Journal of Industrial Information Integration 47 (2025) 100905

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Review article

Explainable AI for industrial fault diagnosis: A systematic review

J. Cação [a,b,*], J. Santos [a,b], M. Antunes [c,d]

[a] TEMA - Centre for Mechanical Technology and Automation, Department of Mechanical Engineering, University of Aveiro, 3810-193, Aveiro, Portugal
[b] LASI - Intelligent Systems Associate Laboratory, Guimarães, Portugal
[c] DETI - Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193, Aveiro, Portugal
[d] IT - Institute of Telecommunications, 3810-193, Aveiro, Portugal

ARTICLE INFO

Keywords:
Explainable AI
Fault detection and diagnosis
Industry 5.0
Interpretability
Transparency
Taxonomy

ABSTRACT

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into industrial environments, particularly for optimising fault detection and diagnosis, has accelerated with Industry 4.0 and 5.0. However, the "black-box" nature of these methods hinders practical implementation, as trust, interpretability, and explainability are crucial for informed decision-making. Furthermore, impending regulatory frameworks like the EU AI Act make directly implementing opaque AI for critical industrial tasks infeasible. Explainable AI (XAI) offers a promising solution by enhancing ML model interpretability and auditability through human-understandable explanations. This review comprehensively analyses recent XAI advancements for industrial fault detection and diagnosis, presenting a novel taxonomy for XAI methods and discussing how XAI outputs are generated, conveyed to end-users, and evaluated. It then systematically reviews real-world industrial XAI implementations, highlighting their applications, methods, and output presentation approaches. Key identified trends include the dominance of post-hoc feature attribution methods, widespread use of SHAP and GradCAM, and a strong reliance on graphical explanation tools. Finally, it identifies current challenges and outlines future research directions to promote the development of interpretable, trustworthy, and auditable AI systems in industrial settings.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have become irreplaceable tools within society, enabling automation across a wide range of tasks, from personalised recommendations and financial services to autonomous vehicles [1–3]. Their efficiency to solve complex problems has also accelerated adoption across several domains, including manufacturing, time series forecasting, and medicine [4–6].

Within the industrial context, AI and ML have driven significant transformation under the paradigm of Industry 4.0. Alongside AI and ML, technologies like the Internet of Things (IoT), Digital Twins, and cloud computing have introduced a widespread digitalisation of industrial production systems [7–9]. The resulting concept of Industrial IoT (IIoT) has motivated the development and deployment of smart sensors and advanced architectures for real-time monitoring, control, and management of large volumes of oncoming data [10–12]. Beyond process monitoring, data processing is also being increasingly integrated in these architectures, revolutionising the way gathered information is used [13,14]. Traditionally, industrial process control relied on knowledge- and rule-based "if-else" systems [15,16]. Although reliable and highly interpretable, such systems are limited: they lack the flexibility to adapt to dynamic and evolving operational conditions, are difficult to scale as complexity grows, and overly rely on domain expertise [17,18]. In response, industry is focused on developing AI and ML-based approaches for data processing. Supporting this trend, a 2023 survey by the Boston Consulting Group (BCG) [19], involving 1800 manufacturing executives from 15 nations, revealed that 89% of respondents were interested in AI implementation, with 68% already exploring AI-driven solutions.

Within this evolving landscape, industrial fault and anomaly detection, often paired with fault diagnosis, arises as a key processing task [20,21]. While fault detection aims to identify deviations from normal behaviour, fault diagnosis contributes to root cause analysis by pinpointing the underlying cause [22]. In manufacturing environments, these tasks are broadly applied across four key areas/domains:

- **Equipment:** monitoring machine data to enable predictive maintenance, detect early tool wear, or identify component defects.

* Corresponding author at: TEMA - Centre for Mechanical Technology and Automation, Department of Mechanical Engineering, University of Aveiro, 3810-193, Aveiro, Portugal.
E-mail addresses: josemaria@ua.pt (J. Cação), jps@ua.pt (J. Santos), mario.antunes@ua.pt (M. Antunes).

https://doi.org/10.1016/j.jii.2025.100905
Received 24 March 2025; Received in revised form 20 June 2025; Accepted 3 July 2025
Available online 17 July 2025
2452-414X/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

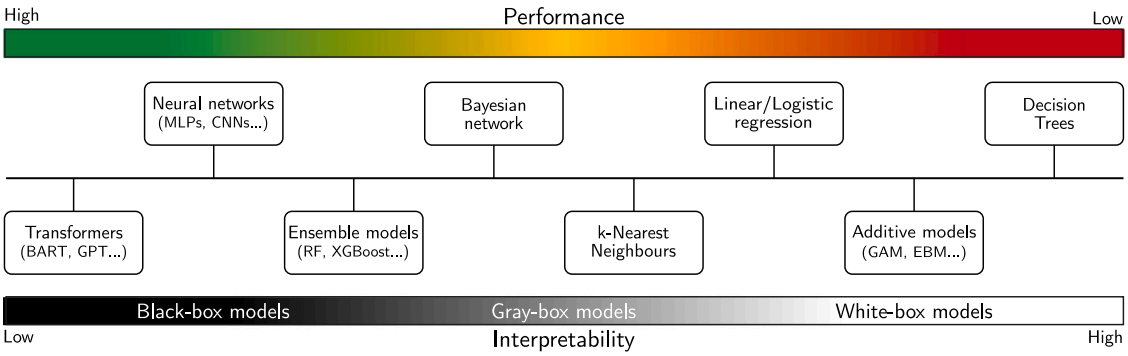


Fig. 1. Trade-off between model interpretability and performance.
Source: Adapted from [42].

- **Process:** leveraging sensor data to flag abnormal stages in production, often linked to malfunctioning equipment or faulty outputs.
- **Product:** applied in post-production stages, often through visual and non-destructive methods, to detect defects in parts.
- **Cybersecurity-wise:** detecting intrusions or anomalies in IIoT systems via network logs, traffic data, and smart device activity.

As industrial automation evolves and data collection intensifies, leveraging ML for real-time industrial fault detection has become essential [18]. Compared to traditional rule-based systems, it offers superior adaptability, real-time responsiveness, and improved performance [16,21]. Consequently, ML methods have been actively explored across equipment [23–25], process [26–28], product [29,30], and cybersecurity fault detection tasks [31–33]. Particularly, Deep Learning (DL) has gained prominence due to its capacity to handle large, complex datasets and model highly non-linear patterns [34–36]. Within industrial applications, convolutional neural networks (CNNs) are commonly employed for image data processing [30,37] and recurrent neural networks (RNNs) for time-series analysis [38,39]. However, the increased performance of DL entails drawbacks regarding model interpretability. DL models often operate as "black-boxes", with very complex internal structures, making their decisions difficult to interpret. This problem raises concerns in industrial settings where explainability is essential for trust, compliance, and effective decision-making [34,35,40,41]. Fig. 1 illustrates the trade-off between model performance and interpretability.

Industries are embracing Industry 4.0's digital transformation, yet Industry 5.0 is already emerging [41,43]. While Industry 4.0 prioritised automation and efficiency, often reducing human involvement, Industry 5.0 aims to restore the human element through sustainability, resilience, and human-centric design [41,43,44]. Here, technology supports human workers, ensuring adaptability to diverse needs and expectations [41]. For safety-critical applications like industrial fault detection, "black-box" ML algorithms are highly undesirable [1,40]. Their lack of transparency hinders AI adoption and conflicts with Industry 5.0 principles [41,43]. The need for transparency, interpretability, and human-centricity in industrial AI is also emphasised by regulatory initiatives such as the **EU AI Act** [45]. This act regulates AI systems based on risk, ensuring high-risk operations, including industrial fault detection, meet transparency, robustness, and safety standards. Relevant articles include Article 10 (data governance), Article 12 (logging), Article 13 (transparency for output interpretation), and Article 14 (human oversight).

Within industrial fault detection, these regulations mandate **auditability**: ML systems must perform well and provide interpretable insights into fault detection and diagnosis [46,47]. Auditability also ensures AI systems are lawful, ethical, robust, and unbiased, a challenge in complex industrial applications [47]. An auditable AI system facilitates collaborative decision-making by enabling stakeholders to understand, justify, and evaluate AI-driven outcomes.

Conventional "black-box" ML and DL lack the necessary interpretability, transparency, and auditability. This gap has spurred interest in **Explainable AI (XAI)**, a research area developing methods to generate human-understandable explanations for AI model decisions [42,48–50]. While XAI has been primarily academic [51], its potential for auditable AI is gaining industry recognition. Siemens, for instance, highlights XAI's benefits across their AI framework, from data preparation to evaluation [51]. Deloitte emphasises XAI's importance for output interpretation and trust with their Lucid tool [52]. Ericsson advocates combining XAI techniques for improved model outputs and interpretation [49].

Fig. 2 contrasts traditional "black-box" AI with XAI-enabled auditable systems. In a "black-box" system, the AI/ML model and process operate interchangeably. Data trains the model, which then deploys to continuously generate outputs. However, there is no visibility into *how* or *why* decisions are made, forcing end-users to "trust" the outputs' correctness. Anomalies require users to interpret and justify the AI's behaviour independently. In contrast, an XAI-based auditable system enhances interaction, interpretability, and control. During development, XAI helps identify data biases. In deployment, XAI provides tailored, understandable explanations (visual, textual, statistical) to different stakeholders. This approach empowers all users, from operators to managers, with customised information, increasing their knowledge and control over the entire AI system.

Overall, the main objective of this work is to provide a comprehensive review of recent advances in XAI specifically applied for industrial fault detection and diagnosis tasks. This paper addresses both the theoretical foundations of XAI and its practical implications within industrial settings. Key contributions include:

1. A systematic literature review of XAI applications in industrial fault detection and diagnosis, focusing on application areas, problem typologies, methodologies, and XAI evaluation approaches and explainability tools.
2. A novel taxonomy for classifying XAI methodologies tailored to industrial use cases, based on multiple literature proposals. It provides various classifications for XAI according to different criteria relevant to an industrial implementation.
3. An in-depth discussion of the current challenges in implementing XAI for industrial fault detection and diagnosis tasks.
4. Proposed guidelines for future research on XAI-based methodologies to support interpretable, trustworthy AI systems in industrial environments.

The remainder of the paper is organised as follows: Section 2 provides a background on XAI, including its theoretical foundations, the proposed taxonomy, and an overview of XAI evaluation strategies. Section 3 describes the research methodology employed for the systematic literature review. Section 4 presents a detailed analysis of the

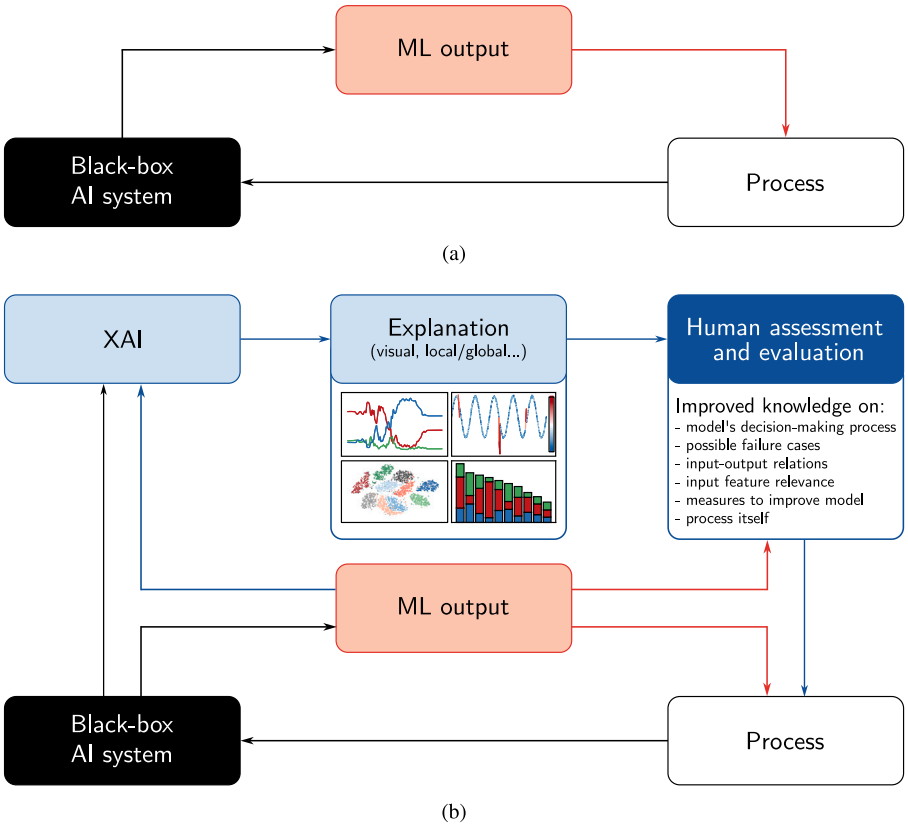


Fig. 2. Comparison between a (a) "black-box" AI system and (b) an auditable AI system.

reviewed studies. Section 5 discusses the key challenges and future research opportunities. And finally, Section 6 summarises the contents of the paper.

2. Theoretical background

This section presents the theoretical foundation for XAI. It offers a brief overview of key concepts discussed in the literature regarding XAI, and it proposes a novel taxonomy for XAI methodologies, emphasising their practical implementation in industrial applications. Finally, it provides an overview of current XAI evaluation methodologies used to assess the quality of their outputs.

2.1. Base concepts

The term **Explainable AI (XAI)**, introduced in 2004 by van Lent et al. [53] for simulation gaming, initially described a system's ability to explain AI unit actions [42]. Now, XAI broadly defines methodologies and tools that provide human-understandable explanations for AI decisions [42,50]. Due to its ties with human interpretability, subjectivity, and trust, XAI often involves philosophical and ethical considerations [54,55]. Key XAI concepts include: **explainability** (why models decide), **interpretability** (how models explain), and **auditability** (overall AI system evaluation, including performance, explanation quality, user impact, and ethics) [42,54,55]. Other crucial concepts, covering both human-centred and technical aspects, are **transparency, fairness, trust, reliability**, and **robustness** [1,2,42].

Regarding practical implications, notable contributions include those by Miller [54] and Dazeley et al. [56], who propose a hierarchical classification for explanations, ranging from zero-order to meta explanations. A graphical illustration of such levels is presented in Fig. 3. Generally, the higher the order, the closer the explanation resembles human reasoning, with meta explanations fully emulating human-like behaviour:

Zero-order/Reactive explanations: focused on a *"single agent's reaction to immediately perceived outputs"* [56]. These serve as the foundation to every other explanation, providing local, instance-specific reasoning through direct input–output analysis.

First-order/Disposition explanations: these reveal an *"agent's underlying disposition towards the environment and other actors that motivated a particular decision"* [56]. These go beyond explaining *why* a decision was made, and address also *how* it was made, reflecting the agent's intentions towards a decision.

Second-order/Social explanations: these imply that an agent recognises the existence of other actors with their own outputs and behaviours. Thus, when providing an explanation, the agent must also perceive and predict the behaviour from other actors.

Nth-order/Cultural explanations: these approximate human-like explanations by considering not only other actors' actions but also their expectations regarding the agent's behaviour.

Meta/Reflective explanations: explanations that fully mimic human explanatory behaviour by incorporating all previous levels of explanations, and providing a step-by-step reasoning of the agent's problem-solving process.

Current industrial XAI implementations still mostly operate at the level of zero-order explanations, *i.e.*, through surface-level analyses mostly based on feature ranking attribution. In only few instances do they move towards first-order explanations, namely through inherent explainable approaches. This reveals a clear gap between academic XAI goals and industrial reality, reinforcing the need for more human-centred, context-aware systems. Beyond this hierarchy, literature primarily categorises XAI methods through taxonomies based on various criteria, commonly including:

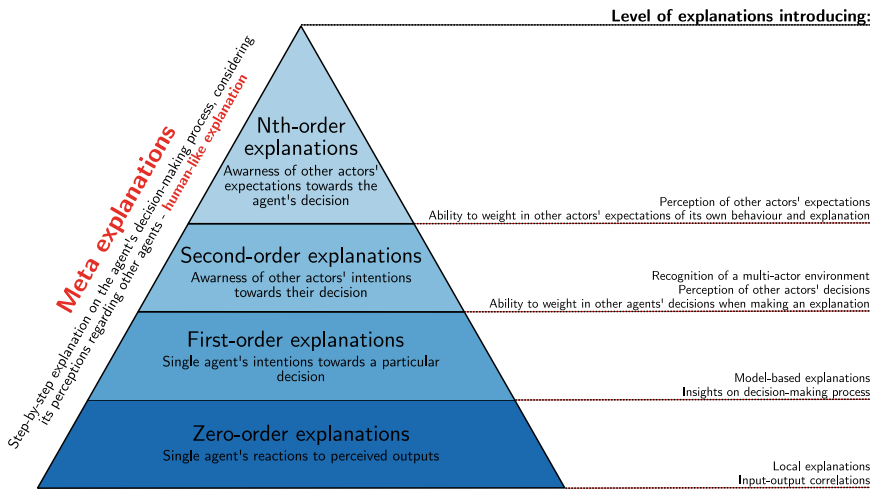


Fig. 3. Hierarchical division of different levels of explanation for an XAI-based system.
Source: Adapted from [56].

- **Local vs. Global:** this distinction, often referred to as the scope of explanation, splits approaches between those which are focused on providing explanations for individual instances of data (local) or for the overall model and/or dataset (global) [2,35,40,42,48,57].
- **Model-specific vs. Model-agnostic:** this separation reflects the versatility of XAI methods. Model-specific approaches are tailored to a particular model or family of models, whereas model-agnostic methods can be applied across different models. In practice, model-agnostic methodologies are generally preferred due to their greater flexibility and adaptability [1,34,35,40,42,48,57]
- **Model-based vs. Post-hoc:** model-based explanations, also known as intrinsic explanations, are integrated within the model's structure, whether through its basic functioning, architecture, or outputs. Such explanations are associated with "transparent models", and are inherently model-specific. Post-hoc explanations, by contrast, are developed after the implementation of a ML classifier or regressor. As a result, post-hoc methods are generally model-agnostic and applicable to a broader range of models [1,2,35,40,42].

Various authors offer specific XAI classifications. Barredo Arrieta et al. [1] identify three approaches for intrinsic explanations in transparent models: *simulatability, decomposability,* and *algorithmic transparency*. They also describe post-hoc, model-agnostic methods, including local, visual, feature relevance, and simplification explanations. Linardatos et al. [57] add *purpose of interpretability* as a criterion, classifying methods beyond intrinsic ("white-box") and post-hoc ("black-box") to include those enhancing fairness or testing prediction sensitivity. They also consider *type of explanation*, such as numerical or graphical outputs. Uniquely, Ali et al. [42] include *input data explainability*, encompassing data pre-processing methods like exploratory data analysis, knowledge graph generation, and dimensionality reduction. For DL explainability, Samek et al. [58] propose a four-way categorisation: interpretable local surrogates, occlusion analysis, gradient-based techniques, and Layerwise Relevance Propagation (LRP). Other sub-classifications in the literature include feature attribution, example-based, and rule-based methods [1,2,34,42].

