Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(10), 340–350
https://doi.org/10.12913/22998624/225622
ISSN 2299-8624, License CC-BY 4.0

Received: 2026.05.26
Accepted: 2026.07.18
Published: 2026.09.01

Explainable CNN-based bottleneck diagnosis in human–cobot manufacturing using process heatmaps: An operations and quality management perspective

Łukasz Kański[1]

[1] Maria Curie-Sklodowska University, Plac Marii Curie-Sklodowskiej 5, 20-031 Lublin, Poland
E-mail: lukasz.kanski@mail.umcs.pl

ABSTRACT

Human–cobot collaborative manufacturing, positioned at the core of the Industry 5.0 paradigm, increases the stochasticity of production systems through human variability, safety-related stops, and resource-sharing constraints. In such environments, rapid identification of dominant bottlenecks and their root causes is critical to operational resilience, production-management responsiveness, and preserving the human-centric character of operations. This study proposes an explainable bottleneck-diagnosis framework based on convolutional neural networks (CNNs) that leverages process heatmaps derived from a discrete-time flow-line simulation model inspired by the digital twin concept. Each simulation run is encoded as a three-channel heatmap (queue length, station busy state, and downtime), yielding a compact spatio-temporal "process fingerprint" suitable for image-based representation learning. We consider a six-class taxonomy: balanced operation, bottleneck caused by human fatigue, bottleneck caused by cobot performance degradation, bottleneck caused by safety stops in a human–cobot cell, bottleneck caused by cobot failure, and supply delays (line starvation). On a balanced dataset of 1,500 synthetic samples, the CNN achieves a test accuracy of 0.938 (macro-F1 = 0.938; 95% Wilson confidence interval for accuracy: 0.898–0.963). An ablation study shows that multi-channel heatmaps improve accuracy by 0.10 compared with queue-only heatmaps. Explainability is operationalised using Gradient-weighted class activation mapping (Grad-CAM), which consistently localises station–time regions associated with queue build-up and downtime patterns, providing process engineers with evidence that can be translated into corrective actions. The results indicate that simulation-generated heatmaps can support transparent and scalable bottleneck analytics for human–cobot manufacturing, enabling rapid prototyping prior to deployment on shop-floor data and providing an evidence-based decision-support instrument for operations and quality management. The framework also clarifies the shop-floor data required for deployment and proposes a staged implementation path for real-factory decision support.

Keywords: digital twin, simulation, Industry 5.0, human–cobot collaboration, bottleneck diagnosis, process heatmaps, convolutional neural networks, explainable artificial intelligence, quality management.

INTRODUCTION

Industry 5.0 reframes advanced manufacturing as a human-centric, sustainable, and resilient socio-technical system in which technological progress is evaluated not only through productivity gains but also through its contribution to workers' wellbeing and overall system resilience [1]. In practice, this paradigm highlights human–cobot collaboration (HRC) and cobot-assisted workstations as key enablers of flexible production. Yet, the very features that make HRC attractive – human variability, shared workspaces, safety interlocks, and task-level autonomy – introduce non-trivial stochasticity into material flow and resource utilisation.

In such dynamic systems, bottlenecks are rarely stationary. They may emerge as a consequence of operator fatigue, drift in cobot performance, micro-stops triggered by proximity-based safety rules, or disruptions in inbound supply. In industry surveys, robotics and automation are consistently ranked among the most relevant technological development directions for Industry 4.0/5.0, complemented by machine connectivity, cloud solutions, and data analytics [2]. At the same time, firms report that resource constraints and implementation complexity remain key barriers [3]. These trends motivate a methodological shift toward scalable decision-support analytics capable of rapid, and, crucially, transparent, bottleneck diagnosis.

Two converging streams are particularly salient. First, simulation-based digital twins are increasingly used for what-if exploration and for generating synthetic data when shop-floor labels are scarce or costly [4]. Second, AI governance is becoming more demanding, emphasising risk management, transparency, and accountability across the lifecycle [5,6]. In industrial contexts with elevated stakes, explainable artificial intelligence (XAI) is therefore not merely an academic preference but a practical condition for adoption.

This paper addresses the following research gap: despite the extensive literature on bottleneck identification using key performance indicators (KPIs), event logs, and machine learning, there is still limited work that (i) explicitly models bottleneck mechanisms specific to HRC, (ii) uses simulation-generated "process fingerprints" for supervised learning, and (iii) provides faithful visual explanations linking model decisions to interpretable process evidence. To close this gap, we propose a CNN-based bottleneck-diagnosis framework trained on simulation-generated process heatmaps and augmented with Grad-CAM explainability [7].

The main contributions of this study are:

1. Simulation model of a human–cobot flow line, inspired by the digital twin concept, together with an explicit taxonomy of HRC-related bottleneck causes;
2. Multi-channel process-heatmap representation that transforms state traces into compact image-like inputs;
3. CNN classifier for multi-class bottleneck diagnosis with an XAI layer based on Grad-CAM;
4. Reproducible synthetic dataset and an empirical evaluation including baseline comparisons and an ablation study.

