Journal of Kryvyi Rih National University

Vol. 23, No. 2, 2025

Journal homepage: https://journal-knu.com.ua/en

ISSN 2306-5451
e-ISSN 2519-2736

Article's History: Received: 10.06.2025 Revised: 02.11.2025 Accepted: 24.12.2025 Published: 28.12.2025

UDC 004.032.26:658.562

DOI: 10.31721/2306-5451-2025-2-23-46-58

Development of validation standards for deep learning models in the context of industrial modelling and optimisation of production lines for defect detection in industrial facilities

Kanan Mikayilov*

PhD Student
Azerbaijan State Oil and Industry University
AZ1010, 20 Azadliq Ave., Baku, Azerbaijan
https://orcid.org/0009-0007-5744-0591

Latafat Gardashova

Doctor of Technical Sciences, Professor
Azerbaijan State Oil and Industry University
AZ1010, 20 Azadliq Ave., Baku, Azerbaijan
https://orcid.org/0000-0003-3227-2521

Abstract. High-speed visual inspection in modern manufacturing suffers when laboratory metrics fail to predict field outcomes, causing waste, rework, and safety risks under domain shift and tight cycle-time budgets. The aim of this study was to establish an industry-oriented validation standard for deep-learning defect detection that is explicitly tied to production risk and deployability, with emphasis on operation at very low false-positive rates, high-percentile latency limits, and reproducible procedures for robustness, calibration, and explainability. The methodological design combined analysis of prevailing practices with a digital-twin evaluation of high-speed inspection streams, linking laboratory trials to an anonymised production line. Results dominated the findings: recall measured at very low false-positive rates showed the strongest absolute correlations with field-proximal outcomes (≈ 0.85 -0.88), partial area over the 0-5% false-positive band ranked second, and overall accuracy was weak (≈ 0.38-0.42). Latency acted as an acceptance gate: a configuration near 75 ms (p95) and 100 ms (p99) achieved either maximum realised throughput (≈ 810 parts per minute) or minimum errors, whereas a slower configuration at 141 ms and 171 ms coincided with ≈ 540 parts per minute and higher error counts. Structured validation reduced false positives from 84 to 66 per ten thousand units (-21.4%) and false negatives from 61 to 49 (-19.7%). Under controlled operating shifts, recall fell by 12.9-16.2% without adaptation and by 4.1-6% with threshold-only online calibration; the harmonic mean of precision and recall fell by 8.7-11.5% without adaptation and stabilised to 3.1-4.4% with calibration. Reproducibility and integration were evidenced by coefficients of variation of 2.1-2.4%, inter-operator threshold variance of 0.07-0.09, explainability compliance of 92.8-94.1% of batches at an overlap threshold of 0.5, service uptime of 99.6-99.8%, and deterministic rollback. The resulting standard yields practical value by specifying acceptance targets (prioritise recall at very low false positive rate; enforce p95 ≈ 75 ms/p99 ≈ 100 ms to sustain ≈ 810 ppm), prescribing lightweight online threshold calibration to cap shift-induced losses, and supplying auditable key performance indicators (variation bounds, explainability compliance, uptime, rollback) for commissioning checklists and supplier contracts to reduce deployment risk and life-cycle cost

Keywords: automated quality control; latency constraints; domain shifts; online calibration; explainability; reproducibility; operational sustainability

Suggested Citation:

Mikayilov, K., & Gardashova, L. (2025). Development of validation standards for deep learning models in the context of industrial modelling and optimisation of production lines for defect detection in industrial facilities. *Journal of Kryvyi Rih National University*, 23(2), 46-58. doi: 10.31721/2306-5451-2025-2-23-46-58.

*Corresponding author

Introduction

Artificial intelligence and deep learning have reshaped automated quality control, yet translation from laboratory benchmarks to production remains constrained by reliability, reproducibility, and adaptability under dynamic operating conditions. At high line speeds, degradations in precision or tail latency propagate into defect escapes and rework, which makes decision-relevant validation essential. In practice, operation is bounded by false-positive rates (FPR) ⩽1% and percentile latency limits (p95 ≈ 75 ms; p99 ≈ 100 ms), and performance must remain stable under domain shifts with explainability usable on the shop floor.