2.2. XAI taxonomy

As previously discussed, a universal taxonomy or standard classification applicable to all **Explainable AI (XAI)** methods does not currently exist. The literature showcases a wide array of taxonomies, which frequently diverge in their categorisation of explanations and XAI methods. Furthermore, few of these existing classifications focus on the practical implementations and real-world implications of applying XAI. In response to this fragmentation, the authors have undertaken a detailed analysis of current taxonomies and subsequently propose a more unified and comprehensive classification for XAI approaches. Distinct from existing taxonomies, this proposed classification specifically considers the practical application of XAI methods, with a particular emphasis on their critical role within industrial applications [59]. This novel taxonomy is visually presented in Fig. 4.

This taxonomy begins with the ML model itself. Unlike other proposals, such as that proposed by Ali et al. [42], this taxonomy focuses exclusively on XAI models and explanations, excluding pre-processing methods. The first distinction concerns *model transparency*, dividing ML models into two categories, transparent and non-transparent models:

- **Transparent models:** these are models which inherently offer some sort of interpretability. According to Barredo Arrieta et al. [1], three levels of model transparency can be distinguished: *simulatability* as the ability of a human to simulate or conceptualise the model; *decomposability* describing the ability to partition a model into its individual components for analysis; and *algorithmic transparency*, where the entire decision-making process from input to output is comprehensible. Examples of transparent models include linear and logistic regressions, Decision Trees (DT), K-Nearest Neighbours (KNN), or rule-based learners [1]. Following the previous discussion on model-intrinsic explainability, transparent models do not require surrogate models to generate explanations.
- **Non-transparent models:** often described as "black-box" models. Due to their increased complexity, they are difficult to interpret and understand without additional tools. Thus, they require surrogate models to provide post-hoc explanations to enhance their interpretability.

Post-hoc explainability, as mentioned above, implies the use of a surrogate model to improve the comprehension of a ML model's decision and outputs. These techniques can be classified according to multiple criteria, and may fall into different XAI categories. As outlined in Fig. 4, four distinct criteria are proposed for distinguishing XAI methods. The following sections briefly introduce and describe each category.

Generalisability

This first criterion, widely recognised throughout the literature, categorises methods based on their applicability to different ML models.

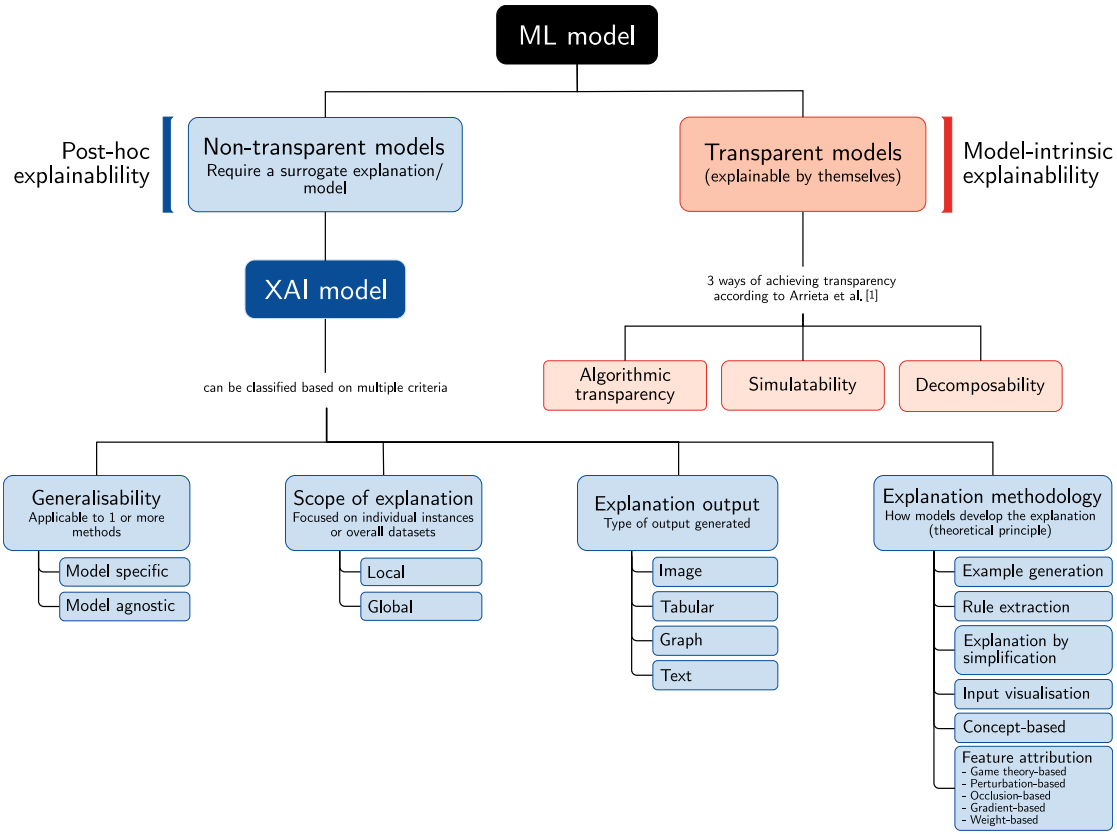


Fig. 4. Proposed taxonomy for XAI and explainability based on the works from [1,2,34,42,50,57].

Model-specific techniques are restricted to particular model architectures, either being tailored to a single model or a narrow group of similar models. In contrast, *model-agnostic* methods are independent of model type and can be applied to any trained model. Such agnostic approaches typically focus on input–output correlation analysis. Furthermore, given their versatility and adaptability across various ML architectures and frameworks, model-agnostic methods are generally more desirable [35].

Scope of explanation

Another common distinction in the literature is the scope of explanation, which refers to the extent/amount of information an explanation provides about a given sample/dataset. *Local* methodologies focus on individual samples or specific predictions, and explain the model's behaviour for a particular instance. Conversely, *global* explanation techniques are generalisable to the entire model and dataset, offering insights into the overall functioning of the model. From an implementation perspective, global explanations are more valuable for extracting process knowledge and understanding model behaviour as a whole. Local explanations, in contrast, are particularly beneficial in production environments, providing quick, instance-specific insights to support real-time decision-making.

Explanation output

Unlike the previous two classifications, widely adopted in the literature for XAI classification, this third criterion is less frequently considered. However, the authors believe it is of upmost importance, as it directly influences how explanations are perceived and conveyed to users. Four distinct types of outputs - explanations - are distinguished in the proposed taxonomy, as also proposed by Linardatos et al. [57]:

- **Image:** for most image classification scenarios, the employed XAI method is based on feature attribution. Explanations are often visualised by superimposing a heatmap over the original image, highlighting the image regions most influential to a model's prediction. Such outputs are generally more intuitive and easily interpreted by users when compared to numerical feature relevance values, for instance.
- **Tabular:** these outputs consist of numerical data. While offering precise and detailed information, they can be more challenging to interpret compared to images and graphical outputs. They often require greater familiarity and knowledge over the specific problem domain and even XAI method employed.
- **Graph:** some XAI methods, including Accumulated Local Effects (ALE) and Partial Dependence Plots (PDP), produce graphical representations as their main outputs. ALE plots display average relevance scores (ALE value) over a feature's range, while PDPs plot model predictions against variations in a particular feature, offering direct input–output correlation analyses. Moreover, clustering-based methods also frequently rely on graphical outputs to convey some sort of explanation and interpretation.
- **Text:** less common, textual explanations are often associated with rule-based methods, or generated through post-XAI processing, translating numerical outputs into text for improved human interpretability. Additionally, self-explanatory neural networks and sentiment analysis techniques can directly output textual explanations using natural language processing [60–62].

Explanation methodology

Finally, the fourth criterion proposed for classifying XAI methods is based on their explanation methodology. This criterion provides a more comprehensive distinction among various methods and their approach to achieving interpretability. While already common in several prior taxonomies, the authors have formalised this category with a specific focus on explainability within industrial scenarios. The following sub-categories are defined:

- **Example generation:** these methods provide representative examples of data samples that intuitively illustrate particular relationships and correlations learned by the model. Such examples can either be drawn from the input dataset or generated through "what-if" scenarios, the latter mostly known as counterfactuals [1,2,42]. A counterfactual example represents the minimal change required in input data to alter a model's prediction to a desired outcome [57].
- **Rule extraction:** these methods provide explainability by deriving rules that describe the model's decision process and output. Typically expressed as "if-then" statements or boolean rules, they are inspired by tree-based methods that use hierarchical decision structures to conduct their predictions [2,42,57]. An example might be *"if feature n is greater than X, then there is a Y% likelihood of class C"*.
- **Explanation by simplification:** these approaches use simpler, transparent models - referred to as surrogate models - to approximate the behaviour of a more complex classifier. Consequently, interpretability is achieved by directly analysing the surrogate "white-box" model rather than the original "black-box" model, mostly locally [1,34].
- **Input visualisation:** these techniques focus on direct visualisation of input–output correlations. After applying the classifier, its output is leveraged and compared in relation to input features. Methods such as PDP and ALE fall within this category, offering graphical insights into feature relevances and predictions.
- **Concept-based:** aims to make ML models more interpretable by linking their decisions to human-understandable concepts. Instead of explaining predictions through raw features or complex internal workings, these methods identify and leverage abstract "concepts" that humans can readily grasp.
- **Feature attribution:** among the most widely implemented methods in the literature, feature attribution techniques assess the contribution of each input feature to the ML model's predictions [1]. Given the variety of methodologies used to compute feature relevance, a further 5-split division can be made:
  - **Game theory-based:** primarily associated with Shapley Additive Explanations (SHAP) and its variants, these methods are grounded in cooperative game theory. Inspired by Lloyd Shapley's work (1953), they evaluate the contribution of each feature (player) to a particular prediction (gain), relative to the average prediction across the dataset [42,63].
  - **Perturbation-based:** typically applied locally, these methods introduce small changes to input feature values, measuring its effects on the model's predictions. As a result, the higher the output change, the higher the relevance of the perturbed feature [42].
  - **Occlusion-based:** often considered a subset of perturbation-based methods, these are frequently used in DL, particularly for image analysis. They involve systematically occluding parts of an input (*e.g.*, image regions or pixels) and observing the impact on the model's output [58]. As a consequence, significant output changes indicate regions of high relevance.
  - **Gradient-based:** also referred to as backpropagation-based methods, such methods are also mostly focused on neural network post-hoc analyses. They calculate the gradients of model predictions with respect to input features, using their values to infer feature relevance.
  - **Weight-based:** typically model-specific, these methods apply to ML algorithms that assign weights to features, similar to linear or logistic regression methods. The learned weights are then interpreted as indicators of feature importance.

Regarding feature attribution, it is also worth noting the existence of a potential sixth group, post-hoc *clustering* approaches. These methods include dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (t-SNE), as well as common unsupervised clustering techniques such as K-means and DBSCAN. However, while these methods can enhance data interpretability, they can be applied either as a pre-processing or post-processing technique. For instance, in fault detection and diagnosis scenarios, clustering is often employed post-hoc, following a binary classifier that distinguishes normal and abnormal samples to further discriminate fault types. In other cases, clustering and dimensionality reduction techniques are used in pre-processing to facilitate process understanding, thereby contributing to intrinsic interpretability. Due to its dual application, clustering approaches are not included in the proposed taxonomy.

Overall, the proposed taxonomy consolidates and extends multiple existing frameworks from the literature. While many such proposals primarily focus on the theoretical principles of **Explainable AI (XAI)** methods, the classification presented here specifically adapts them to practical industrial and stakeholder interests. It offers a clear distinction between the classification of Machine Learning (ML) models and XAI methods, a differentiation that is sometimes unclear in the existing literature. Specifically, this taxonomy differentiates between transparent and non-transparent ML models, with the latter necessitating post-hoc explainability provided by XAI techniques.

Regarding the four categories defined for post-hoc XAI methods, the authors contend that these criteria directly address key industrial priorities, reflecting the design of this taxonomy for evaluating XAI in industrial contexts. **Generalisability** is particularly relevant during model development: while model-specific methods might offer more granular insights, they frequently lack versatility; conversely, model-agnostic approaches provide enhanced applicability, sometimes at the expense of personalisation. The **scope of explanation** plays a crucial role during deployment: local explanations are well-suited for real-time decision-making and analysis on the production floor, whereas global explanations offer high-level insights, proving beneficial for trend analysis and strategic decision-making at the management level. Similarly, the choice of **explanation output** directly influences usability, with image-based and textual explanations often being more accessible to frontline operators, enabling easier and immediate interpretation, while graphical and tabular outputs typically require more detailed and analytical reviews. Finally, the **explanation methodology** directly impacts the nature of the outputs, computational efficiency, and the practical applicability of XAI methods within industrial environments. By integrating all the aforementioned considerations, the proposed taxonomy offers a comprehensive, practical, and user-centred framework that is better aligned with the demanding requirements of industrial fault detection and diagnosis.

2.3. XAI output presentation and evaluation

Human interpretability and understandability are central to the purpose of XAI. Therefore, developing adequate XAI tools and metrics to effectively communicate interpretable outputs is extremely relevant. Furthermore, due to the inherently human-centric nature of XAI, the quality and relevance of explanations are often influenced by user interaction and feedback [42], making users an essential part of the explanation review and refinement process.

Although multiple studies have already explored the topic of XAI evaluation [2,40,42,57,58,64], the literature tends to emphasise theoretical frameworks for evaluating explanations rather than practical approaches that look into how users assess and evaluate XAI outputs themselves. As a result, there is a notable lack of practical, user-centric metrics/indicators for evaluating XAI explanations in applied contexts [57].

Within this theoretical context, several authors have categorised explanations into three primary groups [40,57,64]:

- **Application-grounded:** type of explanation focused on domain-experts. Such explanations provide more detailed insights into specific decisions made by a model. They directly test the system's intended purpose but are typically resource-intensive and costly to implement.
- **Human-grounded:** these explanations are less complex and aimed at a broader non-expert audience. Although they offer less detailed insights than application-grounded explanations, they are easier to achieve and understand and less costly.
- **Functionality-grounded:** this type of evaluation does not require human intervention. Instead, functionality-grounded explanations rely on objective metrics and indicators, such as numerical scores, to evaluate the quality of explanations.

Regarding other literature proposals, Van der Velden et al. [40] propose six evaluation criteria for explanations: (1) ease of use, assessing if an XAI technique can easily generate explanations in multiple different scenarios; (2) validity, assessing if explanations are correct and coherent with end-user beliefs; (3) robustness, analysing if significant changes in the model's architecture accordingly affect the explanation; (4) computational cost, reflecting the computational resources an XAI method requires to provide an explanation; (5) necessity of fine-tuning, assessing if XAI techniques are readily implementable, or if their performance is affected by their parameters and/or the parameters of the target model; and (6) open-source availability.

Carvalho et al. [64] divide interpretability indicators into two groups, qualitative and quantitative, and are among the few authors who focus explicitly on practical interpretability considerations. Their qualitative indicators include the (1) form of explanation (feature relevance score, set of rules extracted, etc.); (2) size of the explanation; its (3) compositionality which is related to how it is organised and how it organises information (if, for instance, it showcases ordered feature rankings); its (4) monotonicity and interactions between units (linear, nonlinear, etc.); its (5) uncertainty, returning some sort of uncertainty measurement; and its (6) stochasticity, assessing if any random processes are part of the explanation. Regarding quantitative indicators, the authors refer to various other works which proposed different explanation axioms:

- Sundararajan et al. [65] proposed the sensitivity and implementation invariance axioms for neural network explanations. Sensitivity states that if a change in a single feature leads to a different predictions, the feature's attribution should be non-zero. Implementation invariance requires that two functionally equivalent models yield identical explanations.
- Honegger [66] proposed three distinct axioms, identity, separability, and stability. Identity ensures that identical inputs produce identical explanations. Separability stipulates that different inputs should yield different explanations. And stability requires similar inputs in similar explanations.
- Silva et al. [67] define the "3 Cs" axioms as completeness, correctness, and compactness. Completeness concerns the coverage of explanations (*i.e.*, the proportion of instances addressed within an explanation). Correctness asserts that instances explained similarly should share the same label. And compactness emphasises the need for concise explanations that are easily understandable by any user.

As discussed previously, the literature appears to be more concerned with the theoretical principles of what constitutes a "good explanation", rather than how explanations are communicated and will influence end users. While the authors agree that compliance with theoretical axioms enhances the quality and interpretability of explanations, they also contend that the way interpretability is delivered - whether through visual, textual, or numerical means - should also be a fundamental component of XAI evaluation. Such indicators act as critical intermediaries between XAI outputs and user comprehension.

In this context, the approach followed in this review aims to bridge this gap by placing a strong emphasis on the practical delivery of explanations. Thus, in Section 4.4, a comprehensive review was conducted on the numerical and graphical tools used by various authors to facilitate user interpretability.

3. Research methodology

This section outlines the research methodology employed for this review paper, which follows the systematic literature review approach based on the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines [68]. This approach has been previously adopted in similar contexts to comprehensively analyse and discuss the literature on XAI [69–71]. However, prior reviews have mostly addressed broader XAI concepts or general industrial applications, rather than focusing specifically on key practical concerns and tasks, such as fault detection and diagnosis. The overall process of identification, screening, and selection of relevant publications is depicted in Fig. 5.

The initial paper identification stage involved conducting a search in the Scopus database, yielding a total of 167 records as the initial number. The search query was the following:

(1) "explainable artificial intelligence" *OR* "explainable AI" *OR* "interpretable AI" *OR* "interpretable artificial intelligence" *OR* "trust* AI" *OR* "trust* artificial intelligence"
AND (2) "industr*" *OR* "manufact*" *OR* "smart manufact*" *OR* "IIoT" *OR* "industrial IoT" *OR* "industrial internet of things"
AND (3) "fault detection" *OR* "anomaly detection" *OR* "failure detection" *OR* "fault classification" *OR* "defect detection" *OR* "failure prediction"

It is relevant to mention that the initial search was conducted on October 10th 2024, resulting in 167 records to be further filtered and analysed.

The second phase focused on a detailed filtering and screening process for the identified papers. As shown in Fig. 5, three initial filters were applied directly in the Scopus database. Firstly, the publication range was limited to 2020 onwards. Given the novelty of the research area, this only excluded two publications. Secondly, a typology filter was employed, accepting only journal and conference papers. This excluded an additional 24 records, mostly review papers on the subject of XAI. Lastly, a language filter eliminated an additional two records which were non-English publications. As a result, this initial filtering process reduced the number of publications to be analysed to 139. Notably, after the initial filtering, 10 additional records screened from other reviews or manually chosen were included, totalling 149 publications. These were further analysed and filtered based on their abstract information.

During the abstract screening stage, two initial and direct filters were applied. These involved excluding publications that were inaccessible to the authors (11) and abstracts related only to conference poster presentations (3). Abstract reading was then conducted, and further exclusions were made based on the following:

- Not fully related to industry (9): publications which lacked industrial applicability were excluded.
- No practical results/implementation (12): any record which focused on theoretical work, and/or lacked practical results and implementation was dropped.
- Little information on XAI (10): records which did not specify the XAI method employed in their abstract, and/or information regarding explainability and interpretability were excluded.

