deep Learning (DL) [4, 5]. This transition is fueled by their enhanced accuracy in handling

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
flora.amato@unina.it (F. Amato); antonino.ferraro@unina.it (A. Ferraro); antonio.galli@unina.it (A. Galli); valerio.lagatta@unina.it (V. L. Gatta); fmoscato@unisa.it (F. Moscato); vincenzo.moscato@unina.it (V. Moscato); marco.postiglione@unina.it (M. Postiglione); carlo.sansone@unina.it (C. Sansone); giancarlo.sperli@unina.it (G. Sperlì)

larger datasets, facilitated by advancements in computing power, particularly attributed to the evolution of Graphics Processing Units (GPUs).

Recent research efforts, exemplified by studies such as [6], underscore a significant industry challenge—the maintenance of technological equipment.

Today, despite the ongoing shift towards Industry 4.0 and the emerging Industry 5.0, many companies still rely on periodic and corrective maintenance strategies [7]. Industry 5.0 introduces a novel manufacturing paradigm emphasizing collaboration between machines and humans to enhance efficiency, productivity, and worker well-being [8]. This paradigm shift involves combining human and equipment capabilities, creating digital twins of entire systems, and implementing artificial intelligence for automatic and efficient industrial processes [9, 10].

In dynamic industrial settings, there is a rising demand for automated predictive maintenance systems analyzing extensive data volumes through condition monitoring [11]. Predictive maintenance aims to optimize costs by maximizing equipment's Remaining Useful Life (RUL), offering a potential return on investment up to 100% and reducing correction costs by up to 60% [12, 13]. Approaches for predictive maintenance are categorized into three groups: *physical model-based*, *data-driven*, and *hybrid* [14]. Physical model-based approaches face challenges in modeling complex systems, while data-driven methods learn system behavior from historical data. Hybrid methods combine both approaches [15].

In recent years, the widespread use of deep learning (DL) models in various industrial applications, such as fault diagnosis [16], classification [17], and predicting industrial Key Performance Indicators (KPIs) [18], has been fueled by increased computing power. Despite their impressive results, these models, often considered as black boxes, face resistance due to the need for interpretability, tractability, and reliability in line with the demand for ethical AI [19]. In the context of Industry 5.0, marked by collaborative efforts between machines and humans [20], the explanation of AI model predictions (explainability) becomes crucial. This has given rise to eXplainable Artificial Intelligence (XAI), defined as systems capable of elucidating decision logic, revealing strengths and weaknesses in decision-making, and offering insights into future behavior [21]. A significant predictive maintenance task involves estimating the Remaining Useful Life (RUL) of Hard Disk Drives (HDDs) [22], crucial for data centers. In this study, we conduct a systematic evaluation of XAI methodologies to explain predictions made by a Long Short-Term Memory (LSTM)-based model assessing HDD health. Despite its superior accuracy, precision, and recall [23], the LSTM model lacks explainability due to its reliance on a three-dimensional dataset $(x_{samples}, y_{timesteps}, z_{features})$, combining spatial and temporal features.

This paper represents an extended abstract of a recent proposal [24], in which the authors present an explanation framework that evaluates the effectiveness of XAI tools, focusing on Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), using the Backblaze dataset. This effort represents one of the first attempts to evaluate the practical utility of XAI tools in real application contexts, both methodologically and operationally.

The structure of the paper is as follows: Section 2 presents a systematic overview of XAI tools. The proposed framework, consisting of two modules (prediction and explanation), is detailed in Section 3. Main findings regarding the explanation of the prediction module on the Backblaze dataset using LIME and SHAP are discussed in Section 4, along with their empirical evaluation. Section 5 concludes the paper and suggests possible future directions.

2. Comprehensive Analysis of XAI

Three fundamental concepts have been introduced to support XAI methodologies: *Interpretability*, which entails the ability to explain in terms understandable to humans [25]; *Explainability*, associated with the role of explanation as a bridge between humans and decision-makers [26]; and *Transparency*, indicating inherent understandability [27]. A clear distinction is evident between models designed for interpretation (*transparent models*) and those necessitating external XAI techniques for explanation (*post-hoc* models).

