https://doi.org/10.1038/s44334-024-00006-9

Towards next-gen smart manufacturing systems: the explainability revolution

Check for updates

Puthanveettil Madathil Abhilash, Xichun Luo ✉, Qi Liu, Rajeshkumar Madarkar & Charles Walker

The paper shares the author's perspectives on the role of explainable-AI in the evolving landscape of AI-driven smart manufacturing decisions. First, critical perspectives on the reasons for the slow adoption of explainable-AI in manufacturing are shared, leading to a discussion on its role and relevance in inspiring scientific understanding and discoveries towards achieving complete autonomy. Finally, to standardize explainability quantification, a new Transparency–Cohesion–Comprehensibility (TCC) evaluation framework is proposed and demonstrated.

Manufacturing processes are becoming increasingly reliant on AI for enhanced quality, productivity, and overall performance. For the past couple of decades, black-box AI models with near-zero feedback on their decision-making rationale have been driving the majority of manufacturing systems[1]. This is due to their superior capability to model intricate, complex, non-linear input–output relationships. While there are no doubts about their capabilities, the mistrust and suspicion surrounding inexplicable systems have significantly hindered or, at the very least, delayed the widespread adoption of AI in several manufacturing domains[2].

Many existing AI-driven manufacturing processes and systems rely on deep learning models due to their superior prediction accuracy[3]. However, equally vital for high-stake decisions, especially for high-value products, is the interpretability of AI-driven decision-making. The lack of transparency and scientific understanding in these systems can lead to regulatory ethical concerns, particularly in applications related to biomedical, nuclear, and aerospace. Blindly trusting a black-box model jeopardizes the autonomy and informed decision-making of experts, seriously limiting troubleshooting and improvement opportunities[4]. As a result, despite the current state of AI's technological maturity, manufacturers are still hesitant to entirely rely on machines, preferring to retain human judgement.

The concept of Explainable AI (XAI) is becoming relevant in this context. XAI refers to a set of techniques and approaches within the field of AI that aim to make the decision-making processes and outcomes of AI systems understandable and interpretable by both experts and non-experts. The primary goal of XAI is to provide insights into why and how AI models arrive at specific decisions or predictions[5].

As the manufacturing industry continues to evolve, embracing XAI will be crucial to advanced innovation and ensuring the responsible and effective adoption of AI technologies. This transition from Industry 4.0, which had a technological focus, to Industry 5.0, emphasizing societal impact, is driven by this high demand for transparency and trust[6]. Industry 5.0 represents a paradigm shift from the purely technology-driven advancements of Industry 4.0 to a more human-centric approach, where the focus is on the collaboration between humans and machines. This new era emphasizes not only efficiency and productivity but also the societal and ethical implications of technological integration. Transparency and trust are essential in Industry 5.0 because they ensure that AI systems are not only effective but also align with human values and ethical standards. By making AI decisions understandable and accountable, we can create a collaborative environment where human expertise and AI capabilities complement each other, leading to innovations that are both technically sound and socially responsible. Following this global trend, regulatory bodies have now prioritized extensive action plans to address the explainability concerns. The principles of openness, fairness and explainability of AI have been underpinned by the US Federal Trade Commission, the General Data Protection Regulation of the European Union, and the UK AI Regulations 2023[7].

As we move towards a more interconnected world, where AI and humans collaborate closely, the ability to comprehend and explain the actions of AI models becomes essential. While efforts to create explainable manufacturing systems are underway, numerous challenges, both explicit and implicit, hinder their widespread adoption. Amongst those, one of the most significant challenges is the standardized evaluation of explainability within manufacturing systems. Current methodologies lack a unified approach to evaluate the explainability of AI models in terms of its key sub-aspects like transparency, cohesion, and comprehensibility, leading to inconsistent and fragmented adoption across the industry. Such inconsistencies not only limits the full integration of AI into existing manufacturing systems, but also increases the lack of trust, where stakeholders are reluctant to embrace AI solutions due to perceived risks and uncertainties.

This perspective paper explores these issues, shedding light on often overlooked obstacles in implementing XAI and exploring the road ahead for its role and applicability in advancing next-generation manufacturing systems. The paper also shares the authors' perspectives on the role of AI in generating scientific understanding within the manufacturing domain, from aiding experts to autonomous discovery, and discusses the implications of moving from weak AI to ultra-strong AI models. To this end, we also

Centre for Precision Manufacturing, DMEM, University of Strathclyde, Glasgow, UK. ✉e-mail: xichun.luo@strath.ac.uk

present an explainability evaluation framework, formalizing and adapting a theory and governing rules, to determine the level of scientific understanding that an explainable smart manufacturing system can provide.

The key contributions of this paper are:

- Critical analysis of slow XAI adoption: provides insights into the reasons for the slow adoption of XAI in manufacturing.
- Perspectives on XAI's role in scientific understanding: discusses the role and relevance of XAI in inspiring scientific understanding and discoveries towards achieving complete autonomy in manufacturing systems.
- A novel framework to evaluate explainability: identifies the challenges and key aspects for the standardized evaluation of explainability in AI-driven manufacturing systems and introduces a framework to address these limitations.

Impact of technological maturity of AI in smart manufacturing

The comprehensibility of AI-driven manufacturing is closely tied to the technical maturity of AI technologies, classified into weak, strong, and ultra-strong ML based on their comprehensibility and contribution to scientific understanding[8]. Here the 'strength' of the algorithm is not based on the prediction accuracy but on the comprehensibility and contribution towards scientific understanding. Despite AI's rapid progress, the adoption of advanced AI in smart manufacturing has been relatively slow. Weak AI models are the traditional black-box models, which may be accurate but are not explainable. Current state-of-the-art XAI models, falling under strong AI, symbolically represent hypotheses as mathematical expressions, allowing deeper AI decision understanding and integration with physics-based models. Ultra-strong AI is an envisioned future technology where advancements in AI enable autonomous domain exploration and knowledge acquisition without human involvement.

Though every type of AI model can be expressed mathematically, the nature of these expressions and their interpretability vary significantly, especially with respect to the final form of the trained model. Weak AI models like deep neural networks, while mathematically expressible, produce highly complex, multi-layered structures that encode relationships in a manner that is not easily understandable by humans. The weights and biases in neural networks do not offer straightforward insights into the model's decision-making process. In contrast, strong AI models, like symbolic regression for instance, generate explicit, human-readable mathematical formulas that directly describe the relationships between variables as the final trained models. These formulas are inherently interpretable because they are composed of familiar mathematical operations and functions that experts can readily understand and analyze.

Currently, most smart manufacturing systems rely on weak 'black-box' algorithms[3]. While effective in various applications, their lack of transparency limits their use in critical domains. The integration of strong AI in smart manufacturing can lead to the discovery of credible physical models that help advance scientific understanding and improve decision-making. However, strong ML has rarely been implemented in the manufacturing domain, barring a few exceptions[9–11].

We believe that the class of ultra-strong ML represents the vision of future autonomous manufacturing systems, where the machine gains understanding by itself without requiring human intervention or oversight. This holds the promise of even greater advancements in manufacturing processes, where the AI system may uncover hidden patterns and relationships in the data that were not evident through traditional physics-based approaches, leading to new insights and manufacturing approaches.

Quickly embracing XAI models and effectively integrating them with physics-based models will be key to unlocking the full potential of AI-driven manufacturing in a new era of innovation and efficiency. However, the adoption of XAI in the manufacturing domain has been slow, significantly hindering the widespread acceptance of AI decisions in manufacturing applications, particularly in the real-world industrial production of high-value products. This represents a missed opportunity[12,13]. Technological advancements in AI, coupled with its rapid adaptation will shape the future of smart manufacturing.

XAI in manufacturing

State-of-the-art XAI techniques

Given below are the definitions of the often interchangeably used XAI-related terminologies:

- Understandability: it refers to the ease with which a person can grasp the overall functioning of a model without needing to investigate its internal mechanisms or the specific algorithms it employs[14].
- Comprehensibility: in the context of AI, comprehensibility is the extent to which a learning model can express its acquired knowledge in a way that is easily understood by humans. This involves creating descriptions that resemble those a human expert would generate, integrating both quantitative and qualitative aspects in a coherent and interpretable manner[15].
- Interpretability: interpretability is the quality of a model that allows it to be explained or its meaning to be conveyed in terms that are clear and understandable to a human observer[16]. Interpretability is about the inherent property of the model itself, where they are designed to be understood directly.
- Explainability: this concept involves providing explanations that serve as a bridge between humans and a decision-making system. These explanations should accurately reflect the decision-making process and be human comprehensible[15]. Unlike interpretability, explainability is not a model's inherent property, rather it often involves additional methods to explain how and why a model behaves in a certain way.
- Transparency: a model is deemed transparent if its workings are clear and understandable on their own. Transparent models are categorized into three types: those that are easily replicated, those that can be broken down into understandable components, and those whose algorithms are inherently clear and straightforward[17].
- Clarity: clarity refers to the quality of being easily understood, and free from ambiguity. In the context of AI, clarity ensures that the information, explanations, and insights provided by the AI models are straightforward, precise, and easy to comprehend by users, including non-experts.

These terms are distinct but related, and they collectively contribute to making AI models more accessible and trustworthy to users in smart manufacturing contexts.

The XAI approaches are classified according to their scope, stage, input, and output as depicted in Fig. 1. The three perspectives with respect to the stage of application are ante-hoc (intrinsic), post-hoc, and real-time approaches. The ante-hoc or intrinsic interpretability approach integrates explanations into the model training phase, ensuring accountability in decision-making[18]. Such explanations offer greater understandability, facilitating comprehension of the model's mathematical underpinnings and decision mechanisms. Here the explainability relies on transparent models, which inherently possess interpretability to some extent. Transparent models, as defined earlier, are 'by-design' interpretable models, whose computational structure and functioning are easily understandable by humans, allowing them to see and comprehend how decisions are made, without the need for additional explanation tools[16]. Examples include decision trees, rule-based systems, and logistic and linear regression.

Intrinsic XAI models can be categorized based on their domain of interpretability, namely algorithmic transparency, decomposability, and simulatability. Algorithmic transparency pertains to the user's ability to understand the process followed by the model to produce outputs from input data. Decomposability involves explaining each part of a model, such as input, parameters, and calculations, enhancing intelligibility and interpretability. Simulatability refers to a model's ability to be simulated or comprehended strictly by a human. For instance, though rule-based systems are generally conceived as comprehensible, once they become very extensive

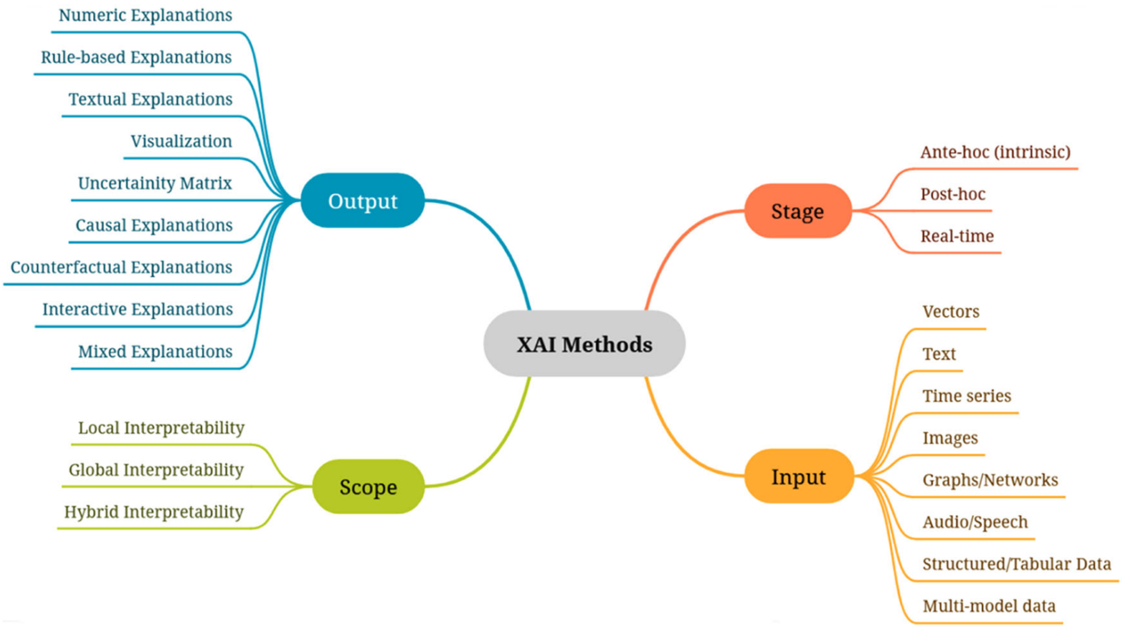

Fig. 1 | Categorization of XAI methods. A comprehensive classification of various XAI methods is given based on four key aspects: stage, scope, inputs, and outputs. 'Stage' illustrates the timing of explanations in the AI process. It includes "Ante-hoc (Intrinsic)" where explanations are built-in before model deployment, and "Post-hoc" where explanations are generated after model deployment. 'Scope' represents the breadth of explanations covered by XAI methods. It ranges from "Local Interpretability" focusing on individual predictions, to "Global Interpretability" which addresses model behaviour on a general level. 'Input' enumerates the types of data that can be processed by XAI systems. It includes diverse data types such as texts, images, time series data, and others, indicating the versatility of XAI in handling various forms of information. 'Output' section lists the possible forms that an explanation might take. It includes "Numerical Explanations", "Rule-based Explanations", "Textual Explanations", and others, highlighting the different ways users can understand AI decisions.

with numerous rules, they may lose their simulatability and become difficult for humans to fully understand and manage.

The explainability aspects of intrinsic XAI models like linear/logistic regression, decision trees, k-nearest neighbours (KNN), rule-based learning, general additive models, and Bayesian models have been studied in the past. In this regard, researchers have investigated the soundness of logistic/linear regression models[19,20] and also reported some critical concerns about its interpretability as well[21]. Although decision trees offer algorithmic transparency[22], they are comprehensible by humans only if their size and features are small. Another challenge for decision trees is their limited generalization capabilities. Through its perception of distance and similarity between the cases, KNN resembles human experts' decision-making rationale and thus was widely adopted in situations demanding interpretability[23,24]. The model's interpretability will be impeded when the number of neighbours increases or when complex features and distance functions are used.

Rule-based learning involves generating rules to characterize data, ranging from simple conditional if-then rules to more complex combinations like fuzzy rule-based systems which allow verbally formulated rules, improving model interpretability and performance in uncertain contexts[25]. While rule-based learners indeed offer transparency[26], challenges arise in balancing rule coverage and specificity, which impact interpretability. Though often not categorized under transparent models, Bayesian[27] and general additive models[28] are sometimes accepted as interpretable modelling choices due to their capacity to offer insights and explanations.