The research questions (RQ) and hypotheses (H) are formulated as follows:

- RQ1: Can a CNN trained on simulation-generated process heatmaps classify dominant bottleneck causes in HRC manufacturing with high accuracy?
- H1: The proposed CNN achieves test accuracy ≥ 0.90 for the considered bottleneck taxonomy.
- RQ2: Does multi-channel heatmap encoding improve diagnostic performance compared with queue-only heatmaps?
- H2: Multi-channel encoding improves accuracy by at least 0.08 relative to queue-only encoding.
- RQ3: Do Grad-CAM explanations localise station–time regions consistent with the bottleneck mechanisms embedded in the simulation model?
- H3: The highest Grad-CAM activations overlap with regions of queue build-up and downtime at the bottleneck station.

BACKGROUND AND RELATED WORK

Digital twins, simulation, and synthetic data in manufacturing

Digital twins are commonly conceptualised as virtual representations of physical systems kept consistent through data and models to support monitoring, prediction, and decision-making. In manufacturing, the digital twin is positioned as a foundational technology for smart production, spanning stages from design and commissioning to operation and maintenance [4]. Simulation plays a critical role because it enables controlled experimentation, robustness testing under disruptions, and the generation of labelled synthetic datasets when real labels are unavailable. Recent studies increasingly exploit this "twin-as-a-data-factory" paradigm to train and validate deep learning models prior to full-scale deployment [8].

In the literature, simulation and digital-twin approaches are frequently used both for system improvement and for educational purposes. For example, Flaga and Pacholczak [9] discuss a web-based digital twin demonstrator, whereas Daneshjo et al. [10] use simulation to redesign a workstation in an assembly line. In the HRC domain, prior work addresses the implementation and deployment of cobots in manufacturing processes [11,12]. Together, these studies support a methodological continuum from process modelling and simulation to data-driven analytics.

Safety constraints in collaborative work, including speed-and-separation monitoring and power-and-force limiting, are formalised in standards such as ISO/TS 15066 [13] and, in the latest revisions, consolidated within the ISO 10218 series for industrial cobots and robot cells [14,15]. This motivates explicit modelling of safety-related micro-stops in bottleneck analyses for HRC environments.

Bottleneck identification and root-cause diagnosis

Bottleneck analysis has a long tradition in production systems engineering, and classical approaches rely on utilisation indicators, queue lengths, and activity/idle-time measures. Classical bottleneck detection methods, particularly the average active-period method proposed by Roser et al. [16], established the principle that bottlenecks can be detected from state durations rather than from static utilisation alone. More recent work has developed dynamic and data-driven bottleneck identification to account for non-stationary behaviour. Su et al. [17] propose an approach to dynamic bottleneck identification in manufacturing systems, emphasising time-varying constraints. A recent systematic review by Skoogh et al. [18] shows that bottleneck-detection methods can be grouped according to queue-state, process-state and combined-state information, and that their operationalisation may rely on gemba walks, discrete-event simulation or data-science approaches. Beyond identification, root-cause diagnosis increasingly uses structured representations of manufacturing knowledge and data. Tang et al. [19] present process-bottleneck identification and root-cause analysis using fusion-based clustering and knowledge-graph embedding, illustrating a shift from detection towards explainable, causality-oriented reasoning. Accordingly, the present study positions bottleneck diagnosis as a dynamic operations-management problem in which the location, timing and managerial response to the constraint are as important as the final class label.

However, two limiting assumptions recur in the literature: (i) bottleneck analytics often assumes a predominance of machine-centric causes, with limited explicit modelling of HRC phenomena such as safety micro-stops and operator fatigue; and (ii) many learning-based approaches provide point predictions without transparent process-level evidence.

Explainable artificial intelligence in human-centric manufacturing analytics

Explainable artificial intelligence (XAI) is an umbrella term for local and global methods that aim to make model behaviour understandable to humans. Popular model-agnostic approaches include LIME [20] and SHAP [21]. For convolutional models, gradient-based visual explanation methods such as Grad-CAM provide class-discriminative localisation maps by combining feature activations with gradient information [7]. Recent reviews emphasise that, despite rapid progress, industrial XAI remains methodologically fragmented and often lacks standardised evaluation protocols [22,23].

In industrial contexts, explainability is coupled with governance and risk management. The NIST AI Risk Management Framework highlights transparency and interpretability as mechanisms for managing AI risks across the lifecycle [5]. Similarly, European AI regulation underscores obligations for certain use cases, strengthening the need for evidence-based documentation and operational transparency [24,25]. These factors are particularly salient when AI systems support decisions that affect production safety and working conditions.

Heatmap representations and vision learning from process traces

Transforming non-image data into image-like representations is an established strategy for leveraging the representational capacity of CNNs. In time-series analytics, such transformations enable the exploitation of spatial locality and hierarchical feature extraction. For example, Pizoń et al. [26] demonstrate an image-based approach to trend classification using candlestick-chart representations and deep learning augmented with gradient explanations.

Building on these advances, this study represents manufacturing-process traces as multi-channel heatmaps that encode queue dynamics, station utilisation, and downtime patterns. This representation supports (i) scalable learning on simulation outputs and (ii) a direct bridge to visual explainability methods.