In the first category, encompassing three levels [25], each level includes its predecessors: *Algorithmic transparency* involves the user's understanding of the model's process to generate output data from its input; *Decomposability* pertains to the ability to explain each component of the model, including input, parameters, and calculations; and *Simulability* refers to the model's ability to be simulated. The *post-hoc* techniques can be categorized as *model agnostic* or *specific*, depending on whether they are model-dependent. The former may involve model simplification, local explanation, feature relevance estimation, and visualization techniques [25].

Most techniques for simplification rely on rule extraction, with notable examples being LIME [28] and Anchors [29]. In particular, LIME builds local linear models around predictions of an opaque model, explaining it.

The second category aims to describe the behavior of black-box models by classifying or measuring the influence, relevance, or importance of each feature in the model's prediction. Noteworthy algorithmic approaches in this category include SHAP [30] and Partial Dependence Plot (PDP). In particular, SHAP computes an additive feature importance score for a particular prediction with desired properties.The third category comprises *visual explanation* techniques, generating visualizations from only the inputs and outputs of a black-box model.

We explore two crucial XAI tools, LIME and SHAP, capable of handling three-dimensional datasets. Our objective is to aid practitioners in the decision-making process, enhancing the comprehension of AI model outputs. Specifically, we seek to elucidate predictions made by the LSTM-based model for assessing HDD health status using these tools, which offer both global and local explanations, highlighting the key features influencing the predictions.

3. Framework

The rapid growth in technology services has escalated the demand for archive space, making Hard Disk Drives (HDDs) the primary storage solution in data centers. This shift has increased the risk of downtime, data loss, and unavailability in data centers. Predicting the health status of HDDs is crucial for optimizing maintenance strategies, reducing costs, and extending the HDDs' Remaining Useful Life (*RUL*). Commonly, health status prediction relies on analyzing *Self-Monitoring, Analysis and Reporting Technology* (S.M.A.R.T.) attributes, often implemented through complex deep-based models. However, their black-box nature poses challenges in understanding predictions.

Our focus is on investigating eXplainable Artificial Intelligence (XAI) techniques for LSTM-based models applied to real-world scenarios, specifically HDDs' health status prediction. The complexity of these models necessitates XAI tools, such as LIME and SHAP, to provide

explanations for predictions. The designed framework for HDDs' Remaining Useful Life (RUL) estimation utilizes an LSTM-based model, focusing on the analysis of dependencies between *S.M.A.R.T.* attributes over time for multi-class health status prediction. In particular, it is composed by two modules: i) Prediction Module; ii) Explanation Module. Finally, the three-dimensional dataset employed is explained solely by LIME and SHAP.

3.1. Prediction module

The prediction module utilizes a LSTM-based model from [23], consisting of two stacked LSTM layers with 128 units, followed by a dense layer with a unit count equal to the number of classes and softmax activation. This model exploits temporal dependencies in *S.M.A.R.T.* features over a time-window to predict HDD health status across four classes (*Alert*, *Warning*, *Very Fair*, and *Good*). The input to each LSTM layer is a three-dimensional data structure with dimensions (z, w, n), where w, z, and n represent the time window size, total number of sequences, and features, respectively. The model predicts HDD health status at time $t + 1$ as a multi-class classification task, assigning each feature sequence to one of the classes (health levels) based on the sequence $(a^{t-w+1}, \cdots, a^{t-1}, a^t)$.

3.2. Explanation module

The explanation module seeks to identify features influencing the model's decision, especially in predicting false positives or misclassifications. Given the multidimensional nature of the problem, two XAI tools (SHAP and LIME) were concurrently applied for the task to compare their explanations. SHAP employs Shapley values, derived from cooperative game theory, to evaluate each feature's contribution to the prediction. Utilizing the DeepExplainer explainer, based on $4,000$ samples and the trained model, SHAP approximates conditional expectations, providing both global and local explanations.

