Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 232 (2024) 3112–3120

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Industry 4.0 and Smart Manufacturing

A methodology to guide companies in using Explainable AI-driven interfaces in manufacturing contexts

Fabio Grandi[a,*], Debora Zanatto[b], Andrea Capaccioli[b], Linda Napoletano[b], Sara Cavallaro[c], Margherita Peruzzini[a]

[a]*Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Via Pietro Vivarelli, 10, Modena, 41125, Italy*
[b]*Deep Blue S.r.l., Via Daniele Manin 53, Rome, 00185, Italy*
[c]*CNH Industrial S.p.A., Viale delle Nazioni 55, Modena, 41122, Italy*

Abstract

Nowadays, the increasing integration of artificial intelligence (AI) technologies in manufacturing processes is raising the need of users to understand and interpret the decision-making processes of complex AI systems. Traditional black-box AI models often lack transparency, making it challenging for users to comprehend the reasoning behind their outputs. In contrast, Explainable Artificial Intelligence (XAI) techniques provide interpretability by revealing the internal mechanisms of AI models, making them more trustworthy and facilitating human-AI collaboration. In order to promote XAI models' dissemination, this paper proposes a matrix-based methodology to design XAI-driven user interfaces in manufacturing contexts. It helps in mapping the users' needs and identifying the "explainability visualization types" that best fits the end users' requirements for the specific context of use. The proposed methodology was applied in the XMANAI European Project (https://ai4manufacturing.eu), aimed at creating a novel AI platform to support XAI-supported decision making in manufacturing plants. Results showed that the proposed methodology is able to guide companies in the correct implementation of XAI models, realizing the full potential of AI while ensuring human oversight and control.

Keywords: explainable AI (XAI); artificial intelligence (AI); manufacturing; Human-Machine Interaction; user interface.

* Corresponding author. Tel.: +39-059-205-6397.
E-mail address: fabio.grandi@unimore.it

1877-0509
10.1016/j.procs.2024.02.127

1. Introduction

Explainable Artificial Intelligence (XAI) refers to the development and implementation of AI systems that can provide understandable and transparent explanations for their decisions or outputs. Traditional AI models, such as deep neural networks, often operate as "black boxes," making it challenging for humans to comprehend the reasoning behind their predictions or actions. XAI techniques aim to address this limitation by providing interpretable explanations, enabling users to understand the internal workings of AI models and the factors influencing their outputs. The goal of XAI is to bridge the gap between the complex nature of AI algorithms and the need for human comprehensibility, especially in critical domains where decisions impact human lives, such as healthcare, finance, and autonomous vehicles. XAI techniques strive to enhance transparency, trustworthiness, and accountability in AI systems by enabling users to understand how and why a particular decision was reached.

Considering the industrial context, XAI finds valuable applications in the manufacturing industry, providing transparency and interpretability to AI systems. By employing XAI techniques, manufacturers can gain insights into the decision-making processes of AI models. This understanding allows them to identify and rectify issues promptly, optimize operational efficiency, and improve overall system performance. XAI also assists in meeting regulatory compliance requirements by providing explanations for AI-driven decisions, ensuring ethical and accountable practices. Moreover, XAI fosters user acceptance and trust in AI systems, facilitating better collaboration between human operators and AI technologies on the shop floor.

Despite this great surge in innovation, the implementation of XAI in the manufacturing context can be challenging due to several barriers and complexities, such as difficulties to understand which XAI explainability visualization types (EVTs) can support the operators during their work at the shop floor in various contexts. EVTs refer to various techniques and visual representations used to enhance the interpretability and understanding of AI models and their decision-making processes. These visualization types aim to provide insights into how AI models arrive at their predictions or classifications, making the decision-making process more transparent and explainable.

Nomenclature

AI	artificial intelligence
XAI	explainable artificial intelligence
EVTs	explainability visualization types
DRs	demonstrator requirements
QFD	quality functional deployment
DRRI	demonstrator requirement relative importance
CM	correlation matrix
EVTI	explainability visualization type importance
EVTRI	explainability visualization types relative importance

To overcome this issue, this work aims at presenting a methodology to find which EVTs identified from literature best fit the Demonstrator Requirements (DRs) for a specific use case. This method is inspired by the Quality Functional Deployment (QFD) approach, originally introduced to match technical requirements of a generic product with customer requirements. Similarly, the proposed method helps designers and engineers to move from the operators' needs to technical requirements, thought the selection of the proper EVTs to fully exploit the potential of AI in manufacturing context. This method has been defined within the H2020 project called XMANAI (G.A. 957362).