The second classification based on the stage of application is post-hoc XAI, which involves analyzing trained or tested AI models to uncover their inner workings and decision logic. Such analyses are often offered as feature significance ratings, rule sets, plots, or human-readable explanations[29]. Recently, there has been a significant surge in post-hoc approaches, aimed at clarifying black-box models[30–33]. Within post-hoc explainability, there exists model-agnostic techniques and model-specific techniques. Model-agnostic techniques offer explanations to any black-box models irrespective of their types and computing architectures. A few modes of model-agnostic explanations are explanation by simplification, feature relevance explanation, and visual explanations.

Explanation by simplification involves approximating the complex decision-making process of a black-box model with a simpler, more interpretable surrogate model. This simplified model aims to capture the essential relationships between the input features and the model's output. It pertains to the overall model, as it seeks to create a simpler representation of the model's behaviour that maintains the input–output relationships. For instance, the very popular local interpretable model-agnostic explanations (LIME)[34] and its variations[35] belong to the explanation by simplification approach, whereas the other widely used Shapley additive explanations (SHAP)[36] method belongs to the feature relevance explanation approach.

The model-specific post-hoc explanation approaches are tailored for specific ML models. Researchers in the past have come up with model-specific explanations for shallow ML models like tree ensembles (in particular, explanation by simplification[37] and feature relevance analysis[38]), multi-class classifiers, SVMs (simplification[39] and visualization[40]), and deep learning models. Within deep learning, specific post-hoc approaches for multi-layer neural networks[41], convolutional neural networks (input–output relationship comprehension[42], network interpretation[43] and visual explanations[44]), and recurrent neural networks[45], have been recently explored. Stacking with auxiliary features (SWAF)[46] and Deep SHAP[47] are widely used model-specific techniques.

In contrast to intrinsic explainability, post-hoc XAI typically uses supplementary models (like SHAP, LIME, and SWAF) to interpret the decisions of the original model, which can introduce additional layers of complexity. These explanations might not always be straightforward, often needing some basic knowledge of the underlying methods and assumptions. Consequently, while they provide insights into model behaviour, the interpretability of these explanations poses challenges for non-experts, potentially hindering their practical application in decision-making processes. In addition, researchers have found a fundamental shortcoming of the existing feature relevance methods for neural networks—the violation of sensitivity and implementation axioms[48].

The third approach, real-time XAI, focuses on providing explanations simultaneously with the AI model's decision-making process, allowing users to understand the model's reasoning as it happens in high-stakes or dynamic environments. It is essential in scenarios where decisions need to be transparent and understandable on the fly, ensuring that users can trust and verify AI decisions as they occur[49]. Examples include attention mechanisms in neural networks and interactive visualization tools like real-time saliency maps.

XAI approaches not only vary according to their stage of operation but also differ based on the forms of explanations, including deep explanations, interpretable models, model induction, explainable human–computer interaction, and psychological grounding. Deep explanations leverage deep learning techniques to render internal model operations more interpretable[50]. Prioritizing clarity and simplicity for interpretable models, such as those based on decision trees or logistic regression, will make them accessible to a wider audience. Model induction techniques aim to derive explainable models from black-box ones[51,52]. Explainable human–computer interaction strategies transform explanations into tangible interfaces for user engagement[53,54]. Explainable human–machine interactions (HMI) describe the act of understandable communication from the machine to humans, particularly in the form of explanations for AI-based decisions. This is a crucial aspect of XAI, since such interactions offer natural language-based, dialogue-based, and virtual reality (VR)-based explanations of AI models to users, enhancing their understanding of the model predictions and guiding them towards better decision-making. Examples include the Bot-X virtual assistant[55] for intelligent manufacturing and a VR-based chatbot[56] for manufacturing services.

Moreover, XAI grounded in psychological theories, known as explanatory ML, enhances interpretability through insights into human cognition[57]. From a scientific point of view, human cognition involves constructing internal models based on perceptual information, which informs decision-making. While the XAI models may be inherently incomplete and imperfect, such methods offer complementary insights into data analysis and decision processes. The principles underlying XAI development include considerations such as stage, scope, input, and output formats, which are summarized in Fig. 1.

XAI approaches in manufacturing

Despite the existence of numerous XAI methodologies as discussed in the previous section, they fall short of delivering comprehensible explanations to humans in smart manufacturing. The documented challenges include data and system complexities, over-reliance on black-box models, and difficulties in offering contextual understanding in the manufacturing domain. Existing research on the integration of XAI in manufacturing aims to address some of these challenges, but the quantity of such study is so far very limited, and some of our perspectives on the notable reasons behind the slow adoption of XAI are shared in the next section. The common application areas of XAI in manufacturing are surface quality prediction, defect detection, condition monitoring, and process control[58].

Goldman et al.[59] employ saliency maps and histograms to explain the functioning of black-box classifiers for ultrasonic weld quality prediction. While visual presentations effectively highlight specific patterns, they often lack the depth required to convey the intricate operations of predictive models. Lee et al.[60] developed an algorithm which converts decision tree-based defect detection logic into human-comprehensible text. Similarly, McLaughlin et al.[61] worked on an XAI approach towards defect detection in lithography coatings defects. A linear graph-based visualization technique was developed by Glock[62] to explain the random forest-based defect detection model.

Another notable area of XAI application in smart manufacturing is in predictive maintenance, monitoring, and prognosis domains. In this regard, Alvanpour et al.[63] worked on developing explanations for black-box predictions of robot grasp failures. They claimed that when the historic data-based prediction was combined with their proposed explanations, it helped in better adaptive control. An explainable interface to score DT-biased machine condition monitoring was developed by Matzka[64], Torcianti and Matzka[65]. Hermansa et al.[66] proposed a feature extraction-based XAI approach for predictive maintenance using time series vibration and temperature data. Wang et al.[67] and Gribbestad et al.[68] worked on XAI-based anomaly detection based on sensor data. An explainable surface quality monitoring system for the grinding process using vibration signals was proposed by Hanchate et al.[10].

XAI has also been used to understand ML-based predictions in additive manufacturing. Guo et al.[9] used a SHAP-based XAI to better comprehend the process physics behind layer-wise emissions during laser powder bed fusion (L-PBF) metal additive manufacturing. Wang and Chen[11] proposed the usage of XAI tools towards 3D printing facility selection. An explainable 3D printer fault diagnosis system, called XAI-3DP, was proposed by Chowdhury et al.[69].

A few studies have also focused on building generic XAI models or frameworks to suit manufacturing applications. In this regard, recently, Kusiak[70] proposed the federated explainable artificial intelligence approach from a digital manufacturing perspective. A human-centric framework named STARdom to implement XAI in manufacturing systems is proposed by Rožanec et al.[71], consisting of an active learning module and a feedback module. Senoner et al.[72] developed a decision-making paradigm, aided by visual explanations, specifically to handle and interpret complex manufacturing data-driven decisions.

While the existing XAI methodologies demonstrate promise, they also exhibit several technical limitations that hinder their broader adoption in manufacturing. One significant limitation is the lack of standardized evaluation criteria for explainable results. This absence makes it challenging to measure the effectiveness and reliability of XAI methods across different applications. Furthermore, many current approaches are highly context-specific, limiting their generalizability to other manufacturing scenarios. For instance, existing XAI approaches are often purely data-driven and work only under specific conditions such as a particular machine, material, and environment, as well as within specific data ranges. A trained XAI-based predictive maintenance model, for example, may fail if the type of machinery or operating environment changes. To address this limitation, XAI needs to incorporate and understand process physics more comprehensively. This would make it more generalizable and usable across various scales, conditions, and contexts. There is also a tendency to oversimplify explanations, which can lead to the omission of critical insights necessary for comprehensive understanding. Additionally, the computational complexity of generating explanations for black-box models often results in increased processing times, which is impractical for real-time applications. Lastly, there is a need for more robust validation frameworks to ensure that the explanations provided are not only accurate but also meaningful and actionable for end users in the manufacturing domain. Addressing these technical limitations is crucial for advancing the integration of XAI in smart manufacturing.

Overall, the extent of XAI integration is very limited in the manufacturing domain and we see it as a lost potential. It is critical to address the reasons for its implementation latency. In this regard, let's explore some implicit, and often overlooked challenges in XAI implementation, both in a general context and specifically within smart manufacturing, which, according to us, have contributed to its slow adaptation.

Key challenges in XAI implementation

Not everyone is concerned

Many people still favour accepting results from black-box models without concern for explainability, as described in refs. 73,74. The following reasons illustrate why not everyone prioritizes explainability. First, many stakeholders currently prioritize immediate results over the potential long-term benefits of deploying interpretable models, particularly if there's a perceived trade-off between transparency and predictive accuracy. Furthermore, in scenarios where the implications of wrong decisions are relatively low, the priority naturally leans towards maximizing predictive accuracy rather than understanding the decision-making rationale. According to Nigam Shah of the Stanford Institute for Human-Centered Artificial Intelligence, if an AI model provides accurate predictions, it's considered useful regardless of our understanding of its workings[75]. This viewpoint is also supported by Marzyeh Ghassemi at the University of Toronto[76]. Manufacturers often prioritize case-specific performance, which black-box models excel at, over repeatability, scalability, and robustness.

In manufacturing, high-stakes decisions often involve safety-critical systems, regulatory compliance, financial implications, and ethical considerations. Examples include decisions related to the production of medical devices, aerospace components, or critical infrastructure, where errors or failures can lead to severe repercussions.In this context, 'stakes' refer to the level of importance, risk, or potential consequences associated with the outcomes of a decision or action. 'High stakes' refers to scenarios where the consequences of a decision or action are significant and can have substantial impacts.

Though presently not many manufacturers are concerned about explainability, we argue that relying on black-box models is problematic for several reasons. They're valuable only when they provide accurate results, especially where the cost and consequences of incorrect predictions are low[73]. Rudin has summarized why black-box ML should not be used for high-stakes decisions[4]. We argue that it's essential to embrace explainability in decision-making for manufacturing high-stakes components like medical devices, aerospace parts, and critical infrastructure. Relying on opaque models for bioimplant manufacturing decisions, for example, contradicts medical ethics principles[77,78]. In such cases, decisions should be based on evidence, reasoning, and a deep understanding of causality—qualities only interpretable models can provide.

The Trust Paradox

While explanations can be valuable, they are not always necessary, and interestingly, there's a downside. There are instances where explanations can inadvertently lead individuals to trust a model even when it's making clear errors. Microsoft Research has successfully demonstrated that people tend to accept obvious mistakes when they receive explanations from an interpretable model, creating a false sense of trust[75]. Consider a smart manufacturing system using AI for quality control in an auto parts factory. If the AI makes an incorrect decision, say, passing a batch of parts with hidden defects as high-quality, human inspectors might naturally question the decision if there are no AI explanations. This might lead to more rigorous inspections. However, if the AI provides explanations that seem reasonable, inspectors might trust the AI's judgement and skip further scrutiny.

The right selection of explainability

While XAI is intended to bring clarity to the use of AI systems, it has, in some cases, introduced additional ambiguity. This ambiguity arises from the need to choose among various types and levels of explainability, in addition to selecting the appropriate AI models from a wide array, which is already a significant challenge.

In general, there exist three distinct types of explainability with respect to a stakeholder's viewpoint, as shown in Fig. 2. First, an engineer's explainability pertains to understanding how an AI model arrives at its decisions. The second type, causal explainability, focuses on understanding how a specific input leads to a particular output. This type of explainability is closely linked to scientific understanding. The third type is local explainability, which is more of a justification for a local prediction. Engineers seek to comprehend how their model functions to facilitate debugging, while stakeholders and end users require causal explanations to trust the model and use it confidently. Presenting an explanation from the wrong category can lead to unexpected consequences, including a loss of trust in an otherwise accurate and beneficial model.

Moreover, the level of explainability required is also contextual. As argued earlier by Arbelaez Ossa[79], low-stakes decisions can accept less XAI, whereas, for high-stakes manufacturing systems, complete transparency is mandatory. Such variable explainability requirements contribute to the existing lack of clarity surrounding interpretability.

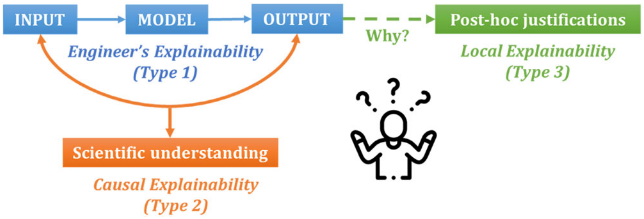
INPUT
MODEL
OUTPUT
Why?
Post-hoc justifications
Engineer's Explainability (Type 1)
Local Explainability (Type 3)
Scientific understanding
Causal Explainability (Type 2)

Fig. 2 | Types of interpretability. The three distinct types of explainability in AI systems from the perspective of different stakeholders are illustrated. Engineer's explainability (Type 1) describes how an AI model arrives at its decisions. Causal explainability (Type 2) focuses on understanding or revealing the fundamental input–output relationships, and finally local explainability (Type 3) clarifies a model's local prediction, often required by end users who need to understand and trust the model's specific decisions.

Consequently, there emerges a need for proper evaluation metrics and a consensus on the types and levels of interpretability. This brings us to the next challenge: the lack of a robust evaluation matrix.

Lack of a robust evaluation matrix

Implementing XAI in smart manufacturing presents multifaceted challenges, with a significant obstacle being the lack of a robust evaluation framework. While the benefits of XAI in enhancing transparency and understanding are clear, determining the required level of explainability is context-dependent. Manufacturing processes vary greatly, demanding distinct degrees of AI interpretability. Therefore, establishing a standardized evaluation framework adaptable to diverse manufacturing contexts is essential. It should consider AI explainability, process complexity, safety, regulations, and human–AI collaboration dynamics. Overcoming this challenge requires collaboration between AI researchers, manufacturing experts, and policymakers to develop criteria that align with smart manufacturing needs.

In summary, the aforementioned challenges underscore the critical need for a strategic vision in the road ahead for explainable manufacturing systems, which is pivotal to the future of smart manufacturing. To address these challenges, this work critically discusses and argues for the necessity of explainability in smart manufacturing decisions, emphasizing why more stakeholders should be prepared to transition and embrace this technology. Additionally, we share our vision for the future role and impact of XAI in smart manufacturing, advocating the urgent need for wider adoption of XAI technologies. Finally, this work proposes the development of a standardized evaluation matrix, enabling manufacturers to evaluate, compare, and implement more explainable systems effectively.

