nature communications

Article

https://doi.org/10.1038/s41467-025-68281-4

Explainable mechanism for production process anomalies based on digital twin

Received: 16 January 2025

Accepted: 30 December 2025

Published online: 10 January 2026

Check for updates

Weiwei Qian[1,2,5] ✉, Litong Zhang[2,3,5], Yu Guo[2,5], Sai Geng[2], Mingjie Jiang[2], Yuhan Zheng[1], Shengbo Wang[2,4] ✉ & Shaohua Huang[2] ✉

In the manufacturing sector, abnormal production can disrupt production schedules, leading to significant economic and reputational losses for manufacturers. To address this issue, in this study, we present an explainable mechanism for production process anomalies (EM2PA) designed to clarify the complex coupling relationships among various manufacturing factors, analyze the impact of these factors on the production process, identify abnormal production, provide explanations for its causes, and enable trace-back analysis. EM2PA consists of three modules: the data augmenter, the influence factor recognizer, and the causal interpreter. Specifically, the data augmenter generates small sample data of abnormal production, the influence factor recognizer decouples the complex coupling relationships and identifies the factors influencing abnormal production, and the causal interpreter provides causal explanations. Furthermore, through a case study based on the actual production process of a discrete manufacturing workshop, we demonstrate the effectiveness of EM2PA in identifying the root causes of problems, while highlighting the importance of explainability and causal analysis of production process anomalies.

With the development of the Internet of Things, big data, and artificial intelligence technologies, the manufacturing industry has entered a new stage of digital competition. The digital twin workshop, as an innovative mode of workshop operation, has been gradually explored and applied within the industry[1,2]. Improving the explainability of the digital twin model (DTM), revealing the mechanism of abnormal production (AP) in the production process, and conducting predictive diagnosis are crucial for enhancing the optimization and precise control capabilities of the production process in the digital twin workshop, as well as for promoting the broader application of the digital twin.

In manufacturing companies, understanding why a disturbance has occurred by identifying its underlying causes is a process of reverse direction. Whether management or production personnel, there is an urgent need for cause analysis of AP modes in the production process, as AP is only a surface-level issue, and the technical and management issues underlying it are the root causes, with many potential risks and problems possibly hidden. Once these causes are identified, countermeasures can be proposed and implemented to eliminate them[3]. The explainable mechanism for production process anomalies (EM2PA) is expected to provide new analytical structures and methods for cause analysis. Cause analysis of the production process can enhance the understanding of DTM prediction results (e.g., production progress, energy consumption, production capacity, and other relevant metrics.), and provide insights into how the model generates predictions by revealing the relationships between model inputs and outputs. It can also help clarify the coupling relationships between various manufacturing factors, analyze their impact on the prediction results, make explicit the potential hidden risks, and establish an early warning mechanism for AP.

[1]Ningbo University of Technology, Ningbo, China. [2]The College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. [3]School of Mechanical Engineering, Shandong University of Technology, Zibo, China. [4]School of Electrical, Electronic, and Mechanical Engineering, University of Bristol, Bristol, United Kingdom. [5]These authors contributed equally: Weiwei Qian, Litong Zhang, Yu Guo.
✉e-mail: qianww@nuaa.edu.cn; shengbo.wang@bristol.ac.uk; shaohuah@nuaa.edu.cn

In recent years, researchers have made great efforts to tackle the explainability issues of models, which is an important prerequisite for promoting the innovative application of artificial intelligence (AI) methods in industrial digital twin and realizing smart manufacturing. With the application of advanced sensing, the Internet of Things and other technologies, the available manufacturing data is evolving across multiple characteristics, commonly referred to as the 6 Vs. of big data: volume, variety, velocity, veracity, value, and variability, thereby facilitating the establishment of more complex DTMs[4]. The increasingly complicated nature of DTMs leads to poor explainability and limits their application scope.

Since Tao et al. proposed the concept of digital twin workshop and its applications, digital twin modeling, verification, and evolution have received extensive attention[2,5]. However, it is still abstract and lacks pertinence in the implementation process, for example, how to achieve the explainability of complex DTMs in industrial environments. Clearly explaining the key question ensures the organizations and end-users know what the DTMs are capable of ref. 6. To explore the explainability of models in industrial environments, some studies focus on explainable frameworks. For example, Wang et al. presented a framework of explainable modeling to enable collaboration and interaction between modelers and stakeholders[7]. Arrieta et al. summarized previous efforts made to define explainability in machine learning and put forward a definition of explainable machine learning that covers prior conceptual propositions with a major focus on the intended audience for explainability[8]. Naser. proposed a framework for integrating explainable and anomalous machine learning into a digital twin to enable fine-tuning of mixtures[9]. Beyond such frameworks and conceptual definitions, model-agnostic explanation methods such as shapley additive explanations (SHAP)[10] and local interpretable model-agnostic explanations (LIME)[11] have also been proposed. LIME explains individual predictions by locally approximating the model with an explainable surrogate model, whereas SHAP assigns feature importance values using Shapley values from cooperative game theory. Although these explainable frameworks play a significant role in guiding, the absence of concrete technical implementations for industrial settings makes them far from practically applicable. Several studies aim to explore more specific, explainable hybrid modeling methods. For example, Blakseth et al. explored an approach to hybrid analysis and modeling to create trustworthy, accurate, and explainable models[12]. Kuhnle et al. investigated methods of explainable reinforcement learning for production control in the context of a real-world task[13]. Moreover, Wehner et al. proposed an explainable lane change prediction method using layer-wise relevance propagation for layer-normalized long short-term memory[14]. Fekete et al. presented a causal AI-based method to uncover cause-and-effect relationships in urban transport data, with an initial application to a bike-sharing public transport system, providing evidence to guide the design of sustainable urban mobility policies[15]. Dibaeinia et al. proposed a causal AI approach called Counterfactual Inference by Machine Learning and Attribution Models to identify gene regulatory differences between biological conditions and to reveal potential regulators of Alzheimer's disease. However, its causal explanation relies on assumptions, which may limit the practical applicability of the method[16].

While the explainability of the model can improve production process management to some extent, simply implementing the explainability of the model is not enough. Explainability tends to focus more on the "forward direction", i.e., how the input variables influence the model output. In contrast, the cause analysis process focuses more on the "reverse direction". Due to the uncertainty and multiple disturbances in the production process of a discrete manufacturing workshop, cause analysis of AP becomes both important and challenging. Cause analysis is a process of understanding the causal mechanisms behind a change from a normal state to an abnormal one, in order to prevent recurring problems.

To explore cause analysis methods in complex manufacturing process under industrial environments, some studies have attempted to use techniques (e.g., association rules) to associate different factors with a problem or fault. Duan et al. proposed a heuristic root cause identification solution based on fuzzy weighted association rule mining for quality accidents[17]. Sun et al. proposed an adaptive fault detection and root cause analysis scheme for complex industrial processes using moving window kernel principle component analysis and information geometric causal inference, which investigates the potential root cause variables and their propagation paths[18]. Xiao et al. introduced a multi-dimensional modeling approach and abnormality handling method for digital twin shop floors[19]. Expert knowledge can also be used to establish a sequence of root causes. For example, Steenwinckel et al. presented a methodology by combining expert knowledge with machine learning for adaptive anomaly detection and root cause analysis on sensor data streams[20]. However, acquiring this causal knowledge and these rules is very difficult in complex industrial environments.

As the sensorization of manufacturing environments expands, the available data has increased, and data-driven root cause analysis methods have developed rapidly. Ma et al. explored a big data-driven root cause analysis method for quality problem solving, which generates features related to the manufacturing process and then analyzes them using machine learning algorithms[21]. Arunthavanathan et al. introduced a methodology for the autonomous diagnosis of the root cause of a detected fault in a complex processing system[22]. Cho et al. presented a quality-discriminative localization method for multisensor signals in root cause analysis, which uses convolutional neural networks to analyze sensor data and determine the location of the root cause within the process[23]. Furthermore, to improve manufacturing processes, Oliveira et al. proposed a two-phase root cause analysis solution that uses factor-ranking algorithms[24]. Causal inference is used to analyze and define overlaps, and interventional probabilities are employed to estimate how likely a given overlapping tuple is the true root cause. Moreover, Zhou et al. introduced a causal- knowledge-based method for cause analysis of equipment spot inspection failures, which fully utilizes causal relationship knowledge to enhance the reliability of cause analysis[25]. Yang et al. put forward a digital twin-driven fault diagnosis method for composite faults that combines virtual and real data to achieve high accuracy in diagnosing composite faults[26]. Specifically, the DTM is used to verify the results of the fault diagnosis model and correct errors in the single fault diagnosis stage. Then, virtual data from the DTM and real data from the physical system are used for diagnosis in the composite fault diagnosis stage. Similarly, Tang et al. proposed an approach that utilizes fusion-based clustering and hyperbolic neural network-based knowledge graph embedding, which can automatically identify bottlenecks and analyze causes from a process perspective[27]. However, these two methods are based on graphs for cause analysis, ignoring the impact of the completeness of the knowledge graph and the quality of causal knowledge, which may reduce the accuracy of cause analysis.

Previous studies have shown the success in applying various techniques, such as association rules, expert knowledge, to cause analysis. However, they mainly focused on specific stages of the production process and relied heavily on extensive rule sets or knowledge bases. In addition, data-driven approaches have been applied to anomaly diagnosis, but they lack the capability for cause analysis. Although mainstream methods can achieve satisfactory performance, there lacks a systematic approach linking anomaly detection and prediction with explanation, which limits the applicability to explainable analysis in dynamic production environments. Moreover, the integration of explainable analysis with digital twin technology

Table 1 | Example of a data sample

Data features	Values	Dimensions
Production task	90,270,300,70,68,102,150,240	8
Statistical information	25,7,0,48,23,7,3,6,...,18,39,15,0,47,17,0,0	16
Type of WIP queue at in-buffer	2,2,2,1,1,6,7,7,...,5,8,8,8,7,3,6,0	20
Waiting time of WIP at in-buffer	55,55,55,48,48,40,35,35,...,26,20,20,20,9,8,4,0	20
Product processing status	0,0,3,6,6,2,5,1,...,0,0,5,3,7,2,1,6	39
Transfer status	8,8,7,7,6,1,1,0,0,0,0,10,15,0,5,9	16
Processing equipment status	2,10,0,0.05,180	5
WIP queue at out-buffer	8,8,7,7,6,1,1,2,2,2	10
Waiting time of WIP at out-buffer	50,44,37,36,31,24,21,18,11,0	10
Time span	800	1
Types of influencing factors	0,1,2,3,4,5	1

remains largely underexplored in the context of intelligent manufacturing, resulting in the DTM with limited capability for traceability.