2. Material and methods

2.1. Research background

XAI is a potential bridge between the user and the AI models adopted, making them transparent about their working to human users. In fact, XAI techniques help make AI system decisions transparent and interpretable [1]. This allows

human operators to understand why a particular decision or recommendation was made by the AI system, enabling them to trust and effectively work with the technology, moving from the black box concept to a transparent glass box [2]. Moreover, XAI allows human operators to monitor the performance of manufacturing systems and identify areas for improvement [3]. By visualizing feature importance, decision rules, or model architecture, operators can gain insights into system behaviours and suggest modifications or enhancements for better alignment with human requirements and preferences [4]. These features helps in the achievement of the Operator 5.0 concept, focusing on bolstering the resilience of manufacturing systems by empowering individuals to collaborate with automated solutions, facilitating swift resolution of production system disruptions [5]. In particular, maintenance practices heavily rely on human skills and experience for the successful execution of the majority of tasks so is crucial to support operators in the interpretation of AI algorithms. The capacity to interpret complex data, often numerically based, and transform it into suitable context-aware narratives and graphical representations is fundamental for the upcoming generation of operator-centric maintenance systems [6].

From a technical point of view there are tools and techniques to make completely black box algorithms explainable, there is a gap in making them usable and understandable by the end-user [7]. All the XAI tools such as SHAP (Shapley Additive Explanations, [8]) and LIME (Local Interpretable Model-Agnostic Explanations, [9]) have specific data representation types for reporting the results of the XAI explanation, but the complexity of the results produced can be too high without specific technical knowledge as demonstrated for LIME [10]. Also, different EVTs could be combined by the users to reach a meaningful explanation for their goals [11], considering at first users' goals and needs and making the implementation of XAI a human-centred concept [12]. By prioritizing human-centeredness in explainability AI, the technology becomes a supportive tool that enhances human decision-making, empowers operators, and aligns AI systems with human values and ethical considerations. This approach recognizes the importance of human agency, responsibility, and well-being in the development and adoption of AI technologies [13].

Matching the technical XAI solutions with the users' requirements is the starting point for this research work to identify the most suitable EVTs addressing industrial manufacturing use case. To achieve this goal, it was considered the extensive review made by Vilone and Longo [14] that mapped the different EVTs categories. From this research work, seven categories of EVTs have been identified: image, heatmap, flowchart, bar plot or line plot, text, scatterplot and mixed. Thereafter, other two categories (i.e., time series and confusion matrix) were added to complete the overview of EVTs, based on a most recent literature analysis of XAI specific application in manufacturing [15,16]. Table 1 describes the various EVTs identified in literature.

Table 1. Explainability visualization types (EVTs) description.

EVTs	Description
Text	involves representing textual data, highlighting important keywords to aid in understanding the model's decision-making process
Bar plot / line plot	used to present quantitative data, displaying categorical data using rectangular bars, where the length of each bar represents the magnitude of a specific attribute
Heatmap	provides a graphical representation of data where different colors or shades indicate the intensity or values of certain features or attributes
Scatterplot	used to visualize the relationship between two or more variables, identifying patterns or correlations between variables. Each data point is represented as a dot on the plot, with the position of the dot determined by the values of the variables
Image	involves displaying visual representations of data, such as images or visual patterns, allowing users to understand how the model processes and classifies images
Flowchart	visual diagrams that represent a series of steps or processes, showcasing the sequence of rules or conditions followed by the model to arrive at a particular outcome
Mixed	combine different types of visual representations to provide a comprehensive understanding of AI model behavior, involving the integration of multiple EVTs
Confusion matrix	is a square matrix with cells representing the counts or percentages of correct and incorrect predictions for each class

Time series	specifically designed to analyze and interpret data that changes over time, displaying patterns, trends, and anomalies in sequential data.

Each EVTs type has been mapped according to different properties: scope (global or local), stage (ante-hoc or post-hoc), applicability (model specific or model agnostic), approach (neural, data-driven or rule-based), type (classification, regression), input data and output data. Overall, the mapping presented a first view on how different kinds of EVTs can address different problems in manufacturing industrial context. These results were used to reflect on the identified explainability requirements that were collected from the manufacturing companies. It was then clear the need for a more specific methodology to support the design of explainable interfaces combining the collected explainability requirements with the EVTs.