After these steps, 104 records remained. To further refine the selection given the high number of remaining papers, publications with no citations (42) were also excluded. This left a total of 62 records to be fully read.

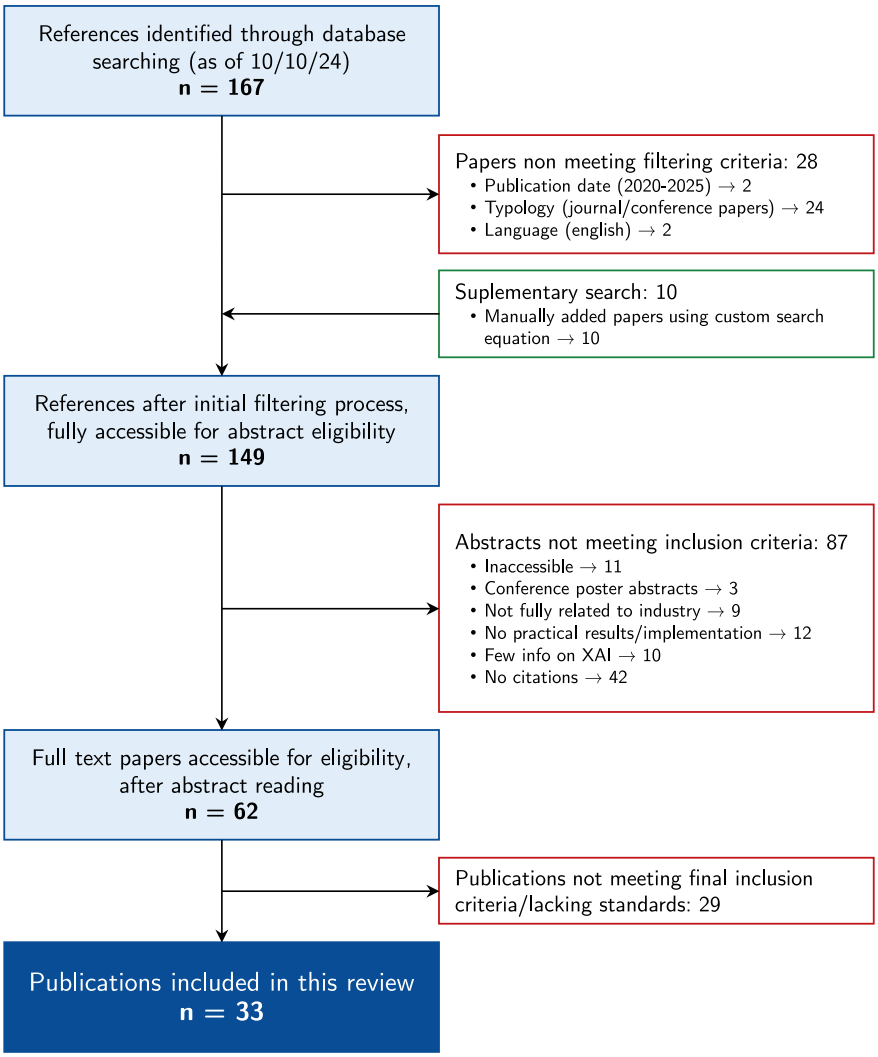


Fig. 5. PRISMA schematic for the article selection process.

The final selection stage involved full paper reading and analysis. Any publication with unclear objectives, limited insightful results, and insufficient XAI-related information was excluded (29). This resulted in a final selection of 33 publications, which were analysed in detail. Fig. 6 provides details regarding publication years and publishers for the remaining records, including journal and conference publications.

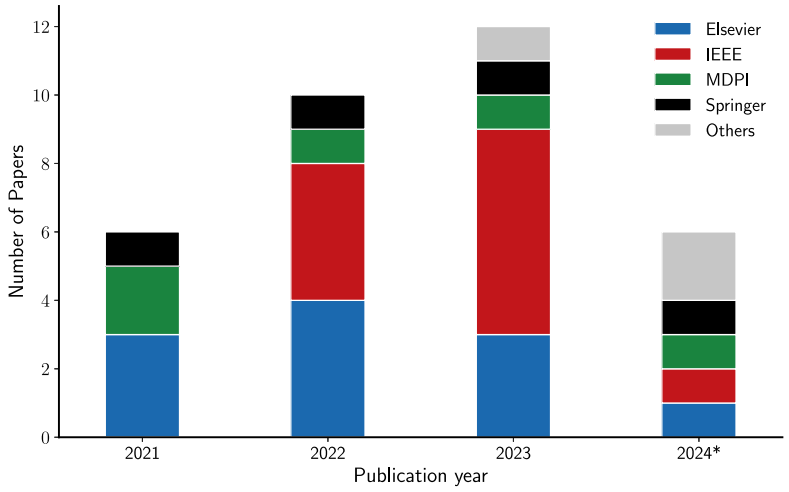
This initial analysis reflects a positive trend in the number of publications per year (Fig. 6(a)), increasing from 6 publications in 2021 to 12 in 2023. The lower number of publications in 2024 may be attributed to several factors: this review process was conducted during 2024, leaving limited time for newer publications to be included; and the filtering process employed by the authors themselves, which excluded at one stage papers with no citations. This may have affected newer publications that had not yet been cited. From the final records, 24 of them belonged to journal publications, while the remaining 9 were presented in conferences. Looking at Fig. 6(b), it can be seen that most papers were published in recognised Q1 journals (a total of 20 papers, 83.3%). As for the remaining journal papers, 3 were part of Q2 journals and the last one to a Q3 journal. As for the conference publications (Fig. 6(c)), 7 of the 9 were organised by IEEE, with the other 2 belonging to Springer and Elsevier.

In summary, the final selection of 33 publications can be considered representative within the context of the review and its current research state. As described in this section, the rigorous filtering process applied throughout each stage of the PRISMA methodology ensured that only high-quality and relevant publications were included. The exclusion criteria, particularly the focus on papers with practical applicability and evidence, and sufficient citation metrics, contributed to refining the results to works that demonstrated relevance and impact within academia. Furthermore, a significant proportion of journal publications (83.3%) were included in high-impact and relevant Q1 journals, with conference publications being organised by recognised publishers - IEEE, Elsevier, and Springer. Given the specific focus of the review - XAI applicability to industrial fault detection and diagnosis tasks - the resulting sample size reflects the novelty of the research area at the time of the search. Therefore, the final number of publications can be regarded as suitable for identifying recent trends in the research area.

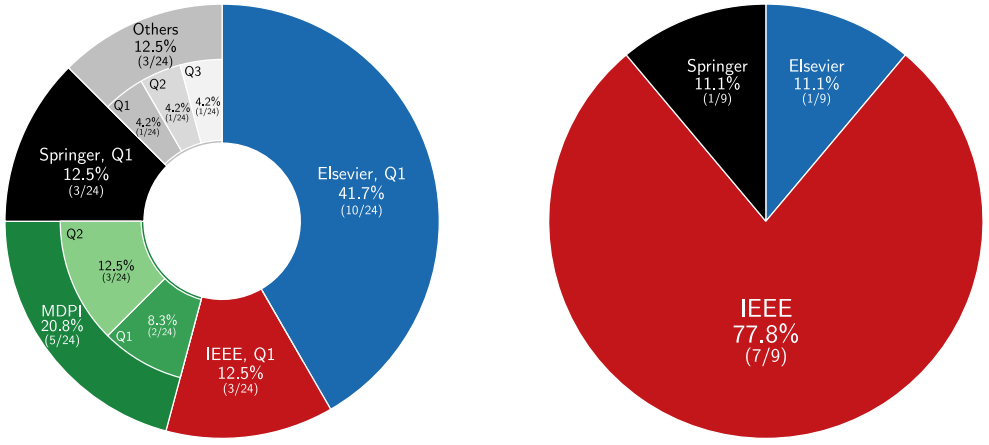
4. Publication detail analysis

This section presents a detailed analysis and discussion of the reviewed papers, addressing the main goals outlined in Section 1. The discussion covers the primary applications and problem typologies identified, the XAI methods employed and their classification (according to the taxonomy provided in Section 2.2), the approaches used to achieve interpretability and explainability in each case, and the tools implemented to support those.

Given the number and complexity of the reviewed works, an external table has been developed to offer a comprehensive overview



(a)



(b) (c)

Fig. 6. (a) Evolution in the number of publications (conference and journals) for the selected records, (b) percentage of final journal papers per publisher and their quartile information, and (c) percentage of conference publications per publisher.

of each paper. The table includes key information regarding application domains, brief description of the proposed solution, ML and XAI techniques employed, XAI-related evaluation metrics, and key insights derived from the results for all individual publications that were analysed.

4.1. Application areas and problem scopes

This analysis begins with an evaluation of each paper's primary investigation area and application, along with insights into how the authors approached their use cases through ML and XAI methodologies. Detailed information on these aspects is provided in Table 1 and Fig. 7.

Starting with the global area of investigation in the reviewed publications, manufacturing industry emerged as the dominant area, with 13 out of 33 papers (39.5%) addressing topics such as anomaly detection and classification for equipment, processes, and products. Cybersecurity, an essential area in modern industries and particularly IIoT environments, was explored in 5 publications (15.2%) focusing on intrusion detection. The remaining papers spanned diverse sectors, including pharmaceutical, healthcare, aerospace, and energy sectors. Additionally, several publications demonstrated versatile applications across multiple case studies. Nor et al. [72] developed a DL approach

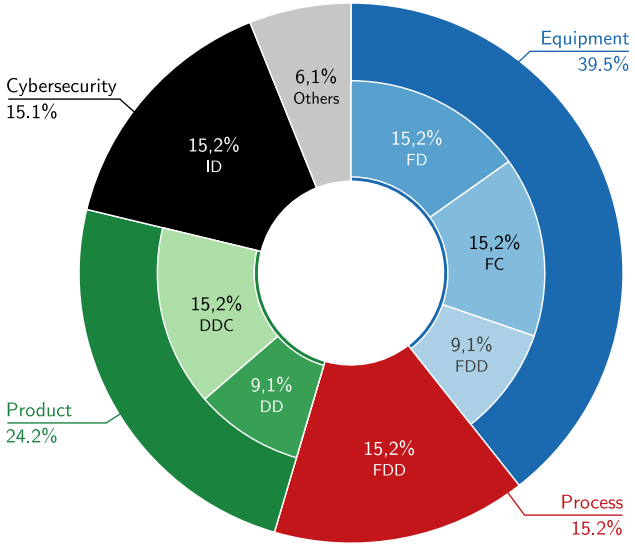


Fig. 7. Problem typologies and focuses of the reviewed papers (FD = fault detection, FC = fault classification, FDD = fault detection and diagnosis, DD = defect detection, DDC = defect detection and classification, ID = intrusion detection).

Table 1
Overview of area, application, and problem scopes of the reviewed publications.

Ref.	Application	Area	Input data	Problem nature
Nor et al. [72]	Equipment fault detection	Energy and aerospace industries	Multivariate time series	Unsupervised binary classification
Anello et al. [73]	Equipment fault detection	Entertainment and manufacturing industries	Multivariate tabular	Unsupervised binary classification
Lorenti et al. [74]	Equipment fault detection	Manufacturing industry	Multivariate tabular	Unsupervised binary classification
Kim et al. [75]	Equipment fault detection	Manufacturing industry	Multivariate tabular	Supervised binary classification
Li et al. [76]	Equipment fault detection	Manufacturing industry	Multivariate time-series + frequency-spectrum	Supervised multilabel classification
Hu et al. [77]	Equipment fault classification	Manufacturing industry	Multivariate time series	Supervised multilabel classification
Wen et al. [78]	Equipment fault classification	Manufacturing industry	Univariate time series + frequency-spectrum	Supervised multilabel classification
Mohan Dash et al. [79]	Equipment fault classification	Energy industry	Multivariate time series	Supervised multilabel classification
Kakavandi et al. [80]	Equipment fault classification	Manufacturing and medical industry	Multivariate time series	Supervised multilabel classification
Brito et al. [81]	Equipment fault classification	Manufacturing industry	Frequency-spectrum	Supervised multilabel classification
Serradilla et al. [82]	Equipment fault detection and diagnosis	Manufacturing industry	Multivariate time series	Unsupervised binary classification (detection) and unsupervised multilabel classification (diagnosis)
Kim et al. [83]	Equipment fault detection and diagnosis	Maritime industry	Multivariate tabular	Unsupervised binary classification (detection) and unsupervised multilabel classification (diagnosis)
Baek and Kim [84]	Equipment fault detection and diagnosis	Manufacturing industry	Multivariate time series	Unsupervised binary classification (detection) and supervised multilabel classification (diagnosis)
Agarwal et al. [85]	Process fault detection and diagnosis	Chemical industry	Multivariate tabular	Supervised binary classification (detection) and supervised multilabel classification (diagnosis)
Harinarayan and Shalinie [86]	Process fault detection and diagnosis	Chemical industry	Multivariate tabular	Supervised multilabel classification
Peng et al. [87]	Process fault detection and diagnosis	Chemical industry	Multivariate tabular	Supervised binary classification (detection) and supervised multilabel classification (diagnosis)
Yang et al. [88]	Process fault detection and diagnosis	Manufacturing industry	Multivariate tabular	Unsupervised binary (detection) classification and unsupervised multilabel classification (diagnosis)
Jang et al. [89]	Process fault detection and diagnosis	Chemical industry	Multivariate tabular	Unsupervised binary classification (detection) and unsupervised multilabel classification (diagnosis)
Gerschner et al. [90]	Product defect detection	Manufacturing industry	Image	Supervised binary classification
Raab et al. [91]	Product defect detection	Pharmaceutical industry	Image	Supervised binary classification
Matos et al. [92]	Product defect detection	Automotive industry	Multivariate tabular	Unsupervised and supervised binary classification
Lee et al. [93]	Product defect detection and classification	Manufacturing industry	Image	Supervised multilabel classification
Kumar et al. [94]	Product defect detection and classification	3D printing industry	Image	Unsupervised binary classification (detection) and unsupervised multilabel classification
Bordekar et al. [95]	Product defect detection and classification	3D printing industry	Image	Supervised multilabel classification
Saleh and Metin Ertunç [96]	Product defect detection and classification	Manufacturing industry	Image	Supervised binary classification (detection) and unsupervised multilabel classification
Meister et al. [97]	Product defect detection and classification	Aerospace industry	Image	Supervised multilabel classification
Wawrowski et al. [98]	Intrusion detection	Cybersecurity	Multivariate tabular	Unsupervised binary classification
Rathod et al. [99]	Intrusion detection	Cybersecurity	Multivariate tabular	Supervised binary classification
Namrita Gummadi et al. [100]	Intrusion detection	Cybersecurity	Multivariate tabular	Supervised multilabel classification
Han and Chang [101]	Intrusion detection	Cybersecurity	Multivariate tabular	Semi-supervised binary classification
Keshk et al. [102]	Intrusion detection	Cybersecurity	Multivariate tabular	Supervised binary classification
Barbado et al. [103]	Fault detection	–	Multivariate tabular	Unsupervised binary classification
Kang and Kang [104]	Process/equipment fault detection	–	Multivariate time series	Unsupervised binary classification

combined with SHAP for unsupervised fault detection in a gas turbine and turbofan. Anello et al. [73] proposed an unsupervised fault detection methodology combined with Accelerated Model-Agnostic Explanations (AcME) for both local and global interpretability, applied to a roller coaster (entertainment) and a compacting machine (manufacturing). Kakavandi et al. [80] integrated gradient-based XAI methods with a DL classifier for fault detection in both a vessel engine and a medical assembly process. Finally, two additional publications did

not specify an a target industrial domain. However, they validated their methodologies using benchmark datasets: Barbado et al. [103] proposed a novel clustering rule extraction methodology, whereas Kang and Kang [104] introduced a DL-based unsupervised fault detection model with inherent interpretability.

As shown in Fig. 7, regarding applications and nature of problems addressed, the reviewed publications fall under four main application areas: equipment, process, product, and cybersecurity. Equipment-related studies were evenly split between fault detection (FD) (typically a binary classification problem), and fault classification (FC), addressing multilabel problems, each with 5 publications (15.1%). In equipment fault detection, Lorenti et al. [74] proposed CUAD-Mo, an unsupervised anomaly detection pipeline for CNC machine tool failure detection. Their approach involved an iForest classifier combined with Depth-based Isolation Forest Feature Importance (DIIFI) for computing local feature importance scores. Moreover, t-SNE was used during preprocessing to assist in input data segmentation, distinguishing between different tool-program combinations. Li et al. [76] applied a 1D-CNN for bearing fault detection using vibration signals and proposed Multilayer Grad-CAM (MLG-CAM), an enhanced interpretability method that extends Grad-CAM by generating activation maps across multiple convolutional layers, capturing features at varying abstraction levels and their evolution across convolutional operations. Regarding equipment fault classification, similar to the previous work, Brito et al. [81] implemented a 1D-CNN model to classify faults in rotating machinery, trained on a synthetic dataset comprised of vibration signals simulating several types of faults. To enhance interpretability, GradCAM was employed as a post-hoc method for visual interpretation of the network's outputs. And Wen et al. [78] introduced a Gradient-based Interpretable Graph Convolutional Network (GIGCN) for bearing fault diagnosis. Their model combines a GCN with GradCAM visualisations, highlighting key signal components in both time and frequency domains, enabling post-hoc visual interpretation of classification results through 1D heatmaps. The remaining three papers (9.1%) focused on fault detection and diagnosis (FDD), combining anomaly detection with fault discrimination, often using unsupervised approaches due to the scarcity of faulty data in industrial settings, leading inadvertently to unbalanced datasets. For example, Serradilla et al. [82] developed a 2D-CNN autoencoder, trained on normal data to detect anomalies in press machines. SHAP was used to locate and visualise anomalous signals, while clustering methods helped diagnosing novel failures by grouping anomalies by similarity. Kim et al. [83] used an unsupervised iForest classifier with hierarchical clustering for unsupervised fault detection in vessel engines, using SHAP for both local and global feature attribution, and Baek and Kim [84] combined a convolutional autoencoder with a Multilayer Perceptron (MLP) classifier to identify various fault types in semiconductor equipment. Once again, SHAP was applied post-hoc to identify sensor interactions that contributed to predicted faults.

In product-related applications, representative of 24.2% of the reviewed papers, studies focused on defect detection (DD) and defect detection and classification (DDC) tasks. Since product inspection is mostly conducted during final inspection stages, most studies rely on image data to identify potential external defects on the finalised products. Furthermore, given the proven effectiveness of CNNs for image-based classification and defect detection, several studies apply these models for product inspection. Saleh and Metin Ertunç [96] developed an Explainable Attention-based Fused CNN (XAFCNN) for tire defect detection and classification using X-ray scans, incorporating a special attention module for improved performance and GradCAM to visualise, through 2D heatmaps, critical defect regions. Gerschner et al. [90] employed convolutional architectures through Transfer Learning - VGG and ResNet - in this case for surface defect detection in mechanical seals, and once again used GradCAM to demonstrate the more defective regions of the part being analysed. Similarly, Raab et al. [91] also leveraged a pre-trained VGG16 CNN through Transfer Learning, targeting visual pharmaceutical capsule inspection and defect detection. Grad-CAM and SHAP were the post-hoc XAI methods applied, highlighting surface defects once again through input heatmaps. An exception to this trend was Matos et al. [92], who used tabular data from assembly and functional tests for binary defect detection in vehicle display systems, and implemented a sensitivity analysis to extract relevant features. Another innovative approach was presented by Bordekar et al. [95], who employed CT scans in a 3D printing process, using pixel-wise classification and clustering (K-means and DBSCAN), to detect and visualise internal defects on printed parts. Local Interpretable Model-Agnostic Explanations (LIME) was also employed as a feature attribution method and to explain individual model predictions regarding specific pixels.