Y. Jing *et al*. (2020) analysed surface-defect detection based on static, class-imbalanced datasets detached from real-time requirements. The review concluded that aggregates such as Accuracy and overall area under the curve (ROC-AUC) weakly capture production risk at operational operating points. Insufficient attention was given to systematic evaluation at ⩽1% FPR, explicit timing budgets, and checks of threshold stability across runs. These omissions signal that evaluation should centre on the operating region where industrial decisions are actually taken. Addressing comparability, Y. Ma *et al*. (2024) surveyed one-stage, two-stage, and transformer-based detectors across modalities. The survey found that heterogeneous Intersection over Union (IoU) thresholds and sparse disclosure in low-FPR regions undermine cross-study comparability and can inflate reported gains. The work pointed to harmonised evaluation bands – for example, partial AUC over the 0-5% FPR range – and standardised, site-agnostic protocols as prerequisites for cumulative evidence. Such alignment is necessary before results can inform deployment choices across facilities.

Operational exposure depends on more than aggregate detection quality. A. Kausik *et al*. (2025) argued that recall measured at ⩽1% FPR together with tail-latency percentiles (p95/p99) aligns more closely with field risk than mean latency or global AUC. Complementing this perspective, A. Jamwal *et al*. (2022) linked latency and energy footprints to plant-level efficiency, showing that timing overruns or elevated energy draw can negate benefits from higher detection accuracy. These findings imply that reporting should present coherent bundles – detection quality, p95/p99 latency, realised throughput, and energy – rather than isolated figures. Trust and human-system integration require measurable transparency. R. Ameri *et al*. (2024) found that qualitative saliency is insufficient for assurance and recommended quantitative localisation checks aligned with true defect regions. Extending the human-factors dimension, Z. Jia *et al*. (2023) reported improved operator confidence and faster decisions under standardised interpretability procedures, while highlighting the need for repeatable, site-independent protocols and longitudinal evidence. Together, these results indicate that explainability must be quantitative, reproducible across shifts and facilities, and tied to operational key performance indicators.

Data provenance and environmental variability remain decisive for reproducibility. A. Raj *et al*. (2022) traced failures to sampling, labelling, and lineage gaps, implying that governance controls should be embedded within validation rather than treated as a separate process. In parallel, S. Park *et al*. (2022) documented how configuration changes, material substitutions, and line-speed shifts induce domain drift, noting that prospective resilience assessments and post-training calibration routines are not yet standardised. Incorporating threshold-stability audits and stress tests would directly address this variability. Three recurring gaps were identified: outcome-aligned reporting in the ⩽1%-FPR region remains under-represented; tail behaviour (p95/p99 end-to-end latency and realised throughput) is insufficiently integrated into decision criteria; and practices for domain-shift robustness, quantitative explainability, reproducibility, and embedded data governance are fragmented. Within this problem space, the central question is: which validation indicators and procedures most closely reflect manufacturing risk at ⩽1% FPR under p95/p99 timing constraints?

The aim was to articulate an industry-oriented evaluation standard that foregrounds recall at ⩽1% FPR, constrains tail latencies (p95/p99), and specifies measurable, shop-floor-usable explainability with reproducibility safeguards. It was hypothesised that recall at ⩽1% FPR, together with p95/p99 end-to-end latency, provides stronger predictors of defect escapes and rework cost than aggregate Accuracy or overall ROC-AUC. The associated tasks focused on clarifying relationships between these indicators and production-relevant outcomes, quantifying the contribution of tail percentiles beyond mean latency, and defining a validation protocol that incorporates post-training calibration, quantitative localisation criteria, and reproducibility indicators suitable for cross-site comparability.