2.2. Methodology

The presented methodology proposes a matrix-based approach to combine the EVTs as identified in literature with the Demonstrators Requirements (DRs) in order to support operators in industrial contexts using XAI and drive companies in selecting the best supporting tools. This method follows a systematic approach and starts from the definition of a set of operators' requirements and needs and maps the correlations with the EVTs using a correlation matrix, similarly to the quality function deployment (QFD) house of quality approach [17,18]. In the same way, the proposed method helps to identify the best EVTs for each industrial scenario starting from the analysis of the operators' needs on the field, moving from a user-centred view to a more technical and functional view.

The proposed methodology consists of five main phases:

- Definition of various scenarios: a set of different scenarios are defined to contextualize the study;
- Definition of DRs: according to the main tasks carried out by the operators in different scenarios, a set of DRs are defined and listed using interviews to various operators;
- Definition weights for each DR: evaluated through the use of focus group (in which are involved a sample of operators) where for each DR is discussed the importance, called Demonstrator Requirement Importance (DRI), and weighted according to a 5-point scale (1 = no importance, 5 = high importance), answering the question: "From 1 to 5, how much important is this requirement for this use case?". The Demonstrator Requirement Relative Importance (DRRI) score (Eq. 1) is then calculated using the formula (with *m* as the number of the DRs, *i* as the number of the generic DR in row from 1 to *m*):

$$DRRI\ i = \frac{DRI\ i}{\sum_1^m DRI} * 100 \tag{1}$$

- Evaluation of the strength of the relationship between the defined DRs and the EVTs categories identified in literature according to a 1-3-9 scale (1 = weak correlation, 3 = medium correlation, 9 = strong correlation). The correlation scores are discussed during the focus group with the operators. As a result of this step, a *m* x *n* matrix called Correlation Matrix (CM), is created and composed as following (Eq. 2):

$$CM = [CM_{11}, \dots, CM_{mn}] \tag{2}$$

with *n* as the number of the EVTs. These values serve as input of the calculation of the Explainability Visualization Type Importance (EVTI), calculated for each EVTs as (Eq. 3):

$$EVTI\ j = DRRI_1 * CM_{1j} + \cdots + DRRI_i * CM_{ij} + \cdots + DRRI_m * CM_{mn} \tag{3}$$

with *j* as the number of the generic EVT from 1 to *n*.

- Results analysis: the result of this methodology is reported as the Explainability Visualization Types Relative Importance (EVTRI), expressed in percentage as follow (Eq. 4):

$$EVTRI\ j = \frac{EVTI\ j}{\sum_1^n EVTI} * 100 \quad (4)$$

The EVTs that have the highest scores are the most suitable for the demonstrator in the case study considered.

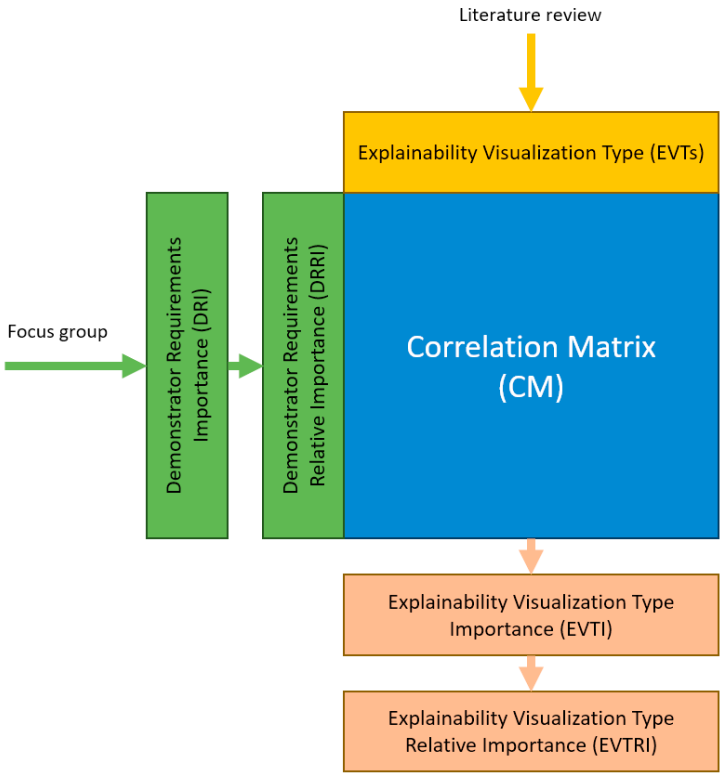

Fig. 1. The proposed methodology.