In this work, we propose an EM2PA based on digital twins to enable trace-back analysis of production process anomalies in the discrete manufacturing workshop. EM2PA reveals the prediction process of the model and the cause analysis of coupling characteristics of AP, decodes the complex underlying relationships between AP modes and input variables, and improves manufacturing operations. The data augmenter (DAr) generates small sample data of AP, the influence factor recognizer (IFRr) decouples complex coupling relationships and identifies factors influencing AP, and then the causal interpreter (CIr) provides causal explanations based on the output of IFRr. EM2PA enhances the dynamic early warning ability for AP and offers a systematic approach to cause analysis, in which specific markers guide the entire process from detection to explanation, enhancing explainability and traceability.

Results

Experimental setup

We use a discrete manufacturing workshop as a representative smart manufacturing system to demonstrate the performance of the EM2PA. This system consists of 13 stations, each engaged in the processing of small structural parts. Over the past few decades, the complexity of structural parts has increased, and the production cycle has been greatly shortened. Any abnormal production (AP) may lead to serious production accidents (e.g., failure to deliver on time). Each station is equipped with an in-buffer area, an out-buffer area, and a processing area. The processing area operates in an orderly manner according to the order in which parts enter the buffer area. Automated guided vehicles (AGVs) are used for transporting materials, work in process (WIP), and finished products between storage areas and stations.

A numerical experiment was conducted using 3900 individual production records, each containing 146 features, to train and test IFRr, as well as to evaluate CIr. The production data examples are shown in Table 1. This dataset included 3000 AP data generated by DAr, with each type of AP containing 600 data. The remaining 900 data came from historical data, consisting of 650 normal data and 250 abnormal data (50 for each type). The algorithm was coded in Python 3.7 and tested in a computer equipped with 8 GB RAM and an Intel Core i7-6700HQ processor running at 2.60 GHz.

The composition of the production tasks reflects the quantities of 8 types of products. The statistical information includes the numbers of completed and in-process products for each type, so the data dimension is 16. The in-buffer includes the WIP queue type and the waiting time of WIP (in minutes), each contributing 20 dimensions. The product processing status indicates the number of incomplete processes for each product type, with the dimensions equal to the number of all processes across 8 types of products. The processing equipment status includes product ID, processing time, equipment status, failure probability, and failure duration, so the data dimension is 5. The transfer status describes the type of WIP being transported and its transport time, contributing a data dimension of 16. The out-buffer has a structure similar to the in-buffer, containing the WIP queue type and WIP waiting time, each with a dimension of 10. Time span denotes the total duration of order processing, and the influencing factors correspond to 6 types of AP modes.

The experimental design comprises two parts, referred to as Experimental Case I and Case II. Case I focuses on predicting production performance using a DTM and explaining key influencing factors through the IFRr. These factors are then used as inputs in Case II to explain production anomalies via the CIr. Case I provides the foundation for explainability that supports the cause analysis conducted in Case II.

Experimental case I: explainability verification of digital twin model

To test the explainability of the IFRr on the digital twin model, a digital twin model (G(.)) for production progress prediction is taken as an example to illustrate. G(.) can be expressed as follows[1]:

$$G(\mathbf{x}) = \sum_{u=1}^{U} \alpha_u g_u(\mathbf{x}) \tag{1}$$

$$\alpha_u = \frac{1}{2}\ln\left[\frac{1-\varepsilon_u}{\varepsilon_u}\right]$$

$$s.t.\begin{cases} \varepsilon_u = \sum_{i^*}^{m^*} \frac{|g_u(x^{i^*}) - y^{i^*}|}{\sup_r[|g_u(x^{i^*}) - y^{i^*}|]} \\ g_u(x^{i^*}) = \begin{cases} g_{u,\mathrm{DNN}}(x^{i^*}), if \sum_{i^*}^{m^*}|g_{u,\mathrm{DNN}}(x^{i^*}) - y^{i^*}| \le \sum_{i^*}^{m^*}|g_{u,\mathrm{LSTM}}(x^{i^*}) - y^{i^*}| \\ g_{u,\mathrm{LSTM}}(x^{i^*}), if \sum_{i^*}^{m^*}|g_{u,\mathrm{DNN}}(x^{i^*}) - y^{i^*}| > \sum_{i^*}^{m^*}|g_{u,\mathrm{LSTM}}(x^{i^*}) - y^{i^*}| \end{cases} \end{cases} \tag{2}$$

Where, U represents the total iteration times, and m^* represents the number of data in the m-th batch. $\omega_{u+1}^{i^*}$ represents the weight of the $u+1$ iteration of i^*-th sample and $l_u^{i^*}$ represents the parameter used to control the sample weight adjustment direction. ε represents the threshold of error. When the error exceeds the threshold, the sample weight is increased, otherwise the sample weight is reduced. α_u represents the weight coefficient of the u-th base learner, which is determined by the accuracy error ε_u of the base learner (DNN or LSTM) in the data set. $\omega_{u+1}^{i^*}$ and $l_u^{i^*}$ can be expressed as follows:

$$\omega_1^{i^*} = \frac{1}{m^*}(i^* = 1, 2, m^*) \tag{3}$$

https://doi.org/10.1038/s41467-025-68281-4

$$\omega_{u+1}^{*i^*} = \omega_u^{i^*} \exp\left[-\alpha_u l_u^{i^*} \exp(\eta t^{i^*} l_u^{i^*})\right] (i^* = 1, 2, m^*; u = 1, 2, U-1) \quad (4)$$

$$l_u^{i^*} = \begin{cases} +1, \ if \left| g_u(x^{i^*}) - y^{i^*} \right| \le \xi \\ -1, \ if \left| g_u(x^{i^*}) - y^{i^*} \right| > \xi \end{cases} \quad (5)$$

The IFRr and Eqs. (14) and (15) are used to identify the important factors for production progress prediction. This method helps in determining the five categories of traceability factors that affect the prediction accuracy of G(.). The SHAP values (impact on G(.) model output) are recorded in Fig. 1. The *y*-axis represents the input variables (features) in order of importance, from top to bottom. Each dot is colored according to the value of the input variable, ranging from low (blue) to high (red). The density represents the distribution of points in the dataset. The *x*-axis represents the SHAP value. The features are sorted according to the absolute value of the sum of SHAP values. A wider position indicates a large number of sample aggregations. Each point represents a sample. A SHAP value less than 0 indicates a negative impact, while a value greater than 0 indicates a positive impact. If the sample distribution is relatively dispersed, it indicates that the feature has a greater impact.

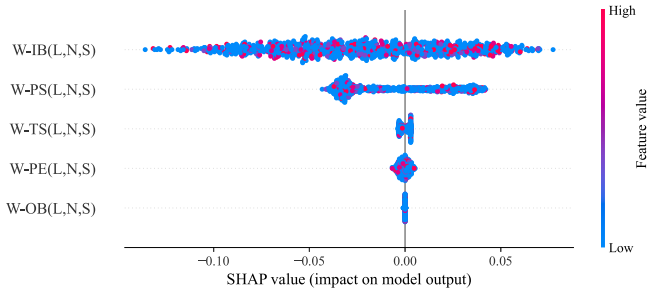

Fig. 1 | SHAP value by G(.). Colors from blue to red denote increasing feature values. Each point corresponds to a sample, with its *x*-axis position indicating the SHAP value and its *y*-axis position reflecting the feature importance ranking. A wider horizontal spread indicates a greater aggregation of samples. Source data are provided as a Source Data file.

From the per global output in Fig. 1, $W-IB_S^{L,N}$(corresponding to W-IB(L,N,S), with others corresponding similarly), the state has the greatest impact on the G(.) model output compared to the other four variables. $W-IB_S^{L,N}$ indicates that an abnormal status occurs in the in-buffer area of the workshop. L, N, and S represent the location, ID number, and state of the workshop, respectively. For example, $W-IB_1^{0,1}$ indicates that the location of AP is zero, the workshop ID is one, and the in-buffer status is one.

The $W-PS_S^{L,N}$ and $W-TS_S^{L,N}$, etc., are the next critical factors. The increase in the actual value of the $W-IB_S^{L,N}$ will cause the output of G(.) to decrease (e.g., the production progress will slow down), and its practical significance may increase in the waiting time for WIPs in the in-buffer area, which results in delayed production progress. Similarly, the increase in the actual value of the $W-PS_S^{L,N}$ will cause the output of G(.) to decrease (e.g., the production progress will slow down), and its practical significance may increase the processing time for WIPs on the machine tools, which in turn slows down the production progress. Thus, it can reveal why a DTM makes predictions and help understand the complex underlying non-linear relationships based on particular input production data. This will assist production staff in analyzing the causes.

Experimental case II: cause analysis ability verification

To elaborate on the practical application process, a part of the case study based on IFRr in the experiment is shown in Fig. 2.

First, the type of AP modes in the production process is analyzed by the workshop-level recognition module. For example, the third number in parentheses of $W-TS_1^{1,1}$ is one, indicating that W-TS is abnormal. Next, the unit-level recognition module is used to

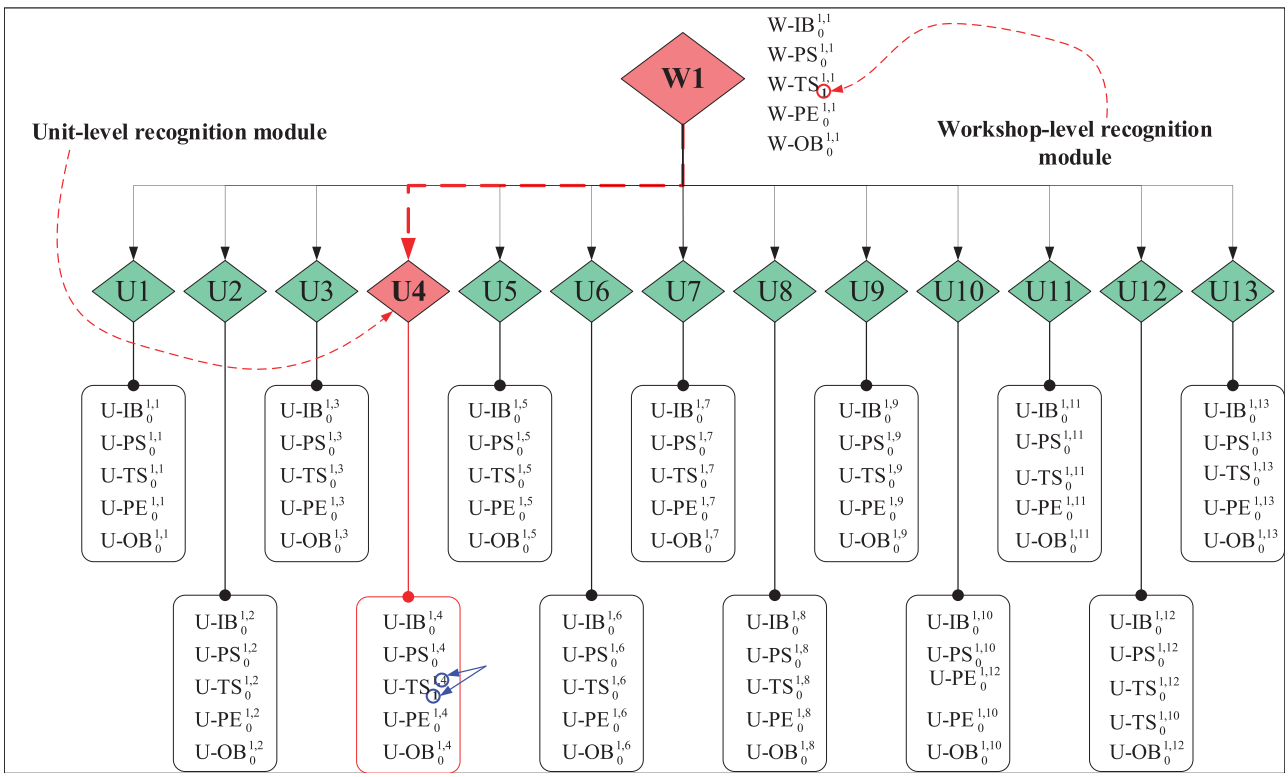

Fig. 2 | A part of the case study based on influence factor recognizer. W-IB, W-PS, W-TS, W-PE, and W-OB represent different workshop-level influencing factors. U-IB, U-PS, U-TS, U-PE, and U-OB represent different unit-level influencing factors (e.g., U-IB: in-buffer, U-PS: processing, U-TS: transfer, U-PE: equipment, U-OB: out-buffer).