Perspectives on the road ahead

We firmly believe that the explainability revolution will cause a paradigm shift in future AI applications in manufacturing systems from predictive modelling towards AI-inspired causal understanding, autonomous explorations of unseen trends, and scientific discoveries. The concept and various methods of using AI towards scientific discovery are reviewed by Wang et al.[80]. Now, we share our perspectives on the various ways in which explainability can potentially affect the future of smart manufacturing.

XAI-driven scientific understanding

Developing interpretable models to capture manufacturing processes or system behaviours can inspire innovative ideas and concepts. One way of interpretable expression is mathematical equations, which is already implemented in material science as a mode for learning from the data[8,81]. Modelling techniques such as symbolic regression hold the potential to act as the 'resource of inspiration' to identify new physics signatures in manufacturing systems, which can then be conceptualized and understood by human experts[82–84]. For instance, in precision machining, accurately computing and accounting for tool centre position (TCP) positioning deviations

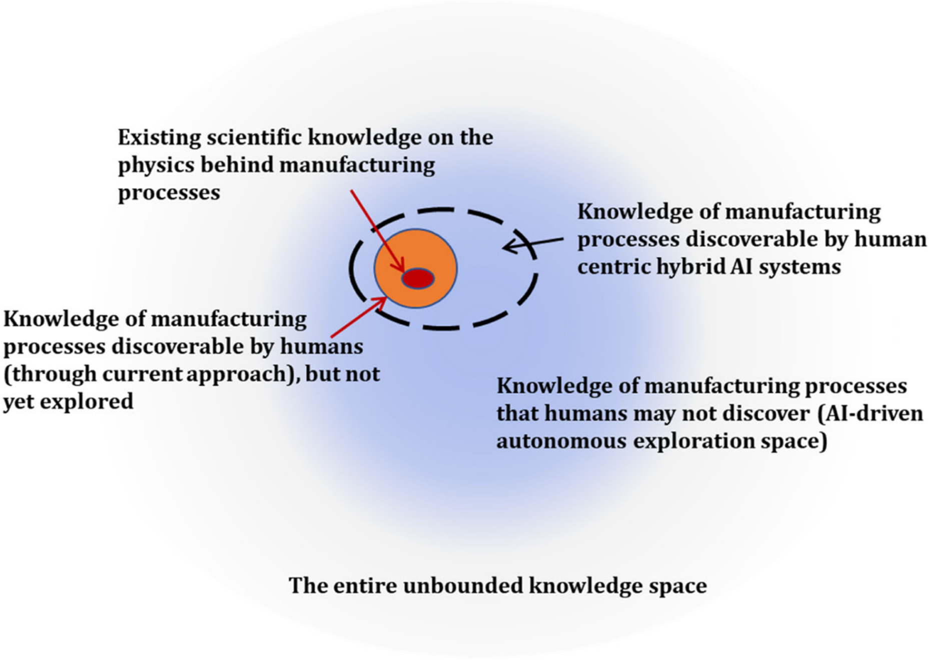
Existing scientific knowledge on the physics behind manufacturing processes
Knowledge of manufacturing processes discoverable by human centric hybrid AI systems
Knowledge of manufacturing processes discoverable by humans (through current approach), but not yet explored
Knowledge of manufacturing processes that humans may not discover (AI-driven autonomous exploration space)
The entire unbounded knowledge space

Fig. 3 | Autonomous AI-driven explorations. The different domains of knowledge related to manufacturing processes and the potential role of AI in expanding these domains are illustrated. Existing scientific knowledge on the physics behind manufacturing processes is shown in red. Knowledge of manufacturing processes discoverable by humans, but not yet explored is shown in orange circular region. This circle represents the knowledge that humans have the potential to discover (through conventional methods) but have not yet explored. Shaded blue region within the dashed line representing the knowledge that can be discovered by AI systems designed to work in conjunction with humans. It includes knowledge that we currently possess. Region outside the dashed line represents the knowledge of manufacturing processes that might be beyond human reach but could potentially be discovered by future AI systems operating autonomously (figure created by the author, taking inspiration from ref. 85).

pose critical challenges. Expressing TCP positioning errors as interpretable mathematical functions of process parameters enables us to identify their root causes and derive effective solutions, thereby enhancing our scientific understanding of these systems.

XAI-driven scientific discovery

In the manufacturing context, scientific discovery refers to the process of identifying new methods, materials, and processes which can enhance efficiency, quality, and innovation in manufacturing systems. This includes the discovery of new manufacturing techniques, optimization of existing processes, and uncovering previously unknown phenomena that can lead to breakthroughs in production technologies and practices. In 2021, a Nobel Turing Challenge was proposed towards using AI for scientific discoveries in the domain of biological sciences[85]. Similarly, we anticipate the utilization of XAI for comprehending and uncovering manufacturing science in the future. In this regard, we propose the multi-stage XAI-driven discovery of manufacturing science.

Stage I: autonomous explorations. In smart manufacturing, the exploration of a data/knowledge space has traditionally been based on global search algorithms like evolutionary and swarm search, especially in the case of well-defined responses. These explorations are usually fully defined by human scientists in terms of the objective function (responses of interest), constraints and search space. However, we anticipate that, in future, identifying exceptional or unique trends from manufacturing datasets through open-ended explorations will have more chances of leading to breakthrough findings, which may be near-impossible for humans to grasp[86]. We argue that explainability in AI-driven explorations is crucial because the path and methodology of scientific discovery must be transparent, enabling further discoveries and applicability across various manufacturing domains.

To push exploration beyond traditional boundaries (as depicted in Fig. 3, inspired by ref. 85), redefining response and objective functions is essential. One strategy involves incentivizing intrinsic responses like curiosity, surprise, and creativity, rather than the typical manufacturing responses. For instance, a curiosity algorithm-driven robot (CA-robot) has explored and revealed unpredictable behaviours in complex chemical systems, leading to the discovery of novel protocell behaviour[87].

Such approaches increase the likelihood of uncovering previously hidden physics signatures in manufacturing, which can subsequently be comprehended and understood by humans. This reveals previously overlooked patterns, fostering fresh conceptual understanding.

Stage II: autonomous hypothesis generation. Next, we anticipate that XAI will advance beyond autonomous data explorations to a stage where it can independently propose hypotheses and plan experiments without human intervention. One potential approach to achieve autonomous hypothesis generation is by leveraging the extensive repository of online publications within a specific manufacturing domain. A similar approach has already proven successful in material science, where it rediscovered existing scientific knowledge[88] and unveiled previously unknown complex scientific inferences[89].

As detailed in Fig. 4, our concept involves extracting a transparent and explainable semantic knowledge graph from online manufacturing literature, with each node representing a manufacturing concept and interconnected links illustrating their relationships. Recent work has demonstrated a framework for knowledge graph extraction from raw data[90,91]. The knowledge space clearly delineates the boundaries of current human knowledge, while outside this space lie current manufacturing challenges (Fig. 4a). AI will autonomously expand manufacturing knowledge by proposing and substantiating hypotheses to address these challenges, bridging the gap between existing domain knowledge space and unsolved problem space (Fig. 4b).

However, we argue that gaining the autonomy to propose a hypothesis will just not be enough to persuade the industries, stakeholders, and regulatory and funding agencies to trust and fund the AI-generated theory. Here, explainability will hold the key to the concept's success, especially in high-stakes manufacturing. The AI hypothesis generation logic and the hypothesis itself need to be human comprehensible, to be judged for bias, fairness, novelty, and feasibility.

Stage III: towards complete manufacturing autonomy. Finally, a stage is reached where robot scientists can autonomously 'experiment and analyze' the AI-generated hypothesis to push the scientific boundaries, a typical example being robot scientists Adam and Eve[92]. We argue that the manufacturing domain has good potential to be one of the earliest adopters of this level of autonomy due to the already existing deep

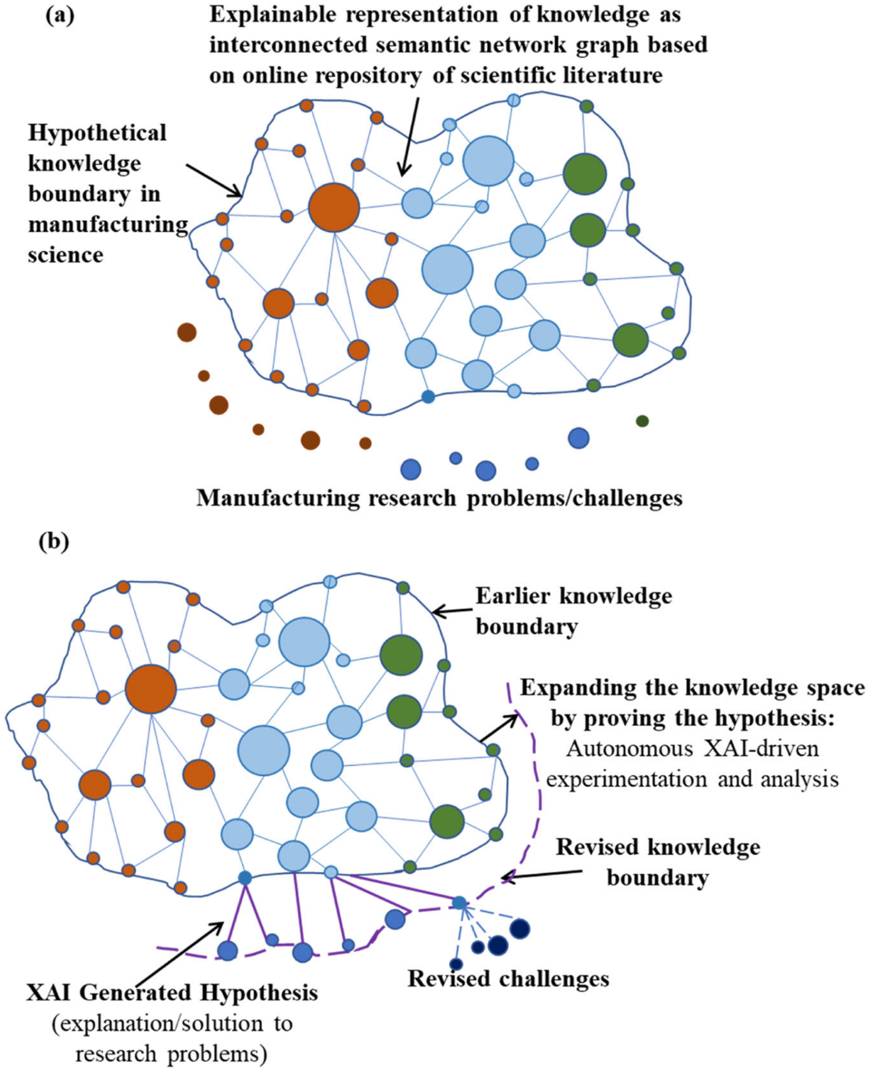
(a)
Explainable representation of knowledge as interconnected semantic network graph based on online repository of scientific literature
Hypothetical knowledge boundary in manufacturing science
Manufacturing research problems/challenges
(b)
Earlier knowledge boundary
Expanding the knowledge space by proving the hypothesis:
Autonomous XAI-driven experimentation and analysis
Revised knowledge boundary
Revised challenges
XAI Generated Hypothesis
(explanation/solution to research problems)

Fig. 4 | Theory of AI-driven hypothesis generation. a A hypothetical representation of scientific knowledge based on an online repository of scientific literature is depicted as a network of interconnected nodes, with each node representing a piece of scientific knowledge. The nodes are interconnected, indicating the relationships between different pieces of knowledge. The boundary surrounding the network represents the current extent of scientific knowledge in the manufacturing domain, with existing research problems and challenges represented outside this knowledge boundary. **b** The potential expansion of the knowledge space through the application of AI is illustrated. Explainable AI (XAI) has the potential to autonomously propose hypotheses, conduct experiments, and perform analyses to validate these hypotheses, thereby further expanding the knowledge space. The revised knowledge boundary represents the potential new extent of knowledge, leading to new challenges represented outside the new boundary.

integration of robots in smart factories. Explainability will still be the key at this stage since reliance on non-explainable systems for autonomous experiments will not be trustworthy due to its high-risk nature.

This stage marks the achievement of fully autonomous scientific discovery in manufacturing, where an XAI-based system explores knowledge space (Stage I), proposes and prioritizes the hypothesis (Stage II), plans and conducts the experiments, and analyses the results (Stage III). With reference to the earlier proposed knowledge graph, solving one challenge not only expands and pushes the knowledge boundary, it also opens up fresh challenges. The semantic graph continuously updates itself by its continuously evolving understanding as shown in Fig. 4b. As this system evolves further, human involvement will be minimal, mostly at the supervisory level.

This autonomous scientific discovery through XAI promises to revolutionize research and development in manufacturing by accelerating the discovery of novel solutions, reducing reliance on human intervention, and facilitating rapid adaptation to evolving scientific understanding within the manufacturing field.

XAI for dialogue-based human–machine interactions (HMI)

Smart manufacturing and digital twins will see significant advancements in terms of HMI, especially in connection with the explainability of decision-making rationale. Advancements in LLM tools such as GPT, BARD, and BERT allow dialogue-based scientific discussions between the machine and the operator/expert. This will be achieved by leveraging LLM's capacity to act as an intermediate layer between operator and machine, converting abstract-level instructions into technical information and vice versa. A chat-based virtual assistant called Bot-X has already been proposed for intelligent manufacturing, which allows natural language communication between man and machine for tasks like ordering, production execution, etc.[55]. Dialogue-based HMI is relevant in the context of XAI, since natural language explanations offer a better understanding to technical non-experts, in comparison to graph-based or plots-based post-hoc explanations which demands a basic level of subject knowledge and expertise.

HMI will advance to a stage where explanations for any AI-based decisions can be connected to or sourced from published literature. This will even enable the machine to propose innovative solutions to previously unseen manufacturing problems. What's even more intriguing is that such suggestions might inspire human scientists towards new scientific insights and innovative solutions, as claimed by Leslie[93].