In contrast, LIME explains the model by observing how predictions change with perturbed data. The RecurrentTabularExplainer explainer, an extension of LimeTabularExplainer for $3D$ data, used the entire training set for input, producing local explanations. Unlike SHAP, LIME allows input datasets larger than $5,000$ elements and calculates feature relevance through locally weighted linear models.

While SHAP offers both global and local explanations, LIME focuses on local explanations. SHAP values can be aggregated (mean or median) for a global representation by comparing features across all dataset instances.

4. Experimental Results

In this section, we explore the evaluation conducted using SHAP and LIME on the LSTM-based model to explain predictions regarding HDDs' health status, as described in Section 3.1. We selected this model due to its superior performance across various metrics in this task. The training process involved a maximum of 150 epochs, a batch size of 500, and a learning rate of 0.001, employing Adam [31] as the optimizer. Detailed results for each class, along with overall outcomes based on Macro averaging, are presented in Table 1.

Metric	*Good*	*Very Fair*	*Warning*	*Alert*	*Overall*
Accuracy	99.21%	87.80%	78.10%	84.42%	98.45%
Precision	99.90%	74.40%	71.80%	75.50%	98.33%
Recall	99.10%	89.60%	85.40%	82.00%	98.34%
F1	99.50%	81.30%	78.00%	78.60%	91.48%

Table 1
Results of the model on the Backblaze dataset detailed by each class. The values in the *Overall* column are computed according to the Macro averaging measure.

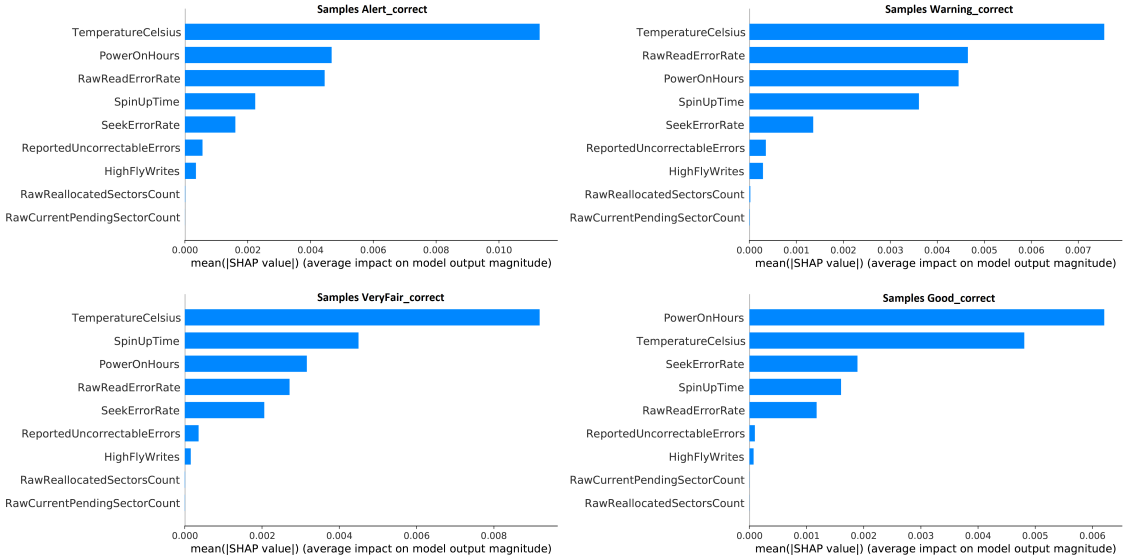


Figure 1: SHAP - Summary Bar Plot

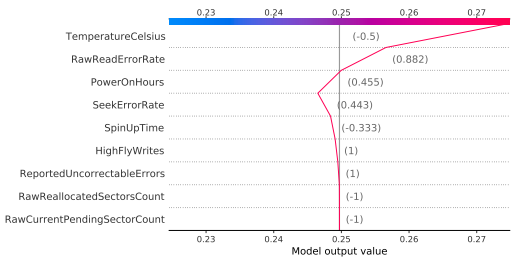
4.1. SHAP

SHAP is the initial framework employed to explore the explanation task regarding HDD health status assessment. It offers diverse analyses, including Summary bar plot, Summary plot, and Dependence plot, providing both global and local explanations.