https://doi.org/10.1038/s41467-025-68281-4

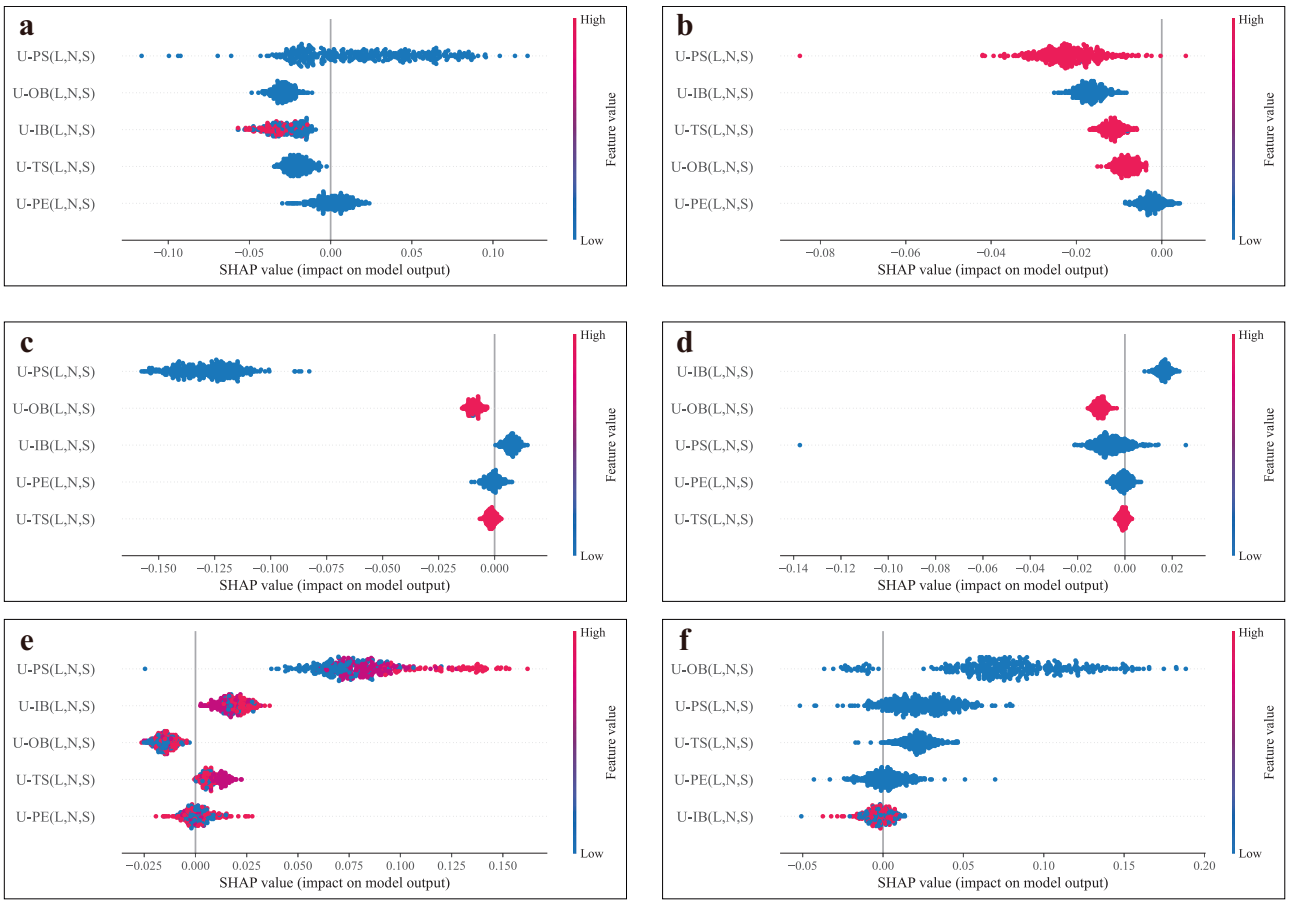

Fig. 3 | Summary plots for various AP modes from output results. a Normal production mode. **b** Abnormal production mode of in-buffer. **c** Abnormal production mode of out-buffer. **d** Abnormal production mode of equipment. **e** Abnormal production mode of work in process (WIP). **f** Abnormal production mode of transfer. Source data are provided as a Source Data file.

determine where the AP of W-TS occurred, and the calculation yields $U-TS_{\mathbf{1}}^{1,\mathbf{4}}$. The second number in parentheses of $U-TS_{\mathbf{1}}^{1,\mathbf{4}}$ is four, indicating that the unit number where the abnormality occurred is four.

The type and location of the AP can be diagnosed by IFRr, part of EM2PA. The other part, CIr, is needed to reveal the cause of the AP. The range and distribution of the impacts of input variables can be revealed through summary plots, which are calculated according to Eqs. (14) and (15), as shown in Fig. 3.

Each point on the plots in Fig. 3 represents a SHAP value for the input variables and an AP type instance. The features represent the input variables. Figure 3b shows that the $U-PS_S^{L,N}$ (corresponding to U-PS(L,N,S), with others corresponding similarly) state has the greatest impact on the AP mode of the in-buffer, compared to the other four variables. The $U-IB_S^{L,N}$ and $U-TS_S^{L,N}$, etc., are the next critical factors. Similarly, the $U-PS_S^{L,N}$ and $U-IB_S^{L,N}$ state exert the strongest influence on the AP mode of the out-buffer and on the AP mode of WIP, compared to the other four variables shown in Fig. 3c, e, respectively. In addition, the other four variables are sorted in a different order in Fig. 3e, f. In terms of the values and sorting of the five variables, there is a clear difference in the influence of different variables on the AP modes. Thus, the cause of the AP can be revealed by properly establishing decision boundaries between normal and abnormal modes.

To further validate the explanation capability of the proposed method, we conducted a comparative experiment using the LIME method. It is worth noting that the explanatory capability of LIME is also built upon the foundation provided by DAr and IFRr. The range and distribution of the impacts of input variables can be revealed through summary plots, as shown in Fig. 4.

Each point on the plots in Fig. 4 represents a LIME value for the input variables and an AP type instance. As shown in Fig. 4b, the $U-TS_S^{L,N}$ (corresponding to U-TS(L,N,S), with others corresponding similarly) state has the greatest impact on the AP mode of the in-buffer, followed by the $U-PS_S^{L,N}$ state. Compared with the states identified in Fig. 3b, specifically $U-PS_S^{L,N}$ and $U-IB_S^{L,N}$, a 50% overlap is observed. Similar results are observed in Fig. 4c–f, indicating a certain degree of consistency and divergence between the two methods.

Next, we compare the two methods in terms of consistency and significance. The comparison of the variance in input variable contributions is shown in Fig. 5. The testing results for the overall comparison, including variance values, are presented in Table 2. For each performance measure, the underlined numbers indicate the optimal values between the two models.

In terms of standard deviation, the overall performance of SHAP is more consistent than that of LIME. Although the raw feature contribution values of SHAP and LIME are not directly comparable, their variance serves as a proxy for stability, with lower values indicating more consistent feature attributions across samples. Specifically, except for Fig. 5b, e, LIME consistently shows higher maximum and average standard deviation values than SHAP across all groups. The most notable differences are observed in Fig. 5f (Max: + 110.43%, Mean: + 48.80%; see Table 2) and Fig. 5d (Max: + 64.29%, Mean: + 16.51%). The only exception occurs in Fig. 5b, e, where the average standard deviation produced by SHAP is slightly higher (+ 10.66% and + 13.35%, respectively). Nevertheless, the maximum standard deviation for this