Research gap and positioning

In summary, the literature provides solid foundations on digital twins, bottleneck analytics, and explainable AI. Nevertheless, a coherent pipeline that addresses HRC-specific bottleneck mechanisms and connects simulation-generated heatmaps with CNN learning and faithful visual explanations remains underdeveloped. The framework proposed in this paper aims to close this methodological gap, while aligning with the current trend towards simulation- and modelling-supported engineering applications.

MATERIALS AND METHODS

Overall methodology

The proposed methodology follows the pipeline shown in Figure 1. A simulation model inspired by the digital twin concept generates station-level state traces for a human–cobot production line under multiple disruption scenarios. The traces are then transformed into multi-channel process heatmaps that serve as inputs to a CNN classifier. Model predictions are explained with Grad-CAM, producing station–time saliency maps that process engineers can interpret as evidence supporting the diagnosed bottleneck cause. In real-factory use, the method would operate as an offline or near-real-time decision-support layer connected to existing data systems rather than as an autonomous controller. Companies would need to collect buffer queue length or work-in-process before each station, station busy/idle states, downtime start and end times, safety-stop events, cobot cycle-time and alarm logs, robot-failure records, and material-arrival or supply-delay timestamps; these data may come from MES/SCADA/PLC logs, cobot controllers, safety PLCs, proximity sensors, buffer counters, RFID/barcode tracking and ERP or warehouse-management systems. A practical implementation roadmap is: (1) identify stations, buffers and decision points; (2) map available shop-floor signals to the three heatmap channels; (3) calibrate the simulator and generate labelled synthetic scenarios; (4) train and validate the classifier against historical or pilot observations; and (5) use the CNN output and Grad-CAM maps in engineering and operations-management meetings. The resulting diagnoses may support corrective actions such as workload balancing, job rotation and ergonomic redesign for C1, cobot recalibration and maintenance review for C2, safety-zone redesign and sensor calibration for C3, preventive maintenance and spare-parts planning for C4, and supplier-reliability or buffer-policy improvement for C5.

Human–cobot manufacturing system model

For the study, a single-product flow line consisting of eight sequential stations (S1–S8) connected by finite-capacity buffers was modelled. Station S3 was defined as a collaborative human–cobot cell, where processing times and micro-stops reflect human variability and safety interlocks. Station S5 represents an automated cobot/machine station that can experience rare but long-lasting failures. The remaining stations model balanced upstream and downstream operations.

The simulator is discrete-time with horizon $T = 512$ time steps. Each station processes at most one unit at a time. Processing times are sampled from a lognormal distribution with mean μ_i and coefficient of variation CV, reflecting right-skewed task-time variability observed in manual and semi-automated operations. Buffer capacity was set to 50 units for all intermediate buffers. Baseline parameters are summarised in Table 1.

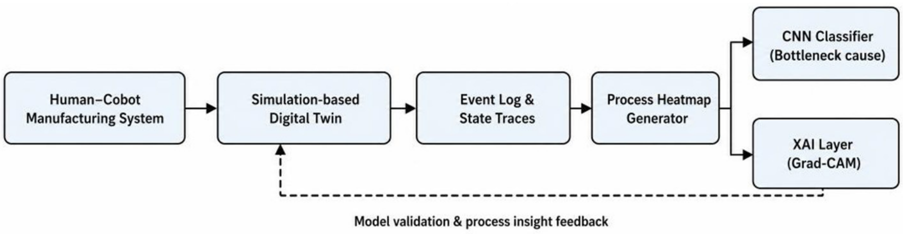

Figure 1. Conceptual pipeline: simulation-based digital twin → process heatmaps → CNN diagnosis with an XAI feedback loop. Source: author's own elaboration

Simulation experiment design and bottleneck taxonomy

To reproduce realistic HRC disruptions and obtain labels for supervised learning, a six-class bottleneck taxonomy was defined. Each class corresponds to a scenario with controlled parameter changes relative to the balanced baseline:

- C0 balanced: stable operation with equalised processing times and no failures.
- C1 HRC–human fatigue: progressive increase in processing time at S3 proportional to cumulative workload.
- C2 HRC–cobot degradation: step increase in processing time at S3 after half the horizon, modelling performance drift.
- C3 HRC–safety stop: frequent short micro-stops at S3, representing interruptions triggered by proximity rules.
- C4 cobot failure: rare but long failures at S5, coupled with slowed processing.
- C5 supply delay / starvation: reduced arrival intensity to S1.

Table 2 specifies key scenario parameters. For each class, 250 Monte Carlo replications were generated, yielding 1500 labelled samples in total. Source: author’s own elaboration. The numerical values in Table 2 should be interpreted as stylised but engineering-plausible scenario generators, not as parameters calibrated to a single factory. They were selected to reproduce qualitatively distinct HRC bottleneck mechanisms under controlled conditions: lognormal processing times and CV = 0.20 represent right-skewed task-time variability in manual and semi-automated work; the C1 fatigue coefficient creates gradual throughput deterioration at the HRC station; the C3 micro-stop probability and duration represent frequent short safety interruptions; the C4 failure probability and duration represent rare but long cobot-related stoppages; and the C5 arrival probability represents line starvation due to supply unreliability. In C1, n_jobs denotes the cumulative number of jobs processed at the HRC station S3 during a simulation run.