Other potential applications

A few other notable applications related to XAI include immersive explanations through Augmented Reality (AR) and VR interfaces, as well as holographic data visualization, offering interactivity, spatial understanding, and scalability for complex manufacturing systems. Explainable digital twin systems, when implemented, enable real-time insights into the functioning of physical assets, enhancing predictive maintenance and overall system reliability. Moreover, stringent regulations will likely lead to dedicated XAI frameworks, and possibly even the establishment of explainability standards for various high-stakes manufacturing sectors (such as medical devices,

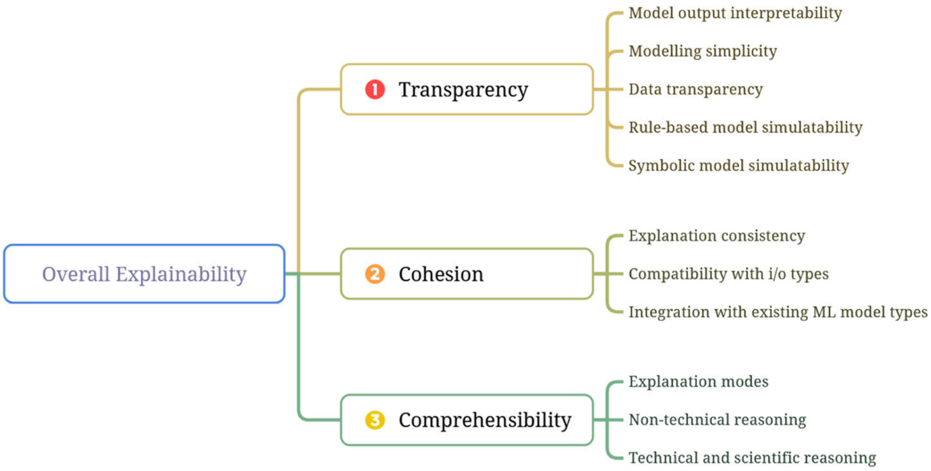
Overall Explainability
1 Transparency
Model output interpretability
Modelling simplicity
Data transparency
Rule-based model simulatability
Symbolic model simulatability
2 Cohesion
Explanation consistency
Compatibility with i/o types
Integration with existing ML model types
3 Comprehensibility
Explanation modes
Non-technical reasoning
Technical and scientific reasoning

Fig. 5 | Evaluation structure of the TCC framework. A structural framework consisting of three main components of overall explainability is illustrated. The first component, 'transparency,' emphasizes the importance of clarity in model outputs and processes. It includes four sub-criteria: model output interpretability, modelling simplicity, data transparency, and rule-based model simulatability. The second component, 'cohesion,' addresses how well the explanations integrate with existing knowledge and systems. It has three sub-components: explanation consistency, input/output compatibility, and compatibility with existing machine learning models. The third and final component, 'comprehensibility,' is evaluated through its sub-components: explanation modes, non-technical modes of explanation, and technical and scientific reasoning.

nuclear, and aerospace components). This will ensure that the required and acceptable level of transparency, as per regulatory requirements, is met.

Based on our perspectives on the role of XAI in advancing the manufacturing domain, we now propose a novel framework for explainability assessment.

Explainability evaluation framework

To address the urgent need for proper evaluation metrics and a consensus on the types and levels of interpretability, we propose a new explainability assessment framework called Transparency–Cohesion–Comprehensibility (TCC) evaluation framework, which includes its constituent elements and its substructure. The overall computational structure of the TCC evaluation framework is given in Fig. 5.

The fundamental constituents of the TCC framework are:

- Transparency: this assesses the accessibility and understandability of an AI-driven manufacturing system's inner workings and decision processes with respect to algorithms, models, data, and reasoning. Transparent systems enable users to trace how decisions are made and understand the factors influencing those decisions. It promotes trust, accountability, and the ability to identify and rectify potential issues or biases.
- Cohesion: cohesion refers to the alignment and consistency between the XAI-driven manufacturing system and established manufacturing principles, guidelines, and ethical standards. It involves ensuring that the logic employed by the AI system resonates with the goals, objectives, and best practices of the manufacturing domain. A cohesive system integrates seamlessly with existing manufacturing processes and practices, maintaining coherence and harmony between human decision-making and AI-driven decision-support.
- Comprehensibility: comprehensibility focuses on the ability of human stakeholders to interpret and understand the explanations provided by the AI-driven manufacturing system. It involves presenting information in a manner that is meaningful and interpretable to domain experts, operators, and decision-makers. Comprehensible systems bridge the gap between the technical complexities of AI and human comprehension, allowing stakeholders to validate, trust, and effectively utilize the outputs and recommendations of the AI system.

The proposed TCC evaluation framework is designed to comprehensively assess the entire explanation process, which inherently includes both the XAI algorithms and the explanation outcomes. It evaluates the transparency and simplicity of the algorithms, ensuring their inner workings are accessible and understandable. Additionally, it assesses the consistency, reliability, and comprehensibility of the explanation outcomes, ensuring they are robust and compatible with various input/output types. By focusing on the entire explanation process, the TCC framework ensures that explanations are integrated and consistent with existing models and systems. This approach facilitates the effective use of AI-driven systems in manufacturing by bridging the gap between technical complexities and human understanding.

Each aspect is evaluated on a numerical scale from 0 to 1, each representing the least and most degree of explainability (DoX), respectively. The explainability scoring scheme in terms of transparency, cohesion, and comprehensibility is detailed in Table 1.

Degree of transparency (DoT)

Transparency is the primary level of explainability that evaluates the extent to which data, training algorithms, computational structure, and prediction rationale are disclosed for subsequent use in understanding and improving the manufacturing system under consideration. The evaluation schemes and scoring criteria for various sub-categories under the transparency section are detailed as follows:

- Model output interpretability: it distinguishes between answering and explaining. In addition to accurate response predictions, an explainable manufacturing system shall answer several archetypal and interrogative questions, such as what, why, or how each decision has been made. Since such explanations are offered after the output prediction, the interpretability of the model output can be quantified by the choices of post-hoc interpretations (PHI) available. Ideally, the more questions are answered, the better performance will be. In Table 1, the explainability score is maximum (score = 1) when at least 3 PHIs are offered by the system under consideration.
- Modelling simplicity: the simplicity of developing (or training) an AI model can be estimated by considering the computational steps/iterations required for model convergence and/or the time taken for this process. These two factors effectively capture the complexity of the task and its computational requirements, respectively. However, since the computational load can often be application-specific, it is preferable to use a predefined reference value, typically a maximum threshold. If the thresholds are not user-defined, then, for model comparisons, the larger value can be considered the threshold.

Table 1 | Evaluation of explainability using the proposed TCC framework

No.	Categories	Sub-category	Evaluation criteria	Explainability score
1	Transparency	Model output interpretability	Types of post-hoc interpretations (PHI) available	$\begin{cases} p \le 3 & 0.25p \\ p > 3 & 1 \end{cases}$ where p is the no. of PHI types available
2		Modelling simplicity	Computational steps (CS), computational time (CT)	Simplicity score, refer Eq. (3)
3		Data transparency	Data transparencies (DT) in: *(1) Data acquisition (DT_{DA})*, *(2) Pre-processing (DT_{PP})*, *(3) Feature extraction (DT_{FE}), &* *(4) Dim. reduction (DT_{DR})*	$\begin{cases} d \ge 4 & 1 \\ 0 \ge d > 4 & 0.25d \end{cases}$ where d is the number of valid DTs
4		Rule-based model simulatability	Number of rules/nodes	Applicable for rule-based/tree-type explainable systems $\begin{cases} n > 100 & 0.25 \\ 50 \le n \le 100 & 0.5 \\ 10 < n < 50 & 0.75 \\ 1 \le n \le 10 & 1 \end{cases}$ where n is the no. of rules/nodes
5		Symbolic model simulatability	MC = Number of operators + Number of features (Eqn)	Complexity score, refer Eq. (5)
1	Cohesion	Explanation consistency	Objective explanations—consistent Generative explanations—inconsistent (LLM based)	$\begin{cases} \text{Computational explanations} & 1 \\ \text{Generative explanations} & 0.5 \\ \textit{Else} & 0 \end{cases}$
2		Compatibility with i/o types	Level of compatibility with input-responses (numeric, categorical, & image)	$\begin{cases} p = 1 & 0.33 \\ p = 2 & 0.66 \\ p > 2 & 1 \end{cases}$ where p is the number of i/o types for which explanations are offered
3		Integration with existing ML algorithms	Level of integration with regression (R), classification (C), and time series (T) ML models	$\begin{cases} m = 1 & 0.33 \\ m = 2 & 0.66 \\ m > 2 & 1 \end{cases}$ where m is the number of ML model types for which explanations are offered
1	Comprehensibility	Explanation modes	Levels of explainability: intrinsic (by-design), post-hoc	$\begin{cases} \textit{Intrinsic AND Posthoc} & 1 \\ \textit{Intrinsic OR Posthoc} & 0.5 \\ \textit{Else} & 0 \end{cases}$
2		Non-technical reasoning	Provision for non-expert's comprehensibility (error debugging and process control)	$\begin{cases} \textit{Yes} & 1 \\ \textit{No/NA} & 0 \end{cases}$
3		Technical & scientific insights	Provision for scientists/expert's comprehensibility for scientific discoveries/insights	$\begin{cases} \textit{Scientific discovery} & 1 \\ \textit{Scientific understanding} & 0.5 \\ \textit{No/NA} & 0 \end{cases}$

For any two arbitrary models S_1 and S_2, the computational steps and time thresholds are given as:

$$\text{Computational steps threshold, } \tau_{cs} = \begin{cases} \hat{\tau}_{cs} & \hat{\tau}_{cs} \ne null \\ \max(cs(S_1), cs(S_2), \ldots) & \hat{\tau}_{cs} = null \end{cases} \tag{1}$$

$$\text{Computational time threshold, } \tau_{cs} = \begin{cases} \hat{\tau}_{cs} & \hat{\tau}_{cs} \ne null \\ \max(cs(S_1), cs(S_2), \ldots) & \hat{\tau}_{cs} = null \end{cases} \tag{2}$$

Where $\hat{\tau}_{cs}$ and $\hat{\tau}_{ct}$ are the predefined computational steps and time thresholds, and cs and ct are functions to extract computational steps and time respectively. Equation (3) computes the simplicity score (ϵ (0,1), 1 being the best).

$$\text{Simplicity score } (S_i) = 1 - \left(w_{cs} \times \left(\frac{cs(S_i)}{\tau_{cs}}\right) + w_{cT} \times \left(\frac{ct(S_i)}{\tau_{ct}}\right)\right) \tag{3}$$

Here w_{cs} and w_{cT} are the relative weights of computational steps and time, respectively.

The underlying concept of calculating an AI model's simplicity by both computational time and steps is derived from the study conducted by Pfitscher et al.[94]. The computational time and steps can either be compared to a predefined threshold or, in the absence of such a threshold, to the largest number of steps observed. Predefined thresholds can be specific to the manufacturing process or application. For instance, real-time predictions would have much lower time thresholds compared to offline predictions. Developing a threshold database based on experimental data or expert knowledge for various manufacturing processes and applications can facilitate this computation.

In the absence of a predefined threshold, we propose using the larger computational time and steps as the threshold to compare the modelling simplicity between two AI models for the same manufacturing process. This approach provides a relative measure of simplicity between the two models, with simpler models often being more explainable. The rationale for this method is that it offers a comparative view of the models' computational efficiency, which is essential in determining the most appropriate model amongst the available alternatives for specific manufacturing applications.

- Data transparency: the data transparency score ($\epsilon \in (0,1)$) indicates the extent to which different aspects of data, including data acquisition (DA), pre-processing (PP), feature extraction (FE), and dimension reduction (DR), are transparently available for the considered manufacturing system.

- Rule-based model simulatability: simulatability measures the human-comprehensibility of a model's computing architecture. Although rule-based systems are generally regarded as simple, simulatability depends on the extent of the system, or in other words, on the number of rules (or number of nodes for a tree-type model). Following this logic, Table 1 presents the criteria for assigning a transparency score to such models.
- Symbolic model simulatability: similar to the previous criteria, the simulatability of a symbolic model, or the complexity of an analytical expression, is defined in terms of the number of features and/or the number of operators[95]. This type of scoring is appropriate for symbolic regression models where the AI model extracts a mathematical expression, correlating inputs, and responses. Surely, since the model complexity is contextual, a threshold τ_c is required to evaluate the complexity score, as given below

$$\text{Model complexity threshold, } \tau_c = \begin{cases} \hat{\tau}_c & \hat{\tau}_c \neq null \\ \max\left(c\left(S_1\right),\ c\left(S_2\right)\right) & \hat{\tau}_c = null \end{cases} \tag{4}$$

where $\hat{\tau}_c$ is a predefined complexity threshold, and c is the complexity function, which sums up the number of features and operators. Now, the complexity score is computed as

$$\text{Complexity score}\left(S_i\right) = 1 - \left(\frac{c\left(S_i\right)}{\tau_c}\right) \tag{5}$$

Once all the five sub-categories of transparency are scored, the total DoT is evaluated as

$$\text{DoT} = w_{T_1} \times T_1 + w_{T_2} \times T_2 + w_{T_3} \times T_3 + w_{T_4} \times T_4 + w_{T_5} \times T_5 \tag{6}$$

Here T_i denotes the transparency scores of each sub-category in the discussed order, and w_{T_i} indicates their corresponding weights.

Degree of cohesion (DoC)

Cohesion indicates the level of integration and consistency of the explanations with respect to existing models, systems, and data types.

- Explanation consistency: it is essential to assess the consistency and reliability of explanations across different scenarios. In smart manufacturing, additive feature attribution models like SHAP and LIME tend to exhibit greater consistency and repeatability compared to data generation-based techniques such as LLM explanations.
- Compatibility with I/O types: explanations should be robust enough to accommodate various input/output types, as shown in Fig. 1. This ensures smooth integration across manufacturing systems and domains. The ability to provide explanations for a higher number of I/O types correlates with a higher explainability score, as outlined in Table 1.

  For example, a monitoring system that employs an XAI model capable of explanations based on both sensor data (e.g., temperature and pressure readings) and visual data (e.g., images of product defects) would receive a high cohesion score, in comparison to a model which is restricted to accommodate just numeric inputs.
- Integration with existing ML model types: another important characteristic is compatibility and consistency across different types of ML models, including regression, classification, and time series models. Greater support for a variety of ML types enhances cohesion.

DoC is computed by the weighted summation of cohesion sub-category scores as follows

$$\text{DoC} = w_{C_1} \times C_1 + w_{C_2} \times C_2 + w_{C_3} \times C_3 \tag{7}$$

Here C_i and w_{C_i} indicate the individual cohesion scores and their corresponding weights in the listed order.

Degree of comprehensibility (DoCm)

Comprehensibility represents the highest level of explainability, evaluating the depth and impact of explanations. This framework assesses the explanation modes, whether explanations enable non-experts to understand and improve AI decisions, help experts grasp underlying physics to enhance similar systems, and inspire scientists towards breakthrough discoveries or autonomous scientific advancements. While current technology may not fully achieve these goals, we believe it represents a future possibility.