https://doi.org/10.1038/s41467-025-68281-4

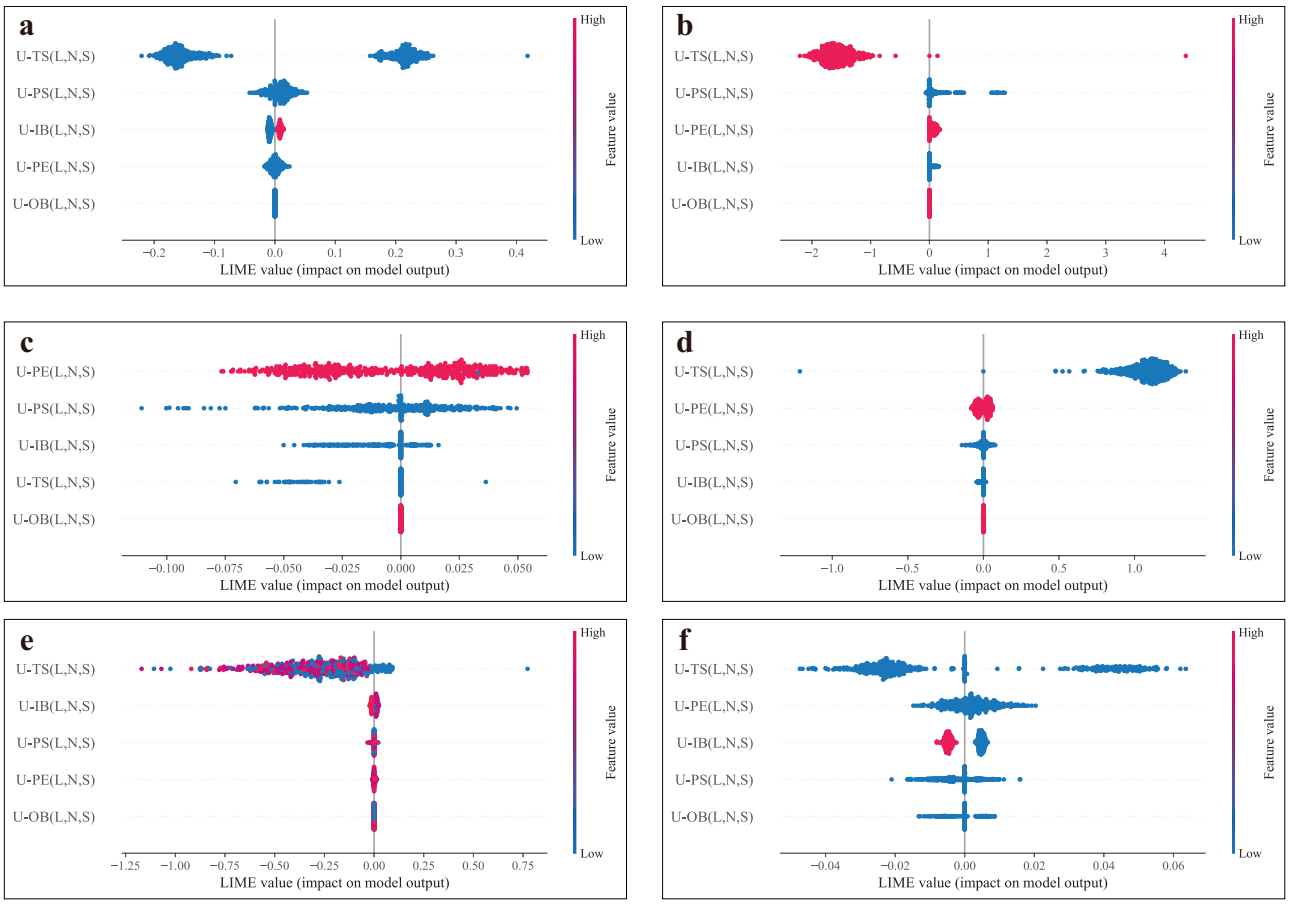

Fig. 4 | Summary plots for various AP modes from output results (LIME). **a** Normal production mode. **b** Abnormal production mode of in-buffer. **c** Abnormal production mode of out-buffer. **d** Abnormal production mode of equipment. **e** Abnormal production mode of work in process (WIP). **f** Abnormal production mode of transfer. Source data are provided as a Source Data file.

group remains higher for LIME (+ 74.13% and + 24.13%, respectively). This indicates that SHAP, based on the foundation provided by DAr and IFRr, is effective in providing stable and consistent explanations of feature contributions.

Based on the above process, the SHAP values of the five AP modes and the normal mode are used to delineate the marginal threshold lines according to Eqs. (17) and (18). The threshold lines and the distribution of input variables for the AP modes are intuitively illustrated with a segment and column chart in Fig. 6.

Figure 6 shows the changes of influencing factors under one normal and five abnormal modes. After the IFRr test, the data are fed into Eqs. (14) and (15) to demonstrate the AP classification and determine the importance of factors. $U-OB_S^{L,N}$ indicates that an abnormal status occurs in the out-buffer area in the unit. L, N, and S represent the location, ID number, and state of the unit, respectively. U represents the unit level. For example, $U-OB_1^{I,4}$ indicates that the location of AP is one, the unit ID is four, and the out-buffer status is one.

The cause of the AP in the in-buffer is defined as $U-OB_S^{L,N}$, $U-PS_S^{L,N}$, $U-PE_S^{L,N}$ and $U-TS_S^{L,N}$ from Fig. 6b, i.e., $U-OB_S^{L,N}$ has the greatest impact on it. Certain coupling interactions exist among $U-PS_S^{L,N}$, $U-PE_S^{L,N}$ and $U-TS_S^{L,N}$. However, since $U-OB_S^{L,N}$ is an intermediate variable, $U-PS_S^{L,N}$ is considered the primary cause of the AP in the in-buffer. Similarly, as shown in Fig. 6f, the cause of the AP for material/WIP transfer is defined as $U-PE_S^{L,N}$, $U-PS_S^{L,N}$, $U-OB_S^{L,N}$ and$U-IB_S^{L,N}$, i.e., $U-PE_S^{L,N}$ is the primary cause.

The causes of the AP can be revealed by comparing the degree of change in SHAP values under normal and abnormal conditions. Specifically, setting marginal threshold lines for AP modes may further reveal the coupling characteristics between the causes of production anomalies. Figure 6a shows the maximum factor's score is defined as the marginal threshold to analyze the cause of the detected AP condition. The purpose of setting the marginal threshold is to analyze multiple factors simultaneously. For example, as illustrated in Fig. 6c, e, if the marginal threshold is set to 8.0, then $U-PS_S^{L,N}$ and $U-OB_S^{L,N}$ are the causes of the AP of the out-buffer and the AP of WIP, respectively. However, if the marginal threshold is set to 7.0, then $U-PS_S^{L,N}$ and $U-PE_S^{L,N}$ are the causes of the AP of the out-buffer, and similarly, the causes of the AP of WIP are $U-OB_S^{L,N}$ and $U-PS_S^{L,N}$.

Based on the research object and issues discussed in this paper, the proposed research approach on the digital twin-based anomaly explainability analysis mechanism lacks comparable methods for reference. Therefore, to further validate the effectiveness of CIr, actual production process data were extracted for analysis to verify whether the derived conclusions align with the real process. A comparison of the actual data under five types of anomalies is shown in Fig. 7.

Figure 7a presents the actual production process data for $U-PS_S^{L,N}$, which reflects the processing time of in-process products, specifically the time for $U-PS_S^{L,N}$ entering the buffer in both normal and abnormal modes. It is evident that there is a significant difference in the $U-PS_S^{L,N}$ data, which supports the conclusion in Fig. 6b that $U-PS_S^{L,N}$ is the primary cause of the buffer status anomaly. Similarly, Fig. 7e shows the actual production process data for $U-PE_S^{L,N}$, which reflects the processing time for the equipment, specifically the processing time for $U-PE_S^{L,N}$ during material/WIPs transfer in normal and abnormal modes. It can be observed that there is a significant difference in the $U-PE_S^{L,N}$ data, which is consistent with the conclusion in

https://doi.org/10.1038/s41467-025-68281-4

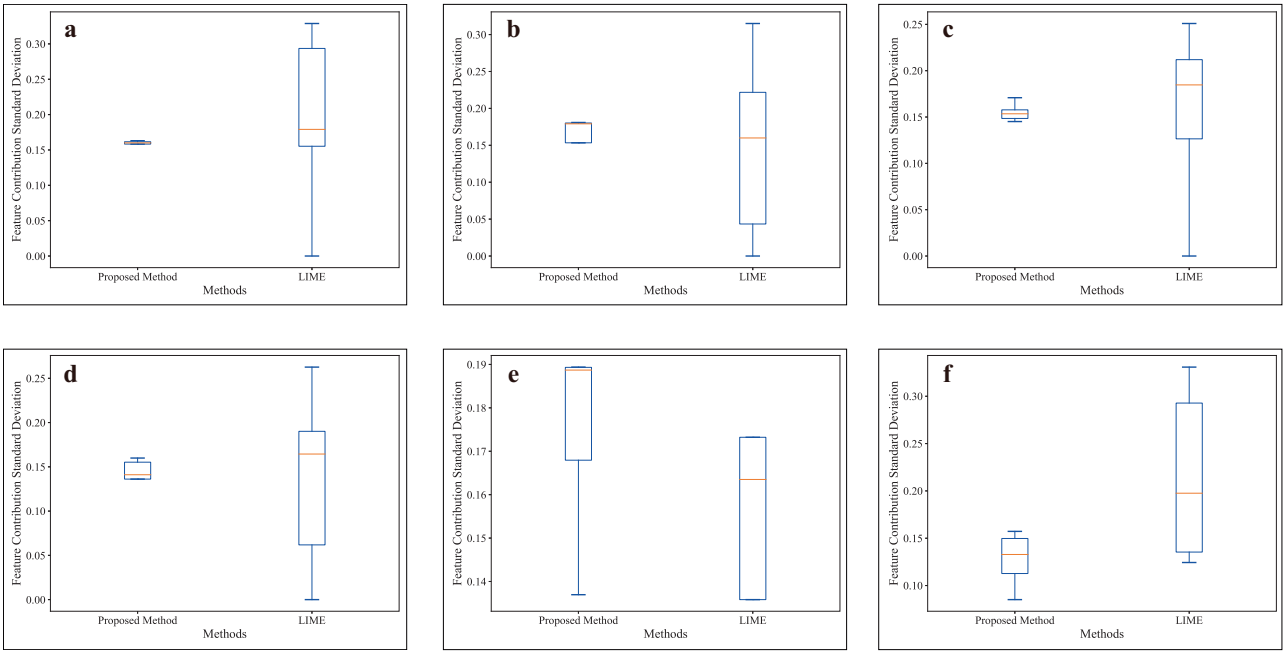

Fig. 5 | Comparison of the variance in input variable contributions. a Normal production mode. **b** Abnormal production mode of in-buffer. **c** Abnormal production mode of out-buffer. **d** Abnormal production mode of equipment. **e** Abnormal production mode of work in process (WIP). **f** Abnormal production mode of transfer. Median represents the central value of the distribution. A lower and stable median indicates more stable explanation results. The box reflects the middle 50% of the data; a narrower box means more concentrated and stable explanations, while a wider box indicates instability. Whiskers show the range of non-outlier values. Shorter whiskers represent smaller variation and more stable explanations, whereas longer whiskers imply unstable behavior in some cases. Source data are provided as a Source Data file.

Table 2 | Testing results of SHAP and LIME

Corresponding figure numbers	Value of the standard deviation					
	Maximum			Mean		
	SHAP	LIME	Percentage difference	SHAP	LIME	Percentage difference
5a	0.2060	0.3288	+59.61%	0.1583	0.1791	+13.14%
5b	0.1809	0.3150	+74.13%	0.1791	0.1600	−10.66%
5c	0.1708	0.2509	+46.90%	0.1535	0.1847	+20.33%
5d	0.1599	0.2627	+64.29%	0.1411	0.1644	+16.51%
5e	0.1894	0.2362	+24.71%	0.1887	0.1635	−13.35%
5f	0.1572	0.3308	+110.43%	0.1328	0.1976	+48.80%

Table 3 | State of resource

Type	State description
Personnel	Available, Unavailable, Assigned
Equipment	Idle, Operating, Faulted, Maintenance
Tools	Available, Unavailable, In Use
WIP	Waiting, Processing, Transferring, Completed

Fig. 6f that $U - PE_S^{L,N}$ is the primary cause of the in-process product transfer anomaly. The conclusions for the remaining Figs, such as Fig. 7b vs. Fig. 6c, Fig. 7c vs. Fig. 6d, and Fig. 7d vs. Fig. 6e, are also consistent.

In summary, the test results clearly show that the EM2PA can effectively reveal the causes of abnormal production processes in the discrete manufacturing workshop.