Process heatmap encoding

For each simulation run, we record station-level traces: input-buffer queue length, busy indicator $b\_i(t)$, and downtime indicator $d\_i(t)$. The traces are aggregated into B = 32 time bins by averaging over successive windows. To obtain a square, image-like representation, each station signal is replicated into a fixed block of rows, forming a B×B heatmap for each channel. The three channels are then stacked to form an input tensor $X \in R^{3 \times B \times B}$, $X\ \{x_{cj}\}$, where $X \in R^{3 \times B \times B}$ denotes a three-channel process heatmap tensor of spatial resolution $B \times B$. Queue lengths are normalised by buffer capacity.

This design deliberately preserves the station–time structure of congestion and stops while keeping the input compact for efficient training. An example of resulting heatmaps is shown in Figure 2.

CNN architecture and training protocol

A lightweight CNN was designed to classify heatmaps into six bottleneck classes. The architecture comprises three convolutional blocks with ReLU activation and max pooling, followed by global average pooling (GAP) and a fully connected classifier. The layer structure is shown in Figure 3.

Table 1. Baseline simulation parameters for the eight-station flow line (T = 512 steps, buffer capacity = 50 units, CV = 0.20)

Station	Type	Mean processing time μ (steps)	Notes
S1	Manual / assembly	6	–
S2	Manual / assembly	7	–
S3	HRC (human–cobot)	8	HRC cell; scenario-dependent variability
S4	Manual / assembly	7	–
S5	Automated: robot/machine	9	Robot station; failure scenario
S6	Manual / inspection	6	–
S7	Packaging	8	–
S8	Packaging	7	–

Table 2. Scenario-based simulation experiment plan (taxonomy of bottleneck causes and parameter modifications)

Class	Arrival probability	S3 processing-time modifier	S3 micro-stop process	S5 failure process
C0	0.18	none	none	none
C1	0.18	$\mu_3 \cdot (1+0.03 \cdot n\_jobs)$	none	none
C2	0.18	$\mu_3 \cdot 2.2$ after $T/2$	none	none
C3	0.18	none	P(stop)=0.20; E[dur]=4	none
C4	0.18	none	none	P(break)=0.01; E[dur]=80; $\mu_5 \cdot 1.6$
C5	0.06	none	none	none

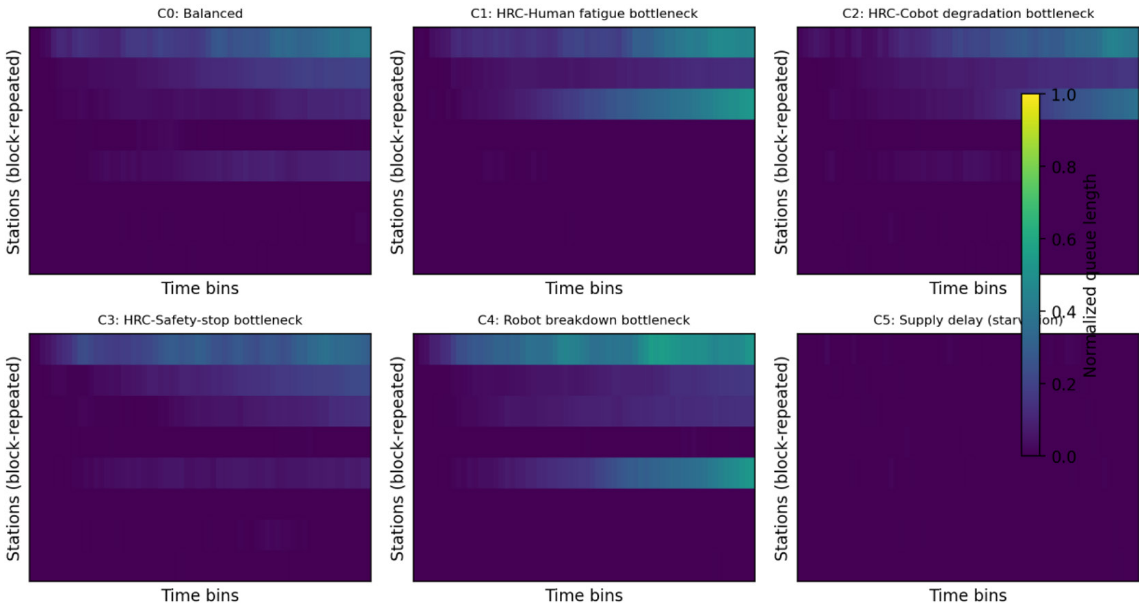

Figure 2. Representative simulation-generated process heatmaps (queue-length channel) for the six bottleneck classes

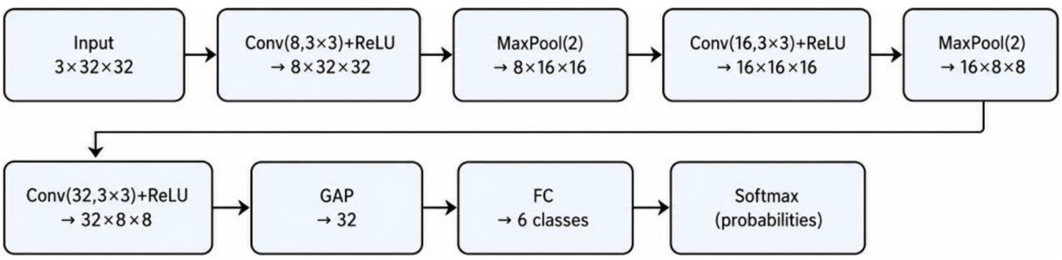

Figure 3. CNN architecture used for heatmap classification (B = 32; three-channel input)

Training uses the Adam optimiser with learning rate 10^{-3}, batch size 256, and up to 60 epochs. Early stopping selects the checkpoint with the best validation accuracy. The dataset is split into stratified train/validation/test sets (70/15/15) with a fixed random seed for reproducibility. Results are reported using accuracy and macro-averaged F1 score.