For process-related use cases, all 5 identified papers focused on process FDD. In most cases, this involved a two-step approach: an initial binary classifier for fault identification, followed by a multilabel model to distinguish potential fault types. Examples include the work from Agarwal et al. [85], who developed two Dynamic Deep Supervised Autoencoders combined with Neural Network classifiers (DDSAE-NN), one for binary fault detection and the other for multilabel fault diagnosis in chemical processes. Notably, their XAI method - LRP - was applied iteratively during training: in each training iteration, LRP relevance scores were used to prune the input features, serving as an embedded feature selection mechanism. Harinarayan and Shalinie [86] also addressed FDD in chemical processes, employing an XGBoost classifier alongside two post-hoc XAI methods: SHAP, providing local and global insights on feature relevance, and Diverse Counterfactual Explanations (DiCE), generating counterfactual samples indicating input changes necessary to achieve a normal operating state. In the context of cybersecurity applications, all 5 papers focused on intrusion detection (ID). Among these, 4 studies developed binary methodologies aimed at distinguishing between normal and anomalous data instances. For instance, Wawrowski et al. [98] employed a Gradient Boosting classifier for intrusion detection in network traffic data, using Shapley Values to provide local post-hoc explanations by identifying input features most indicative of abnormal behaviour. And Keshk et al. [102] proposed an explainable LSTM-based intrusion detection framework for IoT networks, combining local and global interpretability methods - SHAP, Individual Conditional Expectation (ICE), PDP, and Permutation Feature Importance (PFI). Conversely, only Namrita Gummadi et al. [100] extended the analysis to classify different types of intrusions. In this case, a comparative study between multiple multilabel classifiers and post-hoc XAI methods was conducted.

Key takeaways

- **XAI integrated across multiple domains of industrial processes**. Within industrial environments, XAI applications span four core areas. (1) Industrial equipment, mainly employed for fault detection, diagnosis, and predictive maintenance. (2) Industrial processes, for continuous process monitoring and control. (3) Product inspection, at the final production stages, for defect detection and part quality assessment. And for (4) cybersecurity, more focused on data safety and protection in IIoT systems.
- **XAI demonstrates versatility across applications.** While manufacturing industry seems to be the area with more focus from the literature, from the gathered results it can be seen that XAI is increasingly applied in several other areas. In chemical industry, especially for process monitoring and anomaly detection [85–87,89]; within IIoT environments for cybersecurity purposes [98–102]; and for part quality assessment, where critical areas such as 3D printing [94,95], automotive [92], and aerospace industry [97] are employing XAI for testing product quality.
- **XAI enhances unsupervised learning scenarios.** Many studies employ XAI to support fault discrimination in unsupervised contexts [82–85,87–89,94,96]. Especially in combined approaches, where binary and multilabel classification are leveraged, XAI helps interpret different fault types through clustering and validating unsupervised classifiers, facilitating the identification of new fault types.

Table 2
List of common scopes, approaches to achieve interpretability, and outputs identified in the papers.

Identifier	Brief description
Scope of explanation (SE)	
SE_1	Local feature relevance computing
SE_2	Global feature relevance computing
How interpretability is achieved (INT)	
INT_1	Interpretability based on feature relevance scores
INT_2	Interpretability through rule extraction
INT_3	Counterfactual explanation generation
INT_4	Interpretability through a surrogate "white-box" model
INT_5	Interpretability inherent to ML classifier
INT_6	Interpretability directly based on model design, architecture, or output
INT_7	Interpretability through direct input feature - output analysis
Explainability presentation formats (EPF)	
EPF_1	Numerical metric generation and comparison
EPF_2	Graphical tools used for interpretation/explainability
EPF_3	Input heatmap representations aiding interpretability
EPF_4	Clustering graphs for anomaly type discrimination
EPF_5	Elaboration of textual explanations based on XAI outputs
Other indicators (OTH)	
OTH_1	Combining the output of 2 or more methods for more complete interpretations/explanations
OTH_2	Domain/expert-knowledge inclusion for XAI explanation validation
OTH_3	XAI for dimensionality reduction and model optimisation

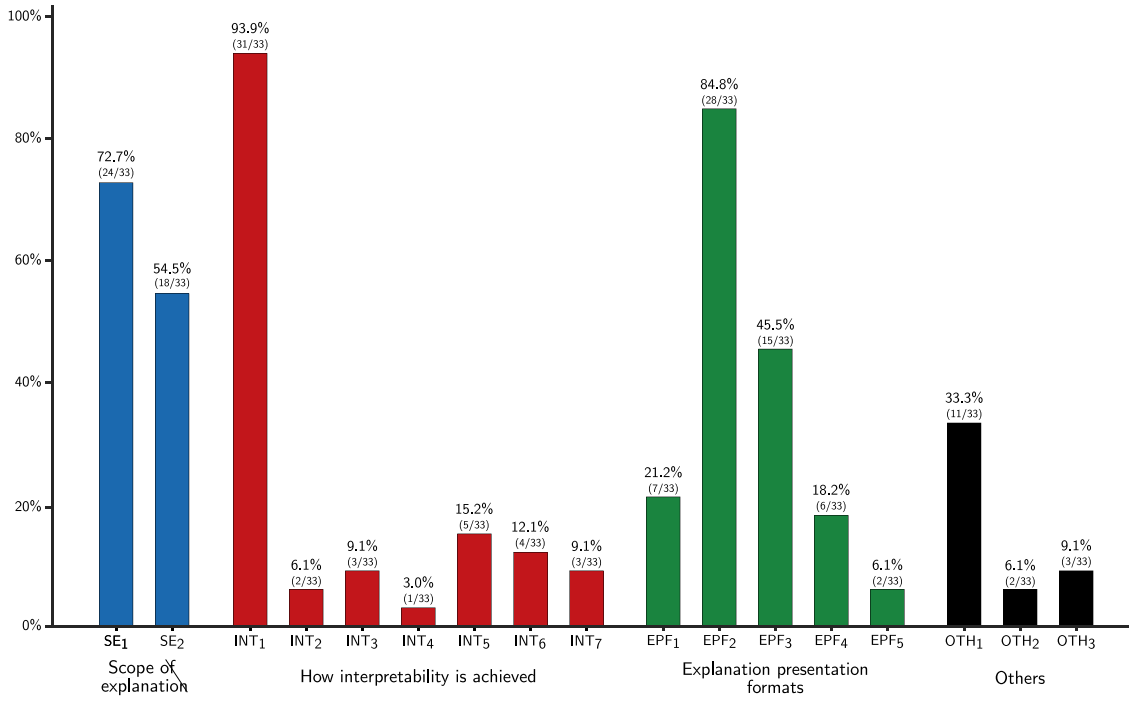


Fig. 8. Graphical overview of the XAI methods, approaches to achieve interpretability, objectives, and outputs of each publication (according to Table 3).

4.2. XAI methods and scopes

Building on the initial analysis, which focused on the areas, applications, and the nature of the ML-based problems, this section focuses on the XAI methods employed in the reviewed papers to promote interpretability and explainability. It explores the specific methods used, the rationale behind their implementation, and the outcomes of their application. To support this analysis, Table 2 provides a list of common objectives, approaches/methodologies to achieve interpretability, and outputs of the application of XAI methods identified in the reviewed papers, with Table 3 mapping those to each individual publication, alongside the specific XAI method(s) used. Finally, Fig. 8 illustrates through a bar graph the numerical distribution of the information detailed in Table 3.

While analysing the different works, several "indicators" were identified concerning the objectives, implementation, and outputs of the XAI implementations (as shown in Table 2). Those were grouped into the following categories:

- **Scope of explanation (SE):** this category defines whether the implementations focused on explaining individual, local instances of data (SE_1), or on promoting global interpretability across an entire dataset (SE_2). In some instances, these were explored simultaneously.
- **How interpretability is achieved (INT):** across the papers, numerous approaches were identified for enhancing model and output interpretability. The most common in the literature involves computing feature relevance scores (INT_1), providing insight into

Table 3
Detail review of the methods, approaches to achieve interpretability, objectives, and outputs of XAI application for each individual paper (based on Table 2).

Ref.	XAI method(s)	Proposed (P)/ Existing (E)	Different objectives for the application of XAI																	Total marks
			SE_1	SE_2	INT_1	INT_2	INT_3	INT_4	INT_5	INT_6	INT_7	EPF_1	EPF_2	EPF_3	EPF_4	EPF_5	OTH_1	OTH_2	OTH_3	
[82]	SHAP (gb), t-SNE, SOM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	✓	–	✓	–	–	6
[85]	LRP	E	–	✓	✓	–	–	–	–	–	–	–	✓	✓	–	–	✓	–	✓	6
[72]	SHAP (gb)	E	✓	✓	✓	–	–	–	–	–	–	–	✓	–	–	–	–	–	✓	5
[83]	SHAP (tb), hierarchical clustering	E	✓	✓	✓	–	–	–	–	–	–	–	✓	✓	✓	✓	✓	–	–	8
[86]	SHAP (tb), DiCE	E	✓	✓	✓	–	✓	–	–	–	–	–	✓	–	–	–	✓	–	–	6
[93]	LRP, DT	E	✓	–	✓	✓	–	✓	–	✓	–	–	✓	✓	–	✓	✓	✓	–	10
[87]	SmoothGrad	E	✓	✓	✓	–	–	–	–	–	–	–	✓	–	–	–	–	–	–	4
[98]	Shapley values	E	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
[89]	SHAP (gb), hierarchical clustering	E	✓	✓	✓	–	–	–	–	–	–	✓	✓	✓	✓	–	✓	–	–	8
[73]	AcME	E	✓	✓	✓	–	–	–	–	–	–	–	✓	–	–	–	–	–	–	4
[96]	GradCAM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
[99]	Eli5	E	–	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	✓	3
[74]	DIFFI, t-SNE	E	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
[77]	SHAP (gb)	E	✓	✓	✓	–	–	–	–	–	–	–	✓	–	–	–	–	–	–	4
[100]	SHAP (tab), LOCO, CEM, ALE, LIME, PFI, ProfWeight	E	✓	✓	✓	–	–	–	–	–	✓	–	✓	–	–	–	✓	–	–	6
[90]	GradCAM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
[78]	GradCAM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
[101]	SHAP (gb), t-SNE	E	–	✓	✓	–	–	–	–	–	–	–	✓	–	✓	–	✓	–	–	5
[84]	SHAP (kb)	E	–	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2
[91]	SHAP (gb), GradCAM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
[92]	Sensitivity analysis	E	–	✓	✓	–	–	–	–	–	✓	–	✓	–	–	–	–	–	–	4
[80]	Occlusion, Gradient, IG, SHAP (kb + gb)	E	✓	✓	✓	–	–	–	–	–	–	✓	✓	✓	–	–	–	✓	–	7
[102]	SHAP (gb), ICE, PDP, PFI	E	✓	✓	✓	–	✓	–	–	–	✓	–	✓	–	–	–	✓	–	–	7
[97]	SmoothGrad, t-SNE	E	✓	–	✓	–	–	–	✓	–	–	✓	✓	✓	✓	–	✓	–	–	8
[81]	GradCAM	E	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
Proposed methodologies																				
[103]	Rule extraction through clustering	P	–	–	–	✓	–	–	–	–	–	✓	✓	–	–	–	–	–	–	3
[88]	Bayesian network	P	–	–	–	–	–	–	✓	✓	–	✓	✓	–	–	–	–	–	–	4
[104]	VTT (w/ attention mechanism)	P	–	✓	✓	–	–	–	✓	–	–	–	✓	–	–	–	–	–	–	4
[94]	Zero-bias CNN	P	–	–	–	–	–	–	✓	✓	–	✓	✓	–	–	–	–	–	–	4
[95]	Cluster-based defect detection and visualisation + LIME	P	✓	–	✓	–	–	–	–	–	–	–	✓	✓	✓	–	✓	–	–	6
[79]	Occlusion-based FDI-X	P	✓	–	✓	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	4
[75]	Frequency- and last condition-based feature correction for sample "normalisation" + SHAP (tb)	P	✓	–	✓	–	✓	–	✓	✓	–	–	–	–	–	–	–	–	–	5
[76]	MLG-CAM	P	✓	–	✓	–	–	–	–	–	–	✓	✓	✓	–	–	–	–	–	5

Note: For SHAP methods, gb = gradient-based, tb = tree-based, tab = tabular-based, kb = kernel-based

the most important input variables both locally and globally. Some instances introduce XAI methods that extrapolate interpretable numerical rules that define particular data classes (INT_2), with others generating counterfactual explanations (INT_3), *i.e.*, determining changes required in input variables to alter model predictions (commonly from anomaly to normal). There is also the possibility of employing a second "white-box" model to provide the explanations of the decisions made by the "black-box" model (INT_4), and of incorporating interpretability directly on the ML classifier (INT_5). Other cases assume interpretability based on the model design or architecture (INT_6), and finally, some authors perform direct (mostly graphical) analysis between input features and their influence on the output of the classifier (INT_7).

- **Explanation presentation formats (EPF):** these represent how authors convey explainability to the end-user. These include the use of numerical indicators and metrics (EPF_1), graphical illustrations (EPF_2), heatmap representations which highlight for the original set of data regions that most influence the output (EPF_3) (mostly used for image data), and clustering representations that help visual identification of several classes of data (EPF_4) (employed in combination with an unsupervised ML classification task). Finally, it is also possible to expose the outputs of XAI techniques through text (EPF_5). A more thorough revision of methods to induce user explainability, as well as numerical evaluation metrics, is provided in Section 4.4.
- **Other indicators (OTH):** other indicators which do not fit any of the previous categories are related to the combination of 2 or more XAI methods (OTH_1), joining their outputs to enhance explanation quality, the inclusion of expert knowledge to either validate or help the process of generating explanations (OTH_2), and finally, using XAI-based feature relevance scores for dimensionality reduction and model pruning tasks (OTH_3).

Aligned with the criteria defined and described in Tables 2, 3 provides a detailed classification of all the reviewed publications according to those, with Fig. 8 graphically illustrating the results of the table.

Depending on the application and complexity of the developed frameworks, different works complied with different scopes, approaches to achieve interpretability and to output results, as represented in Table 3. Of particular emphasis are the works from Lee et al. [93], Jang et al. [89] and Meister et al. [97], as the ones which "ticked" the larger number of boxes. Lee et al. [93] combined two XAI methods for improved interpretability, LRP and a transparent DT model, applying them to a neural network and focusing on product defect detection using image data. The integration of LRP's heatmaps and the DT's decision rules based on the network's outputs provided relevant local insights both at input and output levels. Additionally, their hybrid XAI approach was validated through a structured survey involving domain experts, who reported increased trust, perceived validity, and satisfaction compared to a previous implementation lacking XAI support. Jang et al. [89] combined SHAP, a model-agnostic explainer, with an unsupervised adversarial autoencoder to diagnose faults in an industrial chemical process. By leveraging SHAP's feature relevance rankings in combination with hierarchical clustering, they were able to distinguish different fault types within their unsupervised scenario. Meister et al. [97] explored a model cross-validation strategy using both Support Vector Machines (SVM) and CNN classifiers for composite product defect detection. Their methodology combined hyperplane distance analysis for the SVM method, SmoothGrad for CNN interpretability, and t-SNE clustering in the pre-processing phase to visualise distinct fault groups. These examples demonstrate the potential of combining multiple XAI methods to enhance interpretability, offering richer insights and, in some cases as shown by Lee et al. [93], providing empirical validation of their effectiveness through domain-expert evaluations.

In terms of the scope of explanations, there is no significant prevalence of either local or global approaches among the reviewed works. Of the 33 reviewed implementations, 24 (72.7%) included some form of local explanation, mainly through feature relevance computation, with 18 works (54.5%) including global explanations. Additionally, as it can be seen in Table 3, many publications combined both, providing insights both at the individual sample level as well as broader trends across datasets.

Regarding how interpretability is achieved, there is a clear preference for methods that compute input feature relevance rankings based on the outputs of a ML classifier. This type of approach was explored in 31 out of 33 works (93.9%), either for local analyses, global, or both. The dominance of this strategy aligns with the widespread adoption of SHAP, a well-established, model-agnostic method grounded in solid mathematical principles [63]. In addition to that, gradient-based methods for DL models, leveraging backpropagation to identify influential input features, are also commonly applied. These trends indicate that, although interpretability is actively being explored, most authors still prioritise feature relevance computing over trying to focus on the underlying inference mechanisms of their ML methods. Rather than directly addressing the "black-box" nature of their models, they predominantly focus on input–output interpretation, leaving the models' decision-making processes largely opaque and underexplored.

The second most explored methodology for achieving interpretability involves the implementation of inherently interpretable classifiers, identified in 5 out of the 33 reviewed publications. Among these, three focused on DL approaches. Lee et al. [93] used a DT as a surrogate model to explain the outputs of the neural network, providing clearer insights into output probabilities and uncertainties. Kumar et al. [94] proposed a Zero-bias CNN for defect detection in the 3D printing industry, enhancing interpretability through a cosine similarity layer. Based on its outputs, Mahalanobis distances were calculated to assess the similarity between sample outputs and known class templates. Yang et al. [88] developed a Bayesian Network for industrial process fault detection, with interpretability being achieved through the network's hierarchical structure. Their methodology enabled root cause identification based on parent–child relationships, with network illustrations further providing graphical tools for user interpretability.

Other methodologies include direct input–output analysis, explored in three instances [92,100,102], leveraging methods such as ALE, PDP, and sensitivity analysis. Counterfactual explanations were also explored in three scenarios [75,86,102], providing hypothetical examples that demonstrate how input changes could alter model predictions (commonly from abnormal to normal). Additionally, rule extraction methods were explored twice [93,103]. Barbado et al. [103] combined outputs from a One-class SVM with clustering algorithms (K-Means++ and K-Prototypes) to iteratively generate 2D rules, further evaluating and stating rule complexity, coverage, and overlap through quantitative metrics. Meanwhile, Lee et al. [93] extracted decision rules from a DT model, and were also the only authors to employ a surrogate "white-box" model to explain a "black-box" system. Overall, such alternative approaches highlight a growing interest in moving beyond post-hoc explanations and towards models and techniques that are inherently interpretable. However, it is relevant to state that their adoption remains quite limited compared to feature relevance-based methods.