In this context, the comprehensibility aspects are scored as follows:

- Explanation modes: the framework evaluates the available modes of explainability to ensure maximum comprehensibility. For example, if the system offers both ante-hoc and post-hoc explanations, it is more likely to provide better comprehensibility. The scoring reflects this consideration.
- Non-technical reasoning: next, the framework evaluates whether the model offers sufficient recommendations for a non-expert to improve the manufacturing system under consideration. These recommendations may come in the form of local explanations associated with a predictive decision. Currently, a binary subjective scoring system (1 if it offers, 0 otherwise) is used for this criterion.
- Technical and scientific insights: finally, the framework assesses whether the model stimulates an expert or scientist to comprehend or reveal the physical phenomena behind the manufacturing problem under consideration. Certain models, such as symbolic regressions, are equipped with such capabilities. It also evaluates if the model is independently capable of scientific discoveries.

DoCm is computed as the weighted sum of the aforementioned scores.

$$\text{DoC}_m = w_{cm_1} \times C_{m1} + w_{cm_2} \times C_{m2} + w_{cm_3} \times C_{m3} \tag{8}$$

Here C_{mi} and w_{Cm_i} indicate the individual comprehensibility scores and their corresponding weights in the listed order.

Degree of explainability (DoX)

Once the DoT, DoC, and DoCm are computed, the overall DoX is computed as

$$\text{DoX} = w_T \times \text{DoT} + w_C \times \text{DoC} + w_{Cm} \times \text{DoCm} \tag{9}$$

Here w_T, w_c and w_{cm} represent the relative weights of transparency, cohesion, and comprehensibility, respectively, towards the total explainability of the system. Here, the subjectivity of human evaluators primarily influences the assignment of weights to various evaluation criteria. This subjective weighting reflects the relative importance of different aspects of explainability, which can be highly specific to the application and context of the AI system being assessed. One prominent method for subjective weighting is the Delphi method[96]. Such subjective weight assignment will indeed have an impact on the evaluation outcomes. Different evaluators may prioritize certain aspects of transparency, cohesion, or comprehensibility differently based on their expertise, experience, and the specific needs of the manufacturing process.

While the subjective assignment of weights introduces variability, it also adds valuable flexibility, allowing the framework to be tailored to diverse applications. For example, in some applications like biomedical or nuclear, transparency might be deemed more critical due to regulatory requirements or the necessity of understanding decision-making processes for safety reasons. In other scenarios, comprehensibility might take precedence to ensure that non-expert users can effectively interact with and trust the AI system.

Apart from the three fundamental aspects of explainability, an additional four desirable characteristics are proposed herewith. These are accountability, fairness, regulatory compliance, and user centricity. These aspects are currently not included in the scoring but can be potentially considered based on the context and application.

- Accountability: there should be mechanisms in place to track and audit the AI system's actions, allowing manufacturers to understand why specific decisions were made and facilitating post-hoc analysis, debugging, and error correction.
- Fairness: explainable systems should be designed to mitigate biases and ensure fair treatment of all stakeholders. For example, the AI model for personalized bioimplant recommendation should mitigate training data bias towards certain demographic groups.
- Regulatory compliance: the AI-driven manufacturing system should align with guidelines and standards set forth by regulatory bodies to ensure safety, privacy, data protection, and ethical considerations in the manufacturing domain.
- User centricity: the explanations provided by the system should be tailored to the knowledge level and context of the intended audience, facilitating effective communication and understanding.

A case study: explainability of fingerprint development systems

Concept of fingerprints

In the context of smart manufacturing, the process and product fingerprints (FP) represent the core contributing process parameters (including sensor features) and surface geometric characteristics respectively, towards a manufactured product's functional performance. FP development is a relatively recent concept which identifies the most significant features of a process/product (suppressing the other irrelevant ones) with respect to its end functionality. This approach ensures that controlling the FPs guarantees design compliance and functional performance, thereby substantially reducing the time and effort needed for offline metrology and separate process optimization[97].

Given that manual, statistical, and physics-based FP extraction methods have been demonstrated to be time-consuming, resource-intensive, and prone to inaccuracies, researchers have increasingly turned to ML-based FP extraction with notable success[98]. In addition to computational efficiency and other discussed merits, an implicit advantage of the FP concept is its ability to explain the underlying physical phenomena of a manufacturing process. Therefore, it is crucial for an ML-based FP extraction framework to be explainable; otherwise, it fails to provide insights into the underlying physics of the process, rendering it non-generalizable and challenging to debug. These shortcomings could cumulatively have a significant negative impact on the performance of a manufacturing system.

ML-based approach for FP development

For this case study, we have chosen two different FP extraction frameworks recently developed by us: one is random forest regression (RFR) based[98], and the other is XAI-based[79], to demonstrate the TCC framework. Both frameworks use a common dataset—nanosecond laser structuring for superhydrophobic surface fabrication.

Superhydrophobic surfaces, characterized by their extreme water repellence, have significant relevance and applications across various industries, including self-cleaning materials, anti-icing coatings, and fluid transport systems. The design and manufacturing of such surfaces are complex due to the precise control required over surface textures at the micro- and nanoscale. Laser texturing, a sophisticated manufacturing process, can generate the intricate patterns needed for superhydrophobicity, but it involves numerous process parameters and interactions that are challenging to optimize. The FP approach is particularly relevant here as it identifies the key features and parameters that most significantly impact the functional performance of the surfaces, thereby streamlining the design and manufacturing process. ML can greatly aid in this task by efficiently handling large datasets and uncovering complex relationships between process parameters and surface properties.

Figure 6 shows the details of the experiment, the correlations probed, and the resultant surfaces. Both the frameworks are shown in Fig. 7. Here the process fingerprint FP candidates are process parameters—laser power, exposure time, and pitch distance; and the product FP candidates are surface geometric candidates—S_a, S_z, S_{dr}, S_{ku}, S_{dq}, and R_{sm}. The product functional performance, the hydrophobicity of the surface, is measured in terms of contact angle (CA).

The RFR and XAI frameworks are from hereon called S_1 and S_2, respectively. The fundamental difference between the two FP development frameworks is that, while framework S_1 has chosen an exhaustive search approach combined with a black-box model for its predictions, framework S_2 uses a multi-level interpretable approach for both FP development and its subsequent predictions. Due to these fundamental differences, the final extracted FPs are also different for both frameworks as given in Eqs. (10–13). The differences also exist in the computational steps, time, and dimensionality reduction.

$$\text{Process FP}_{\text{RFR}} = \frac{\text{Power}}{\text{Pitch}^{0.75}} \tag{10}$$

$$\text{Process FP}_{\text{XAI}} = \frac{\text{Pitch}^{0.029}}{\text{Power}^{0.0315}} \tag{11}$$

$$\text{Product FP}_{\text{RFR}} = \frac{S_z^{\,0.5} \times S_{dq} \times R_{hy}}{S_{dr}^{\,2}} \tag{12}$$

$$\text{Product FP}_{\text{XAI}} = \frac{1}{S_z^{0.058} \times R_{hy}^{0.07}} \tag{13}$$

The two approaches are further discussed in the following subsections.

RFR-based approach. The overall RFR approach has the following steps:

- Parameter identification: first, identify the product functional performance, process parameters, and surface characterization parameters relevant to the manufacturing process and product requirement.
- Experimentation: conduct experiments by changing process parameters and levels, measuring and recording the functional performance and surface characterization parameters.
- FP candidate generation: use the experimental data to identify potential process and product fingerprint candidates. Consider cross-terms (products of first-order parameters) and exponential forms to represent non-linear relationships.
- ML model selection: the framework utilizes an RFR model due to its high predictive performance. The RFR model combines predictions from multiple decision trees for better accuracy.
- ML model training: train the selected ML model utilizing the leave-one-out cross-validation (LOOCV) approach. In this approach, each experiment's data is used as a test set while the rest serve as the training set. This ensures all datasets are used for both training and validation, addressing the challenges of small datasets.
- FP candidate evaluation: evaluate the FP candidates by calculating the testing error using the LOOCV approach. Use the correlation and testing error ratio (CTER) to evaluate candidates, aiming for a high correlation with product characteristics and low testing error. The 'best FP' is determined based on the highest CTER value, balancing the number of parameters used and optimization efforts.
- Final FP selection: apply a threshold (0.90 × maximum CTER) to select candidates with the minimum number of parameters. The candidate with the highest CTER above this threshold is chosen as the final process/product FP. If no other candidate meets the threshold criteria, then the 'best FP' having the highest CTER (from the previous step) is selected as the final FP.

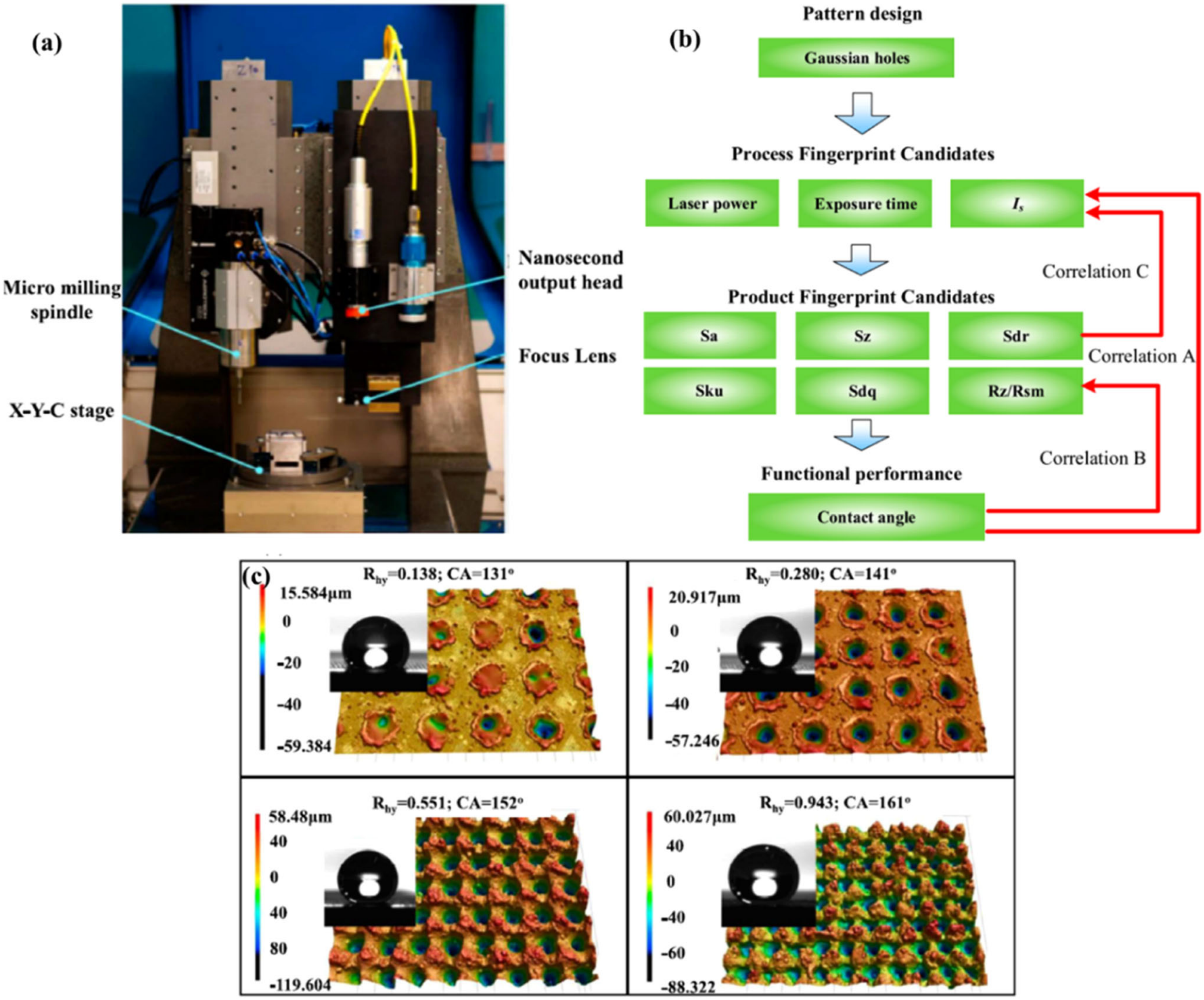

Fig. 6 | Fingerprint extraction for laser textured superhydrophobic surfaces. **a** The experimental setup of the micro-machining system, which consists of micro-milling and nanosecond laser texturing capabilities, is illustrated. **b** The concept of process and product fingerprints is presented, focusing on three specific correlations. Correlation A is between the process parameters and the end functionality, the contact angle. Correlation B connects the geometric characteristics of the product, such as S_a, S_z, S_{ku}, S_{dq}, and R_z, with the contact angle. Correlation C links the process parameters with surface features. **c** Four 3D surface topography images, obtained from an optical profilometer and produced by different process parameter combinations, are shown. The resultant contact angles shown within the topography images are 131°, 141°, 152°, and 161°.

XAI-based approach. The methodology of XAI-based approach is briefly summarized as follows:

- Process and parameter selection: this step involves identifying the manufacturing process, along with relevant process parameters or product functional performance (PF) indicators.
- Selection of FP expression: an interpretable FP expression is chosen to be transparent and simple, capable of capturing cross-interactions and non-linearities between process parameters/surface characteristics and product functionality.
- Optimization algorithm: a global optimization algorithm, such as MaxLIPO, is selected to extract the best process and product FP. This algorithm ensures the best correlation between the FP and product functionality by iteratively adjusting the unknown parameters in the FP expression.
- Dimensionality reduction: recurrent feature elimination (RFE) is used to reduce the dimensionality of the FP by eliminating the minimum contributing features in iterative steps until a termination criterion, typically a drop-in correlation coefficient below a certain threshold, is reached.
- Response prediction: for continuous responses, an ML model, particularly a stacked ensemble model, is trained to predict responses based on the extracted process/product FP. Stacked ensemble models combine the predictions of multiple base models to improve accuracy.
- Probabilistic model for event detection: for categorical responses, a probabilistic model is developed using process FP values to detect manufacturing events. This involves mapping FP values to probabilities using decision boundaries and training stacked ensemble models for event prediction.
- Post-hoc explanations: SHAP is used to provide post-hoc explanations for stacked ensemble model predictions. These explanations offer insights into both global and local interpretations of the model's decisions. Various post-hoc explanations are offered, including local and global feature importance and partial dependency analysis.