Baseline models and evaluation metrics

To contextualise CNN performance, two feature-based baselines were considered, trained on engineered summary features (mean and maximum queue values, mean utilisation, mean downtime per station, and an approximate throughput indicator). Baselines include multinomial logistic regression and a random-forest classifier. All models are evaluated on the same stratified test split. Statistical uncertainty of CNN accuracy is reported as a 95% Wilson confidence interval.

Explainability using Grad-CAM

To obtain visual explanations of CNN decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was used (Selvaraju et al. 2017). For a given class c, Grad-CAM computes weights a_k^c by globally averaging gradients of the class score y^c with respect to feature maps A^k of the last convolutional layer:

$$\alpha_k^{(c)} = \frac{1}{H\,x\,W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^{(c)}}{\partial A_{ij}^{(k)}} \quad (1)$$

The class activation map is then given by $L^{(c)}_{Grad-CAM} = ReLU(\sum_k \alpha_k^{(c)} A^{(k)})$ where $A(k) \in R^{H\,x\,W}$ denotes the $k-th$ feature map of the last convolutional layer, H and W are its spatial dimensions, and $y^{(c)}$ represents the pre-softmax class score for class c, which is upsampled to the input resolution. In the proposed application, high activations in L^c indicate station–time regions that contributed most to the diagnosed bottleneck cause, supporting interpretation that can be translated into engineering actions. They are treated as saliency-based decision evidence and as hypotheses for engineering verification, not as direct proof of physical causality.

RESULTS

Qualitative simulation behaviour

Figure 2 shows representative queue-length heatmaps for all scenarios. The balanced variant (C0) is characterised by moderate queues without persistent congestion. Human fatigue in the HRC cell (C1) generates gradual queue build-up before S3, consistent with progressively increasing processing time. Cobot degradation (C2) produces a distinct "phase-shift" signature where congestion rises clearly after half the horizon. Safety micro-stops (C3) lead to intermittent congestion and pronounced downtime traces. Cobot failures (C4) manifest as strong congestion correlated with long downtime episodes at S5. Supply delays / starvation (C5) result in low queues throughout the line, reflecting insufficient inflow.

Diagnostic performance of the CNN

The CNN trained on three-channel heatmaps achieves a test accuracy of 0.938 and macro-F1 = 0.938. For n = 225 test samples, the 95% Wilson confidence interval for accuracy is [0.898, 0.963]. Figure 4 presents the confusion matrix. Most classes are well separated; the dominant confusion occurs between the two HRC throughput-drift mechanisms (human fatigue vs cobot degradation), which intentionally share similar congestion signatures but differ in temporal dynamics. The confusion between C1 (human fatigue) and C2 (cobot degradation) was analysed as a boundary case: both mechanisms reduce the effective throughput of the same HRC station S3 and therefore produce similar queue build-up upstream of the cell. The discriminating information is mainly temporal, because fatigue produces gradual deterioration, whereas cobot degradation produces a clearer phase change after T/2. In practice, process heatmaps alone may be insufficient to separate these two causes with full certainty; operator shift rosters, ergonomic observations, cobot maintenance logs and controller diagnostics should therefore be included during real-factory calibration.

Comparison with baseline models

Table 3 compares the CNN with feature-based baselines. Logistic regression and random forest achieve very high accuracy on engineered summary features, reflecting the discriminative power of mean/max queue statistics and downtime measures in controlled synthetic scenarios. In contrast, the CNN operates directly on raw heatmaps without manual feature engineering and provides spatio-temporal explainability via Grad-CAM, which is not natively available in classical models. Therefore, the CNN is not presented as the most accurate model under all synthetic-data conditions. Its value lies in learning directly from raw spatio-temporal process fingerprints and in producing localised visual evidence that can support managerial and engineering root-cause triage when manually engineered features are unavailable, unstable across plants, or difficult to maintain.

Table 3. Test-set results: proposed CNN vs feature-based baselines

Model	Test accuracy	Macro-F1
CNN (multi-channel heatmaps)	0.938	0.938
Logistic regression (summary features)	0.991	0.991
Random forest (summary features)	0.996	0.996

Ablation — single-channel vs multi-channel heatmaps

To assess the value of multiple process signals, an ablation variant of the CNN was trained using only the queue-length channel. The multi-channel representation improves accuracy by 0.102 (absolute). This indicates that utilisation and downtime information provides complementary evidence, particularly for distinguishing safety micro-stops and cobot failures (Table 4).