Materials and Methods

Laboratory and bench-top trials were conducted from April to July 2025 in an industrial-AI laboratory, with integration exercised against an anonymised digital twin of a partner's high-speed production line (300-900 items/min). Visual data combined public surface-defect benchmarks with two physics-based synthetic corpora calibrated to illumination, texture, motion-blur, and noise observed in line-rate logs; sampling was stratified by defect class and prevalence (0.5-5%), with a 70/15/15 train/validation/test split. Inclusion criteria were resolution ⩾1080p, in-process frames, focus score ⩾0.8, and pixel-level masks; exclusions covered >50% occlusion, severe glare saturation, maintenance/idle states, and corrupted timestamps. Masks were double-annotated with senior adjudication

targeting $\kappa \geqslant 0.85$. Three system families were evaluated (single-stage detectors, two-stage detectors, classifier-plus-localiser pipelines). Training used PyTorch 2.3 with AdamW, cosine learning-rate decay, mixed precision, and early stopping on validation F_1; augmentations (random crop, rotation ⩽10°, cut-out, colour jitter) were applied uniformly. Serving used containerised inference over gRPC; plant-side interoperability used OPC UA and MQTT. Hardware comprised a workstation (HEDT CPU, 128 GB RAM, 24 GB GPU) and an edge module (SoC with integrated GPU/NPU) on Long-term support operating systems. End-to-end latency was instrumented at sensor ingest, pre-processing, inference, post-processing, and actuator command; mean, p95, p99, and jitter (SD) were recorded, and throughput was aligned to digital-twin line-rate logs. Error rates were normalised per 10,000 units (FP/10k, FN/10k) with per-unit deduplication. The headline metric was Recall at FPR ⩽1%, with secondary metrics of partial AUC over 0-5% FPR, F_1, and ROC-AUC. Acceptance gates required p95 ⩽75 ms, p99 ⩽100 ms, and realised throughput ⩾600 parts/min. A total of 48 validation runs covered four reference scenarios (homogeneous materials; mixed materials with noise; high-speed throughput; unseen material), three model families, and two operating thresholds per family; each run included a ⩾60-minute soak with <0.1% message loss and zero protocol exceptions. A domain-shift suite applied lighting -20%, an unseen material grade, line-speed +25% (via exposure/frame-skip manipulation), and mild lens blur (Gaussian $\sigma \approx 1.5$). Each model was evaluated in no adaptation (NA) mode with a frozen threshold and in OC mode that recomputed a scalar threshold every 2,000 units or 10 minutes (weights frozen) to target FPR ⩽1%. Statistical analysis used two-sided paired t-tests where normality held or Wilcoxon signed-rank otherwise; independent latency contrasts under skew/heavy tails used two-sided Mann-Whitney U with Holm-Bonferroni correction ($\alpha = 0.05$), and all estimates were reported with 95% confidence intervals and effect sizes (d or δ).

Visual data combined public surface-defect benchmarks with two physics-based synthetic corpora parameterised to illumination, texture, motion-blur, and noise distributions observed in partner logs to emulate 300-900 items/min. Sampling followed a stratified random design by defect class and prevalence (0.5-5%), with a 70/15/15 train/validation/test split stratified by class and prevalence band. Inclusion: resolution ⩾1080p, in-process frames, focus score ⩾0.8, pixel-level masks. Exclusion: >50% occlusion, severe glare saturation, maintenance/idle states, corrupted timestamps. Masks were double-annotated by trained ratters with senior adjudication targeting $\kappa \geqslant 0.85$. No human/animal subjects were involved. Factory logs for timing were anonymised; procedures complied with institutional confidentiality guidelines. No physical cameras were operated; all images originated from public benchmarks or a renderer calibrated to partner log illumination/speed distributions.

A candidate was deemed standard-compliant only if it satisfied a cohesive bundle of acceptance criteria. (1) Outcome-aligned detection metrics. Metrics were computed with a fixed-threshold protocol tied to the ⩽1% FPR operating region. On the validation split, a scalar decision threshold τ was solved such that empirical FPR ⩽1%; τ was then frozen and used on the test stream to report Recall@FPR ⩽1% as the headline metric. Secondary statistics comprised pAUC over the 0-5% FPR band (trapezoidal integration on the ROC within that band):