3. Results and Discussion

3.1. Use case description

The use case has been developed in collaboration with CNH Industrial, a global leader in design and manufacturing agricultural machines, buses, and trucks. In particular, the collaboration was developed within the Modena plant, located in Italy, in which the main tractors' components are built and where medium-sized tractor transmissions are produced. The case study stems from a real problem frequently occurring in the Modena production plant. Indeed, the production plant faces frequent unexpected failures of the production machine, causing prolonged stoppages in the production line. Downtimes are typically due to defective parts or maintenance. Operators need to quickly restore the machine operation to minimize production losses. However, machine interfaces are complex and not fully manageable from plant operators, so that usually expert technicians are required, prolonging the downtime. XAI can serve as a bridge between machines and operators, enabling faster understanding of machine status and improving productivity through prompt interventions. XAI models can provide simplified suggestions to plant operators, optimizing the production line as well as reducing time and cost. To implement XAI within the CNHi plant, the developed platform will take data from the current systems, carrying information about the status of a specific CNC machine selected for

the use case (i.e., Heller MCS-H400). More specifically, in the use case XAI is crucial to provide the worker with the necessary explanations and clarity to understand which part of the machine may have caused the failure.

3.2. Methodology application to the considered case study

From the analysis of the real industrial context, a set of different scenarios were defined with a focus group, involving four operators with at least 10-years expertise in this field. For this specific use case, operators highlighted three main scenarios: alarms and event handling, production restoring and production monitoring. For each scenario a set of DRs were discussed, as reported in Table 2. For each DRs, the Demonstrator Requirement importance from 1 to 5 were defined.

Table 2. Demonstrator Requirements (DRs) and Demonstrator Requirements Importance (DRIs) for the use case.

Scenarios		Demonstrator Requirements (DRs)	Demonstrator Requirements Importance (DRIs, from 1 to 5)
1.	Alarms and event handling	To show the status of the machine	5
		To display the features with related trends	3
		To rank the most critical features	4
2.	Production restoring	To suggest and rank possible troubleshooting procedures	4
		To display a detailed step-by-step guide in AR	4
3.	Production monitoring	To optimize the production based on operator feedback	5

Then, according to the proposed methodology, the correlation matrix was filled discussing each correlation between DRs and EVTs in the focus group. In order to calculate the scores in a quicker way, an Excel worksheet was developed for this scope.

3.3. Results of the proposed approach

Figure 2 shows the results of the proposed approach for the considered case study. The higher score of EVTI is reached by text visualization type (20,4%), followed by other four EVTs such as bar/line plot, heatmap and time series with the same score (15,4%). In the next section these results are critically discussed to evaluate how the EVTs suggestions are in line with the case study expectations.

Scenarios	Row #	DRRI	DRI (from 1 to 5)	Explainability visualization types (EVTs) / Demonstrators requirements (DRs)	text	bar/line plot	heatmap	scatterplot	images	decision tree (flowchart)	mixed	confusion matrix	time series
	Column #				1	2	3	4	5	6	7	8	9
1. Handle Alarms and events	1	20.0	5	show the status of the machine	O	O	⊙	▲	▲	▲	▲	⊙	O
	2	12.0	3	interface displays the features with related trends	▲	O	▲	▲	▲	▲	▲	⊙	O
	3	16.0	4	rank the most critical features	▲	⊙	⊙	▲	▲	▲	▲	O	⊙
2. Restore production	4	16.0	4	suggest and rank possible troubleshooting procedure	⊙	⊙	O	▲	▲	▲	▲	▲	⊙
	5	16.0	4	displays a detailed step-by-step guide in AR	⊙	▲	▲	▲	⊙	▲	▲	▲	▲
3. Monitor Production	6	20.0	5	optimize production based on operator feedback	⊙	▲	▲	▲	▲	▲	▲	▲	▲
	100.0			Max Relationship Value in Column	9	9	9	1	9	1	1	9	9
				EVTI	556.0	420.0	420.0	100.0	228.0	100.0	100.0	388.0	420.0
				EVTRI (%)	20.4	15.4	15.4	3.7	8.3	3.7	3.7	14.2	15.4

Legend		
⊙	Strong Relationship	9
O	Moderate Relationship	3
▲	Weak Relationship	1

Fig. 2. Combination between DRs and EVTs according to the proposed methodology.

3.4. Preliminary validation and discussion

Considering the results obtained from the previous steps, the text plot collected the higher score. Textual instructions allowed the explanation of the output of the AI algorithm in the simplest and clearest way, allowing the operator to understand the model's decision-making process. In the considered use case, textual representation is the best solution in order to rank the most critical features and to suggest the possible troubleshooting procedures, meeting two of the more significant DRs described above.