Explainability analysis

Figure 5 shows Grad-CAM explanations for selected scenarios. The saliency maps consistently highlight blocks corresponding to stations implicated by the bottleneck mechanism (e.g., concentrated activations around S3 for HRC cases and around S5 for cobot failures). Notably, explanations emphasise the time segments in which congestion builds up or downtime episodes occur, enabling engineers to link predictions to tangible process evidence rather than treating the output as a black-box classification. The maps should therefore be interpreted as model-attention evidence indicating where the classifier found discriminative information, while the final causal interpretation must remain with process engineers and managers who can compare the saliency map with physical observations and operational logs.

DISCUSSION

Empirical results support the central hypothesis that CNNs can learn discriminative spatio-temporal patterns from simulation-generated process heatmaps and deliver high diagnostic accuracy. The most pronounced confusion emerges between the human-fatigue and cobot-degradation scenarios, which is expected because both mechanisms reduce the effective throughput of the same HRC station. In a real plant, separating these causes with higher confidence may require additional context (e.g., operator shift rosters, cobot maintenance logs, proximity-sensor traces). Importantly, Grad-CAM explanations remain informative even under uncertainty, as they localise the station–time regions that warrant further investigation.

Beyond predictive performance, the contribution of the proposed pipeline should be read against the broader Industry 5.0 agenda, which foregrounds human-centredness, resilience, and safe collaboration. Recent HRC scholarship emphasises that performance limitations in hybrid teams are rarely purely “technical”: they arise from the coupled dynamics of humans, robots, safety constraints, and organisational arrangements [27,28]. This broader organisational perspective is also consistent with recent analyses of paradoxical tensions in the implementation of human–cobot collaboration under Industry 5.0 conditions [29]. The presented taxonomy explicitly elevates human fatigue and safety micro-stops to first-class bottleneck mechanisms rather than treating them as unstructured noise. This modelling stance is particularly relevant when HRC deployment is pursued not only for throughput but also for ergonomics and risk reduction.

A second alignment with current research concerns synthetic data and digital-twin-enabled learning. There is growing evidence that digital twins can serve as controllable “data factories” for training and validating deep learning models before shop-floor instrumentation and labelling mature [30,31]. This is increasingly coupled with standardisation efforts that aim to move digital

Table 4. Ablation study – effect of multi-channel heatmap encoding

Encoding	Test accuracy	Macro-F1
Queue-only (1 channel)	0.836	0.821
Queue + utilisation + downtime (3 channels)	0.938	0.938

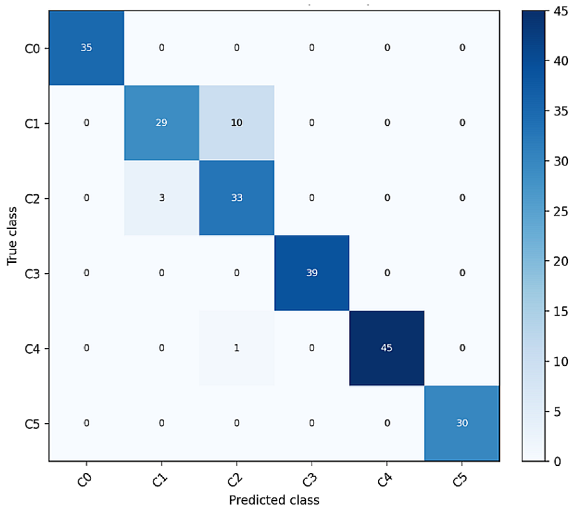

Figure 4. CNN confusion matrix on the test set (classes C0–C5)

twins beyond bespoke, isolated implementations towards interoperable industrial ecosystems [32]. In this context, heatmap-based process fingerprints can be interpreted as a pragmatic interface layer: they compress heterogeneous traces into a stable representation that can be generated consistently across scenarios and, in future work, across plants. For transfer to real-world conditions, synthetic-to-real adaptation should begin with calibration of arrival rates, processing-time distributions, buffer capacities, downtime distributions and safety-stop patterns using a limited set of historical shop-floor logs. A short pilot phase should then compare model diagnoses with expert assessments on real shifts; if discrepancies are observed, domain randomisation, fine-tuning on labelled pilot samples, or extension of the heatmap channels (e.g., energy consumption, proximity-sensor states or robot alarms) can reduce the simulation-to-reality gap.

Explainability is likewise situated in an active and rapidly consolidating research stream. Recent reviews note that, despite accelerating adoption, XAI in manufacturing remains fragmented and often lacks standardised evaluation protocols for explanation quality and human factors [22,23]. From a governance perspective, this matters because explainability increasingly intersects with compliance-oriented documentation, human oversight, and risk management [5]. The EU AI Act has entered into force and is being rolled out progressively, with key milestones for general-purpose AI governance and high-risk systems coming into application in 2025–2027 [6,25]. Although the present study focuses on a decision-support setting, the broader regulatory trajectory reinforces the value of evidence-producing diagnostics, where model outputs can be audited and discussed by process and safety stakeholders. In this study, governance is operationalised narrowly through traceable input signals, documented scenario assumptions, preserved heatmap evidence, human review of Grad-CAM outputs and pilot validation before operational use; the framework is not claimed to constitute a complete regulatory-compliance system.