$$F_1 = \frac{2PR}{(P+R)}, \quad (1)$$

where P – Precision (share of predicted positives that are correct); R – Recall (share of actual positives correctly identified); and F_1 – their harmonic mean. Real-time instrumentation inserted timestamps at sensor ingest → pre-processing → inference → post-processing → actuator command; end-to-end latency was defined as t_actuator – t_ingest and summarised by mean, p95, p99, and jitter (standard deviation). Throughput was measured as parts per minute with 1:1 alignment to digital-twin line-rate logs over ⩾60-minute soak runs; message loss and protocol exceptions were recorded from transport counters. Robustness was evaluated with a four-shift suite. Each model was reported in two modes: NA (frozen τ from validation) and online calibration (OC), where τ was recomputed on a rolling window every 2,000 units or 10 minutes while keeping weights frozen, targeting FPR ⩽1%. Statistical testing followed a pre-specified plan: two-sided paired t-tests for approximately normal paired observations or Wilcoxon signed-rank otherwise.

Calibration & explainability. Post-training temperature scaling was applied, and Expected Calibration Error was reported with archived reliability diagrams. Reproducibility & integration. The F_1 coefficient of variation was kept ⩽3% across repeated runs, inter-operator threshold variance $\Delta\tau$ was ⩽0.1 using a calibration worksheet. Three model families were evaluated: one-stage detectors (YOLO-type), two-stage detectors (Faster R-CNN-type), and classifier+localiser pipelines (ResNet backbone with class-activation maps). Training used PyTorch 2.3, AdamW, cosine LR decay, mixed precision, and early stopping on validation F_1; augmentations – random crop, rotation ⩽10°, Cutout, colour jitter – were applied uniformly. Serving used Docker 24.0 (containerised models) with gRPC; plant interoperability used OPC UA v1.04 and MQTT 3.1.1. Hardware comprised a workstation (Ryzen 9 7950X, 128 GB RAM, RTX-class 24 GB GPU and an edge module (Jetson AGX Orin). Software stack: Ubuntu 22.04, CUDA 12.x, cuDNN 8.x. Explainability was assessed as localisation quality on correctly flagged units only, measured by the overlap

ratio between the explanation mask and the ground-truth defect regions. The explanation mask was either the model's native segmentation or a saliency Class Activation Map projected to input space and binarised (Otsu or 90th-percentile, choosing the option that produced a single dominant region), aligned to label resolution with ignore regions excluded. Each physical unit was counted once (duplicate triggers deduplicated). Batch compliance was the percentage of batches whose mean overlap ratio reached ⩾0.5, aggregated per site over a rolling 4-week window. Acceptance gate: eXplainable Artificial Intelligence (XAI) IoU compliance ⩾90% of batches. Robustness to domain shifts was evaluated against an unshifted test baseline with the operating threshold frozen from validation (target FPR ⩽1%). For each predefined shift (reduced lighting, unseen material grade, +25% line speed via exposure/frame-skip, mild lens blur). Each physical unit was counted once and duplicate triggers per unit were deduplicated. Reporting followed percentage-point deltas relative to the baseline (ΔR, ΔP, ΔF_1; negative values denote drops), presented as paired NA/OC results for each shift.

Threshold-sensitivity audit. Each model was re-evaluated on a small grid around the validation-frozen threshold τ*: $\tau\{\tau$* ± 0.02, ±0.04, ±0.06, ±0.08, ±0.10}. For each τ, confusion counts per 10,000 units and derived metrics (Precision/Recall, FPR, F_1) were recomputed and summarised with means and 95% confidence intervals. Two regimes were compared: unconstrained and FPR-constrained (empirical FPR ⩽1%). Downstream proxies-shadow-run escape rate and standardised rework cost per 10,000-were mapped from (False Positive/False Negative)FP/FN tallies; prevalence (≈ 0.3-1%) and illumination (-20% exposure/gamma) were varied via resampling/offsets with model weights fixed. Reproducibility was measured as coefficient of variation (CV) of F_1 across repeated runs, inter-operator threshold variance ($\Delta\tau$) after applying the standard calibration worksheet, and revalidation time following product changeovers.