Discussion

Cause analysis is the process through which we find the true cause of a problem. It is a crucial step in manufacturing, as only by identifying and addressing the root cause can improvements be made to the manufacturing operation. For example, identifying which features are important for the decision-making strategy of the digital twin model (DTM) and understanding why these features drive specific decisions made by the model during prediction. This article proposes the concept of the explainable mechanism for production process anomalies (EM2PA) and establishes the data augmenter (DAr) to generate small sample data, the influence factor recognizer (IFRr) to identify the influencing factors, and the causal interpreter (CIr) to provide causal explanation. Notably, the proposed framework is abstract, serving primarily as a conceptual and methodological foundation. EM2PA exhibits generality and is suitable for domains that necessitate the explanation of underlying causes. However, transferability within the industrial manufacturing domain requires scenario-specific adaptation, industrial domain expertise, and technological integration.

An empirical experiment on five abnormal production (AP) modes is conducted to illustrate the entire process. Using 3900 data with 146 features, either collected from actual production progress or generated by DAr, IFRr identifies the unknown test sets for AP modes. CIr is then used to decode the complex underlying relationships between AP

https://doi.org/10.1038/s41467-025-68281-4

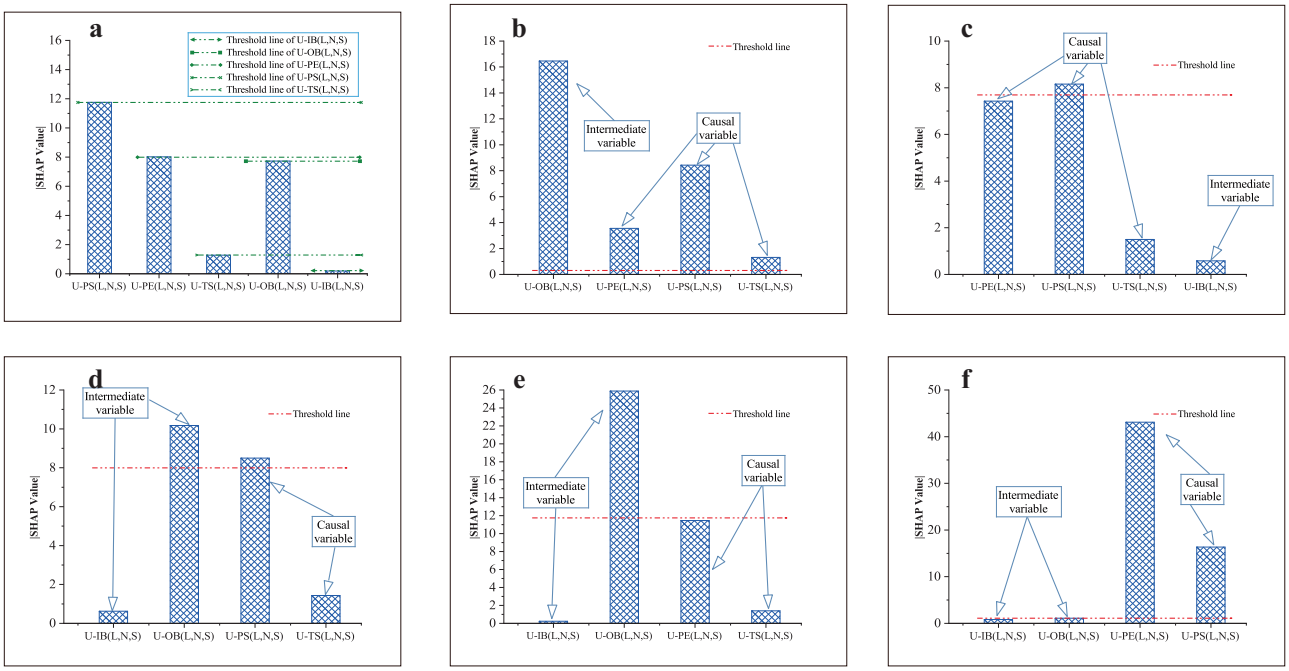

Fig. 6 | Analysis of the five AP modes using the causal interpreter. a Normal production status. **b** Abnormal status of the in-buffer. **c** Abnormal status of the out-buffer. **d** Abnormal status of equipment. **e** Abnormal status of work in process (WIP). **f** Abnormal status of transfer. Source data are provided as a Source Data file.

mode and input variables. Experimental results demonstrate that EM2PA performs well.

The limitation of this approach, with respect to its transferability, is the requirement for a computing terminal with sufficient computational power, particularly during the training of models across multiple scenarios. For example, calculating SHAP values for 1000 data with 146 features takes at least 3 hours. Therefore, the real-time performance limitations of CIr may affect the cause analysis efficiency of EM2PA. IFRr is constructed using SVM; however, it can also be built using other classification algorithms, such as Gradient Boosting Decision Trees, Decision Trees, and so on. The average accuracy of influencing factor identification by the workshop-level recognition module is 95.28%, with a 95% confidence interval of [93%, 97%]. For the unit-level recognition module, the average accuracy reaches 99.66%, with a narrow confidence interval of [98.8%, 100%]. Both results are obtained using SVM. Since SVM is a well-established algorithm, its validation experiments are not included in this paper.

Considering the research content and the development of this field, extreme conditions may occur in the manufacturing process. More small and unbalanced datasets can be further explored in future work, so as to propose more effective approaches for cause analysis of additional AP modes. Moreover, the threshold ($\mathbf{U}_c^{thr}$) of each AP needs to be adaptive and update dynamically according to the production environment in practical applications. Furthermore, to support real-time cause analysis of AP modes, it is urgent to explore methods for cause analysis under a distributed computing framework.

Methods

Problem description

Abnormal production (AP) refers to an event where the production process does not comply with the pre-designed working mode, resulting in a significant deviation of production performance metrics from the production plan. This may be caused by factors such as the production process, operation methods, machinery and equipment, materials, etc. Therefore, it is broadly defined as a set of five traceability factors, including personnel status, equipment status, material status, method status, and environment status.

However, due to the complex production process and frequent disturbances in the discrete manufacturing workshop, the causes of AP are also diverse. Therefore, in order to ensure the rigor and operability of this study, the scope of AP is further narrowly defined as the status of the in-buffer area, processing equipment, WIP processing, material/WIP transfer and out-buffer area. This set of factors can be expressed as follows:

$$U_c = \{IB_S^{L,N}, PE_S^{L,N}, PS_S^{L,N}, TS_S^{L,N}, OB_S^{L,N}\} \quad (6)$$

Where $IB_S^{L,N}$ and $PE_S^{L,N}$ represent the in-buffer area status and processing equipment status, respectively; $PS_S^{L,N}$ and $TS_S^{L,N}$ represent the product processing status and material/WIP transfer status, respectively; $OB_S^{L,N}$ represents out-buffer area status. $IB_S^{L,N}$ and $OB_S^{L,N}$ are intermediate variables, while $PE_S^{L,N}$, $PS_S^{L,N}$ and $TS_S^{L,N}$ are causal variables. The purpose of this study is to identify the factors from $\mathbf{U_c}$ and explain the results by analyzing $\mathbf{Da}(k)$. The $\mathbf{Da}(k)$ can be collected using various types of sensors (e.g., RFID, UWB, etc.).

A digital model is generally a representation without automated or real-time data interaction, while a digital shadow extends this by enabling unidirectional data flow from the physical to the digital space[28]. In contrast, a DTM supports bidirectional data exchange and provides advanced functions such as simulation and optimization. As a virtual counterpart to physical systems, the DTM facilitates the analysis and tracing of AP, while also supporting functions such as simulation, prediction, and explainability. By leveraging bidirectional data flows between the physical workshop and its virtual model, the DTM ensures continuous synchronization of operational data and provides actionable feedback in the form of anomaly explanation and decision-making support for management or production personnel. An idealized digital twin model is defined as follows.

Definition 1. DTM is a virtual representation of a product or complex system, replicating its characteristics, behaviors, logic, and performance. It integrates key functions including description, monitoring, simulation, prediction, explainability, optimization, and control. Maintaining "form-and-spirit consistency" with its physical counterpart, DTM leverages historical and real-time data along with

https://doi.org/10.1038/s41467-025-68281-4

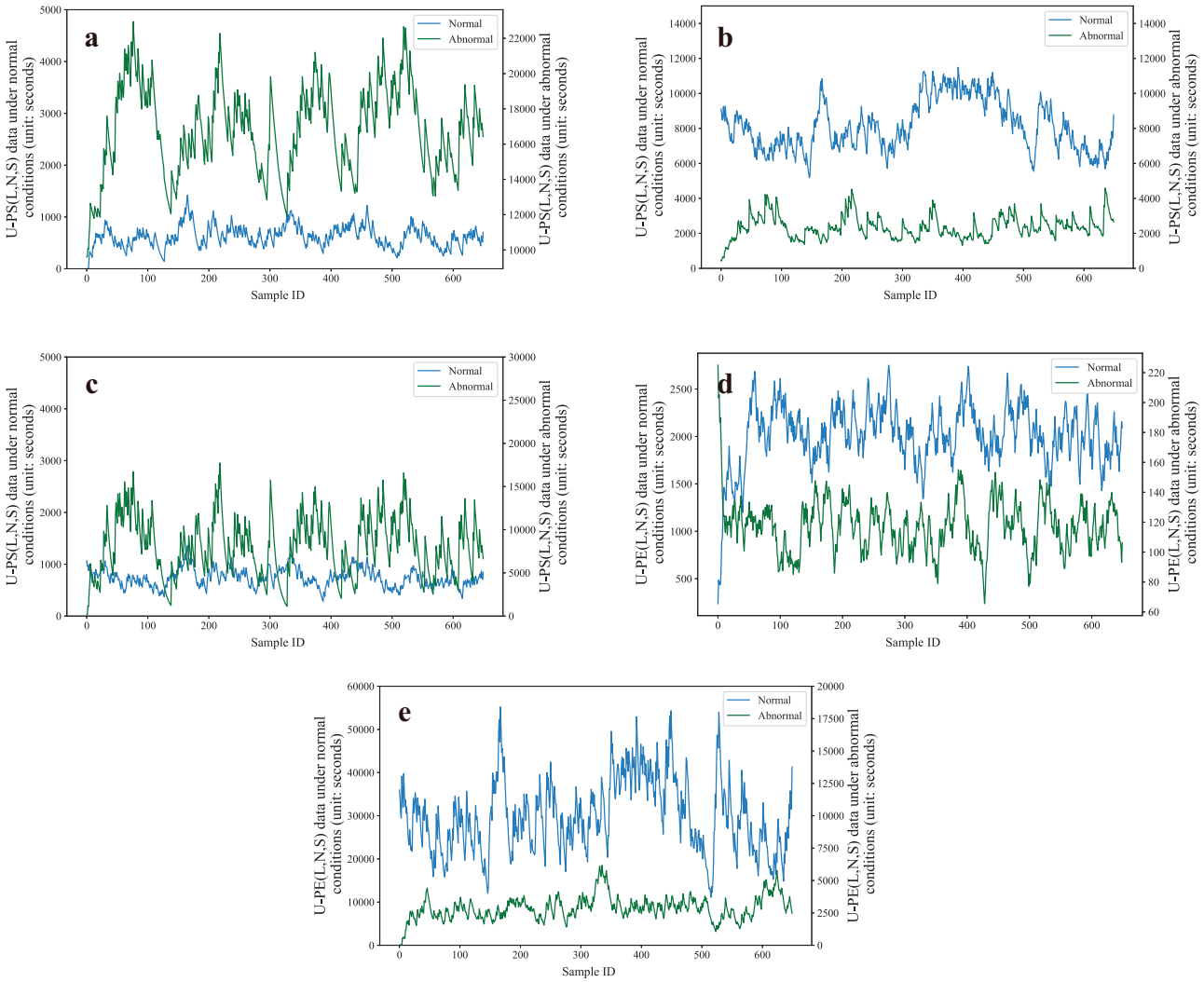

Fig. 7 | The comparison of actual data under five AP modes. a U-PS(L,N,S) data under normal/abnormal mode of in-buffer. **b** U-PS(L,N,S) data under normal/abnormal mode of out-buffer. **c** U-PS(L,N,S) data under normal/abnormal mode of processing equipment. **d** U-PE(L,N,S) data under normal/abnormal mode of WIP processing. **e** U-PE(L,N,S) data under normal/abnormal mode of material transfer. L, N, and S represent the location, ID number, and state of the unit, respectively. Source data are provided as a Source Data file.

algorithmic models to enable lifecycle-wide state visibility, performance assessment, solution optimization, and process control.