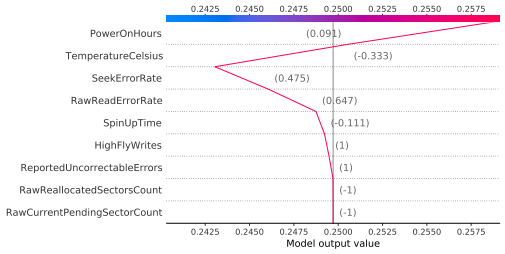
4.1.1. Global explanation

This analysis concentrates on the Summary Bar Plot, offering a global explanation to discern features influencing the model's performance based on their Shapley values. The absolute Shapley values per feature (I_j), representing S.M.A.R.T. attributes for a single HDD within a time window, are summed over n samples and sorted by decreasing importance.

Figure 1 illustrates the importance of SHAP features for the four predicted classes. Notably, *Power On Hours* (*POH*) emerges as the most critical feature, followed by *Temperature Celsius* (TC), *Seek Error Rate* (SER), and *Spin Up Time* (SUT). The analysis highlights how *TC* becomes increasingly significant as HDD status deteriorates, particularly in alert and warning classes. This investigation focuses on correctly classified samples to ensure the accuracy of the analysis.

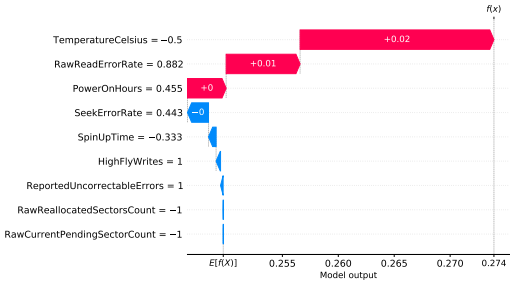


(a) Decision Plot: $Alert_{correct}$, element 8223

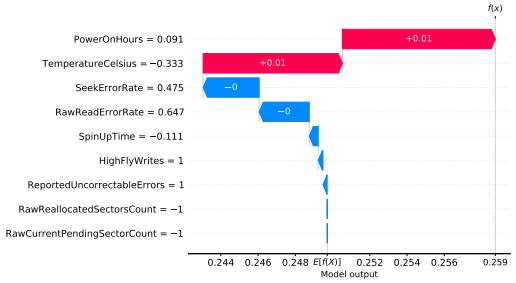


(b) Decision Plot: $Alert_{misclassified}$, element 8371

Figure 2: Decision plot for Alert class by using SHAP



(a) Waterfall Plot: $Alert_{correct}$, element 8223



(b) Waterfall Plot: $Alert_{misclassified}$, element 8371

Figure 3: Waterfall plot for Alert class by using SHAP

4.1.2. Local explanation

For local explanations, SHAP provides different types of plots (Single element Decision Plot, Waterfall Plot), that have been applied to each HDDs' health level status for explaining prediction module's output. The plots related to the class Alert are reported and discussed below.