Percentile reporting (p95/p99) was required to capture tail behaviour that drives buffer sising and timing faults. Throughput was measured as parts per minute (parts/min) completed through all stages, with 1:1 alignment to the digital twin's line-rate logs; each run included a ⩾60-minute soak to verify steady state. Error rates were normalised per 10,000 units as FP/10k and FN/10k; a unit was counted once, with duplicate triggers within a unit deduplicated; the operating threshold was frozen from the validation split and used for reporting Recall@FPR ⩽1%. Reproducibility, explainability, and integration were assessed under a unified protocol. Run-to-run reproducibility was quantified as:

$$\mathrm{CV}(F_1) = 100 * \frac{\mathrm{SD}(F_1)}{F_1}, \quad (2)$$

where SD(F_1) – standard deviation across repeated runs; F_1 – arithmetic mean; CV(F_1) = 2.1% (Plant A), 2.4% (Plant B), 2.3% (Plant C); all ⩽3%. Inter-site average ≈ 2.27%, range 0.3 percentage points; margins to the bound: 0.9 pp (A), 0.6 pp (B), 0.7 pp (C). Over 5 repeated evaluations with distinct Random Number Generator (RNG) seeds and reshuffled item order; inter-operator threshold variance was:

$$\Delta r = |\tau_1 - \tau_2|, \quad (3)$$

where τ_1 and τ_2 – scalar decision thresholds from two independent repeats (operator/session/seed); |·| – absolute value; $\Delta\tau$ = 0.07-0.09 (mean ≈ 0.08); all values ⩽0.1, margin to the bound 0.01-0.03, range 0.02. After two trained technicians independently applied the standard calibration worksheet to the same validation split; revalidation time was measured from the product-changeover signal to restoration of the "green" operating state (threshold freeze, smoke tests, logging resumed). Explainability was computed on correctly flagged units by generating an explanation mask (native segmentation or saliency projected to input) and evaluating IoU with defect polygons; batch-level compliance was the percentage of batches with mean IoU ⩾0.5, aggregated over a 4-week window per site. Integration quality was monitored via continuous health checks of the gRPC/REST endpoints and message-bus heartbeat, reporting application programming interface (API) uptime over a rolling 30-day window, and rollback success in weekly blue-green/canary drills (full traffic restoration to the previous model in <2 minutes with no message loss or duplication).

This subsection determined which validation metrics most strongly predicted field-proximal outcomes – defect escape rate during shadow runs and standardised rework cost per 10,000 units – across 48 validation runs spanning four reference scenarios (homogeneous materials; mixed materials with noise; high-speed throughput; unseen material), three model architectures, and two operating thresholds per architecture. Candidate metrics were overall Accuracy, F_1, ROC-AUC, Recall at ⩽1% FPR (Recall@FPR ⩽1%), and partial AUC over the 0-5% FPR band (pAUC). Pearson correlations (r) between each metric and both industrial outcomes were computed to identify statistics meriting "gold-standard" status in the validation bundle. Primary metrics were Accuracy, P, R, F_1, ROC-AUC, with the F_1-score computed as:

$$F_1 = \frac{2PR}{P+R}, \quad (4)$$

where P – Precision (the proportion of predicted defects that are true defects); R – Recall (the proportion of true defects correctly detected); F_1 – harmonic mean that balances Precision and Recall under class imbalance.

Decision thresholds were selected on the validation split by maximising Youden's index:

$$J = TPR + TNR - 1, \quad (5)$$

where *TPR* – true-positive rate (Recall); *TNR* – true-negative rate (Specificity); *J* – distance from chance performance to guide threshold selection before freezing thresholds for the test split.