Influence factor

An order in the discrete manufacturing workshop often involves a variety of products, and each product contains multiple processes. Each process is carried out at a workstation with the required processing capacity. The performance of the workshop at time k is expressed by the running state and input parameters. The workshop can be split at the unit-level (workstation) according to the system's operation, which can be expressed as follows:

$$\mathbf{Da}(k) = \{\mathbf{da}_{s1}(k), \mathbf{da}_{s2}(k), \ldots, \mathbf{da}_{sm}(k), \ldots, \mathbf{da}_{sM}(k)\} \tag{7}$$

$$\mathbf{da}_{sm}(k) = \{\mathbf{da}^{1}_{sm}(k), \mathbf{da}^{2}_{sm}(k), \ldots, \mathbf{da}^{h}_{sm}(k), \ldots, \mathbf{da}^{H}_{sm}(k)\} \tag{8}$$

Where, M is the number of workstations. $\mathbf{da}_{sm}(k)$ represents the manufacturing information of the m-th workstation at time k. H represents the identification number of manufacturing resources (e.g., people, equipment, tooling, materials, WIP etc.) of the m-th workstation, and $\mathbf{da}^{h}_{sm}(k)$ represents the information of the h-th manufacturing resource at time k, which can be expressed as follows:

$$\mathbf{da}^{h}_{sm}(k) = \{k, l^{h}_{sm}(k), ID^{h}_{sm}(k), con^{h}_{sm}(k)\} \tag{9}$$

Where, $l^{h}_{sm}(k)$ represents the location of the h-th resource at time k. $ID^{h}_{sm}(k)$ represents the identification number of the h-th resource. $con^{h}_{sm}(k)$ denotes the state of the h-th resource (e.g., waiting, processing, and completed state of WIP) as shown in Table 3.

The spatio-temporal data chain of $\mathbf{Da}^{h}_{sm}$ in the manufacturing process is formed by combining the data at each behavior change moment in chronological order, which can be expressed as follows:

$$\mathbf{Da}^{h}_{sm} = \begin{bmatrix} k_1 & l^{h}_{sm}(k_1) & ID^{h}_{sm}(k_1) & con^{h}_{sm}(k_1) \\ k_2 & l^{h}_{sm}(k_2) & ID^{h}_{sm}(k_2) & con^{h}_{sm}(k_2) \\ \vdots & \vdots & \vdots & \vdots \\ k_i & l^{h}_{sm}(k_i) & ID^{h}_{sm}(k_i) & con^{h}_{sm}(k_i) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{10}$$

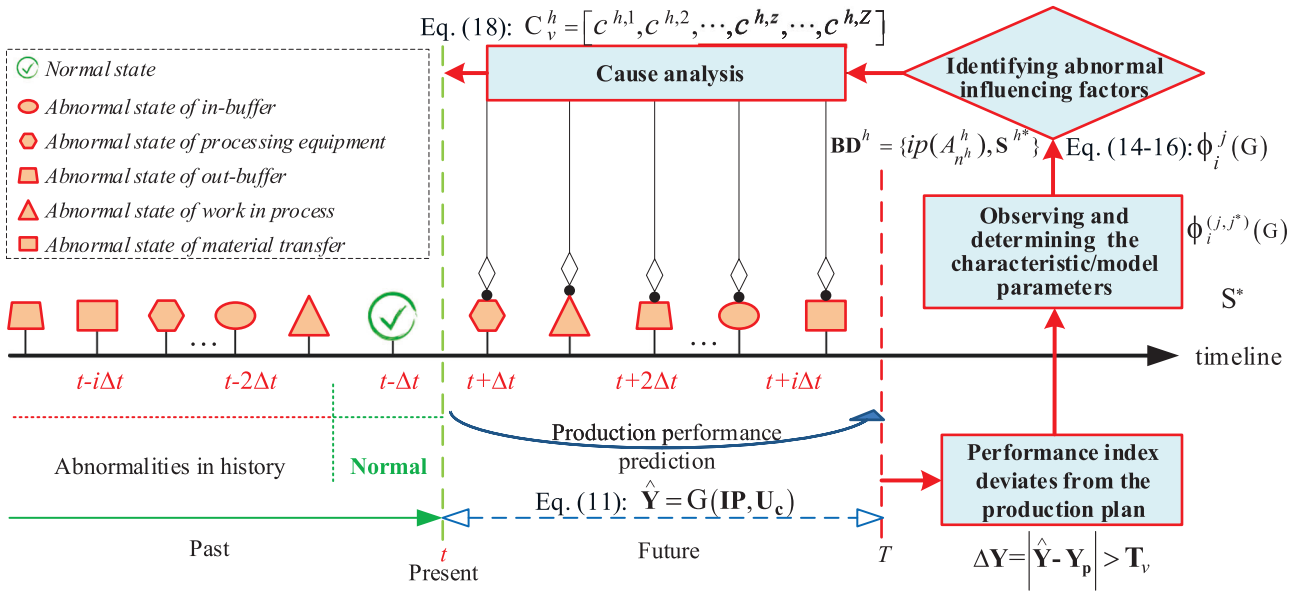

Fig. 8 | Cause analysis diagram of abnormal production in discrete manufacturing workshop. The analysis investigates the deviation between the production performance index and the planned target, identifies the key influencing factors, and traces their origins to reveal the root causes.

Explanation process

The inability of the production plan is mainly manifested in the deviation between the planned and actual performance indices, with this deviation exhibiting dynamic characteristics as production status changes. Identifying the factors that cause AP based on manufacturing resource data and analyzing the underlying influence mechanisms are the core aspects of EM2PA. When various uncertain disturbance events (e.g., equipment degradation, machine failure, untimely material distribution, etc.) affect the production process in a discrete manufacturing workshop and cause the production plan to fail to meet requirements, it is necessary to observe and diagnose via the EM2PA. The schematic diagram is shown in Fig. 8.

The EM2PA is used to observe and determine the characteristic parameters or model parameters that cause the deviation between the production performance index and the planned value. Its purpose is to reveal the causes of the deviation of the performance index from the production plan at time T in the future. Unlike the "forward direction" of AP prediction, the cause analysis of abnormal production is based on the deviation between the performance index and the production plan at time T, and it "reversely" reveals the causes of the deviation in the performance index.

If the network architecture of the workshop system is determined, and the input parameters (**IP**), such as production process data, are given and operate under certain operating conditions ($\mathbf{U_c}$), then the corresponding production performance value (**Y**) can be obtained. **Y** represents a multidimensional construct encompassing various aspects of production performance, such as production progress, energy consumption, and production capacity. In practice, however, Y is operationalized as a single quantitative indicator, functioning as a scalar in specific analytical contexts. Based on the historical dataset $\{\mathbf{IP}_t, \mathbf{U_c}_t, \mathbf{Y}_t\}$ with $t < 0$, a predictive model G(.) is established using data-driven methods. This model is subsequently applied to estimate the production performance $\mathbf{Y}^\wedge$ for future operations with $t > 0$. The relationship between **IP** and $\mathbf{Y}^\wedge$ can be expressed as follows:

$$\mathbf{Y}^\wedge = G(\mathbf{IP}, \mathbf{U_c}) \tag{11}$$

Where, G(.) represents the performance prediction DTM. Considering the **IP** and $\mathbf{U_c}$, along with the planned production progress value ($\mathbf{Y_p}$), the causer analysis definition of $\mathbf{Y}^\wedge$ can be expressed as follows when condition $\Delta\mathbf{Y} = |\mathbf{Y}^\wedge - \mathbf{Y_p}| > \mathbf{T}_v$ is satisfied:

$$\mathbf{U_c}^* = G^*(\Delta\mathbf{Y}, \mathbf{IP}^*) \tag{12}$$

Where, $\Delta\mathbf{Y}$ represents the deviation between $\mathbf{Y}^\wedge$ and $\mathbf{Y_p}$. A threshold $\mathbf{T}_v$ is introduced to filter out minor fluctuations and focus on significant deviations in production performance. $G^*(.)$ is the performance cause analysis DTM. Specifically, when $\Delta\mathbf{Y} > \mathbf{T}_v$, it indicates a notable deviation from the plan, thereby triggering the causer analysis process to identify potential root causes of the abnormal performance. By inputting $\mathbf{IP}^*$ into $G^*(.)$, the model is expected to output the name, number, value, or value range of the influencing factors of $\mathbf{U_c}^*$ under the given operating conditions. For example, the reason why the production progress (one of the performance index) does not meet the production plan is that the transferring status of WIP numbered 3 is abnormal. **IP** includes historical data, real-time data, and digital twin data. $\mathbf{U_c}$ includes information such as in-buffer area, processing equipment, WIP processing, WIP transferring status, and out-buffer area, all of which are influencing factors that help identify deviations in production performance. The core of EM2PA is to establish the model $G^*(.)$, which can be applied to reveal the mechanism of the influence of manufacturing factors change on subsequent production performance **Y** in the discrete manufacturing workshop.

Various factors influencing production performance are highly complex and nonlinear, making it difficult to establish an accurate performance cause analysis DTM for the discrete manufacturing workshop. In addition, some performance prediction models established by machine learning algorithms exhibit a black-box characteristic. Based on the problem description, $G^*(.)$ should include a performance prediction module (G(.)) and an explanation module (E(.), F(.) and J(.)) for AP. Therefore, $G^*(.)$ can be expressed as follows:

$$G^*(.) = \{G, E, F, J, U_c\} \tag{13}$$

Where, E(.) represents DAr, F(.) represents IFRr, and J(.) represents CIr.