Bar/line plots were used in XAI to present quantitative data: in particular, bar plots showed the individual contribution of each sensor in the specific selected anomaly (Figure 3a). The values were arranged in descending order based on their impact, with empty values placed at the bottom. Red bars represented the algorithm's confidence in a sensor's contribution to a specific anomaly, while blue bars indicate non-contribution. The width of the bar indicates the level of algorithm confidence. If data is missing, it meant either the model is uncertain about the sensor's contribution, or no data was available for that sensor regarding the considered anomaly. This enabled maintainers to initiate suitable maintenance procedures, prioritizing the most likely components first and excluding those for which the algorithm strongly believes they are not involved.

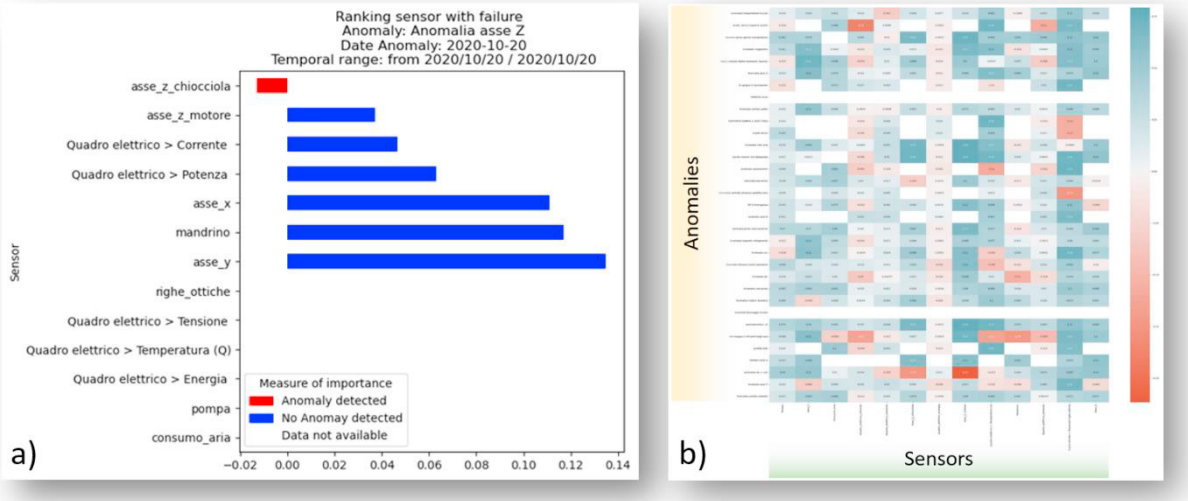

Fig. 3. Examples of EVTs for the use case: (a) bar plot; (b) heatmaps

Unlike the previous plot that focused on a single anomaly, heatmaps can correlate all possible anomalies with various sensors to provide a comprehensive overview (Figure 3b). In this direction, each anomaly was compared to every sensor's value, and a relationship score was assigned to each pair. A lower score (indicated by a red colour) suggested a higher correlation, while a higher score (indicated by a blue colour) indicated a lower correlation. This visualization was intended to be used in later stages to identify patterns in the correlations between sensors and the occurrences of anomalies. It enabled experienced maintainers to plan periodic checks and proactively order replacement parts based on the identified patterns. Bar plots and heatmaps are particularly effective to show the operator the status of the machine: the first graph is more suitable to analyse the single anomaly while the second one allows a holist evaluation of the entire system.

To conclude, the time series was a plot that allows operators to understand the trend over time of the various sensors by highlighting when the value deviates from the standard value and when it approaches the limit threshold, understanding which sensors is the most critical one. In Figure 4 is shown the implementation in the considered use case so it addresses the need to visualize the various features with related trends.