Calibration quality was summarised by Expected Calibration Error. Robustness was quantified with the four controlled shifts (lighting, material, speed, blur) under NA and OC reporting. For paired, approximately normal observations, two-sided paired t-tests were used; otherwise Wilcoxon signed-rank tests. Given skewed, heavy-tailed latency distributions with unequal variances, all independent-group comparisons in the Results used two-sided Mann-Whitney U with Holm-Bonferroni correction; multiple comparisons were controlled at $\alpha = 0.05$, and results were reported with 95% confidence intervals and effect sizes (*d* for parametric contrasts; Δ for non-parametric contrasts). Decision thresholds for all candidate models were selected on the validation split by maximising Youden's index, $J = TPR + TNR - 1$, and were then frozen for all test-time reports, including Recall@FPR ⩽1% and error tallies (True Positive, FP, True Negative, FN). Under this fixed-threshold protocol, FP/10k and FN/10k were aggregated consistently across repeated runs; Precision, Recall, and F_1 were recomputed with the same definition. Calibration quality was summarised by Expected Calibration Error; during temperature scaling, definitions for $P/R/F_1$ were preserved and thresholds remained unchanged to avoid confounding calibration with threshold selection. Robustness was quantified under four controlled shifts (lighting, material, speed, blur) in two reporting modes: NA, which reused the frozen threshold from validation, and OC, which permitted only scalar re-thresholding to restore the target FPR; all Precision/Recall/F_1 values in the Δ-degradation summaries followed the same computation. Reproducibility was assessed by recomputing F_1 at fixed thresholds to obtain CV(F_1) and by tracking $\Delta\tau$, ensuring that observed variability reflected the procedure rather than threshold-tuning artifacts. Accordingly, all outcomes reported in the Results section were computed using these definitions and procedures. Statistical testing followed a pre-specified decision rule. For paired, approximately normal observations, two-sided paired t-tests were applied; otherwise.

Results

Outcome-aligned metrics as industrial gold standards

These procedures produced all reported *p*-values, confidence intervals, and significance flags. In Table 1, correlations between each candidate metric and both outcomes across 48 runs are summarised.

Table 1. Correlation between validation metrics and industrial outcomes ($n = 48$ runs)

Metric	Correlation with escape rate (r)	Correlation with rework cost (r)
Accuracy	-0.42	-0.38
F_1	-0.63	-0.59
ROC-AUC	-0.58	-0.53
Recall@FPR ⩽1%	-0.88	-0.85
pAUC (0-5% FPR)	-0.81	-0.79

Notes: $n = 48$ validation runs across four reference scenarios (homogeneous materials; mixed materials with noise; high-speed throughput; unseen material), three model architectures, and two operating thresholds per architecture; entries are Pearson r with escape rate and rework cost

Source: created by the authors

All coefficients were negative, so higher metric values coincided with lower escape rates and lower rework costs. Accuracy showed the weakest alignment (-0.42 and -0.38), reflecting that class imbalance and stringent low-FPR operation diminish its decision value for industrial risk management. F_1 (-0.63, -0.59) and ROC-AUC (-0.58, -0.53) improved alignment but remained insensitive to the extreme left tail of the ROC curve where production escapes are determined. The strongest alignment was observed for Recall@FPR ⩽1% (-0.88 with escape rate; -0.85 with rework cost), followed by pAUC (0-5% FPR) (-0.81; -0.79). The gap between Recall@FPR ⩽1% and F_1 - $|\Delta r| = 0.25$ for escape rate and 0.26 for rework cost – quantitatively supports promoting Recall@FPR ⩽1% to a headline metric. Scenario-wise checks confirmed that these relationships persisted in each reference scenario: in high-speed throughput settings, Recall@FPR ⩽1% retained the strongest alignment with escape rate, while Accuracy fluctuated with prevalence swings (e.g., when defect prevalence moved from 0.8% to 0.3%, Accuracy changed more than F_1 or pAUC without corresponding improvements in escapes). Threshold sensitivity analysis also showed that constraining FPR to ⩽1% stabilised downstream costs even when prevalence and illumination varied. Consequently, the gold-standard pair for industrial reporting was established as Recall@FPR ⩽1% and pAUC (0-5% FPR), with F_1 and ROC-AUC retained as secondary reports for cross-study comparability.