Regarding the explainability presentation formats, there is an evident preference for graphical tools, employed in 28 out of 33 scenarios (84.8%). These include input heatmaps, implemented in 45.5% of studies, mainly focusing on image data analysis and used typically in conjunction with backpropagation-based methods for DL classifiers, and cluster heatmaps, used in 18.2% of cases, mostly in unsupervised scenarios, to identify and distinguish different fault groups. The widespread adoption of graphical tools reflects their effectiveness in promoting human interpretability, as visual representations tend to be more accessible and intuitive for humans. In contrast, numerical indicators were only applied in 7 situations (21.2%), likely due to their increased complexity and reduced versatility across different scenarios and applications. Textual explanations were the least explored, in

just two instances [83,93]. Despite their limited use, textual outputs may offer benefits within industrial environments, where easily understandable, descriptive explanations can aid in decision-making, report generation, and post-analysis processes. A more detailed discussion of graphical and numerical XAI indicators and metrics is provided in Section 4.4.

Focusing on other relevant criteria for the analysis, 11 studies (33.3%) explored two or more XAI methods for improved interpretability. These combinations typically resulted in broader coverage of explainability objectives, as reflected by their higher number of indicators in Table 3. Additionally, three works [72,85,99] leveraged XAI methods beyond their interpretability benefits, demonstrating their versatility and benefits towards dimensionality reduction, dataset simplification, and model optimisation. Finally, only two works [80,93] were successful in incorporating human domain knowledge to support the development and/or validation of explanations. Although relatively underexplored, the integration of expert knowledge is most certainly crucial for building trust in ML-based systems, and tailoring explanations to user-specific requirements. Future research should focus on expanding this human-centred approach to achieve more effective and user-aligned explainability frameworks.

Key takeaways

- **Feature relevance scoring dominates interpretability methods.** The majority of works (93.9%) rely on input feature relevance rankings as one of their XAI tools. While this indicates interest in exploring input interpretability, it may also suggest that authors focus mainly on input–output correlations rather than elucidating the underlying decision-making processes of ML models.
- **Graphical outputs are the preferred explanation tool.** Graphical indicators, namely input relevance heatmaps and cluster visualisations, are employed in 84.8% of works. These tools are preferred due to their versatility and easy interpretation by human users. In contrast, numerical indicators are explored only in 21.2% of cases, likely due to their complexity and lower applicability.
- **Combining XAI tools for enhanced interpretability.** One-third of studies combined two or more XAI techniques within their frameworks. This combination is beneficial, providing complementary analyses and insights, simultaneously allowing the identification of influential features, data cluster distinction (*e.g.*, different fault groups), decision rule extraction, and counterfactual explanations with sample generation. Such hybrid approaches offer a more comprehensive understanding of model behaviour, which is particularly valuable for real-world industrial applications.
- **Existing approaches are prioritised.** Not discussed above, but as seen in Table 3, the majority of works leverage established XAI methods, mostly related to input feature relevance computing, reflecting the maturity and wide availability of such methods. Conversely, only 24.2% of the reviewed studies aim to propose novel and customised approaches for improved interpretability.

4.3. Classification according to proposed taxonomy

Following the previous analysis of the main objectives and scopes of XAI implementation in several scenarios, this section complements it by identifying all distinct XAI methods employed and classifying them according to the proposed taxonomy in Section 2.2. In that sense, Tables 4 and 5 classify all distinct pre-existing and proposed methods employed, respectively, while Fig. 9 graphically summarises the information exposed in both tables.

Within the reviewed works, a total of 30 distinct XAI methods were identified, along with 3 clustering-based approaches (which as previously mentioned, due to their duality in implementation, were left out of the taxonomy). From these 30 methods, 26 (86.7%) are post-hoc, indicating a strong preference for employing a surrogate model to improve interpretability. Within those, there appears to be a slight tendency towards implementing model-agnostic methods, accounting for 17 of the 26 (65.4%) approaches, with the remaining 9 (34.6%) being model-specific. Most employed model-specific approaches were designed for neural networks, such as GradCAM and MLG-CAM for CNNs. One method [75] targeted tree-based models, and the remaining 3 were developed for more specific scenarios (as demonstrated in Table 5). While model-agnostic methods may have some disadvantages, including bad model generalisation and potential disregard for feature dependencies [63], they seem to be preferred due to their larger versatility and applicability to different ML approaches. Nonetheless, interest in inherently interpretable methods is growing, as these provide more transparency regarding decision-making processes and computational structure behind a model's rationale [50,105,106]. Among the reviewed works, 4 out of 30 (13.3%) methods were intrinsically interpretable. Those included a DT model and its rule-based inner structure, and 3 DL approaches: a Bayesian Network [88], a Variable Temporal Transformer [104], and a Zero-bias CNN [94]. Overall, as shown in Table 4, SHAP emerged as the most frequently employed method, used in 13 out of the 33 reviewed publications. GradCAM followed, being employed in 5 occasions. Both methods belong to the feature attribution category, reflecting a clear trend towards input feature relevance analyses, as previously highlighted in Section 4.2.

The next two classifications, regarding scope of explanation and explanation methodology, are perfectly aligned with the insights provided in Section 4.2. There is a balanced distribution of methods that output only local (11 of 26) and local and global (11 of 26) explanations, with 4 methods (ALE and PDP with two proposed approaches [75,103]) only capable of delivering global explanations. Furthermore, there is a clear dominance of methods whose methodology is based on feature attribution (69.2%), with the remaining 30.8% being distributed across example generation (2 methods), rule extraction (1 method), explanation by simplification (1 method), and input visualisation (4 methods). Among the 18 feature attribution methodologies, there are then several perturbation-based (5), gradient-based (6) and occlusion-based approaches (3), 2 game theory-based methods - SHAP and Shapley Values - a single weight-based method (ProfWeight) and also a custom attribution method, proposed by Kim et al. [75].

Finally, although graphical outputs are frequently preferred in the literature (see Section 4.2), the analysis in Fig. 9 reveals that 73.1% of post-hoc methods identified in the reviewed publications predominantly produce tabular outputs. This prevalence is likely tied to the dominance of feature attribution methods, which typically present results as numerical rankings of input features. In contrast, there are only 9 methods (34.6%) that typically produce image-based outputs, 4 methods (15.4%) producing graphical outputs, and just two methods - DiCE and Contrastive Explanation Method (CEM) - which are often directly implemented in combination with textual outputs.

Key takeaways

- **Large dominance of post-hoc implementations.** From the reviewed publications, 86.7% implement post-hoc approaches, typically applied as a surrogate model to the ML classifier. While useful and valuable for generating base explanations, they still often lack transparency on providing insights on the model's decision-making process. Inherently interpretable models, although still relatively underexplored compared to post-hoc methods, are starting to gain some interest, with 3 of the 4 identified approaches being novel proposals.
- **SHAP is the preferred method in the literature.** Along the 33 reviewed publications, SHAP was implemented in 13 of them, with the second most implemented method being GradCAM, with 5 implementations. Its popularity reflects what was previously shown, a strong preference for feature relevance analyses, though still offering limited insights into internal model reasoning.

Table 4
Overview of all pre-existing XAI techniques identified in the reviewed papers and their classification according to the proposed taxonomy in Section 2.2. Clustering techniques are separated due to their possible duality in classifications.

Method	# times used	Classification according to				
		Transparency	Generalisability	Scope	Methodology	Output
SHAP [72,75,77,80,82–84,86,89,91,100–102]	13	Post-hoc	Model-agnostic	Local and global	Feature attribution (gtb)	Tabular
GradCAM [78,81,90,91,96]	5	Post-hoc	Model-specific (CNNs)	Local	Feature attribution (gb)	Image
LIME [95,100]	2	Post-hoc	Model-agnostic	Local	Explanation by simplification (leading to feature attribution)	Tabular
LRP [85,93]	2	Post-hoc	Model-specific (DL)	Local	Feature attribution (gb)	Image
PFI [100,102]	2	Post-hoc	Model-agnostic	Local and global	Feature attribution (pb)	Tabular
SmoothGrad [87,97]	2	Post-hoc	Model-specific (DL)	Local	Feature attribution (gb)	Image
AcME [73]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (pb) (leading to example generation)	Tabular
ALE [100]	1	Post-hoc	Model-agnostic	Global	Input visualisation	Graph
CEM [100]	1	Post-hoc	Model-agnostic	Local	Example generation	Tabular or text
Decision Tree [93]	1	Model-intrinsic	(interpretability through tree architecture analysis and rule extraction from the tree nodes)			
DiCE [86]	1	Post-hoc	Model-agnostic	Local	Example generation	Tabular, text or image
DIFFI [74]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (pb)	Tabular
Eli5 [99]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (pb)	Tabular
Gradient [80]	1	Post-hoc	Model-specific (DL)	Local	Feature attribution (gb)	Tabular and image
ICE [102]	1	Post-hoc	Model-agnostic	Local and global	Input visualisation	Graph
Integrated Gradients [80]	1	Post-hoc	Model-specific (DL)	Local	Feature attribution (gb)	Tabular and image
LOCO [100]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (ob)	Tabular
Occlusion [80]	1	Post-hoc	Model-agnostic	Local	Feature attribution (ob)	Tabular and image
PDP [102]	1	Post-hoc	Model-agnostic	Global	Input visualisation	Graph
ProfWeight [100]	1	Post-hoc	Model-specific (models with weight attributions)	Local and global	Feature attribution (wb)	Tabular
Shapley values [98]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (gtb)	Tabular
Sensitivity analysis [92]	1	Post-hoc	Model-agnostic	Local and global	Feature attribution (pb)	Tabular
Clustering methods employed						
t-SNE [74,82,97,101]	4	Used either post-hoc or as a pre-processing step				
Hierarchical clustering [83,89]	2	Post-hoc	Model-agnostic	Global	–	Graph
SOM [82]	1	Post-hoc	Model-agnostic	Global	–	Graph

Note: For methods belonging to the feature attribution (methodology) class, gtb = game theory-based, pb = perturbation-based, ob = occlusion-based, gb = gradient-based, wb = weight-based.

Table 5
Overview of the novel proposed XAI methodologies identified in the review papers and their classification according to the proposed taxonomy in Section 2.2.

Method	Classification according to				
	Transparency	Generalisability	Scope	Methodology	Output
Rule extraction through clustering [103]	Post-hoc	Model-agnostic[a]	Global	Rule extraction	Tabular and graph
Cluster-based defect detection and visualisation [95]	Post-hoc	Model-specific (pixel-wise image classifiers)	Local and global	Input visualisation	Image
Occlusion-based FDI-X [79]	Post-hoc	Model-agnostic	Local	Feature attribution (ob)	Tabular
Bayesian network [88]	Model-intrinsic	(interpretability through network architecture representation and direct output analysis)			
Frequency- and last condition-based feature correction for sample "normalisation" [75]	Post-hoc	Model-specific (tree-based methods)	Global	"Custom" feature attribution leading to example generation	Tabular
MLG-CAM [76]	Post-hoc	Model-specific (CNNs)	Local	Feature attribution (gb)	Tabular and image
VTT (w/ attention mechanism) [104]	Model-intrinsic	(interpretability based on the network's direct output analysis)			
Zero-bias CNN [94]	Model-intrinsic	(interpretability based on the network's direct output analysis)			

[a] While model-agnostic, it is focused on binary classification situations.

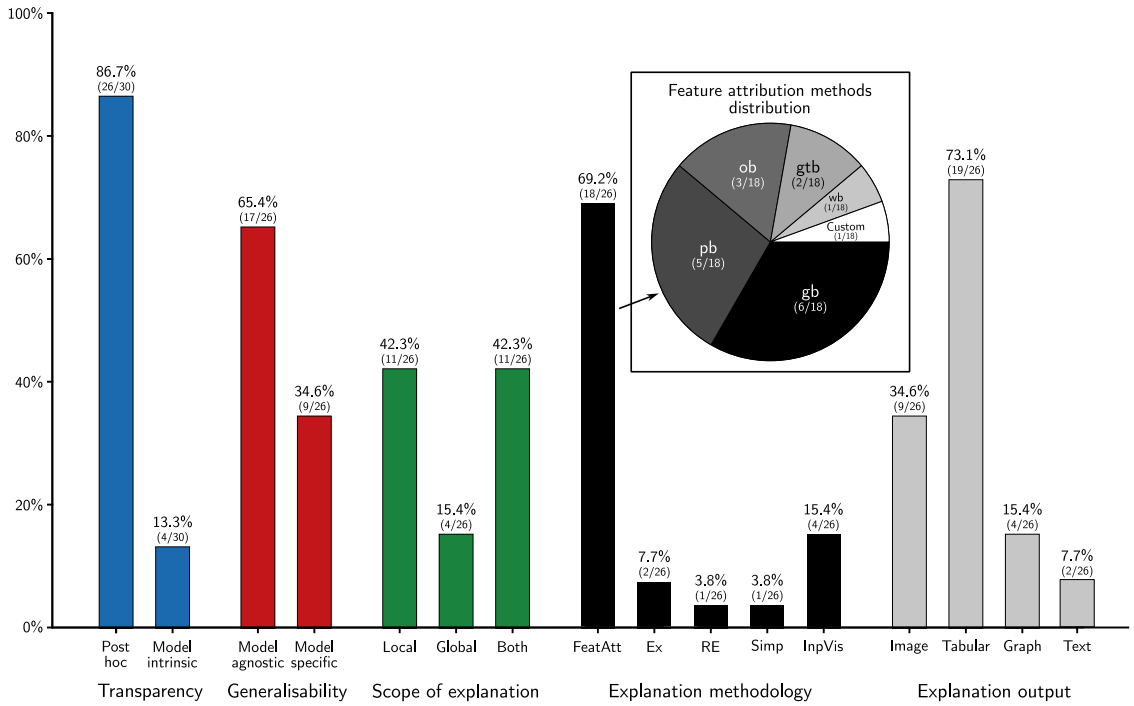


Fig. 9. Graphical analysis of the XAI methods and their classification according to the taxonomy provided in Section 2.2 and the information from Tables 4 and 5 (note that Ex = example generation, RE = rule extraction, cluster = clustering, Simp = explanation by simplification, InpVis = input visualisation, FeatAtt = feature attribution). Furthermore, the generalisability, scope of explanation, explanation methodology, and output analyses only apply to post-hoc methods.

- **Preference for model-agnostic methods.** 65.4% of the individual methods identified are model-agnostic approaches, with the remaining 34.6% being model-specific. The larger versatility of model-agnostic methods, as they can be applied to multiple groups and families of ML methods, may justify their preferred implementation. Furthermore, within model-specific approaches, most of them target DL classifiers and specific families, such as CNNs.
- **Dominance of feature attribution methods.** As depicted in Fig. 9, from the 26 different XAI approaches identified in the reviewed papers, 18 (69.2%) rely on feature attribution, most of them conducting it through gradient-based and perturbation-based methodologies. Once again, this is indicative of a clear interest in the literature in leveraging XAI for input–output correlation analyses. Other approaches, such as rule extraction and input visualisation, are much less common.
- **Tabular outputs are the most common.** For all post-hoc methods, 73.1% often produce tabular outputs (including feature relevance scores, counterfactual sample examples, etc.). Image and graphical outputs, which according to the analysis in Section 4.2 are the most explored within the literature to output XAI results, are only commonly output by 34.6% and 15.4% methods, respectively. Text-based explanations are rare, with only two methods commonly offering this output.

4.4. XAI output presentation and interpretation tools

Having analysed the multiple XAI methods employed by different authors, this section focuses on how XAI outputs are conveyed to the end-users. Throughout the reviewed papers, there were two main groups of outputs identified, graphical and numerical. Graphical outputs are employed in the form of several types of illustrations and in various formats (images, graphics, text, etc.), while numerical outputs serve as quantitative indicators, often used as metrics for direct comparison of different aspects of ML explainability.

Several graphical tools were identified within the reviewed publications. Due to their large variety, similarly to the scope and application analysis conducted in Section 4.2, these were classified according to a set of criteria, as outlined in Table 6. The classifications include:

- **Visualisation type (VT):** refers to the method of conveying explanations to the users. Heatmaps (Heat) highlight areas of interest using colour schemes. Clustering (Cluster) plots represent groups of samples, often in 2D space, valuable for unsupervised learning scenarios. Input–output (InOut) visualisations focus on relationships between inputs and outputs, being mostly independent from an XAI method. And feature relevance (Rel) visualisations display rankings or scores for input variables.
- **Scope of interpretability (SI):** similar to the XAI methods classification, this categorisation groups graphical interpretability tools according to their ability to either provide local (L), individual sample analysis, or global (G) analysis for entire datasets. It is important to mention that certain methods are capable of both.
- **Scope of application (SA):** once again related to the XAI post-hoc approaches "generalisability" criterion, it distinguishes between model-specific (MS) outputs, tailored to specific XAI methods and outputs, and model-independent (MI) outputs, broadly applicable.
- **Interpretability level (IL):** more focused on the implementation side, this category splits approaches which are easily graspable and understandable by any user - user-friendly (UF) - and those which require expert knowledge for one to achieve interpretability - expert-oriented (EO).
- **End-user purpose (EU):** similar to the previous category in terms of practical applicability, this classification focuses on the end-users, classifying graphical XAI explainability tools based on their practical use: for real-time monitoring and decision-making (Mon); for model validation and debugging (Val); or for root-cause analysis and diagnosis (Root).

According to these criteria defined in Tables 6, Table 7 summarises all graphical explainability tools identified in the process of reviewing the publications, with a graphical breakdown presented in Fig. 10.

Table 6
List of separation criteria defined for graphical XAI output representations, alongside a brief description.