The G(.) is a machine learning model established by a data-driven modeling method and can be expressed as $\mathbf{Y}^\wedge = G(\mathbf{IP}, \mathbf{U_c})$, where, $\mathbf{IP}, \mathbf{U_c}$ and $\mathbf{Y}^\wedge$ are all known. Then, the EM2PA problem is transformed into solving $\mathbf{U_c}^* = G^*(\Delta\mathbf{Y}, \mathbf{IP}^*)$, i.e., under the condition that the $\mathbf{IP}^*$, $\Delta\mathbf{Y}$, $G^*(.)$ are known. The $\mathbf{IP}^*$ can be obtained through data acquisition systems

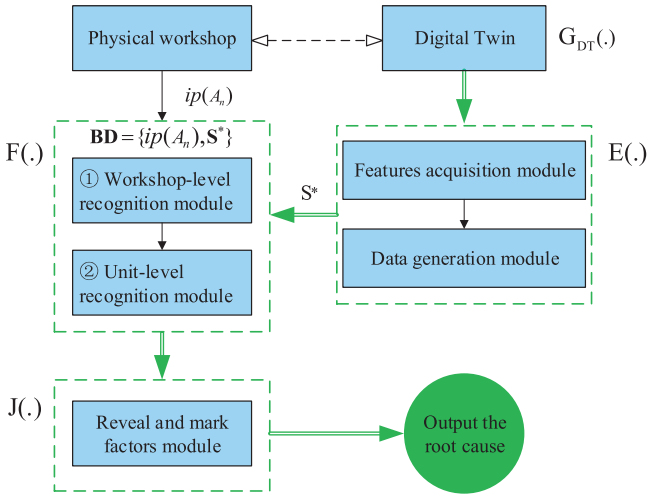

Fig. 9 | The logical relationship. E(.) represents the data augmenter, F(.) represents the influence factor recognizer, and J(.) represents the causal interpreter.

(e.g., RFID, UWB, industrial bus). Since $\mathbf{Y}^{\wedge}$ and $\mathbf{Y_p}$ are known, $\Delta\mathbf{Y}=|\mathbf{Y}^{\wedge}-\mathbf{Y_p}|$ is also known. Therefore, the EM2PA problem is to model $G^*(.)$ under the condition that both the input and output are known.

The logical relationship between E(.), F(.), and J(.) is shown in Fig. 9.

Where, $ip(A_n)$ represents the production process data corresponding to the first n features. $\mathbf{S}^*$ represents the data generated by E(.), and $\mathbf{BD}=\{ip(A_n),S^*\}$ represents a balanced dataset, which is used to train F(.).

Data augmenter

The production process in the workshop is influenced by various factors, such as equipment failures, material transfer status, etc. If each factor is abstracted as a team member and the performance prediction result is considered as the total revenue, the contribution of each factor to the prediction result can be measured using the Shapley value. For the historical data of production performance influencing factors $\boldsymbol{ip}_{\text{history}}$, $\boldsymbol{ip}_{\text{history}}\subseteq\boldsymbol{IP}$, the Shapley value $\phi_i^j(\mathrm{G})$ of the j-th influencing factor $ip_{\text{history}}^{(i,j)}$ for one sample $ip_{\text{history}}^{i}=\left\{ip_{\text{history}}^{(i,1)},ip_{\text{history}}^{(i,2)},ip_{\text{history}}^{(i,j)},ip_{\text{history}}^{(i,j^*)}ip_{\text{history}}^{(i,J)}\right\}$ can be expressed as follows:

$$\phi_i^j(\mathrm{G})=\sum_{S\subseteq ip_{\text{history}}^{i}\setminus ip_{\text{history}}^{(i,j)}}\frac{|S|!\,(J-|S|-1)!}{J!}\left[\mathrm{G}\left(\mathrm{S}\left\{ip_{\text{history}}^{(i,j)}\right\}\right)-\mathrm{G}(S)\right] \tag{14}$$

The Shapley value $\phi_i^{(j,j^*)}(\mathrm{G})$ for the interaction effect between the j-th and j^*-th influencing factors in sample ip_{history}^{i} can be expressed as follows:

$$\phi_i^{(j,j^*)}(G)=\sum_{S^-\subseteq ip_{\text{history}}^{i}\setminus\left\{ip_{\text{history}}^{(i,j)},ip_{\text{history}}^{(i,j^*)}\right\}}\frac{|S^-|!\,(J-|S^-|-2)!}{J!}\left[G\left(S^-\left\{ip_{\text{history}}^{(i,j)},ip_{\text{history}}^{(i,j^*)}\right\}\right)-G\left(S^-\left\{ip_{\text{history}}^{(i,j)}\right\}\right)-G\left(S^-\left\{ip_{\text{history}}^{(i,j^*)}\right\}\right)+G(S^-)\right] \tag{15}$$

Where, S represents the subset of the remaining factors after excluding the j-th influencing factor $ip_{\text{history}}^{(i,j)}$ in ip_{history}^{i}; S represents the subset of the remaining factors after excluding both the j-th and j^*-th factors $ip_{\text{history}}^{(i,j)}$ and $ip_{\text{history}}^{(i,j^*)}$ [29]. $|S|$ represents the number of influencing factors in S.

The importance of the feature variables in G(.) is quantified based on the SHAP method[30], and the prediction results are explained both globally and locally. The relative contribution of each feature variable to the prediction results is clarified, which provides valuable information for feature ranking, the selection of the top n ($n\in N$) most relevant features, and the expansion of small and unbalanced data in a digital twin environment.

In most cases, steady-state data (e.g., normal operation state data) is sufficient, while oscillation data (e.g., abnormal operation state data) is scarce, leading to a serious data imbalance and small sample problem in the discrete manufacturing workshop. $G_{DT}(.)$ is used to expand the sample data, and the expanded dataset is recorded as S^*. The DAr adopts a strategy based on feature transformation. The AP small sample data generation method is shown in Fig. 10.

When obtaining the feature ranking of factors affecting production performance (e.g., production progress) based on SHAP, the factors that have less impact on performance can be removed. Assuming that $\boldsymbol{A}=\{a_1,a_2,\cdots,a_n,\cdots,a_N\}$ is the feature affecting production performance, a_n is the n-th feature sorted by decreasing importance, and $\boldsymbol{x}^*$ is the actual input parameter of the workshop (e.g., $\mathbf{Da}_{sm}^h$, etc.), the calculation equation for augmenting the dataset can be expressed as follows:

$$\mathrm{S}^*=\mathrm{G_M}\left[\mathrm{G_{DT}}\left(n|\boldsymbol{x}^*=x_1,x_2,\cdots,xq,\cdots,x_Q\right),ip_{\text{history}},ip_{\text{real}}\right] \tag{16}$$

Select n ($n<Q$) features corresponding to unbalanced data from $G_{DT}(.)$, which can be transformed into $\boldsymbol{x}^*=\{x_1,x_2,\cdots,x_q,\cdots,x_Q\}$ in $G_{DT}(.)$, where Q is the data dimension, to obtain a small-sample unbalanced dataset, recorded as $\mathrm{DT}(\mathbf{U_c})$,$\mathrm{DT}(\mathbf{U_c})\subseteq\mathbf{S}^*$.

By taking out the production process data $ip(A_n)$, $ip(A_n)\in\{ip_{\text{history}},ip_{\text{real}}\}$, and combining it with $\mathbf{S}^*$ to form a balanced dataset (BD), $\mathbf{BD}=\{ip(A_n),\mathbf{S}^*\}$, the **BD** is used to train IFRr.

Influence factor recognizer

In the process of identifying influencing factors, the type (mode) of the AP is first recognized at the workshop level, followed by recognition at the unit level. In this step, information such as the location and ID number of the unit can be obtained. Based on this information, a cause analysis of the AP (at the unit-level) is performed to identify the influencing factors.

Due to the high dimensionality of production process data in the discrete manufacturing workshop, using clustering algorithms to process this data is not effective, as clustering algorithms inherently perform dimensionality reductions. The support vector machine (SVM) is a machine learning algorithm that solves non-linear classification problems by mapping the original data into high-dimensional spaces, helping to alleviate the problem of local optimal and the dimensional disaster in non-linear spaces[31].

Therefore, a workshop-level recognition module is established using SVM and the **BD** dataset, which is used to identify the type of AP, i.e., the type of event that occurred in the workshop.

The unit-level recognition module is used to identify the location of the AP and the influencing factors, such as which factors affect the occurrence of the AP at the workstation. Assuming that $\mathbf{BD}^h=\{ip(A_{n^h}^h),\boldsymbol{S}^{h^*}\}$ is a balanced dataset containing data $ip(A_{n^h}^h)$, $ip(A_{n^h}^h)\in ip_{\text{history}},ip_{\text{real}}$, corresponding to the n^h features of the h-th workstation, and $\boldsymbol{S}^{h^*}$ ($\mathbf{S}^{h^*}\in\mathbf{S}^*$). Similarly, the unit-level recognition module is established using SVM and $\mathbf{BD}^h$.

Causal interpreter

This module is established to identify and mark the factors that lead to the AP. When an AP is detected in a time window (t^w), the degree of change in the input feature importance $\boldsymbol{W}_{t^w,\text{unit}}^h$ is compared with the $\boldsymbol{W}_{t^w-1,\text{unit}}^h$ under the previous normal condition. The index $\Delta\boldsymbol{W}^h\in$