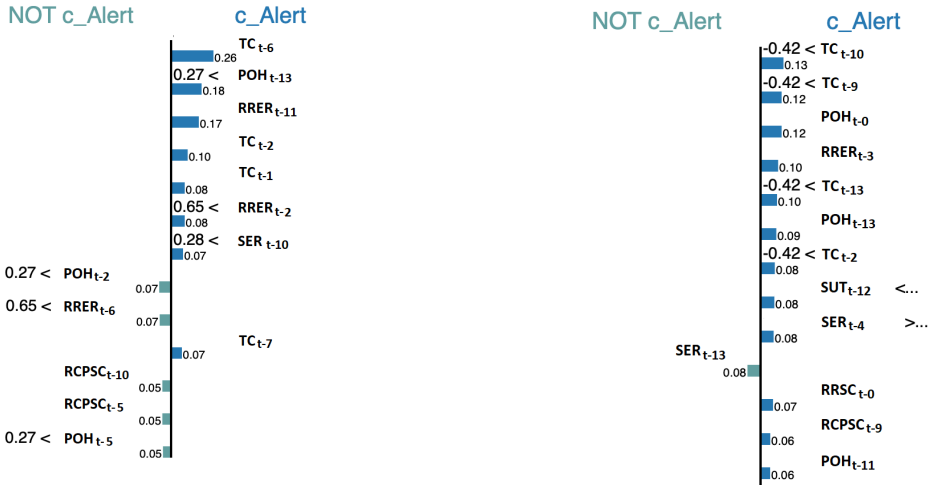
Illustrated in Figure 2a, the central vertical line in the Decision plot signifies the model's base value. From the plot's bottom, the prediction line depicts the aggregation of Shapley values (i.e., feature effects) from the base value to the ultimate model score at the top. Each feature is denoted with its value in brackets, and the slope represents the contribution of that feature to the prediction.

Comparing Figure 2a and 2b, it's evident that sample 8371 is misclassified as belonging to the Alert class due to the model heavily relying on *TC* and *POH* features for this classification.

Moreover, Waterfall plots (refer to Figure 3a and 3b) are tailored for individual prediction explanations, expecting a single row of an Explanation object as input. The bottom of the Waterfall plot starts with the model's expected output, and each subsequent row illustrates how the positive (red) or negative (blue) contribution of each feature shifts the value from the expected output to the model's actual output. In this context, the Waterfall plot not only provides more information but also enhances clarity regarding the contributions of each feature.

4.2. LIME

In this section we investigate the prediction of HDD's health status by using LIME explainability framework. In Figure 4a, we employ LIME to analyze model prediction for sample 8223, displaying the contribution of all features at each time instant. The features positively influencing Alert class prediction include TC_{t-6}, POH_{t-13}, $RRER_{t-11}$, TC_{t-2}, TC_{t-1}, and $RRER_{t-2}$. The explanation for misclassification of sample 8371 in Figure 4b reveals *POH* and *TC* across different time instants as the most confusing features.



(a) LIME Plot 1: $Alert_{correct}$, element 8223

(b) LIME Plot 1: $Alert_{misclassified}$, element 8371

Figure 4: LIME plot for Alert class.

4.3. Quantitative Evaluation

In this section, we conduct an empirical evaluation using the axiomatic explanation consistency framework [32]. The framework consists of two steps: (1) axiomatic and (2) explanation consistency. This involves computing metrics such as *Identity*, *Stability*, and *Separability* on test sets by explaining different objects with their corresponding predictions multiple times.

Table 2 displays the results for each metric on the test sets, representing the percentage of instances satisfying each defined metric. Green highlights the highest performance, while red indicates the lowest. LIME shows poor performance in the *Identity* metric due to the uniform and random sample technique, unlike SHAP, which satisfies the identity metric for all instances. LIME outperforms SHAP in the *Stability* metric with 95.5% compared to 85.5%. Both tools achieve the maximum result (100%) for the *Separability* metric, though this axiom may not be significant due to the non-linear nature of the problem.

Table 3 evaluates the tools' performance in terms of confidence intervals, employing a bootstrap procedure. The analysis includes investigating feature contributions to model predictions and comparing results with a white-box model's ground truth.