Indicator	Criterium	Brief description
Visualisation Type (VT)		
VT_{Heat}	Heatmap	Colour schemes used to express XAI outputs (feature relevance, reconstruction errors, etc.)
$VT_{Cluster}$	Clustering	Sample clusters/groups representation
VT_{InOut}	Input-Output visualisations	Express data behaviour over time, independent of an XAI method
VT_{Rel}	Feature relevance visualisations	Quantify and rank feature importance through graphs
Scope of interpretability (SI)		
SI_{Loc}	Local	Focused on individual samples
SI_{Glob}	Global	Focused on groups of samples and overall datasets
Scope of Application (SA)		
SA_{MS}	Model-specific	Specific to a particular XAI technique
SA_{MI}	Model-independent	Does not rely on a specifc technique and/or model
Interpretability Level (IL)		
IL_{UF}	User-friendly	Easily understandable by any user
IL_{EO}	Expert-oriented	May require domain/expert knowledge to be well interpreted
End-user purpose (EU)		
EU_{Mon}	Monitoring and decision-making	Enable easy and real-time analysis, by operators
EU_{Val}	Model validation and debugging	Useful for model development and optimisation, by data scientists and model developers
EU_{Root}	Root-cause analysis and diagnosis	Enable post-processing and gathering process knowledge, by process engineers

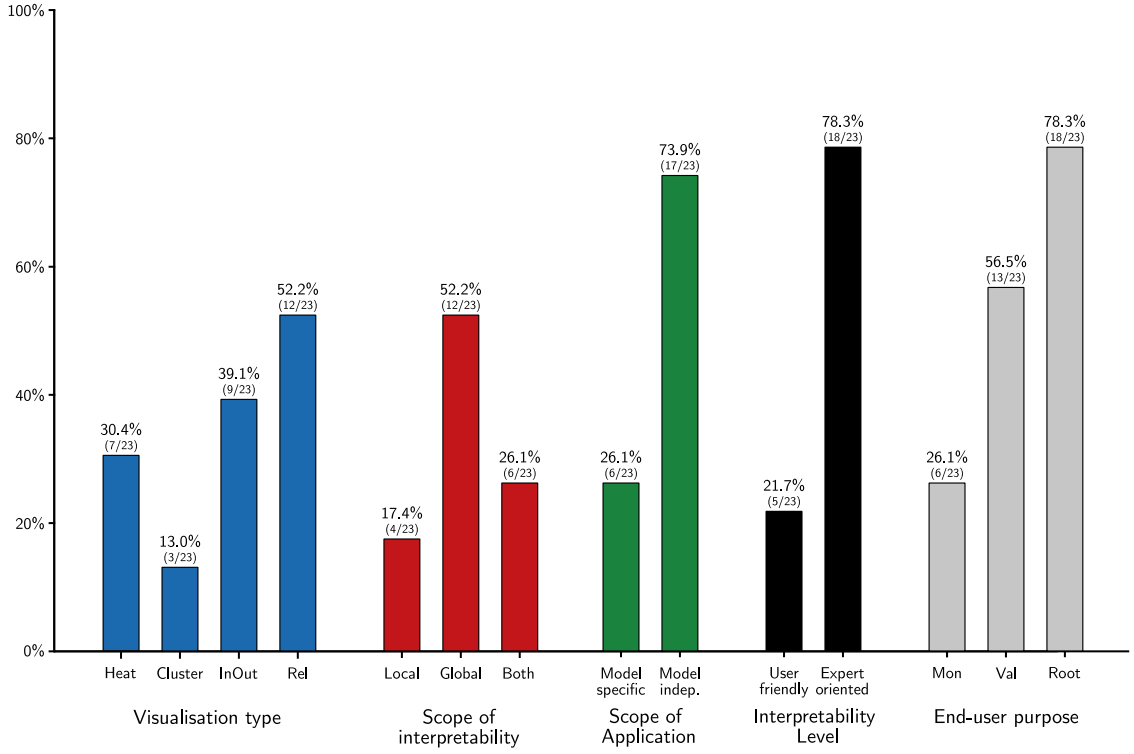


Fig. 10. Graphical analysis of all graphical explanation tools identified in the reviewed publications, according to the criteria defined in Table 6 and results exposed in Table 7.

Among the 23 graphical tools identified, the most frequently used were 1D/2D input heatmap representations, and class clustering plots, each being employed in 6 occasions. As it can be seen in Table 7, heatmaps are particularly versatile. They combine feature relevance rankings with a heatmap representation, for time series (1D) or image (2D) data, easily representing input regions of interest to users and enabling direct root cause analysis. Furthermore, they are compatible with any feature relevance ranking method, making them widely applicable. In contrast, class clustering plots, typically employed in unsupervised scenarios, often require domain-expertise. They are valuable in exploratory data analysis stages, and for the identification of fault types and data groups. As for the tool which complies with the largest number of classifications, it is the LIME plots. Although specific to the XAI LIME algorithm and its Python implementation, such representations combine multiple simple explanatory components: (1) class probabilities, demonstrating the degree of certainty of a classifier, (2) top influential features, exposing locally the most important features to a specific output, and (3) corresponding feature values. Despite their simplicity, LIME plots exemplify the potential of integrating multiple explainability elements to enhance user understanding.

Further analysing the results from Table 7 and Fig. 10, regarding the visualisation type (VT) criterion, the most dominant set of graphical/visualisation tools are those related to feature relevance ranking representations, with 12 out of 23 (52.2%) enabling such analysis. This is consistent with the widespread use of feature attribution methods, notably SHAP. Direct input–output visualisation tools are also effective methodologies employed (39.1%), being versatile approaches which

Table 7
Overview of the identified graphical XAI output representations and their classification based on the criteria from Table 6.

Metric	Refs.	Brief description	Classification criteria for graphical metrics													Total marks
			VT_{Heat}	$VT_{Cluster}$	VT_{InpOut}	VT_{Rel}	SI_{Loc}	SI_{Glob}	SA_{MS}	SA_{MI}	IL_{UF}	IL_{EO}	EU_{Mon}	EU_{Val}	EU_{Root}	
1D input heatmap	[82], [78], [79], [80], [76,81]	Represent specific regions of interest of the input in 1D data (time series)	✓	–	–	✓	✓	–	–	✓	✓	–	✓	–	✓	7
2D input heatmap	[93], [95], [96], [90], [91], [97]	Represent specific regions of interest of the input in 2D data (images)	✓	–	–	✓	✓	–	–	✓	✓	–	✓	–	✓	7
Class cluster plots	[82], [103], [83], [95], [101], [97]	Represents input data divided in 2D/3D clusters	–	✓	–	–	–	✓	–	✓	–	✓	–	✓	✓	6
Violin plot	[72,83], [73], [77], [102]	Combination of feature relevance and value representation	–	–	–	✓	✓	✓	–	✓	–	✓	–	✓	✓	7
Force plot	[72], [86], [102]	Local feature relevance representation of top features	–	–	–	✓	✓	–	–	✓	✓	–	✓	–	–	5
Classifier arhictecture representation	[88,93], [95]	Direct representation of the model's architecture	–	–	–	–	–	–	✓	–	–	✓	–	✓	✓	4
Summary plot	[86], [77], [100]	Accumulated bar plot representation of feature relevancies per output class	–	–	–	✓	✓	✓	–	✓	–	✓	–	✓	✓	7
Waterfall plot	[72], [83]	Local feature relevance representation of top features	–	–	–	✓	✓	–	–	✓	✓	–	✓	–	–	5
Time series graph analysis	[85], [72]	Direct time series comparison of normal and anomalous class sequences	–	–	✓	–	✓	✓	–	✓	–	✓	–	–	✓	6
LIME graph[a]	[95], [100]	Specific to the LIME method, represents class probabilities, top influential features, and features' values	–	–	✓	✓	✓	✓	✓	–	✓	–	✓	–	✓	8
Relevance score-based heatmap	[83], [89]	Heatmap showcasing relevance score evolution for each data sample	✓	–	–	✓	–	✓	–	✓	–	✓	–	✓	✓	7
Control chart	[88]	Direct input analysis of the evolution of an anomaly score per sample	–	–	✓	–	–	✓	–	✓	–	✓	–	✓	✓	6

(*continued on next page*)

Table 7 (*continued*).

Metric	Refs.	Brief description	Classification criteria for graphical metrics													Total marks
			VT_{Heat}	$VT_{Cluster}$	VT_{InpOut}	VT_{Rel}	SI_{Loc}	SI_{Glob}	SA_{MS}	SA_{MI}	IL_{UF}	IL_{EO}	EU_{Mon}	EU_{Val}	EU_{Root}	
Contribution plot	[88]	Input feature representation of anomalous variables per sample	–	–	✓	–	✓	✓	–	✓	–	✓	✓	–	✓	7
Voronoi diagram	[94]	Data cluster representation with cluster frontier definition	–	✓	–	–	–	✓	–	✓	–	✓	–	✓	✓	6
Reconstruction error visualisation	[101]	Per sample/time-step representation of an ML model's reconstruction error	–	–	✓	–	✓	✓	–	✓	–	✓	–	✓	–	6
Variable contribution heatmap	[85]	Average relevance/-contribution score of each input to each output class	✓	–	–	✓	–	✓	–	✓	–	✓	–	✓	✓	7
Top variables correlation matrix	[104]	Correlation matrix representation for top influential features, assessing their independence	✓	–	–	✓	–	✓	–	✓	–	✓	–	✓	–	6
VEC graph	[92]	VEC = Variable Effect Characteristic, direct individual feature input–output analysis	–	–	✓	–	–	✓	–	✓	–	✓	–	✓	–	5
SOM map	[82]	SOM = Self Organising Maps, 2D clustering + heatmap representation of each class samples	✓	✓	–	–	–	✓	✓	–	–	✓	–	–	✓	6
Time-series feature relevance	[87]	Time-series representation of each feature's relevance across sequences	–	–	–	✓	–	✓	–	✓	–	✓	–	✓	✓	6
ALE plot	[100]	Specific to the ALE method, represents ALE score across the feature's range of values	–	–	✓	✓	–	✓	✓	–	–	✓	–	✓	✓	7
PDP plot	[102]	Specific to the PDP method, represents target output response across an input feature's range of values	–	–	✓	–	–	✓	✓	–	–	✓	–	–	✓	5
Change of correlation map	[104]	Specific to [104], representing the correlation between attention maps of input and reconstructed data	✓	–	✓	–	–	✓	✓	–	–	✓	–	–	✓	6

[a] LIME graphs are specific to the 'lime' Python package (https://lime-ml.readthedocs.io/en/latest/).

may induce large benefits regarding gathering process knowledge. And finally, the least explored tools appear to be clustering methods (13%), methods that may require additional knowledge over a process, and therefore not as compatible with real-time production analysis.

In terms of scope of interpretability (SI), global-scoped explainability tools were more prevalent, represented by 12 out of the 23 (52.2%) identified graphical tools, whereas only 4 (17.4%) were limited to offering local insights. From the remaining 7 approaches, 6 are versatile enough to enable both local and global analyses, with the remaining one (classifier architecture representation) being a special case more focused on inferring the model's decision-making system. Overall, this dominance of global approaches may reflect a focus on holistic model evaluation over instance-level decision analysis. Moreover, this trend may be related to the last criterion analysed, the end-user purpose (EU). Explainability tools which induce local explainability can be considered much more useful for quick and direct analysis of specific behaviours of a model towards a particular decision. As a result, they are mostly related to directly exposing the model's decision to any user. Conversely, employing a global approach, focused on groups of samples and entire collected datasets, may not be directly representative of a particular decision. Instead, it enables a more complex and deeper analysis to the overall behaviour of a particular classifier. Therefore, such tools are useful for model development and optimisation stages (*e.g.*, for dimensionality reduction and performance optimisation tasks), and for post-processing analyses of the entire process behaviour, conducted by domain-experts. As a result, it can be seen that for the EU criterion, only 6 graphical indicators (26.1%) are more focused on direct analysis of results, while 13 tools (56.5%) focus on providing insights valuable for model development and optimisation, and 18 tools (78.3%) induce post-processing analyses, the latter two requiring therefore a deeper understanding of the processes.

The scope of application (SA) classification showcases the versatility of the various graphical XAI output representations employed. 17 out of 23 (73.9%) were labelled as model independent, with the remaining 6 (26.1%) being model-specific, meaning they are limited to the implementation of a particular XAI technique. Having model-independent tools, including input heatmaps and clustering plots, and feature relevance-based tools such as summary and waterfall plots, indicates that those types of interpretability can be easily achieved independent of the XAI methodology employed.

Finally, in terms of the interpretability level (IL), only 5 out of 23 (21.7%) graphical explanation representations were actually classified as user-friendly, *i.e.*, easily understandable by any user. The remaining 18 (78.3%) were considered to be more difficult to interpret, requiring some degree of expertise and knowledge over a modelled process. Such results are compatible with other criteria already discussed, mainly the scope of interpretability and end-user purpose. User-friendly tools are most beneficial for real-time analyses, where one needs to quickly understand the output of the ML classifier and the decision made to eventually directly act upon a specific decision. As a result, the lower number of local analysis tools is aligned with the lower amount of user-friendly methods. Conversely, the focus of the literature regarding global tools, requiring some degree of domain-knowledge, indicates a trend towards more complex analytical tools aimed at developers and process engineers rather than operators.

In contrast to graphical representations, numerical XAI output indicators were less frequently explored. Perhaps due to their larger complexity, harder interpretability, and higher specificity, there is a lower number of authors focused on developing and employing such indicators. Table 8 describes all relevant numerical XAI output indicators and metrics identified in the reviewed publications. It is important to point out that although feature relevance rankings are numerical indicators explored in a large number of works, the authors opted to leave them out of the analysis and focus on those which were less common.

As already mentioned, it is evident that numerical XAI outputs are much less explored in the literature, with only 9 out of the 33 reviewed publications (27.3%) developing and exploring such tools to improve user interpretability. In contrast to graphical tools, numerical outputs were often tailored to specific XAI techniques or problem types, including rule extraction [103] and counterfactual sample generation [75,86], to a particular method or groups of methods, such as SVM [97] and tree-based classifiers [83], or even to particular types of input data and problems, such as those involving frequency-domain signals [76]. Moreover, in most cases, these numerical indicators are mostly seen as performance metrics, used to assess and compare the performance of a particular XAI approach in a quantitative manner.

This specificity of numerical metrics may be double-edged. On the one hand, it limits their applicability, particularly concerning end-users. Compared to some graphical tools, such as input relevance heatmaps which can directly provide information to any end-user regarding model behaviour and its decision, most numerical metrics require a more thorough analysis. They appear most beneficial for model developers, providing several important insights that reflect an ML and/or XAI model's performance, or to any process engineer aiming to gather useful knowledge about a modelled process. Examples of such metrics include the correct isolation rate [88] for a classifier (useful to assess model effectiveness), the comprehensibility, representativeness, stability, and diversity metrics [103] focused on rule extraction evaluation and assessment, or the SVM hyperplane distance [97] which reflects class separation. Possible numerical indicators that could be more easily interpreted include counterfactual and "normalised" samples [80,86], directly exposing required changes on a sample to the user. On the other hand, the implementation of such numerical indicators may provide an additional degree of explainability to the overall ML and XAI framework that graphical tools are unable to achieve and transmit. Especially regarding the model building and optimisation process, having the ability to, for instance, analyse the complexity of extracted rules by a rule extraction method can help to develop simpler approaches and consequently improve those rules and how they are conveyed to users. Additionally, it also allows one to assess, prior to the implementation step, the impacts of a particular XAI methodology on the user.

Overall, numerical indicators are often more challenging to define and, in some instances, more difficult to interpret than graphical tools. As a result, they remain less explored in the literature. However, when effectively developed and used alongside graphical tools, numerical metrics can significantly enhance interpretability and provide deeper insights into model behaviour.

Key takeaways

- **Graphical XAI explanation tools dominate XAI outputs.** Numerous graphical tools are explored within the literature, including input heatmaps, cluster plots, and variations of feature relevance scoring graphs. Such tools offer more intuitive, versatile representations which support local and global interpretability, with many being independent of the XAI method chosen.
- **Strong focus on global insights and model optimisation.** Most graphical tools (78.3%) are used for post-processing analysis and model validation, offering high-level insights into model behaviour and decision-making. Conversely, tools designed for operator-level real time monitoring remain relatively underexplored.
- **Lower applicability of numerical indicators.** Only 27.3% of reviewed publications employed and explored numerical indicators. These metrics are often complex and tailored to specific applications, limiting their generalisability. Nonetheless, they provide useful information for the model and explanation development stages.

Table 8
Specific numerical metrics employed by some authors (feature relevance scores are excluded given they are employed in many situations).

Ref.	Metrics	Brief description on their usage
Barbado et al. [103]	Comprehensibility	How well can a human understand the explanation (examples include number of rules or size of rules)
	Representativeness	Number of instances covered by a specific rule
	Stability	Representative of whether the rules actually approximate the underlying model
	Diversity	Degree of hyperspace overlapping the rules
	Weighted final metric	Weighted metric combining the previous four, to provide an overall comparison metric
Harinarayan and Shalinie [86]	Counterfactual samples	Direct analysis of counterfactual explanations, and specific input variable changes, compared to the original samples
Yang et al. [88]	Correct isolation rate	Percentage of observations where the faulty variables are successfully isolated
Jang et al. [89]	Adjusted mutual information score	Measurement of the agreement between clustering results and ground truth (specific to clustering methods)
Kumar et al. [94]	Mahalanobis distance values	Measurement of the multivariate separation of a point from a distribution. Large values indicate anomalous behaviour
Kim et al. [75]	Feature frequencies on tree nodes	Specific to a tree-based method, consisting on analysing tree node rules and the frequency each input feature appears on those
	Feature differences compared to node conditions	Specific to a tree-based method, consisting on analysing the feature values for output conditions and comparing them to the respective input features' values
	"Normalised" samples	Similar to Harinarayan and Shalinie, authors presented artificial samples that showcased necessary feature value alterations to change the model output to normal
Kakavandi et al. [80]	Infidelity	Expected difference between the dot product of the input perturbation to the explanation and the output perturbation
	Sensitivity	Extent of explanation change by insignificant perturbations from the test point
	Cosine similarity score	Measurement of similarity between two vectors (in the specific case, measuring the similarity between a method's relevance score attribution and one from a domain expert)
Meister et al. [97]	SVM hyperplane distance	Specific to the SVM classifier, used to separate data groups and identify different data classes
Li et al. [76]	Relative amplitude to maximum amplitude (RATM)	Specific to frequency-domain signals, used to describe the ability of a network to extract characteristic frequency features
	Relative amplitude to average amplitude (RATA)	Specific to frequency-domain signals, used to describe the eminence of the characteristic frequency feature that the network has learned
	Comprehensive explainable indicator (CEI)	Specific to frequency-domain signals, represents the difference between RATA and RATM

- **Combining graphical and numerical indicators can enhance interpretability.** Leveraging the intuitive understanding of visual tools with the deeper, more complex analytical power of numerical indicators can lead to improved explanations, bridging the gap between user-friendly decision support at operator-level and in-depth analysis at both development and post-processing analysis stages.

5. General discussion

Building upon the comprehensive review of existing literature concerning XAI in the scope of industrial fault detection and diagnosis, this section outlines and discusses the main findings of the review, compares them with prior studies, and also addresses key challenges and possible directions for future research.

Although practical implementations of XAI in industrial fault detection and diagnosis remain relatively limited, this review revealed its exploration across several critical industrial domains. These included equipment monitoring, process control, product inspection, and cybersecurity. Moreover, while the manufacturing sector remains the primary area of focus, other fields such as chemical process industry, automotive engineering, and aerospace industry, are also beginning to benefit from XAI applications. Implementation-wise, this review confirms a primary trend highlighted in other works, the domination of feature attribution methods [57,107–109]. Of the 33 reviewed publications, 31 (93.9%) employed at least one input feature ranking method. SHAP was the most widely used method (13 times), followed by GradCAM (5 times). These findings align with the literature: SHAP is constantly regarded as the leading technique for both local and global interpretability [34,35], while the increasing popularity of CNNs for image data processing [110–112], particularly in product defect detection and classification in the scope of this review, favours the implementation of GradCAM as the preferred explainability method [57,111,112].

Across the review, a total of 30 distinct XAI methods were identified. Among these, 26 methods (86.7%) are post-hoc approaches, with only 4 offering inherent interpretability. This dominance may be justified by the simplicity of such methods and their wider availability in the literature [42,50,69]. However, post-hoc methods - especially those focused on feature attribution - only mainly offer input–output correlation analysis, often failing to reveal and explain the underlying decision-making mechanisms of the models [106,113]. Within these post-hoc methods, model-agnostic approaches were the most explored, in 17 out of 26 (65.4%) occasions, possibly due to their flexibility and applicability across different ML models and architectures.