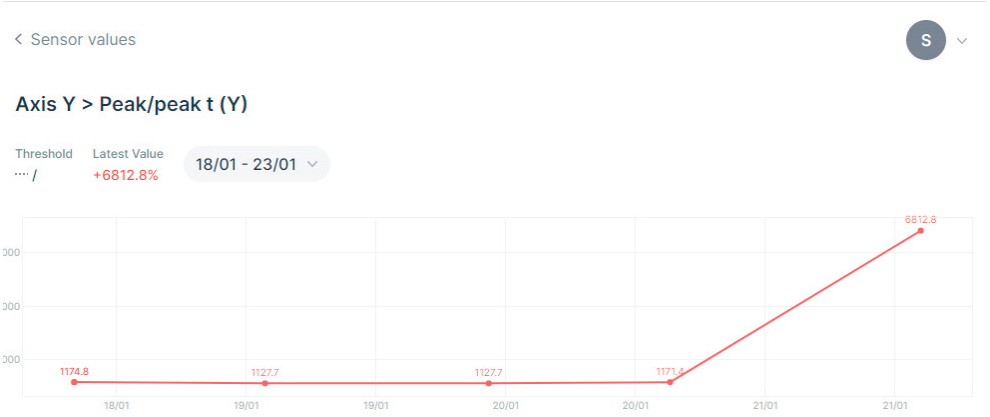

Fig. 4. Example of EVTs for the use case: time series.

4. Conclusions

The paper investigated the application of XAI in manufacturing context, trying to support the operators at the shop floor introducing a set of EVTs to increase trustworthiness in AI algorithm, moving from a black box concept to a glass one. Incorporating XAI into systems is vital for enhancing user interaction and comprehension of algorithm-derived data. Equally important is the selection of appropriate EVTs, as they profoundly impact how users perceive and interpret the AI-generated insights. Thus, it becomes crucial to develop tools and frameworks that actively encourage the utilization of suitable XAI visualizations, enabling users to harness the full potential of the information provided. To achieve this goal, this research work proposed a methodology to guide companies in the selection of the best EVTs for different manufacturing context scenarios in which operators can be directly involved, using an approach similar to the QFD house of quality. The methodology was applied to a real industrial case, within the H2020 project called XMANAI (G.A. 957362), concerning the use of CNC machines in manufacturing plant, involving four operators in the selection of the EVTs using a matrix-based approach. Using a focus group were defined some scenarios and demonstrator requirements (DRs) and these are correlated with the different categories of EVTs identified in literature. The results of this methodology drove the company involved in the project in the selection of the EVTs that best fits the operators' requirements, promoting a user-centred approach in the use of XAI in manufacturing. Future works will focus on the application of the proposed methodology to other case study in different manufacturing sectors.

Acknowledgments

The authors wish to acknowledge CNH Industrial Italia for the precious collaboration.

Funding

This research was funded by the European Community under the HORIZON 2020 programme, grant agreement No. 957362 (XMANAI).

References

[1] Sanneman L, Shah JA. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. Int J Hum Comput Interact 2022;38:1772–88.

[2] Rai A. Explainable AI: From black box to glass box. J Acad Mark Sci 2020;48:137–41.

[3] Chen T-CT. Explainable Artificial Intelligence (XAI) in Manufacturing. Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications, Springer; 2023, p. 1–11.

[4] Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A. Explainable AI in industry: practical challenges and lessons learned: implications tutorial. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, p. 699.

[5] Romero D, Stahre J. Towards the resilient operator 5.0: The future of work in smart resilient manufacturing systems. Procedia CIRP 2021;104:1089–94.

[6] Turner C, Okorie O, Oyekan J. XAI Sustainable Human in the Loop Maintenance. IFAC-PapersOnLine 2022;55:67–72.

[7] Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, p. 1–18.

[8] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.

[9] Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, p. 1135–44.

[10] Dieber J, Kirrane S. Why model why? Assessing the strengths and limitations of LIME. ArXiv Preprint ArXiv:201200093 2020.

[11] Jin W, Fan J, Gromala D, Pasquier P, Hamarneh G. EUCA: The end-user-centered explainable AI framework. ArXiv Preprint ArXiv:210202437 2021.

[12] Vaughan JW, Wallach H. A human-centered agenda for intelligible machine learning. Machines We Trust: Getting Along with Artificial Intelligence 2020.

[13] Liao QV, Varshney KR. Human-centered explainable ai (xai): From algorithms to user experiences. ArXiv Preprint ArXiv:211010790 2021.

[14] Vilone G, Longo L. Explainable artificial intelligence: a systematic review. ArXiv Preprint ArXiv:200600093 2020.

[15] Heydarian M, Doyle TE, Samavi R. MLCM: Multi-label confusion matrix. IEEE Access 2022;10:19083–95.

[16] Veerappa M, Anneken M, Burkart N, Huber MF. Validation of XAI explanations for multivariate time series classification in the maritime domain. J Comput Sci 2022;58:101539.

[17] ISO 16355-1:2015 2015. https://www.iso.org/standard/62626.html (accessed May 3, 2021).

[18] Hauser JR, Clausing D. The house of quality 1988.