LSTM - Backblaze data-set		
	LIME	**SHAP**
Identity	0%	100%
Stability	95.5%	85.5%
Separability	100%	100%

Table 2
Evaluation of interpretability frameworks on Backblaze data-set

Features	**LIME**	**SHAP**
SpinUpTime	0.58±0.009	0.001±0.005
RawReallocatedSectorsCount	0.345±0.007	-0.001±0.008
RawReadErrorRate	0.234±0.006	-0.001±0.005
HighFlyWrites	0.082±0.004	0.001±0.007
RawCurrentPendingSectorCount	0.058±0.004	-0.001±0.023
SeekErrorRate	0.052±0.004	-0.0004±0.0172
PowerOnHours	0.028±0.038	0.001±0.009
ReportedUncorrectableErrors	0.035±0.027	0.001±0.005
TemperatureCelsius	0.04±0.003	0.001±0.004

Table 3
Mean and deviation standard of LIME and SHAP explanations.

5. Conclusions

The widespread use of deep neural networks presents challenges in result interpretation due to their complex structures. Despite this, their high performance in critical applications like predictive maintenance, necessitates eXplainable AI (XAI). LSTM-based models, designed for learning long-term dependencies, are ideal for predictive maintenance tasks. This study focuses on explaining predictions of a multi-class LSTM model assessing HDD health. With the three-dimensional input data, LIME and SHAP were chosen as the primary XAI tools, handling such data effectively. Comparison using invariance, separability, and stability metrics showed LIME and SHAP reaching 0% and 100% for invariance, and both achieving 100% for separability. LIME excelled in stability over SHAP (95% vs. 85.5%). While SHAP provides comprehensive explanations, LIME's *RecurrentTabularExplainer* specializes in recurrent networks, detailing feature contributions across all time instances within a window. Yet, limitations in XAI tools' completeness and correctness measures need addressing. Continuous user engagement is crucial for evaluation, especially in tailoring explanations for different users. Concerns also exist regarding model confidence and potential biases in the learning process.

Future work will explore explanations for predictions from different deep networks in various industrial applications, using diverse real-world datasets. Validating results with field experts remains crucial for enhancing confidence in AI models through XAI.

Acknowledgments

We acknowledge financial support from the PNRR project "Future Artificial Intelligence Research (FAIR)" – CUP E63C22002150007

References

[1] D. Markudova, S. Mishra, L. Cagliero, L. Vassio, M. Mellia, E. Baralis, L. Salvatori, R. Loti, Preventive maintenance for heterogeneous industrial vehicles with incomplete usage data, Computers in Industry 130 (2021) 103468. doi:https://doi.org/10.1016/j.compind.2021.103468.

[2] M.-C. Chiu, J.-H. Huang, S. Gupta, G. Akman, Developing a personalized recommendation system in a smart product service system based on unsupervised learning model, Computers in Industry 128 (2021) 103421. doi:https://doi.org/10.1016/j.compind.2021.103421.

[3] L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, Meet-lm: A method for embeddings evaluation for taxonomic data in the labour market, Computers in Industry 124 (2021) 103341. doi:https://doi.org/10.1016/j.compind.2020.103341.

[4] B. Mao, Z. M. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, K. Mizutani, Routing or computing? the paradigm shift towards intelligent computer network packet transmission based on deep learning, IEEE Transactions on Computers 66 (2017) 1946–1960. doi:10.1109/TC.2017.2709742.

[5] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0, Information Fusion 50 (2019) 92–111. doi:https://doi.org/10.1016/j.inffus.2018.10.005.

[6] R. De Luca, A. Ferraro, A. Galli, M. Gallo, V. Moscato, G. Sperli, A deep attention based approach for predictive maintenance applications in iot scenarios, Journal of Manufacturing Technology Management 34 (2023) 535–556.

[7] V. J. Ramírez-Durán, I. Berges, A. Illarramendi, Towards the implementation of industry 4.0: A methodology-based approach oriented to the customer life cycle, Computers in Industry 126 (2021) 103403. doi:https://doi.org/10.1016/j.compind.2021.103403.

[8] X. Xu, Y. Lu, B. Vogel-Heuser, L. Wang, Industry 4.0 and industry 5.0—inception, conception and perception, Journal of Manufacturing Systems 61 (2021) 530–535. doi:https://doi.org/10.1016/j.jmsy.2021.10.006.