Baseline comparisons indicate that feature-engineered classical models can achieve near-perfect accuracy in controlled synthetic scenarios, reflecting the discriminative power of aggregated queue and downtime statistics. This does not diminish the relevance of CNNs; rather, it clarifies a practical division of labour. Feature-based monitors can be deployed as lightweight early-warning systems, whereas the CNN+Grad-CAM pipeline provides localised station–time evidence that directly corresponds to observable process behaviour – valuable when an alarm requires root-cause triage and action prioritisation. This distinction also clarifies why higher baseline accuracy does not invalidate the proposed approach: the intended managerial use case is transparent diagnosis and action prioritisation, not only maximum classification accuracy on pre-engineered features.

Several threats to validity remain. Internal validity is supported by controlled scenario definitions; however, the simplified flow-line model cannot capture all complexities of real HRC cells (e.g., multi-skilled operators, parallel stations,

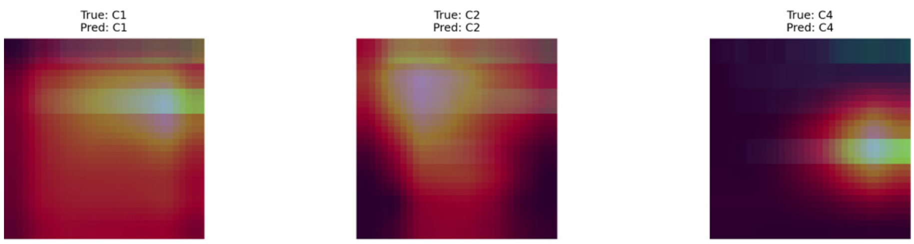

Figure 5. Grad-CAM overlays on queue heatmaps for representative, correctly classified samples

ergonomic constraints, adaptive task allocation, or multimodal safety perception). External validity is limited because the dataset is synthetic; transferring models to real data will likely require digital-twin calibration, domain randomisation, and/or fine-tuning on a small set of labelled shop-floor observations. Construct validity depends on the adopted taxonomy: other bottleneck causes (e.g., rework loops, intralogistics constraints, quality inspection policies) can be incorporated by expanding scenarios and heatmap channels. Finally, future work should examine the robustness of explanations (e.g., stability under perturbations) and integrate complementary causal structures (e.g., knowledge graphs) to strengthen root-cause reasoning beyond classification [19].

CONCLUSIONS

This article introduced an explainable CNN-based bottleneck-diagnosis framework for human–cobot manufacturing settings. A simulation-based digital twin was used to generate labelled, multi-channel process heatmaps encoding queues, station utilisation, and downtimes. On 1,500 synthetic runs, the CNN achieved a test accuracy of 0.938 and a macro-F1 of 0.938, while multi-channel encoding improved diagnostic performance relative to queue-only heatmaps. Grad-CAM explanations consistently highlighted station–time regions aligned with the simulated bottleneck mechanisms, enabling transparent diagnostics that can be translated into corrective actions.

The practical implication is a deployable design-to-deployment workflow: engineers and operations managers can prototype diagnostic models in silico, where bottleneck causes are controllable and labels are unambiguous, and then adapt them to shop-floor data as instrumentation and governance mature. The study also shows that analytics supporting Industry 5.0 human–cobot collaboration should be not only accurate, but also interpretable to the human stakeholders responsible for safety, work organisation, process improvement, and quality performance. From an operations and quality management perspective, the framework can support faster prioritisation of corrective actions, more transparent cross-functional communication, and better-informed decisions regarding workload balancing, maintenance, and safety-related process adjustments. For industrial adoption, companies should first ensure reliable collection of queue, utilisation, downtime, safety-stop, cobot-performance and supply-delay data, then validate the model in a small pilot before using it as routine decision support. Grad-CAM outputs should be used to guide investigation and communication, not to replace causal engineering analysis.

Acknowledgements

The author declares that no external funding, institutional support requiring formal acknowledgment, or other acknowledgments need to be reported.

REFERENCES

1. M B, L DN, A P. Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry. Innovation in the Digital Economy. Luxembourg: Publications Office of the European Union; 2021. Available from: https://op.europa.eu/en/publication-detail/-/publication/468a892a-5097-11eb-b59f-01aa75ed71a1
2. BearingPoint. Industrie 4.0 – Relevanz technologischer Entwicklungen der Industrie 4.0 in 2024. Statista; 2025.
3. BearingPoint. Industrie 4.0 – Herausforderungen durch Industrie 4.0 in 2024. Statista; 2025.
4. Lu Y, Liu C, Wang KIK, Huang H, Xu X. Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robot Comput Integr Manuf. 2020;61:101837. https://doi.org/10.1016/j.rcim.2019.101837
5. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (NIST AI 100-1). Artificial Intelligence Risk Management Framework (NIST AI 100-1). NIST; 2023.
6. Directorate-General for Communication. AI Act enters into force. European Commission [Internet]. 2024 Feb 23;1–3. Available from: https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en
7. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision. 2017; 618–26. https://doi.org/10.1109/ICCV.2017.74
8. Dachepalli V, Gavini S. A Virtually assisted digital twin enabled object detection in smart industrial manufacturing. Expert Syst Appl. 2025 Dec;294:128574. Available from: https://doi.org/10.1016/j.eswa.2025.128574
9. Flaga S, Pacholczak K. Demonstrator of a digital twin for education and training purposes as a web