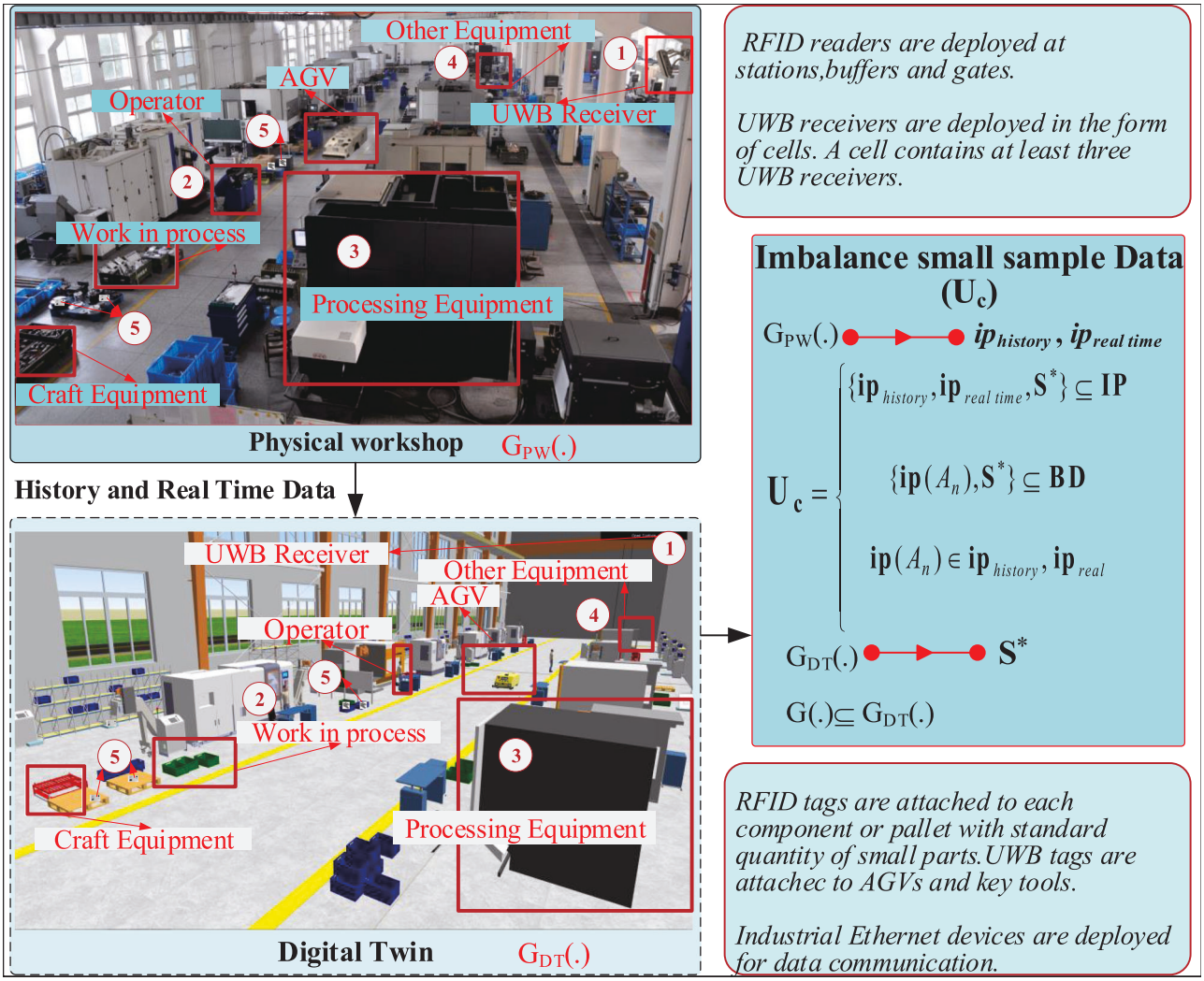

Fig. 10 | Small sample data generation based on feature transformation. In most cases, steady-state data (e.g., normal operation data) is sufficient, while oscillation data (e.g., abnormal operation data) is scarce, leading to data imbalance. GDT(.) is used to expand the sample data, and the expanded dataset is recorded as S*.

$R^{1\times Z}$ to measure the degree of change can be expressed as follows:

$$\begin{aligned}\Delta \mathbf{W}^h &= \left[\Delta w^{h,1}, \Delta w^{h,2}, \cdots, \Delta w^{h,z}\right] \\ &= \left[\frac{w_{t^w}^{h,1} - w_{t^w-1}^{h,1}}{w_{t^w-1}^{h,1}}, \frac{w_{t^w}^{h,2} - w_{t^w-1}^{h,2}}{w_{t^w-1}^{h,2}}, \cdots, \frac{w_{t^w}^{h,z} - w_{t^w-1}^{h,z}}{w_{t^w-1}^{h,z}}\right]\end{aligned} \tag{17}$$

Where, $\Delta w^{h,z}$ represents the degree of change of the z-th feature under AP and normal conditions. $\Delta \mathbf{W}^h$ can be calculated using the function $\phi(.)$, as mentioned earlier. The feature with the largest contribution value in $\Delta \mathbf{W}^h$ is selected and expressed as follows:

$$\begin{aligned}C_v^h &= \left[c^{h,1}, c^{h,2}, \cdots, c^{h,z}, \cdots, c^{h,Z}\right] \\ c^{h,z} &= \frac{\Delta w^{h,z}}{\sum_{z=1}^{Z} \Delta w^{h,z}}\end{aligned} \tag{18}$$

Where, C_v^h represents the contribution value of the feature, with the values sorted in descending order, i.e., $c^{h,z} \geq c^{h,Z}$. $c^{h,z}$ represents the contribution value of the z-th feature to AP. Thus, the cause analysis problem is transformed into assessing the contribution degree of each feature to AP. Then, the features are mapped to the five types (modes) of AP, respectively. Assuming that $\mathbf{U}_c^{thr}$ is the threshold for each AP, $\mathbf{U}_c^{thr}$ is obtained by analyzing the normal data. When the contribution of features to AP exceeds the threshold, it can be diagnosed as the cause of the AP problem. Otherwise, the features are considered irrelevant variables for AP.

Data availability

The source data are provided with this paper. The datasets supporting the findings of this study have been deposited in a public repository and are publicly accessible via the following https://doi.org/10.24433/CO.2416108.v1[32]. Source data are provided in this paper.

Code availability

The core code is provided with this paper. The code supporting the findings of this study has been deposited in a public repository and is publicly accessible via the following https://doi.org/10.24433/CO.2416108.v1[32].

References

1. Qian, W. et al. Digital twin driven production progress prediction for discrete manufacturing workshop. *Robot. Comput. Integr. Manuf.* **80**, 102456 (2023).
2. Tao, F. & Qi, Q. Make more digital twins. *Nature* **573**, 490–491 (2019).
3. Adriana, I. et al. Improved root cause analysis supporting resilient production systems. *J. Manuf. Syst.* **64**, 468–478 (2022).

4. Wang, J., Xu, C., Zhang, J. & Zhong, R. Big data analytics for intelligent manufacturing systems: A review. *J. Manuf. Syst.* **62**, 738–752 (2022).
5. Lei, Z. et al. Digital twin based monitoring and control for DC-DC converters. *Nat. Commun.* **14**, 5604 (2023).
6. Fuller, A., Fan, Z., Day, C. & Barlow, C. Digital twin: enabling technologies, challenges and open research. *IEEE Access* **8**, 108952–108971 (2020).
7. Wang, L., Deng, T., Zheng, Z. & Shen, Z. Explainable modeling in digital twin. In *Winter Simulation Conference*. 1–12 (2021).
8. Arrieta, A. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
9. Naser, M. Digital twin for next gen concretes: On-demand tuning of vulnerable mixtures through Explainable and Anomalous Machine Learning. *Cem. Concr. Compos.* **131**, 104640 (2022).
10. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
11. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144 (2016).
12. Sindre, B., Adil, R., Trond, K. & Omer, S. Deep neural network enabled corrective source term approach to hybrid analysis and modeling. *Neural Netw.* **146**, 181–199 (2022).
13. Andreas, K., Marvin, M., Louis, S. & Gisela, L. Explainable reinforcement learning in production control of job shop manufacturing system. *Int. J. Prod. Res.* **60**, 5812–5834 (2022).
14. Wehner, C., Powlesland, F., Altakrouri, B. & Schmid, U. Explainable online lane change predictions on a digital twin with a layer normalized LSTM and layer-wise relevance propagation. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. 621–632 (2022).
15. Fekete, T., Mengistu, D. & Wicaksono, H. Leveraging causal AI to uncover the dynamics in sustainable urban transport: A bike sharing time-series study. *Sustain. Cities Soc.* **122**, 2210–6707 (2025).
16. Dibaeinia, P., Ojha, A. & Sinha, S. Interpretable AI for inference of causal molecular relationships from omics data. *Sci. Adv.* **11**, 1–18 (2025).
17. Duana, P. et al. Root cause analysis approach based on reverse cascading decomposition in QFD and fuzzy weight ARM for quality accidents. *Comput. Indust. Eng.* **147**, 106643 (2020).
18. Sun, Y., Qin, W., Zhuang, Z. & Xu, H. An adaptive fault detection and root-cause analysis scheme for complex industrial processes using moving window KPCA and information geometric causal inference. *J. Intell. Manuf.* **32**, 2007–2021 (2021).
19. Xiao, B., Qi, Q. & Tao, F. Multi-dimensional modeling and abnormality handling of digital twin shop floor. *J. Indust. Inf. Integr.* **35**, 100492 (2023).
20. Steenwinckel, B. et al. FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning. *Future Gener. Comput. Syst.* **116**, 30–48 (2021).
21. Ma, Q., Li, H. & Thorstenson, A. A big data-driven root cause analysis system: Application of machine learning in quality problem solving. *Comput. Ind. Eng.* **160**, 107580 (2021).
22. Arunthavanathan, R., Khan, F., Ahmed, S. & Imtiaz, S. Autonomous fault diagnosis and root cause analysis for the processing system using one-class SVM and NN permutation algorithm. *Ind. Eng. Chem. Res.* **61**, 1408–1422 (2022).
23. Cho, Y. & Kim, S. Quality-discriminative localization of multisensor signals for root cause analysis. *IEEE Trans. Syst. Man Cybern. Syst.* **52**, 4374–4387 (2022).
24. Oliveira, E., Miguéis, V. & Borges, J. On the influence of overlap in automatic root cause analysis in manufacturing. *Int. J. Prod. Res.* **60**, 6491–6507 (2022).
25. Zhou, B., Li, J., Li, X., Hua, B. & Bao, J. Leveraging on causal knowledge for enhancing the root cause analysis of equipment spot inspection failures. *Adv. Eng. Inf.* **54**, 101799 (2022).
26. Yang, C. et al. Digital twin-driven fault diagnosis method for composite faults by combining virtual and real data. *J. Ind. Inf. Int.* **33**, 100469 (2023).
27. Tang, J., Liu, Y., Lin, K. & Li, L. Process bottlenecks identification and its root cause analysis using fusion-based clustering and knowledge graph. *Adv. Eng. Inf.* **55**, 101862 (2023).
28. Ladj, A. sma. et al. A knowledge-based digital shadow for machining industry in a digital twin perspective. *J. Manuf. Syst.* **58**, 168–179 (2021).
29. Wen, X., Xie, Y., Wu, L. & Jiang, L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid. Anal. Prev.* **159**, 106261 (2021).
30. Mangalathu, S., Hwang, S. & Jeon, J. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng. Struct.* **219**, 110927 (2020).
31. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019).
32. Qian, W. et al. Explainable mechanism for production process anomalies based on digital twin. Code Ocean, https://doi.org/10.24433/CO.2416108.v1 (2025).

Acknowledgements

This work was supported in part by the Scientific Research Foundation of NBUT under Grant 24KQ008 (W.Q.), in part by the National Natural Science Foundation of China under Grant 52105522 (S.H.) and 52575578 (Y.G.), and in part by the China Postdoctoral Science Foundation under Grant 2022M721597 (S.H.).

Author contributions

Y.G. and S.H.H. initiated and supervised the project with input from S.B.W. W.W.Q., L.T.Z., S.G., M.J.J., and Y.H.Z. performed the experiments, developed the code, performed data analysis, prepared illustrations, and wrote the manuscript with input and approval from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-68281-4.

Correspondence and requests for materials should be addressed to Weiwei Qian, Shengbo Wang or Shaohua Huang.

Peer review information *Nature Communications* thanks Matteo de Marchi, Rakesh Kumar Phanden, Jyrki Savolainen and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.