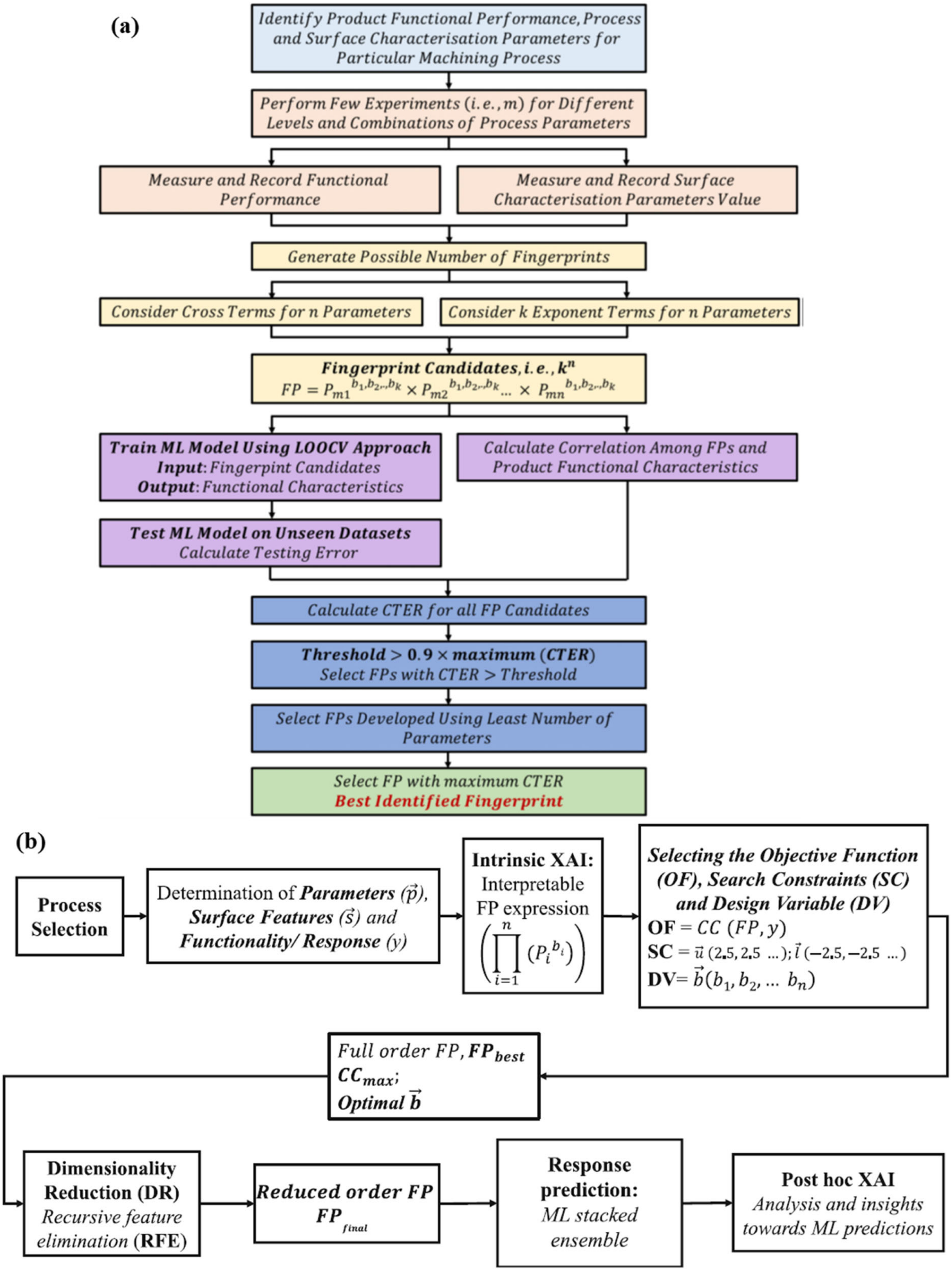
(a)
Identify Product Functional Performance, Process and Surface Characterisation Parameters for Particular Machining Process
Perform Few Experiments (i.e., m) for Different Levels and Combinations of Process Parameters
Measure and Record Functional Performance
Measure and Record Surface Characterisation Parameters Value
Generate Possible Number of Fingerprints
Consider Cross Terms for n Parameters
Consider k Exponent Terms for n Parameters
Fingerprint Candidates, i.e., k^n
$FP = P_{m1}^{b_1,b_2,..,b_k} \times P_{m2}^{b_1,b_2,..,b_k} \ldots \times P_{mn}^{b_1,b_2,..,b_k}$
Train ML Model Using LOOCV Approach
Input: Fingerpint Candidates
Output: Functional Characteristics
Calculate Correlation Among FPs and Product Functional Characteristics
Test ML Model on Unseen Datasets
Calculate Testing Error
Calculate CTER for all FP Candidates
Threshold > 0.9 × maximum (CTER)
Select FPs with CTER > Threshold
Select FPs Developed Using Least Number of Parameters
Select FP with maximum CTER
Best Identified Fingerprint
(b)
Process Selection
Determination of Parameters ($\vec{p}$), Surface Features ($\vec{s}$) and Functionality/ Response (y)
Intrinsic XAI: Interpretable FP expression
$\left(\prod_{i=1}^{n} (P_i^{b_i})\right)$
Selecting the Objective Function (OF), Search Constraints (SC) and Design Variable (DV)
OF = $CC\ (FP, y)$
SC = $\vec{u}$ (2.5, 2.5 …); $\vec{l}$ (−2.5, −2.5 …)
DV= $\vec{b}(b_1, b_2, \ldots\ b_n)$
Full order FP, FP_{best}
CC_{max};
Optimal $\vec{b}$
Dimensionality Reduction (DR)
Recursive feature elimination (RFE)
Reduced order FP
FP_{final}
Response prediction:
ML stacked ensemble
Post hoc XAI
Analysis and insights towards ML predictions

Fig. 7 | Fingerprint generation frameworks. **a** The overall RFR approach, as illustrated, involves identifying relevant parameters, conducting experiments, generating potential fingerprint candidates, selecting and training an RFR model using LOOCV, and evaluating candidates based on testing error and correlation to select the final process/product fingerprint with the highest CTER value and minimum parameters[98]. **b** The XAI-based approach, as shown, involves selecting process parameters, extracting an optimal, interpretable FP expression with MaxLIPO, and reducing dimensionality using recurrent feature elimination. It uses stacked ensemble models for continuous response prediction and probabilistic models for event detection, with SHAP providing post-hoc explanations for model insights[99].

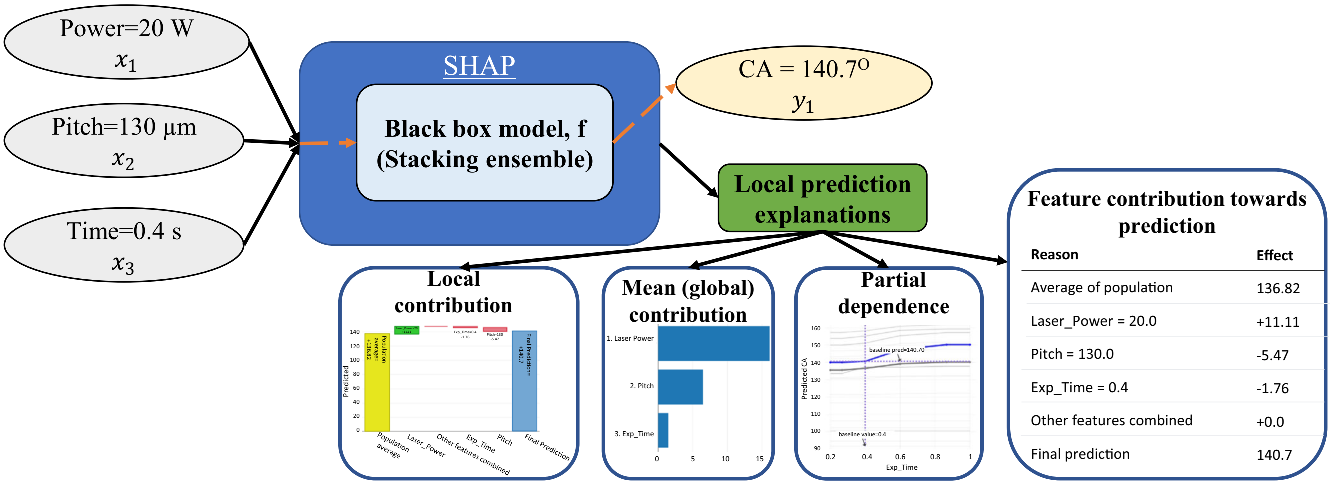

Fig. 8 | Post-hoc explanations offered by XAI approach[99]**.** SHAP-based post-hoc explanations for the stacked ensemble prediction of CA for a given process parameter combination (power = 20 W; pitch = 130 µm; time = 0.4 s) are illustrated. Various forms of explanations offered include local and global feature contribution visualizations, as well as partial dependence plots.

Table 2 | Explainability evaluation using the TCC framework

TCC categories	Sub-category	Explainability evaluation details		Scores		Overall scores		Total score	
		RFR based (S_1)	XAI based (S_2)	S_1	S_2	S_1	S_2	S_1	S_2
Transparency	Model output interpretability	NA	Local feature contribution Global contribution Partial dependence	0	0.75	0.22	0.63	0.14	0.74
	Modelling simplicity	CS: 531,441 steps CT: 19.19 h	CS: 2000 steps CT: <1 min	0	0.99				
	Data transparency	Transparent	Transparent	1	1				
	Rule transparency	NA	NA	0	0				
	Model complexity	Operators: 5 Features: 4	Operators: 4 Features: 2	0.1	0.4				
Cohesion	Explanation consistency	NA	Computational explanations	0	1	0.22	0.77		
	Compatibility with I/O types	I/O: numeric	I/O: numeric, categorical	0.33	0.66				
	Integration with existing ML algorithms	Explanation support for ML types: R	Explanation support for ML types: R, C	0.33	0.66				
Comprehensibility	Explanation modes	NA	Intrinsic & post-hoc	0	1	0.00	0.83		
	Non-technical reasoning	NA	Supported	0	1				
	Technical & scientific insights	NA	Supported	0	0.5				

The explainability of the RFR-based and XAI-based approaches varies significantly due to their inherent methodologies and the transparency of their processes. The RFR-based approach leverages a black-box RFR, whose overall complexity and the aggregation of multiple decision trees make it challenging to derive clear insights into the underlying physical phenomena of the manufacturing process. On the other hand, the XAI-based approach is specifically designed for interpretability, utilizing a transparent FP expression and optimization algorithms to capture cross-interactions and nonlinearities. As shown in Fig. 8, post-hoc explanations provided by SHAP enhance this approach by offering detailed insights into how each feature influences the model's predictions, both globally and locally.

It will indeed be an interesting endeavour to quantify and compare the overall explainability of these frameworks, with respect to a common dataset, as discussed in the next section.

Explainability of the FP development approach

Now, the proposed TCC evaluation framework is used to quantify and compare the explainability of both systems in terms of transparency, cohesion, and comprehensibility, as detailed in Table 2. Throughout this case study, uniform weighting is applied where applicable. The explainability score computations are further elaborated in the below subsections.

DoT calculation.

- Model output interpretability (T_1): here the evaluation criteria are the number of PHIs offered. RFR approach (S_1) does not offer any post-hoc explanations and hence is scored 0. XAI approach (S_2) offers three PHIs: partial dependency analysis and global and local feature importance, as given in Fig. 8. Since, the number of PHI types available, $p = 3$, the XAI approach gets a score of $0.25 \times 3 = 0.75$, as given in Table 1.
- Modelling simplicity (T_2): the simplicity score is computed based on computational steps and time. As given in ref. 98, computational steps

and time for S_1 are 531,441 steps and 1151.4 min, respectively. On the other hand, for S_2, the computational steps and time for convergence are <2000 steps and 1 min[99].

Based on Eq. (3), the simplicity score is computed for S_1 and S_2

$$\text{Simplicity score}\left(S_1\right) = 1 - \left(0.5 \times \left(\frac{531441}{531441}\right) + 0.5 \times \left(\frac{1151.4}{1151.4}\right)\right) = 0$$

$$\text{Simplicity score}\left(S_2\right) = 1 - \left(0.5 \times \left(\frac{2000}{531441}\right) + 0.5 \times \left(\frac{1}{1151.4}\right)\right) = 0.99$$

Here the threshold for computational steps (τ_{cs}) and time (τ_{ct}) is not predefined. Thus, following Eqs. (1) and (2), the steps and time threshold are considered as the maximum value among S_1 and S_2 as 531,441 steps and 1151.4 min, respectively.

- Data transparency (T_3): data are transparent for both S_1 and S_2 during all the stages: acquisition, processing, feature extraction, and dimensionality reduction. Hence both S_1 and S_2 are assigned the maximum score of 1.
- Rule-based simulatability (T_4): both S_1 and S_2 are not rule-based approaches and hence this criterion is not applicable.
- Symbolic model simulatability (T_5): the complexity function, $c(S_i)$, sums up the number of features and operators, as explained in the 'Degree of transparency (DoT)' section. From final FP Eqs. (12) and (13), it can be noted that the complexity function values for S_1 and S_2 are 9 and 6, respectively.

Now, based on Eq. (5), the complexity score is calculated as follows:

$$\text{Complexity score}\left(S_1\right) = 1 - (9/10) = 0.1$$

$$\text{Complexity score}\left(S_2\right) = 1 - (6/10) = 0.4$$

Contrary to the previous case, the threshold (τ_c) is predefined as $\hat{\tau}_c = 10$.

Based on the individual transparency scores, the overall DoT of the models is computed using Eq. (6) as

$$\text{DoT}\left(S_1\right) = 0.2 \times 0 + 0.2 \times 0 + 0.2 \times 1 + 0.2 \times 0 + 0.2 \times 0.1 = 0.22$$

$$\text{DoT}\left(S_2\right) = 0.2 \times 0.75 + 0.2 \times 0.99 + 0.2 \times 1 + 0.2 \times 0 + 0.2 \times 0.4 = 0.63$$

DoC calculation.

- Explanation consistency: post-hoc explanations like feature importance and partial dependency are objective computations and hence consistent. In comparison, generative explanations based on LLMs will be less consistent. S_1 does not offer any explanations either objective or generative, and hence is scored 0. On the other hand, since S_2 uses intrinsic XAI with SHAP, its explanations are objective and thus consistent. S_2 is thus scored 1.
- I/O compatibility: this criterion evaluates the robustness to handle various I/O types. S_1 can accommodate only numeric data ($p = 1$) and is thus assigned a score of 0.33 as given in Table 1. S_2 follows a generic architecture and offers explanations to numeric and categorical inputs (hence, $p = 2$) and is assigned a score of 0.66. To handle categorical responses, a probability mapping scheme is embedded within S_2, as detailed in the 'XAI-based approach' section.
- ML model integration: compatibility with various types of ML models is scored under this criterion. S_1 is limited to handling just regression tasks and is given a score of 0.33. S_2 can handle regression and classification, owing to its integrated probability mapping scheme, and is thus scored 0.66.