Given the inherently human-centric goals of XAI, the final part of the review focused on assessing how different authors are conveying XAI outputs to users. Two primary explanation/output modalities were identified, graphical and numerical. Graphical tools are predominant, and feature in 28 studies (84.8%) while spanning across 23 different visualisation methods. This diversity reflects the adaptability of XAI to various end-user roles - from real-time decision-making support for

operators with local, instance-based analyses, model validation and optimisation for data scientists, to root-cause analysis for engineering and management teams. Such flexibility supports the development of multimodal explanation frameworks, hybrid methodologies combining two or more methods and XAI outputs, and capable of meeting the interpretability needs of different stakeholders. Notably, Lee et al. [93] demonstrated through domain-expert surveys that combining multiple XAI outputs enhanced user trust and satisfaction. In contrast, numerical explanation tools are much less explored, appearing in only 9 of the reviewed papers. Their limited adoption may be due to their larger complexity and specificity to particular applications, hindering interpretability for non-expert users. Nevertheless, they may have important roles for model development, offering quantitative tools to compare and benchmark different XAI methods.

Overall, this review confirms several trends reported in the general XAI literature, namely the dominance of post-hoc feature attribution methods, the widespread use of SHAP, and the limited exploration of inherently interpretable approaches. Such findings also enable the identification of both challenges and opportunities for XAI applied research to industrial fault detection and diagnosis, including the need for hybrid approaches which address varying interpretability needs across different end-user profiles. The remainder of the section outlines some key challenges identified by the authors and proposes directions of future research along those lines.

Performance vs. Interpretability trade-off

Despite the growing application of XAI techniques in industrial environments, the trade-off between performance and interpretability remains a real challenge. As discussed in this review, most XAI implementations adopt post-hoc approaches that offer zero-order explanations. These methods typically focus on demonstrating correlations between input features and outputs or presenting feature relevance analyses. While providing a certain level of interpretability, they fail to offer transparent insights into the decision-making rationale of complex models, particularly DL architectures.

This superior performance of DL approaches often comes at the expense of interpretability. As these methods are increasingly explored in the literature for critical industry-related applications, this lack of transparency may undermine user trust. As a result, future research should prioritise the development of explainability strategies which offer clear, human-understandable insights and accounts of DL models' decision-making processes. Without such effort, this performance-interpretability trade-off will most likely persist, and limit the practical viability and implementation of advanced AI systems in safety-critical industrial domains.

Transparent models and model-intrinsic explainability

Continuing upon the previous point, one of the most promising strategies to address the interpretability gap is the development and adoption of transparent methodologies with model-intrinsic explainability, with such techniques aiming to embed interpretability directly into the model architecture and outputs, moving beyond the limitations of post-hoc explanations. Current intrinsically explainable models include rule-based and decision tree-based systems, as well as probabilistic graphical models such as Bayesian Networks. These models inherently provide a level of interpretability through rule extraction, or by explicit representation of dependencies between variables. However, their applicability in high-dimensional complex industrial scenarios is limited by their low scalability and computational complexity. As the number of features and relationships increases, maintaining transparency without compromising performance becomes challenging.

Future research could explore hybrid approaches that combine the strengths of intrinsically interpretable models and post-hoc explanation methods. By doing so, it may be possible to deliver more comprehensive explanations, with such strategy helping to balance the transparency, scalability, and performance demands.

Integrating explainability at different stages of an AI-based framework

Another possible option for enhancing explainability in industrial AI systems is to integrate interpretability mechanisms throughout the entire AI framework, from data pre-processing tasks to model development, optimisation, and deployment.

At the data pre-processing stage, techniques such as data normalisation to reduce model bias, input data visualisation for identifying specific data trends, clustering approaches for unsupervised learning scenarios, and dimensionality reduction for feature simplification can contribute to reduced problem and model complexities. Several works in this study have demonstrated the utility of XAI methods in guiding these tasks. For instance, relevance-based feature selection has been explored to simplify complex datasets and optimise model structures [72,85,99]. During model training and validation, incorporating explainability mechanisms - attention mechanisms, feature relevance tracking, and interpretable loss functions - can aid in identifying model biases, overfitting, and other undesirable behaviours. Notably, the work of Lee et al. [93] illustrated how LRP can be integrated within the model building and fitting process, iteratively eliminating irrelevant variables. In the deployment stages, post-hoc explanation methods can complement intrinsic interpretability, offering instance-specific insights and real-time decision support. These explanations may take the form of feature importance scores, counterfactual examples, or rule-based summaries tailored to different stakeholder needs and interests.

Exploring multimodal explanations

When adopting XAI methods, it is insufficient to rely solely on feature attribution techniques that provide input feature rankings, or on input visualisation methods that illustrate direct input–output relationships. While these approaches provide important insights, they offer limited perspectives on the broader decision-making process.

On this note, future research should explore multimodal explanation strategies that combine complementary XAI techniques to deliver richer, more comprehensive explanations. Such approaches could integrate global explanations - offering insights into the model's and problem's overall behaviour and trends - with local explanations providing instance-level, particular interpretations. For example, tabular feature relevance outputs from a SHAP or LIME method could be combined with visual heatmaps, clustering representations, or counterfactual scenarios that offer multiple interpretability modalities tailored to different stakeholders. Production operators benefit from simple and direct visual explanations, such as highlighting anomalies in real time, while process engineers and data scientists may benefit from complete reports that guide model refinement and system/process optimisation. This multimodal approach not only enhances interpretability, but also supports different user roles and expertise levels, fostering collaboration and trust across the human-AI interface.

Conveying XAI outputs

A significant challenge identified throughout the review is the lack of standardised frameworks for evaluating and conveying XAI outputs to users. Much of existing work focuses on theoretical and philosophical principles of explainability, overlooking the practical utility of explanations to end-users. To bridge this gap, future research should focus on developing frameworks that incorporate both qualitative and quantitative indicators. On the one hand, graphical tools such as input relevance heatmaps, clustering visualisation, and feature relevance-based plots have proven effective in translating interpretability to users. On the other hand, numerical indicators have proven effective particularly for model development phases, providing objective benchmarks for comparing different XAI methods and the effectiveness of explanations.

Furthermore, the inclusion of domain-expert and worker feedback is also critical for validating explanations in industrial settings, given they are the ones which will be most in contact with the AI systems. From the reviewed publications, only two studies [80,93] have successfully attempted to incorporate human feedback in their frameworks,

ensuring XAI outputs aligned with their knowledge and operational requirements. As a result, future research should also focus on prioritising the development of user-centric evaluation platforms. Such platforms should support multimodal explanations, as previously discussed, and enable different stakeholders to customise and influence the level and type of information they receive, according to their requirements and necessities.

Regulatory alignment

Ensuring that AI-based systems comply with emerging regulatory frameworks, such as the EU AI Act [45], represents a pressing challenge for industrial AI-based operations and systems. The most recent legislation imposes strict requirements for transparency, robustness, and human oversight and control, particularly in high-risk AI systems such as those in industrial fault detection and diagnosis systems.

As evidenced in this review, most current XAI implementations rely on post-hoc methods that offer limited insight into the internal logic of ML models. As a result, these approaches may fall short of the auditability and accountability standards set by regulatory bodies. The inability to provide comprehensive, traceable explanations of AI system decisions undermines efforts to ensure legal compliance and build stakeholder trust.

Future research should focus on developing XAI frameworks that inherently support regulatory compliance by design. This includes promoting transparency in model architectures, enabling traceable decision-making processes, and providing verifiable audit trails of AI system behaviour. A move towards inherently interpretable models, combined with robust explanation mechanisms, is essential to meeting the evolving regulatory landscape and ensuring the ethical deployment of AI in industrial settings.

6. Conclusion

This paper has presented a comprehensive systematic literature review on the role of XAI in industrial fault detection and diagnosis, an area of growing significance as industries transition towards human-centric and trustworthy AI systems under the Industry 5.0 paradigm. With the increasing reliance on ML and DL models to automate and optimise critical industrial tasks, the need for transparent, interpretable, and auditable AI-based systems is critical, particularly in light of emerging regulatory frameworks such as the EU AI Act.

Within the scope of XAI applications for industrial fault detection and diagnosis, this review has examined in detail 33 high-quality publications. In addition, it proposed a taxonomy for classifying XAI methods, with a focus on their industrial applicability. The findings reveal that XAI is actively being explored and successfully implemented across several relevant industrial domains, ranging from equipment monitoring and process control to product inspection and cybersecurity. In terms of the implementations, the literature shows a clear dominance of feature attribution methods, with 31 out of 33 (93.9%) works employing some form of input feature relevance scoring. Among these, SHAP (with 13 implementations) and GradCAM (with 5 implementations) emerge as the most widely adopted techniques, both being feature attribution methods. Post-hoc explainability remains the most popular approach, accounting for 26 out of the 30 distinct identified XAI methods (86.7%), while only 4 (13.3%) explored intrinsic explainability. In terms of how explainability is conveyed to the end-users, the literature prioritises graphical tools, with 28 works (84.8%) incorporating at least one form of visual explanation, most often input heatmaps (feature relevance based) and class clustering plots.

While these advancements mark important progress, key challenges still persist in the literature. These include the enduring trade-off between model performance and interpretability, the scarcity of intrinsically explainable models suited for complex industrial environments, and the absence of standardised frameworks for evaluating XAI outputs. Moreover, the inclusion of human feedback into the design and validation of interpretable AI-based systems still remains underexplored. This is despite its importance in ensuring that explanations are actionable, comprehensible, and aligned with the diverse needs of stakeholders ranging from data scientists to production-floor operators.

Despite the structured and comprehensive nature of the review, some limitations should also be acknowledged. First, the scope of the review may have been constrained by the filtering criteria applied. Mainly, the minimum citation threshold may have excluded some relevant or emergent works, especially given the novelty of the research area. Second, while the review provided a systematic classification of XAI methods used for industrial fault detection and diagnosis, through a proposed taxonomy based on the literature, it ended up offering limited insight into the effectiveness and comparative performance of XAI methods. This limitation is primarily due to the lack of evaluation frameworks and benchmark metrics in the reviewed publications. Finally, perspectives and opinions from end-users regarding XAI outputs and how they are conveyed to them were also largely missing in the reviewed literature, restricting the ability to assess how explanations are actually interpreted, trusted, or acted upon in real industrial settings.

Future research should prioritise the practical implementation and implications of XAI within industrial settings. This includes the development of hybrid XAI frameworks that offer multiple layers of explanation and integrate these across the full AI lifecycle, from data pre-processing and model development to validation, deployment, and monitoring stages. The exploration of multimodal explanation strategies and the design of user-centric evaluation and visualisation platforms is essential for bridging the current gap between theoretical advancements and practical deployment. In parallel, ensuring regulatory compliance will require that AI systems provide comprehensive, trustworthy, and auditable explanations that meet the demands of increasingly stringent legislative frameworks.

In conclusion, XAI holds the potential to promote trust, transparency, and accountability in AI-driven industrial fault detection and diagnosis systems, and is most certainly a key area of research for future AI development and deployment in industrial scenarios. However, significant efforts are still required to transform XAI from a promising research topic into a reliable solution capable of supporting critical industrial applications in alignment with Industry 5.0 principles.

CRediT authorship contribution statement

J. Cação: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **J. Santos:** Writing – review & editing, Supervision, Funding acquisition. **M. Antunes:** Writing – review & editing, Supervision, Conceptualization.

Funding

This work is supported by the project UID/00481 – Centro de Tecnologia Mecânica e Automação (TEMA) – Fundação para a Ciência e a Tecnologia and by PRR - Plano de Recuperação e Resiliência under the Next Generation EU from the European Union, Project "Agenda ILLIANCE" [C644919832-00000035 | Project no. 46].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jii.2025.100905.

Data availability

No data was used for the research described in the article.

References

[1] 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, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[2] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy artificial intelligence: A review, ACM Comput. Surv. 55 (2) (2022) 1–38, http://dx.doi.org/10.1145/3491209.

[3] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (3) (2021) http://dx.doi.org/10.1007/s42979-021-00592-x.

[4] J.F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, A. Troncoso, Deep learning for time series forecasting: A survey, Big Data 9 (1) (2021) 3–21, http://dx.doi.org/10.1089/big.2020.0159.

[5] A. Dogan, D. Birant, Machine learning and data mining in manufacturing, Expert Syst. Appl. 166 (2021) 114060, http://dx.doi.org/10.1016/j.eswa.2020.114060.

[6] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, Cancer Lett. 471 (2020) 61–71, http://dx.doi.org/10.1016/j.canlet.2019.12.007.

[7] Y. Lu, C. Liu, K.I.-K. Wang, H. Huang, X. Xu, Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues, Robot. Comput.-Integr. Manuf. 61 (2020) 101837, http://dx.doi.org/10.1016/j.rcim.2019.101837.

[8] Y. Cui, S. Kara, K.C. Chan, Manufacturing big data ecosystem: A systematic literature review, Robot. Comput.-Integr. Manuf. 62 (2020) 101861, http://dx.doi.org/10.1016/j.rcim.2019.101861.

[9] Z. Jan, F. Ahamed, W. Mayer, N. Patel, G. Grossmann, M. Stumptner, A. Kuusk, Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities, Expert Syst. Appl. 216 (2023) 119456, http://dx.doi.org/10.1016/j.eswa.2022.119456.

[10] M. Antunes, A.R. Santiago, S. Manso, D. Regateiro, J.P. Barraca, D. Gomes, R.L. Aguiar, Building an IoT platform based on service containerisation, Sensors 21 (19) (2021) 6688, http://dx.doi.org/10.3390/s21196688.

[11] Z. Jiang, Y. Guo, Z. Wang, Digital twin to improve the virtual-real integration of industrial IoT, J. Ind. Inf. Integr. 22 (2021) 100196, http://dx.doi.org/10.1016/j.jii.2020.100196.

[12] M. Younan, E.H. Houssein, M. Elhoseny, A.A. Ali, Challenges and recommended technologies for the industrial internet of things: A comprehensive review, Measurement 151 (2020) 107198, http://dx.doi.org/10.1016/j.measurement.2019.107198.

[13] J. Cação, M. Antunes, J. Santos, D. Gomes, Intelligent assistant for smart factory power management, Procedia Comput. Sci. 232 (2024) 966–979, http://dx.doi.org/10.1016/j.procs.2024.01.096.

[14] P. Pravin, J.Z.M. Tan, K.S. Yap, Z. Wu, Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems, Digit. Chem. Eng. 4 (2022) 100047, http://dx.doi.org/10.1016/j.dche.2022.100047.

[15] D. Küpper, K. Triantafyllidis, Harnessing the AI Revolution in Industrial Operations: A Guidebook, World Economic Forum, 2023.

[16] R. Rai, M.K. Tiwari, D. Ivanov, A. Dolgui, Machine learning in manufacturing and industry 4.0 applications, Int. J. Prod. Res. 59 (16) (2021) 4773–4778, http://dx.doi.org/10.1080/00207543.2021.1956675.

[17] C. Grosan, A. Abraham, Rule-based expert systems, in: Intelligent Systems: A Moderns Approach, Springer Berlin Heidelberg, 2011, pp. 149–185, http://dx.doi.org/10.1007/978-3-642-21004-4_7.

[18] W. Yan, J. Wang, S. Lu, M. Zhou, X. Peng, A review of real-time fault diagnosis methods for industrial smart manufacturing, Processes 11 (2) (2023) 369, http://dx.doi.org/10.3390/pr11020369.

[19] Boston Consulting Group, AI-powered industrial operations: AI survey results, 2023, https://www.bcg.com/about/partner-ecosystem/world-economic-forum/ai-project-survey.

[20] A. Angelopoulos, E.T. Michailidis, N. Nomikos, P. Trakadas, A. Hatziefremidis, S. Voliotis, T. Zahariadis, Tackling faults in the industry 4.0 era—A survey of machine-learning solutions and key aspects, Sensors 20 (1) (2019) 109, http://dx.doi.org/10.3390/s20010109.

[21] S.R. Saufi, Z.A.B. Ahmad, M.S. Leong, M.H. Lim, Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review, IEEE Access 7 (2019) 122644–122662, http://dx.doi.org/10.1109/access.2019.2938227.

[22] D. Leite, E. Andrade, D. Rativa, A.M.A. Maciel, Fault detection and diagnosis in industry 4.0: A review on challenges and opportunities, Sensors 25 (1) (2024) 60, http://dx.doi.org/10.3390/s25010060.

[23] X. Bampoula, G. Siaterlis, N. Nikolakis, K. Alexopoulos, A deep learning model for predictive maintenance in cyber-physical production systems using LSTM autoencoders, Sensors 21 (3) (2021) 972, http://dx.doi.org/10.3390/s21030972.

[24] A. Choudhary, T. Mian, S. Fatima, Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images, Measurement 176 (2021) 109196, http://dx.doi.org/10.1016/j.measurement.2021.109196.

[25] P.F. Orrù, A. Zoccheddu, L. Sassu, C. Mattia, R. Cozza, S. Arena, Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry, Sustainability 12 (11) (2020) 4776, http://dx.doi.org/10.3390/su12114776.

[26] I. Lomov, M. Lyubimov, I. Makarov, L.E. Zhukov, Fault detection in Tennessee Eastman process with temporal deep learning models, J. Ind. Inf. Integr. 23 (2021) 100216, http://dx.doi.org/10.1016/j.jii.2021.100216.

[27] L. Feng, C. Zhao, Fault description based attribute transfer for zero-sample industrial fault diagnosis, IEEE Trans. Ind. Inform. 17 (3) (2021) 1852–1862, http://dx.doi.org/10.1109/tii.2020.2988208.

[28] S.-K.S. Fan, C.-Y. Hsu, D.-M. Tsai, F. He, C.-C. Cheng, Data-driven approach for fault detection and diagnostic in semiconductor manufacturing, IEEE Trans. Autom. Sci. Eng. 17 (4) (2020) 1925–1936, http://dx.doi.org/10.1109/tase.2020.2983061.

[29] W. Li, H. Zhang, G. Wang, G. Xiong, M. Zhao, G. Li, R. Li, Deep learning based online metallic surface defect detection method for wire and arc additive manufacturing, Robot. Comput.-Integr. Manuf. 80 (2023) 102470, http://dx.doi.org/10.1016/j.rcim.2022.102470.

[30] J.-C. Chien, M.-T. Wu, J.-D. Lee, Inspection and classification of semiconductor wafer surface defects using CNN deep learning networks, Appl. Sci. 10 (15) (2020) 5340, http://dx.doi.org/10.3390/app10155340.

[31] Y. Li, Y. Xu, Z. Liu, H. Hou, Y. Zheng, Y. Xin, Y. Zhao, L. Cui, Robust detection for network intrusion of industrial IoT based on multi-CNN fusion, Measurement 154 (2020) 107450, http://dx.doi.org/10.1016/j.measurement.2019.107450.

[32] S. Latif, Z. Zou, Z. Idrees, J. Ahmad, A novel attack detection scheme for the industrial internet of things using a lightweight random neural network, IEEE Access 8 (2020) 89337–89350, http://dx.doi.org/10.1109/access.2020.2994079.

[33] P. Ruzafa-Alcazar, P. Fernandez-Saura, E. Marmol-Campos, A. Gonzalez-Vidal, J.L. Hernandez-Ramos, J. Bernal-Bernabe, A.F. Skarmeta, Intrusion detection based on privacy-preserving federated learning for the industrial IoT, IEEE Trans. Ind. Inform. 19 (2) (2023) 1145–1154, http://dx.doi.org/10.1109/tii.2021.3126728.