application. Adv Sci Technol Res J. 2022;16(5):110–9. https://doi.org/10.12913/22998624/152927
10. Daneshjo N, Mareš A, Malega P, Župčan V. Proposal of workplace modification in the assembly line in automotive production. Adv Sci Technol Res J. 2023;17(3):88–100. https://doi.org/10.12913/22998624/163423
11. Pizoń J, Cioch M, Kański Ł, Garcia ES. Cobots implementation in the era of Industry 5.0 using modern business and management solutions. Adv Sci Technol Res J. 2022;16(6):166–178.
12. Pizoń J, Kanski Ł, Kulisz M, Chadam J, García ES. Using business process model and notation 2.0 to deploy cobots in a manufacturing system – Case study. Adv Sci Technol Res J. 2024;18(7):150–62. https://doi.org/10.12913/22998624/192491
13. ISO. ISO/TS 15066:2016 – Robots and robotic devices — Collaborative robots [Internet]. International Organization for Standardization (ISO). Geneva: ISO; 2016. Available from: https://www.iso.org/standard/62996.html
14. ISO. ISO 10218-1 Robotics - Safety requirements Part 1: Industrial robots. Geneva: International Organization for Standardization; 2025. Available from: https://www.iso.org/standard/73933.html
15. ISO. ISO 10218-2 Robotics - Safety requirements Part 2: Industrial robot applications and robot cells [Internet]. Geneva: International Organization for Standardization; 2025. Available from: https://www.iso.org/obp/ui/en/#iso:std:iso:10218:-2:ed-2:v1:en
16. Roser C, Nakano M, Tanaka M. A practical bottleneck detection method. In: Proceedings of the 2001 Winter Simulation Conference. 2001. p. 949–53.
17. Su X, Lu J, Chen C, Yu J, Ji W. Dynamic bottleneck identification of manufacturing resources in complex manufacturing system. Appl Sci. 2022;12(9):4195. https://doi.org/10.3390/app12094195
18. Skoogh A, Thürer M, Subramaniyan M, Matta A, Roser C. Throughput bottleneck detection in manufacturing: a systematic review of the literature on methods and operationalization modes. Prod Manuf Res. 2023;11(1):2283031. https://doi.org/10.1080/21693277.2023.2283031
19. Tang J, Liu Y, Lin K-Y, Li L. Process bottlenecks identification and its root cause analysis using fusion-based clustering and knowledge graph. Adv Eng Inform. 2023;55:101862. https://doi.org/10.1016/j.aei.2022.101862
20. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;1135–44.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017.
22. Puthanveettil Madathil A, Luo X, Liu Q, Walker C, Madarkar R, Qin Y. A review of explainable artificial intelligence in smart manufacturing. Int J Prod Res [Internet]. 2025;63(23):8654–97. Available from: https://doi.org/10.1080/00207543.2025.2513574
23. Tzionis G, Mouratidis P, Kougka G, Gialampoukidis I, Vrochidis S, Kompatsiaris I, et al. A review of explainable AI methods and their application in manufacturing systems. Discov Appl Sci [Internet]. 2026;8(1):52. Available from: https://doi.org/10.1007/s42452-025-07908-z
24. Commission E. AI Act: Timeline. 2024. Available from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
25. Commission E, Desk AIAS. Timeline for the Implementation of the EU AI Act | AI Act Service Desk. 2026. Available from: https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act?
26. Pizoń J, Kański Ł, Chadam J, Pęk B. Image-based time series trend classification using deep learning: A candlestick chart approach. Adv Sci Technol Res J. 2025;19(11):45–58. https://doi.org/10.12913/22998624/208472
27. Roveda L. Editorial: Human–robot collaboration in Industry 5.0: a human-centric AI-based approach. Front Robot AI. 2024;11:1511126. Available from: https://doi.org/10.3389/frobt.2024.1511126
28. Dhanda M, Rogers BA, Hall S, Dekoninck E, Dhokia V. Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0. Robot Comput Integr Manuf. 2025;93:102937. Available from: https://doi.org/10.1016/j.rcim.2024.102937
29. Callari TC, Curzi Y, Lohse N. Realising human-robot collaboration in manufacturing? A journey towards industry 5.0 amid organisational paradoxical tensions. Technol Forecast Soc Change. 2025;219:124249. Available from: https://doi.org/10.1016/j.techfore.2025.124249
30. Urgo M, Terkaj W, Simonetti G. Monitoring manufacturing systems using AI: A method based on a digital factory twin to train CNNs on synthetic data. CIRP J Manuf Sci Technol. 2024;50:249–68. Available from: https://doi.org/10.1016/j.cirpj.2024.03.005
31. Lopes PV, Silveira L, Guimaraes Aquino RD, Ribeiro CH, Skoogh A, Verri FAN. Synthetic data generation for digital twins: enabling production systems analysis in the absence of data. Int J Comput Integr Manuf. 2024;37(10–11):1252–69. Available from: https://doi.org/10.1080/0951192X.2024.2322981
32. Technology NI of S and. Digital Twin Standardization. 2026. Available from: https://www.nist.gov/digital-twins/digital-twin-standardization