Based on the individual cohesion scores, the overall DoC of the models is computed using Eq. (7) as

$$\text{DoC}\left(S_1\right) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0.33 + \frac{1}{3} \times 0.33 = 0.22$$

$$\text{DoC}\left(S_2\right) = \frac{1}{3} \times 1 + \frac{1}{3} \times 0.66 + \frac{1}{3} \times 0.66 = 0.77$$

DoCm calculation.

- Explanation modes: here the capacity to offer various modes of explanations, like intrinsic and post-hoc, are scored. S_1 does not offer any modes of explanation and is scored 0. S_2, however, offers both intrinsic and post-hoc explainability and is scored 1.
- Non-technical reasoning: the ability to provide clarity in the model's local predictions aiding model debugging and process control is assessed. S_1 with no capability for local prediction explanations is scored 0. S_2, however, could assist non-experts with non-technical reasoning through its partial dependency and local feature contribution analytics and is scored 1.
- Technical and scientific insights: S_1 does not offer any technical or scientific insights due to its black-box nature, thus scored 0. S_2 can provide a further scientific understanding of the manufacturing process, however, is not yet capable of discoveries, and is scored 0.5.

Based on the individual comprehensibility scores, the overall DoCm of the models is computed using Eq. (8) as

$$\text{DoCm}\left(S_1\right) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

$$\text{DoCm}\left(S_2\right) = \frac{1}{3} \times 1 + \frac{1}{3} \times 1 + \frac{1}{3} \times 0.5 = 0.83$$

Finally, based on the DoT, DoC, and DoCm values of the models S_1 and S_2, DoX is computed using Eq. (9) as

$$\text{DoX}\left(S_1\right) = \frac{1}{3} \times 0.22 + \frac{1}{3} \times 0.22 + \frac{1}{3} \times 0 = 0.146$$

$$\text{DoX}\left(S_2\right) = \frac{1}{3} \times 0.63 + \frac{1}{3} \times 0.77 + \frac{1}{3} \times 0.83 = 0.74$$

The comparison of the explainability of the considered FP development approaches in terms of their explainability score, as per the proposed TCC framework is summarized as:

- S_2 is 3 times more transparent than S_1.
- In terms of cohesion, S_2 is 3.5 times better than S_1.
- S_1 does not offer any scientific insights, while S_2 has a high degree of comprehensibility.
- Overall, S_2 is 5 times more explainable than S_1.

The case study presented here employed a uniform weighing scheme for explanation categories and sub-categories. However, non-uniform subjective weighting methods like Delphi[96] can be applied based on the relative importance of different aspects of explainability and on the application and context of the AI system being assessed. This ensures a more precise evaluation of explainability tailored to specific applications. For instance, in manufacturing contexts, such as in biomedical or nuclear industries, transparency can be especially crucial. This is often due to regulatory demands and the need for clear decision-making processes to ensure safety. Conversely, in other situations, ensuring the AI system is comprehensible might be more important. This helps non-expert operators to interact with the system effectively and build trust in its operations.

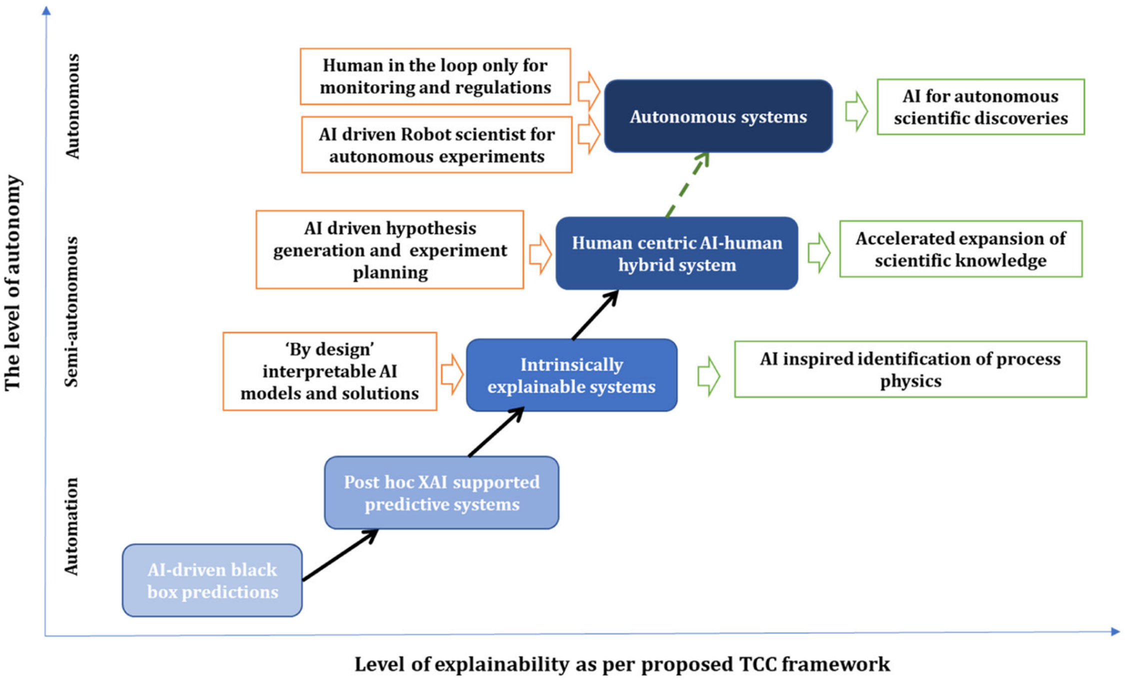
The level of autonomy
Autonomous
Semi-autonomous
Automation
Human in the loop only for monitoring and regulations
AI driven Robot scientist for autonomous experiments
Autonomous systems
AI for autonomous scientific discoveries
AI driven hypothesis generation and experiment planning
Human centric AI-human hybrid system
Accelerated expansion of scientific knowledge
'By design' interpretable AI models and solutions
Intrinsically explainable systems
AI inspired identification of process physics
Post hoc XAI supported predictive systems
AI-driven black box predictions
Level of explainability as per proposed TCC framework

Fig. 9 | The proposed explainability–autonomy correlation. This figure presents the proposed correlation between explainability and autonomy. The diagram demonstrates how the level of autonomy increases in relation to the increase in explainability, according to the TCC evaluation framework. Both explainability and the level of autonomy are at their lowest for black-box model predictions. Advancements in AI have resulted in various levels of explainability, namely, post-hoc explainability, intrinsic explainability (model is by-design explainable and can reveal process physics), hybrid human–AI systems (where AI can suggest/assist in hypothesis generation and the results can enhance scientific understanding), and autonomous AI systems (AI can independently solve hypotheses, leading to scientific discoveries). These advancements are illustrated to enhance both explainability and autonomy in the listed order.

Discussion and insights

The TCC framework presents a structured and systematic approach to evaluate the explainability of AI-driven manufacturing systems, addressing a critical gap in the current landscape of explainability metrics. By decomposing explainability into three fundamental components—transparency, cohesion, and comprehensibility—the framework provides a comprehensive assessment that encompasses not only the technical transparency of models but also their alignment with domain-specific principles and their interpretability for various stakeholders. The proposed approach ensures that the evaluation is holistic, taking into account different aspects that contribute to the overall explainability of AI systems in complex environments like manufacturing. It is essential to highlight that the framework is both generalizable and flexible, allowing for the incorporation of multiple scoring, ranking, and weighting schemes tailored to diverse manufacturing applications and systems. The case study can be viewed as a successful initial step, and the framework holds the potential for significant expansion through additional case studies and discussions.

However, the framework has several opportunities for enhancement through future investigations and improvements. One of the present challenges is the subjective nature of weighting and scoring within the TCC categories. For instance, assigning weights to different sub-categories (e.g., modelling simplicity, explanation modes) can introduce biases based on the evaluator's perspective or the specific application context. Additionally, the binary scoring for certain aspects, such as non-technical reasoning and scientific insights, might oversimplify the nuances of explainability. Finally, accommodating and quantifying explainability in the latest generative AI-based manufacturing systems poses a potential future challenge. While not currently a significant issue due to the limited use of generative AI in manufacturing, this could become more relevant as the adoption of generative AI increases, necessitating further adaptations of the framework.

Despite these challenges, the TCC framework represents a significant step forward in the systematic evaluation of AI explainability, offering a robust foundation for enhancing trust, accountability, and usability of AI systems across various applications.

Manufacturing autonomy through explainability

We argue that explainability is vital to achieve autonomy in smart manufacturing. While some of the proposed routes may initially be resource-intensive and operationally expensive, we believe that in the long run, trustworthy and explainable systems will only be able to attract wider acceptance among stakeholders, regulatory bodies, and policymakers. This, in turn, will lead to better investments in this direction and cover up for the initial expenses.

Until now, there existed a false perception that an AI model's predictive performance is the key to trusting it with autonomous decisions. However, we argue that this is not the case, especially for high-stakes manufacturing applications. From now on, more than accuracy, explainability will determine the level of autonomy in manufacturing systems. The proposed explainability–autonomy correlation is shown in Fig. 9. Currently, we are at the stage of intrinsically explainable systems, which can inspire an understanding of process physics.

The integration of explainability will progress towards novel modes of collaboration between humans and computers, leading to hybrid intelligence. At this level, scientific understanding will accelerate, but unexpected discoveries are still unlikely. This is because human scientists still hold the autonomy of defining problems, prioritizing them, and predefining them from the set of potential AI-suggested hypotheses. While the AI–human collaborative approach is indeed an improvement over traditional methods, there is still some lost potential since seemingly less relevant research problems, which could have later led to major discoveries, may be screened out by human scientists.

The synergies between human intelligence and AI intelligence are crucial in this context. Human intelligence excels in creative problem-solving, contextual understanding, and ethical judgement, while AI intelligence is characterized by its ability to process vast amounts of data, identify patterns, and perform repetitive tasks with high precision and consistency. By integrating XAI, we can ensure that AI systems not only provide accurate insights but also offer transparency and interpretability in their decision-making processes. This combination can lead to transformative advancements in manufacturing systems. Collaborative decision-making, where human intuition and contextual knowledge combine with XAI's data-driven insights, can lead to higher efficiency and innovation. Hybrid intelligence systems, where humans and AI work together seamlessly, create more robust and adaptable manufacturing processes. XAI can augment human capabilities, providing tools and insights that were previously inaccessible and can continuously learn and adapt, guided by human expertise. Moreover, human supervision ensures that the deployment of XAI in manufacturing is ethical and fair, fostering trust and facilitating the adoption of AI-driven solutions across the industry.

As the level of explainability progresses, a stage is reached where the attained trust, acceptability, and investments make it feasible to explore even seemingly low-stakes hypothesis spaces. This is done with the anticipation that it will ultimately yield high-value outcomes. At this point, AI can engage in an unrestricted exploration of the generated hypothesis space, subsequently planning and executing experiments autonomously with robot scientists. Human involvement mainly revolves around monitoring and overseeing the entire process. Thus, at the highest level AI will grow into a stage where it can be implemented for the discovery of novel unseen concepts. Ultimately, a series of new discoveries will be integrated into an integrated model that is large-scale, high-precision, and in-depth.

In summary, this perspective paper highlights the evolving landscape of XAI in smart manufacturing, emphasizing transparency and interpretability in AI-driven decision-making. We have shared some implicit challenges that might be causing the slow adaptation of XAI in smart manufacturing. We have explored AI's transformative potential in generating scientific understanding within manufacturing, envisioning AI as an autonomous investigator driving innovation and informed decision-making. The transition from weak AI to ultra-strong AI heralds a new era of seamless collaboration between machines and humans, bridging data-driven AI and domain expertise.

Furthermore, a TCC evaluation framework is proposed, which offers a structured and systematic approach to the evaluation of the explainability of smart manufacturing systems. The framework underscores the importance of transparency, cohesion, and comprehensibility of explanations. It introduces a detailed evaluation criterion which has been demonstrated through a case study.

The road ahead for AI in manufacturing is paved with challenges, but also with immense possibilities. By embracing explainability, manufacturers can not only enhance the quality and productivity of their products but also gain the trust of regulators, stakeholders, and the public. The transition from Industry 4.0 to Industry 5.0 is driven by the demand for transparency and accountability, and regulatory bodies are taking steps to ensure responsible AI integration. The responsible integration of XAI into manufacturing processes is not just a technological advancement but a paradigm shift that will shape the future of smart manufacturing, ensuring that it is both technically advanced and morally sound.

Data availability

No datasets were generated or analysed during the current study.

Received: 11 April 2024; Accepted: 2 September 2024;
Published online: 06 November 2024

References

1. Wang, J., Ma, Y., Zhang, L., Gao, R. X. & Wu, D. Deep learning for smart manufacturing: methods and applications. *J. Manuf. Syst.* **48**, 144–156 (2018).
2. Ahmed, I., Jeon, G. & Piccialli, F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Trans. Ind. Inform.* **18**, 5031–5042 (2022).
3. Tercan, H. & Meisen, T. Machine learning and deep learning based predictive quality in manufacturing: a systematic review. *J. Intell. Manuf.* **33**, 1879–1905 (2022).
4. Rudin, C. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nat. Rev. Methods Prim.* **2**, 1–2 (2022). *2022 21*.
5. Gunning, D. et al. XAI-Explainable artificial intelligence. *Sci. Robot.* **4**, eaay7120 (2019).
6. Rožanec, J. M. et al. Human-centric artificial intelligence architecture for industry 5.0 applications. *Int. J. Prod. Res.* **2023**, 6847–6872 (2022).
7. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328 (2021).
8. Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A. & Besold, T. Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Mach. Learn.* **107**, 1119–1140 (2018).
9. Guo, W., Gawade, V., Zhang, B. & Guo, Y. Explainable AI for layer-wise emission prediction in laser fusion. *CIRP Ann.* **72**, 437–440 (2023).
10. Hanchate, A., Bukkapatnam, S. T. S., Lee, K. H., Srivastava, A. & Kumara, S. Explainable AI (XAI)-driven vibration sensing scheme for surface quality monitoring in a smart surface grinding process. *J. Manuf. Process.* **99**, 184–194 (2023).
11. Wang, Y. C. & Chen, T. New XAI tools for selecting suitable 3D printing facilities in ubiquitous manufacturing. *Complex Intell. Syst.* **9**, 6813–6829 (2023).
12. Chen, T. C. T. Explainable artificial intelligence (XAI) in manufacturing. in *SpringerBriefs in Applied Sciences and Technology* 1–11 (Springer Science and Business Media Deutschland GmbH, 2023). https://doi.org/10.1007/978-3-031-27961-4_1.
13. Baum, D., Baum, K., Gros, T. P. & Wolf, V. XAI requirements in smart production processes: a case study. in *Communications in Computer and Information Science CCIS* Vol. 1901, 3–24 (Springer Science and Business Media Deutschland GmbH, 2023).
14. Montavon, G., Samek, W. & Müller, K. R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **73**, 1–15 https://doi.org/10.1016/j.dsp.2017.10.011 (2018).
15. Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surv*. **51**, (2018).
16. Barredo Arrieta, A. et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
17. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61**, 35–43 (2018).
18. Luo, X., Liu, Q., Madathil, A. P. & Xie, W. Predictive digital twin-driven dynamic error control for slow-tool-servo ultraprecision diamond turning. *CIRP Ann.* **73**, 377–380 (2024).
19. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* **96**, 3–14 (2002).
20. Bursac, Z., Gauss, C. H., Williams, D. K. & Hosmer, D. W. Purposeful selection of variables in logistic regression. *Source Code Biol. Med.* **3**, 17 (2008).
21. Mood, C. Logistic regression: why we cannot do what we think we can do, and what we can do about it. *Eur. Sociol. Rev.* **26**, 67–82 (2010).
22. Rokach, L & Maimon, O. *Data Mining with Decision Trees: Theory and Applications* 2nd edn, Vol. 81, 1–305 (2014).
23. Li, L., Umbach, D. M., Terry, P. & Taylor, J. A. Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics* **20**, 1638–1640 (2004).
24. Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. An kNN model-based approach and its application in text categorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. **2945**, 559–570 (2004).
25. Angelov, P. & Yager, R. A new type of simplified fuzzy rule-based system. *Int. J. Gen. Syst.* **41**, 163–185 (2012).