[34] P.P. Angelov, E.A. Soares, R. Jiang, N.I. Arnold, P.M. Atkinson, Explainable artificial intelligence: an analytical review, WIREs Data Min. Knowl. Discov. 11 (5) (2021) http://dx.doi.org/10.1002/widm.1424.

[35] R. Machlev, L. Heistrene, M. Perl, K. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, Energy AI 9 (2022) 100169, http://dx.doi.org/10.1016/j.egyai.2022.100169.

[36] G. Pang, C. Shen, L. Cao, A.V.D. Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. 54 (2) (2021) 1–38, http://dx.doi.org/10.1145/3439950.

[37] A.-A. Tulbure, A.-A. Tulbure, E.-H. Dulf, A review on modern defect detection models using DCNNs – deep convolutional neural networks, J. Adv. Res. 35 (2022) 33–48, http://dx.doi.org/10.1016/j.jare.2021.03.015.

[38] D. Cabrera, A. Guamán, S. Zhang, M. Cerrada, R.-V. Sánchez, J. Cevallos, J. Long, C. Li, Bayesian approach and time series dimensionality reduction to LSTM-based model-building for fault diagnosis of a reciprocating compressor, Neurocomputing 380 (2020) 51–66, http://dx.doi.org/10.1016/j.neucom.2019.11.006.

[39] H. Liu, R. Ma, D. Li, L. Yan, Z. Ma, Machinery fault diagnosis based on deep learning for time series analysis and knowledge graphs, J. Signal Process. Syst. 93 (12) (2021) 1433–1455, http://dx.doi.org/10.1007/s11265-021-01718-3.

[40] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470, http://dx.doi.org/10.1016/j.media.2022.102470.

[41] X. Xu, Y. Lu, B. Vogel-Heuser, L. Wang, Industry 4.0 and industry 5.0—Inception, conception and perception, J. Manuf. Syst. 61 (2021) 530–535, http://dx.doi.org/10.1016/j.jmsy.2021.10.006.

[42] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, http://dx.doi.org/10.1016/j.inffus.2023.101805.

[43] M.C. Zizic, M. Mladineo, N. Gjeldum, L. Celent, From industry 4.0 towards industry 5.0: A review and analysis of paradigm shift for the people, organization and technology, Energies 15 (14) (2022) 5221, http://dx.doi.org/10.3390/en15145221.

[44] J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis, L. Wang, Industry 5.0: Prospect and retrospect, J. Manuf. Syst. 65 (2022) 279–295, http://dx.doi.org/10.1016/j.jmsy.2022.09.017.

[45] European Union, EU artifical intelligence act, 2024, https://artificialintelligenceact.eu/.

[46] J. Qadir, M.Q. Islam, A. Al-Fuqaha, Toward accountable human-centered AI: rationale and promising directions, J. Inf. Commun. Ethics Soc. 20 (2) (2022) 329–342, http://dx.doi.org/10.1108/jices-06-2021-0059.

[47] A. Bourgais, I. Ibnouhsein, Ethics-by-design: the next frontier of industrialization, AI Ethics 2 (2) (2021) 317–324, http://dx.doi.org/10.1007/s43681-021-00057-0.

[48] I. Ahmed, G. Jeon, F. Piccialli, From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where, IEEE Trans. Ind. Inform. 18 (8) (2022) 5031–5042, http://dx.doi.org/10.1109/tii.2022.3146552.

[49] R. Inam, A. Terra, A. Mujumdar, E. Fersman, A. Vulgarakis Feljan, Explainable AI - How Humans Can Trust AI, Ericsson, 2021.

[50] Z. Alexander, D.H. Chau, C. Saldaña, An interrogative survey of explainable AI in manufacturing, IEEE Trans. Ind. Inform. 20 (5) (2024) 7069–7081, http://dx.doi.org/10.1109/tii.2024.3361489.

[51] M. Lebacher, R. Gross, S. Hagen Weber, The rise of industrial explainable artificial intelligence - Insights across the AI life cycle, Siemens, 2024.

[52] D. Thogmartin, Bringing Transparency to Machine Learning Models & Predictions, Deloitte, 2021.

[53] M. van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence, IAAI '04, AAAI Press, 2004, pp. 900–907, http://dx.doi.org/10.5555/1597321.1597342.

[54] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.

[55] S. Larsson, F. Heintz, Transparency in artificial intelligence, Internet Policy Rev. 9 (2) (2020) http://dx.doi.org/10.14763/2020.2.1469.

[56] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, F. Cruz, Levels of explainable artificial intelligence for human-aligned conversational explanations, Artificial Intelligence 299 (2021) 103525, http://dx.doi.org/10.1016/j.artint.2021.103525.

[57] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (1) (2020) 18, http://dx.doi.org/10.3390/e23010018.

[58] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, K.-R. Muller, Explaining deep neural networks and beyond: A review of methods and applications, Proc. IEEE 109 (3) (2021) 247–278, http://dx.doi.org/10.1109/jproc.2021.3060483.

[59] T.R. Gadekallu, P. Kumar Reddy Maddikunta, P. Boopathy, N. Deepa, R. Chengoden, N. Victor, W. Wang, W. Wang, Y. Zhu, K. Dev, XAI for industry 5.0—Concepts, opportunities, challenges, and future directions, IEEE Open J. Commun. Soc. 6 (2025) 2706–2729, http://dx.doi.org/10.1109/ojcoms.2024.3473891.

[60] S. Velampalli, C. Muniyappa, A. Saxena, Performance evaluation of sentiment analysis on text and emoji data using end-to-end, transfer learning, distributed and explainable AI models, J. Adv. Inf. Technol. 13 (2) (2022) http://dx.doi.org/10.12720/jait.13.2.167-172.

[61] L. Bacco, A. Cimino, F. Dell'Orletta, M. Merone, Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach, Electronics 10 (18) (2021) 2195, http://dx.doi.org/10.3390/electronics10182195.

[62] D. Rajagopal, V. Balachandran, E.H. Hovy, Y. Tsvetkov, SELFEXPLAIN: A self-explaining architecture for neural text classifiers, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 836–850, http://dx.doi.org/10.18653/v1/2021.emnlp-main.64.

[63] C. Molnar, Interpretable Machine Learning, second ed., Seiten 309-318, Christoph Molnar, Munich, Germany, 2022, Literaturverzeichnis.

[64] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832, http://dx.doi.org/10.3390/electronics8080832.

[65] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017, http://dx.doi.org/10.48550/ARXIV.1703.01365.

[66] M. Honegger, Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions, 2018, http://dx.doi.org/10.48550/ARXIV.1808.05054.

[67] W. Silva, K. Fernandes, M.J. Cardoso, J.S. Cardoso, Towards complementary explanations using deep neural networks, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer International Publishing, 2018, pp. 133–140, http://dx.doi.org/10.1007/978-3-030-02628-8_15.

[68] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ (2021) n71, http://dx.doi.org/10.1136/bmj.n71.

[69] M. Saarela, V. Podgorelec, Recent applications of explainable AI (XAI): A systematic literature review, Appl. Sci. 14 (19) (2024) 8884, http://dx.doi.org/10.3390/app14198884.

[70] A. Saranya, R. Subhashini, A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends, Decis. Anal. J. 7 (2023) 100230, http://dx.doi.org/10.1016/j.dajour.2023.100230.

[71] K. Kalasampath, K.N. Spoorthi, S. Sajeev, S.S. Kuppa, K. Ajay, A. Maruthamuthu, A literature review on applications of explainable artificial intelligence (XAI), IEEE Access 13 (2025) 41111–41140, http://dx.doi.org/10.1109/access.2025.3546681.

[72] A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data, Mathematics 10 (4) (2022) 554, http://dx.doi.org/10.3390/math10040554.

[73] E. Anello, C. Masiero, F. Ferro, F. Ferrari, B. Mukaj, A. Beghi, G.A. Susto, Anomaly detection for the industrial internet of things: an unsupervised approach for fast root cause analysis, in: 2022 IEEE Conference on Control Technology and Applications, CCTA, IEEE, 2022, pp. 1366–1371, http://dx.doi.org/10.1109/ccta49430.2022.9966158.

[74] L. Lorenti, G. De Rossi, A. Annoni, S. Rigutto, G.A. Susto, CUAD-Mo: Continuos unsupervised anomaly detection on machining operations, in: 2022 IEEE Conference on Control Technology and Applications, CCTA, IEEE, 2022, pp. 881–886, http://dx.doi.org/10.1109/ccta49430.2022.9966138.

[75] S. Kim, H. Seo, E.C. Lee, Advanced anomaly detection in manufacturing processes: Leveraging feature value analysis for normalizing anomalous data, Electronics 13 (7) (2024) 1384, http://dx.doi.org/10.3390/electronics13071384.

[76] S. Li, T. Li, C. Sun, R. Yan, X. Chen, Multilayer grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis, J. Manuf. Syst. 69 (2023) 20–30, http://dx.doi.org/10.1016/j.jmsy.2023.05.027.

[77] J. Hu, Y. Zhang, W. Li, X. Zheng, Z. Tian, Trustworthy artificial intelligence based on an explicable temporal feature network for industrial fault diagnosis, Cogn. Comput. 16 (2) (2023) 534–545, http://dx.doi.org/10.1007/s12559-023-10218-4.

[78] K. Wen, R. Huang, D. Li, Z. Chen, W. Li, Gradient-based interpretable graph convolutional network for bearing fault diagnosis, in: 2023 IEEE International Instrumentation and Measurement Technology Conference, I2MTC, IEEE, 2023, pp. 1–6, http://dx.doi.org/10.1109/i2mtc53148.2023.10175946.

[79] B. Mohan Dash, B. Ould Bouamama, K. Midzodzi Pekpe, M. Boukerdja, FDI-X: An occlusion-based approach for improving the explainability of deep learning models in fault detection and isolation, in: 2023 International Conference on Control, Automation and Diagnosis, ICCAD, IEEE, 2023, pp. 01–06, http://dx.doi.org/10.1109/iccad57653.2023.10152392.

[80] F. Kakavandi, P. Han, R. de Reus, P.G. Larsen, H. Zhang, Interpretable fault detection approach with deep neural networks to industrial applications, in: 2023 International Conference on Control, Automation and Diagnosis, ICCAD, IEEE, 2023, pp. 1–7, http://dx.doi.org/10.1109/iccad57653.2023.10152435.

[81] L.C. Brito, G.A. Susto, J.N. Brito, M.A.V. Duarte, Fault diagnosis using explainable AI: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data, Expert Syst. Appl. 232 (2023) 120860, http://dx.doi.org/10.1016/j.eswa.2023.120860.

[82] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, U. Zurutuza, Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data, Appl. Sci. 11 (16) (2021) 7376, http://dx.doi.org/10.3390/app11167376.

[83] D. Kim, G. Antariksa, M.P. Handayani, S. Lee, J. Lee, Explainable anomaly detection framework for maritime main engine sensor data, Sensors 21 (15) (2021) 5200, http://dx.doi.org/10.3390/s21155200.

[84] M. Baek, S.B. Kim, Failure detection and primary cause identification of multivariate time series data in semiconductor equipment, IEEE Access 11 (2023) 54363–54372, http://dx.doi.org/10.1109/access.2023.3281407.

[85] P. Agarwal, M. Tamer, H. Budman, Explainability: Relevance based dynamic deep learning algorithm for fault detection and diagnosis in chemical processes, Comput. Chem. Eng. 154 (2021) 107467, http://dx.doi.org/10.1016/j.compchemeng.2021.107467.

[86] R.R.A. Harinarayan, S.M. Shalinie, XFDDC: explainable fault detection diagnosis and correction framework for chemical process systems, Process. Saf. Environ. Prot. 165 (2022) 463–474, http://dx.doi.org/10.1016/j.psep.2022.07.019.

[87] P. Peng, Y. Zhang, H. Wang, H. Zhang, Towards robust and understandable fault detection and diagnosis using denoising sparse autoencoder and smooth integrated gradients, ISA Trans. 125 (2022) 371–383, http://dx.doi.org/10.1016/j.isatra.2021.06.005.

[88] W.-T. Yang, M.S. Reis, V. Borodin, M. Juge, A. Roussy, An interpretable unsupervised Bayesian network model for fault detection and diagnosis, Control Eng. Pract. 127 (2022) 105304, http://dx.doi.org/10.1016/j.conengprac.2022.105304.

[89] K. Jang, K.E.S. Pilario, N. Lee, I. Moon, J. Na, Explainable artificial intelligence for fault diagnosis of industrial processes, IEEE Trans. Ind. Inform. (2024) 1–8, http://dx.doi.org/10.1109/tii.2023.3240601.

[90] F. Gerschner, J. Paul, L. Schmid, N. Barthel, V. Gouromichos, F. Schmid, M. Atzmueller, A. Theissler, Domain transfer for surface defect detection using few-shot learning on scarce data, in: 2023 IEEE 21st International Conference on Industrial Informatics, INDIN, IEEE, 2023, pp. 1–7, http://dx.doi.org/10.1109/indin51400.2023.10217859.

[91] D. Raab, E. Fezer, J. Breitenbach, H. Baumgartl, D. Sauter, R. Buettner, A deep learning-based model for automated quality control in the pharmaceutical industry, in: 2022 IEEE 46th Annual Computers, Software, and Applications Conference, COMPSAC, IEEE, 2022, pp. 266–271, http://dx.doi.org/10.1109/compsac54236.2022.00045.

[92] L.M. Matos, A. Domingues, G. Moreira, P. Cortez, A. Pilastri, A comparison of machine learning approaches for predicting in-car display production quality, in: Intelligent Data Engineering and Automated Learning – IDEAL 2021, Springer International Publishing, 2021, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-91608-4_1.

[93] M. Lee, J. Jeon, H. Lee, Explainable AI for domain experts: a post hoc analysis of deep learning for defect classification of TFT–LCD panels, J. Intell. Manuf. 33 (6) (2021) 1747–1759, http://dx.doi.org/10.1007/s10845-021-01758-3.

[94] D. Kumar, Y. Liu, H. Song, S. Namilae, Explainable deep neural network for in-plain defect detection during additive manufacturing, Rapid Prototyp. J. 30 (1) (2023) 49–59, http://dx.doi.org/10.1108/rpj-05-2023-0157.

[95] H. Bordekar, N. Cersullo, M. Brysch, J. Philipp, C. Hühne, eXplainable artificial intelligence for automatic defect detection in additively manufactured parts using CT scan analysis, J. Intell. Manuf. (2023) http://dx.doi.org/10.1007/s10845-023-02272-4.

[96] R.A.A. Saleh, H. Metin Ertunç, Explainable attention-based fused convolutional neural network (XAFCNN) for tire defect detection: an industrial case study, Eng. Res. Express 6 (1) (2024) 015090, http://dx.doi.org/10.1088/2631-8695/ad23c8.

[97] S. Meister, M. Wermes, J. Stüve, R.M. Groves, Cross-evaluation of a parallel operating SVM – CNN classifier for reliable internal decision-making processes in composite inspection, J. Manuf. Syst. 60 (2021) 620–639, http://dx.doi.org/10.1016/j.jmsy.2021.07.022.

[98] Ł. Wawrowski, M. Michalak, A. Białas, R. Kurianowicz, M. Sikora, M. Uchroński, A. Kajzer, Detecting anomalies and attacks in network traffic monitoring with classification methods and XAI-based explainability, Procedia Comput. Sci. 192 (2021) 2259–2268, http://dx.doi.org/10.1016/j.procs.2021.08.239.

[99] T. Rathod, N.K. Jadav, S. Tanwar, Z. Polkowski, N. Yamsani, R. Sharma, F. Alqahtani, A. Gafar, AI and blockchain-based secure data dissemination architecture for IoT-enabled critical infrastructure, Sensors 23 (21) (2023) 8928, http://dx.doi.org/10.3390/s23218928.

[100] A. Namrita Gummadi, J.C. Napier, M. Abdallah, XAI-IoT: An explainable AI framework for enhancing anomaly detection in IoT systems, IEEE Access 12 (2024) 71024–71054, http://dx.doi.org/10.1109/access.2024.3402446.

[101] Y. Han, H. Chang, XA-GANomaly: An explainable adaptive semi-supervised learning method for intrusion detection using GANomaly, Comput. Mater. Contin. 76 (1) (2023) 221–237, http://dx.doi.org/10.32604/cmc.2023.039463.

[102] M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, A.Y. Zomaya, An explainable deep learning-enabled intrusion detection framework in IoT networks, Inform. Sci. 639 (2023) 119000, http://dx.doi.org/10.1016/j.ins.2023.119000.

[103] A. Barbado, Ó. Corcho, R. Benjamins, Rule extraction in unsupervised anomaly detection for model explainability: Application to OneClass SVM, Expert Syst. Appl. 189 (2022) 116100, http://dx.doi.org/10.1016/j.eswa.2021.116100.

[104] H. Kang, P. Kang, Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism, Knowl.-Based Syst. 290 (2024) 111507, http://dx.doi.org/10.1016/j.knosys.2024.111507.

[105] A. Puthanveettil Madathil, X. Luo, Q. Liu, C. Walker, R. Madarkar, Y. Cai, Z. Liu, W. Chang, Y. Qin, Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes, J. Intell. Manuf. 35 (8) (2024) 4159–4180, http://dx.doi.org/10.1007/s10845-023-02266-2.

[106] V. Swamy, J. Frej, T. Käser, The future of human-centric explainable artificial intelligence (XAI) is not post-hoc explanations, 2023, http://dx.doi.org/10.48550/ARXIV.2307.00364, arXiv:2307.00364.

[107] S. Moosavi, M. Farajzadeh-Zanjani, R. Razavi-Far, V. Palade, M. Saif, Explainable AI in manufacturing and industrial cyber–physical systems: A survey, Electronics 13 (17) (2024) 3497, http://dx.doi.org/10.3390/electronics13173497.

[108] V. Yepmo, G. Smits, O. Pivert, Anomaly explanation: A review, Data Knowl. Eng. 137 (2022) 101946, http://dx.doi.org/10.1016/j.datak.2021.101946.

[109] J. Tritscher, A. Krause, A. Hotho, Feature relevance XAI in anomaly detection: Reviewing approaches and challenges, Front. Artif. Intell. 6 (2023) http://dx.doi.org/10.3389/frai.2023.1099521.

[110] T. Liu, P. Zheng, J. Bao, Deep learning-based welding image recognition: A comprehensive review, J. Manuf. Syst. 68 (2023) 601–625, http://dx.doi.org/10.1016/j.jmsy.2023.05.026.

[111] R. Hoffmann, C. Reich, A systematic literature review on artificial intelligence and explainable artificial intelligence for visual quality assurance in manufacturing, Electronics 12 (22) (2023) 4572, http://dx.doi.org/10.3390/electronics12224572.

[112] E. Mohamed, K. Sirlantzis, G. Howells, A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation, Displays 73 (2022) 102239, http://dx.doi.org/10.1016/j.displa.2022.102239.

[113] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C.A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: XxAI - beyond Explainable AI, Springer International Publishing, 2022, pp. 39–68, http://dx.doi.org/10.1007/978-3-031-04083-2_4.