[9] P. K. R. Maddikunta, Q.-V. Pham, P. B, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, M. Liyanage, Industry 5.0: A survey on enabling technologies and potential applications, Journal of Industrial Information Integration 26 (2022) 100257. doi:https://doi.org/10.1016/j.jii.2021.100257.

[10] A. Du, Y. Shen, Q. Zhang, L. Tseng, M. Aloqaily, Cracau: Byzantine machine learning meets industrial edge computing in industry 5.0, IEEE Transactions on Industrial Informatics 18 (2022) 5435–5445. doi:10.1109/TII.2021.3097072.

[11] L. Silvestri, A. Forcina, V. Introna, A. Santolamazza, V. Cesarotti, Maintenance transformation through industry 4.0 technologies: A systematic literature review, Computers in Industry 123 (2020) 103335. doi:https://doi.org/10.1016/j.compind.2020.103335.

[12] C. Colemen, S. Damodaran, M. Chandramoulin, E. Deuel, Making maintenance smarter, Deloitte University Press (2017).

[13] Y. Lavi, The rewards and challenges of predictive maintenance, InfoQ(jul2018) (2018).

[14] L. Liao, F. Köttig, A hybrid framework combining data-driven and model-based methods for system remaining useful life prediction, Applied Soft Computing 44 (2016) 191–199.

doi:https://doi.org/10.1016/j.asoc.2016.03.013.
[15] Z. Gao, C. Cecati, S. X. Ding, A survey of fault diagnosis and fault-tolerant techniques—part i: Fault diagnosis with model-based and signal-based approaches, IEEE transactions on industrial electronics 62 (2015) 3757–3767.
[16] S. Ma, F. Chu, Ensemble deep learning-based fault diagnosis of rotor bearing systems, Computers in Industry 105 (2019) 143–152. doi:https://doi.org/10.1016/j.compind.2018.12.012.
[17] Z. Li, Y. Wang, K. Wang, A deep learning driven method for fault classification and degradation assessment in mechanical equipment, Computers in Industry 104 (2019) 1–10. doi:https://doi.org/10.1016/j.compind.2018.07.002.
[18] Q. Sun, Z. Ge, Deep learning for industrial kpi prediction: When ensemble learning meets semi-supervised data, IEEE Transactions on Industrial Informatics 17 (2020) 260–269.
[19] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. Youngblood, Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation, 2018 IEEE Conference on Computational Intelligence and Games (CIG) (2018) 1–8.
[20] S. Zeb, A. Mahmood, S. A. Khowaja, K. Dev, S. A. Hassan, N. M. F. Qureshi, M. Gidlund, P. Bellavista, Industry 5.0 is coming: A survey on intelligent nextg wireless networks as technological enablers, arXiv preprint arXiv:2205.09084 (2022).
[21] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, L. Chen, A survey of data-driven and knowledge-aware explainable ai, IEEE Transactions on Knowledge and Data Engineering (2020) 1–1. doi:10.1109/TKDE.2020.2983930.
[22] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, G. P. Li, Predictive maintenance in the industry 4.0: A systematic literature review, Computers & Industrial Engineering 150 (2020) 106889. doi:http://doi.org/10.1016/j.cie.2020.106889.
[23] A. De santo, A. Galli, M. Gravina, V. Moscato, G. Sperli, Deep learning for hdd health assessment: an application based on lstm, IEEE Transactions on Computers (2020) 1–1. doi:10.1109/TC.2020.3042053.
[24] A. Ferraro, A. Galli, V. Moscato, G. Sperlì, Evaluating explainable artificial intelligence tools for hard disk drive predictive maintenance, Artificial Intelligence Review 56 (2023) 7279–7314.
[25] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.
[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). doi:10.1145/3236009.
[27] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57. URL: https://doi.org/10.1145/3236386.3241340. doi:10.1145/3236386.3241340.
[28] M. T. Ribeiro, S. Singh, C. 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, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778.

doi:10.1145/2939672.2939778.
[29] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018).
[30] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 4768–4777.
[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[32] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).