26. Núñez, H., Angulo, C. & Català, A. Rule-based learning systems for support vector machines. *Neural Process. Lett.* **24**, 1–18 (2006).
27. Synnaeve, G. & Bessière, P. A Bayesian model for opening prediction in RTS games with application to StarCraft. In *2011 IEEE Conference on Computational Intelligence and Games, CIG* 281–288 https://doi.org/10.1109/CIG.2011.6032018 (IEEE, 2011).
28. Taylan, P., Weber, G. W. & Beck, A. New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and techology. *Optimization* **56**, 675–698 (2007).
29. Moradi, M. & Samwald, M. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Syst. Appl.* **165**, 113941 (2021).
30. de Sousa, I. P., Vellasco, M. M. B. R. & da Silva, E. C. Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors* **19**, 2969 (2019).
31. Ramamurthy, K. N., Vinzamuri, B., Zhang, Y. & Dhurandhar, A. Model agnostic multilevel explanations. In *Advances in Neural Information Processing Systems* Vol. 2020-Decem 5968–5979 (2020).
32. Zafar, M. R. & Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach. Learn. Knowl. Extr.* **3**, 525–541 (2021).
33. Plumb, G., Molitor, D. & Talwalkar, A. Model agnostic supervised local explanations. in *Advances in Neural Information Processing Systems* Vol. 2018-Decem 2515–2524 (2018).
34. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Vol. 13-17-Augu, 1135–1144 (Association for Computing Machinery, 2016).
35. Ribeiro, M. T., Singh, S. & Guestrin, C. Nothing else matters: model-agnostic explanations by identifying prediction invariance (2016).
36. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* 2017-Decem, 4766–4775 (2017).
37. Deng, H. Interpreting tree ensembles with inTrees. *Int. J. Data Sci. Anal.* **7**, 277–287 (2019).
38. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (CRC Press, 2017). https://doi.org/10.1201/9781315139470.
39. Fu, X., Ong, C., Keerthi, S., Hung, G. G. & Goh, L. Extracting the knowledge embedded in support vector machines. In *IEEE International Conference on Neural Networks—Conference Proceedings* Vol. **1**, 291–296 (IEEE, 2004).
40. Üstün, B., Melssen, W. J. & Buydens, L. M. C. Visualisation and interpretation of support vector regression models. *Anal. Chim. Acta* **595**, 299–309 (2007).
41. Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K. R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **65**, 211–222 (2017).
42. Zeiler, M. D., Taylor, G. W. & Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In *Proc. IEEE International Conference on Computer Vision* 2018–2025 https://doi.org/10.1109/ICCV.2011.6126474 (2011).
43. Zhang, Q., Wu, Y. N. & Zhu, S. C. Interpretable convolutional neural networks. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 8827–8836 (IEEE Computer Society, 2018). https://doi.org/10.1109/CVPR.2018.00920.
44. Xiao, T. et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Vol. 07-12-June 842–850 (IEEE Computer Society, 2015).
45. Arras, L., Montavon, G., Müller, K. R. & Samek, W. Explaining recurrent neural network predictions in sentiment analysis. In *EMNLP 2017 - 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2017—Proc. Workshop* 159–168 (Association for Computational Linguistics (ACL), 2017). https://doi.org/10.18653/v1/w17-5221.
46. Rajani, N. F. & Mooney, R. J. Stackingwith auxiliary features for visual question answering. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proc. Conference* Vol. 1, 2217–2226 (Association for Computational Linguistics (ACL), 2018).
47. Chen, H., Lundberg, S. & Lee, S. I. Explaining models by propagating Shapley values of local components. in *Studies in Computational Intelligence* Vol. 914, 261–270 (Springer Science and Business Media Deutschland GmbH, 2021).
48. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML* Vol. 7, 5109–5118 (PMLR, 2017).
49. Alzetta, F., Giorgini, P., Najjar, A., Schumacher, M. I. & Calvaresi, D. In-time explainability in multi-agent systems: challenges, opportunities, and roadmap. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence, LNAI and Lecture Notes in Bioinformatics)* Vol. 12175, 39–53 (Springer, 2020).
50. Gunning, D. & Aha, D. W. DARPA's explainable artificial intelligence program. *AI Mag.* **40**, 44–58 (2019).
51. Hagras, H. Toward human-understandable, explainable AI. *Computer* **51**, 28–36 (2018).
52. Hussain, F., Hussain, R. & Hossain, E. Explainable artificial intelligence (XAI): an engineering perspective. Preprint at https://doi.org/10.48550/arXiv.2101.03613 (2021).
53. Chromik, M. & Butz, A. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Lecture Notes in Computer Science, LNCS (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 12933, 619–640 (Springer Science and Business Media Deutschland GmbH, 2021).
54. Hendricks, L. A. et al. Generating visual explanations. In *Lecture Notes in Computer Science*, *LNCS (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 9908, 3–19 (Springer Verlag, 2016).
55. Li, C. & Yang, H. J. Bot-X: an AI-based virtual assistant for intelligent manufacturing. *Multiagent Grid Syst.* **17**, 1–14 (2021).
56. Trappey, A. J. C., Trappey, C. V., Chao, M. H. & Wu, C. T. VR-enabled engineering consultation chatbot for integrated and intelligent manufacturing services. *J. Ind. Inf. Integr.* **26**, 100331 (2022).
57. Islam, M. R., Ahmed, M. U., Barua, S. & Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* **12**, 1353 (2022).
58. Walker, C. et al. Digital twin of dynamic error of a collaborative robot. In *European Society for Precision Engineering and Nanotechnology, Conference Proceedings—23rd International Conference and Exhibition, EUSPEN* 309–312 (2023).
59. Goldman, C. V., Baltaxe, M., Chakraborty, D., Arinez, J. & Diaz, C. E. Interpreting learning models in manufacturing processes: towards explainable AI methods to improve trust in classifier predictions. *J. Ind. Inf. Integr.* **33**, 100439 (2023).
60. Lee, M., Jeon, J. & Lee, H. Explainable AI for domain experts: a post Hoc analysis of deep learning for defect classification of TFT–LCD panels. *J. Intell. Manuf.* **33**, 1747–1759 (2022).
61. McLaughlin, M. P. et al. Enhanced defect detection in after develop inspection with machine learning disposition. In *ASMC (Advanced Semiconductor Manufacturing Conference) Proceedings* Vol. 2021-May (Institute of Electrical and Electronics Engineers Inc., 2021).
62. Glock, A. C. Explaining a random forest with the difference of two ARIMA models in an industrial fault detection scenario. In *Proc. Computer Science* Vol. 180, 476–481 (Elsevier, 2021).

63. Alvanpour, A., Das, S. K., Robinson, C. K., Nasraoui, O. & Popa, D. Robot failure mode prediction with explainable machine learning. *IEEE Int. Conf. Autom. Sci. Eng.* **2020-Augus**, 61–66 (2020).
64. Matzka, S. Explainable artificial intelligence for predictive maintenance applications. In *Proc.—2020 3rd International Conference on Artificial Intelligence for Industries AI4I* 69–74 https://doi.org/10.1109/AI4I49448.2020.00023 (2020).
65. Torcianti, A. & Matzka, S. Explainable artificial intelligence for predictive maintenance applications using a local surrogate model. In *Proc.—2021 4th International Conference on Artificial Intelligence for Industries AI4I* 86–88 https://doi.org/10.1109/AI4I51902.2021.00029 (2021).
66. Hermansa, M. et al. Sensor-based predictive maintenance with reduction of false alarms—a case study in heavy industry. *Sensors* **22**, 226 (2022).
67. Wang, J., Liu, C., Zhu, M., Guo, P. & Hu, Y. Sensor data based system-level anomaly prediction for smart manufacturing. In *Proc.—2018 IEEE International Congress on Big Data, BigData Congress 2018—Part of the 2018 IEEE World Congress on Services* 158–165 (Institute of Electrical and Electronics Engineers Inc., 2018). https://doi.org/10.1109/BigDataCongress.2018.00028.
68. Gribbestad, M., Hassan, M. U., Hameed, I. A. & Sundli, K. Health monitoring of air compressors using reconstruction-based deep learning for anomaly detection with increased transparency. *Entropy* **23**, 83 (2021).
69. Chowdhury, D., Sinha, A. & Das, D. XAI-3DP: diagnosis and understanding faults of 3-D printer with explainable ensemble AI. *IEEE Sensors Lett.* **7**, 1–4 (2023).
70. Kusiak, A. Federated explainable artificial intelligence (fXAI): a digital manufacturing perspective. *Int. J. Prod. Res.* **62**, 171–182 (2023).
71. Rožanec, J. M. et al. STARdom: an architecture for trusted and secure human-centered manufacturing systems. In *IFIP Advances in Information and Communication Technology* Vol. 633, 199–207 (Springer Science and Business Media Deutschland GmbH, 2021).
72. Senoner, J., Netland, T. & Feuerriegel, S. Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Manage. Sci*. https://doi.org/10.1287/mnsc.2021.4190 (2021).
73. Holm, E. A. In defense of the black box. *Science* **364**, 26–27 (2019).
74. Mohammadi, B., Malik, N., Derdenger, T. & Srinivasan, K. Sell Me the Black Box! Regulating eXplainable AI (XAI) May Harm Consumers. *arXiv* 1–17.
75. Katharine Miller. *Should AI Models Be Explainable? That Depends* https://hai.stanford.edu/news/should-ai-models-be-explainable-depends (Stanford University Human-Centered Artificial Intelligence, 2021).
76. Wald, B. *Making AI More 'Explainable' in Health-Care Settings May Lead to More Mistakes: U of T Researcher* https://www.utoronto.ca/news/making-ai-more-explainable-health-care-settings-may-lead-more-mistakes-u-t-researcher (University of Toronto, 2020).
77. Kulkarni, P. G. et al. Overcoming challenges and innovations in orthopedic prosthesis design: an interdisciplinary perspective. *Biomed. Mater. Devices* **1**, 1–12 (2023).
78. Farah, L. et al. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. *Mayo Clin. Proc. Digit. Health* **1**, 120–138 (2023).
79. Arbelaez Ossa, L. et al. Re-focusing explainability in medicine. *Digit. Health* **8**, https://doi.org/10.1177/20552076221074488 (2022).
80. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
81. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS Commun.* **9**, 793–805 (2019).
82. Udrescu, S. M. & Tegmark, M. AI Feynman: a physics-inspired method for symbolic regression. *Sci. Adv.* **6**, eaay2631 (2020).
83. Wilstrup, C. & Cave, C. Combining symbolic regression with the Cox proportional hazards model improves prediction of heart failure deaths. *BMC Med. Inform. Decis. Mak.* **22**, 1–7 (2022).
84. René Broløs, K. et al. *An Approach to Symbolic Regression Using Feyn* (2021).
85. Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Syst. Biol. Appl.* **7**, 1–12 (2021).
86. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769 (2022).
87. Grizou, J., Points, L. J., Sharma, A. & Cronin, L. A curious formulation robot enables the discovery of a novel protocell behavior. *Sci. Adv.* **6**, eaay4237 (2020).
88. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
89. Olivetti, E. A. et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **7**, 41317 (2020).
90. Shi, Y. et al. ChatGraph: interpretable text classification by converting ChatGPT knowledge to graphs (2023).
91. Adesso, G. Towards the ultimate brain: exploring scientific discovery with ChatGPT AI. *AI Mag*. https://doi.org/10.1002/AAAI.12113 (2023).
92. Sparkes, A. et al. Towards robot scientists for autonomous scientific discovery. *Autom. Exp.* **2**, 1–11 (2010).
93. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
94. Pfitscher, R. J., Rodenbusch, G. B., Dias, A., Vieira, P. & Fouto, N. M. M. D. Estimating code running time complexity with machine learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence, LNAI and Lecture Notes in Bioinformatics)* Vol. 14196, 400–414 (Springer Science and Business Media Deutschland GmbH, 2023).
95. Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl (2023).
96. Okoli, C. & Pawlowski, S. D. The Delphi method as a research tool: an example, design considerations and applications. *Inf. Manag.* **42**, 15–29 (2004).
97. Cai, Y. et al. Product and process fingerprint for nanosecond pulsed laser ablated superhydrophobic surface. *Micromachines* **10**, 177 (2019).
98. Kundu, P., Luo, X., Qin, Y., Cai, Y. & Liu, Z. A machine learning-based framework for automatic identification of process and product fingerprints for smart manufacturing systems. *J. Manuf. Process.* **73**, 128–138 (2022).
99. Abhilash, P. M. et al. Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. *J. Intell. Manuf*. https://doi.org/10.1007/s10845-023-02266-2 (2024).

Acknowledgements

The authors gratefully acknowledge the financial support from the UK Engineering and Physical Sciences Research Council (EPSRC, EP/T024844/1, EP/V055208/1, and EP/X033686/1).

Author contributions

A.P.M.: conceptualization, methodology, formal analysis, investigation, visualization, writing—original draft. X.L.: conceptualization, methodology, supervision, project administration, funding acquisition, writing—review & editing. Q.L.: writing—review & editing. R.M.: writing—review & editing. C.W.: writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Xichun Luo.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.