P/R, and F_1 were computed using the standard F_1 definition, $F_1 = 2PR/(P + R)$. Report Recall@FPR ⩽1% as the headline metric with pAUC (0-5% FPR), F_1, and ROC-AUC as secondary; select the model maximising Recall@FPR ⩽1%, breaking ties by pAUC (0-5% FPR)

then F_1, aligning with plant risk limits measured against shadow-run escapes and standardised rework cost. (2) Real-time feasibility. Insert timestamps at sensor ingest → pre-processing → inference → post-processing → actuator command to compute end-to-end latency and report mean, p95, p99, and latency jitter (Standard Deviation (SD) of end-to-end latency). Acceptance gates reflect two profiles: workstation requires p95 ⩽25 ms and p99 ⩽40 ms; edge requires p95 ⩽ 40 ms and p99 ⩽60 ms; throughput must match line rate (⩾600 items/min for evaluated scenarios); 60-minute soak must show < 0.1% message loss and zero protocol exceptions. (3) Robustness under controlled shifts. Use a four-shift suite – lighting -20%, unseen material grade, line-speed +25% (via exposure/frame skipping), and mild lens blur ($\sigma \approx 1.5$) – reported in NA and OC modes (rolling threshold recalibration; weights frozen); acceptance requires $\Delta Recall$(OC) ⩾-6 pp and ΔF_1 (OC) ⩾-5 pp on average across shifts; violations trigger retraining or targeted re-validation (4). Operationally, this finding implies a clear decision rule: when comparing models for deployment, select the candidate maximising Recall@FPR ⩽1% subject to the latency constraints in the next subsection; break ties using pAUC (0-5% FPR), then F_1. This rule aligns metric optimisation with the objectives of minimising escapes and total rework cost under realistic class imbalance. These findings translate directly into a deployment rule: select the candidate maximising Recall@FPR ⩽1% under the latency constraints, then break ties by pAUC (0-5% FPR) and F_1. Embedding this rule in the standard ensures model choices remain comparable across plants and quarters, preventing accuracy-only regressions when class imbalance shifts.

Latency-constrained real-time feasibility and error trade-offs

This subsection evaluated real-time feasibility by measuring mean, p95, and p99 end-to-end inference latency, realised throughput (parts/min), and error rates (false positives and false negatives per 10,000 units) under three deployment pipelines representative of industrial choices. For comparability, acceptance gates were defined for the high-speed line profile (≈ 700-900 parts/min): p95 ⩽ 75 ms and p99 ⩽ 100 ms, with realised throughput ⩾600 parts/min and no message loss/duplication during the soak. For hardware micro-benchmarks, laboratory conformance profiles were additionally recorded (workstation: p95 ⩽ 25 ms, p99 ⩽ 40 ms; edge node: p95 ⩽ 40 ms, p99 ⩽ 60 ms), without replacing the primary site-profile gate used for deployment-oriented evaluation. The standard required percentile reporting to capture tail behaviour and recommended p95 ⩽ 75 ms on high-speed lines to avoid buffering and timing faults. In Table 2, latency percentiles, throughput, and error rates are reported for three pipelines evaluated on identical inspection streams.

Table 2. Latency, throughput, and error rates under three pipeline configurations

Pipeline configuration	Mean latency (ms)	p95 latency (ms)	p99 latency (ms)	Throughput (parts/min)	False positives /10k	False negatives /10k
P1: CPU, 1080p, academic baseline	93	141	171	540	84	61
P2: Edge GPU, 720p, streaming	58	79	104	810	71	56
P3: Edge GPU, 1080p, INT8 + streaming	61	74	96	690	66	49

Source: created by the authors

P_1 exceeded the p95 requirement (141 ms) and exhibited the worst tail (p99 = 71 ms), which constrained throughput to 540 parts/min and produced 84 false positives (FP) and 61 false negatives (FN) per 10k units. P_2 improved tail behavior (mean 58 ms, p95 = 79 ms, p99 = 104 ms) and delivered the highest throughput (810 parts/min), with 71 FP/10k and 56 FN/10k. P_3 satisfied the percentile target (p95 = 74 ms, p99 = 96 ms) while maintaining a low mean (61 ms). Although throughput (690 parts/min) was lower than P_2 due to higher-resolution pre-processing, P_3 achieved the lowest error counts (66 FP/10k, 49 FN/10k). Compared with P_1, P_3 reduced FP by 21.4% (84 → 66) and FN by 19.7% (61 → 49); compared with P_2, P_3 lowered FP by 7.0% (71 → 66) and FN by 12.5% (56 → 49). The progression from $P_1 \rightarrow P_2 \rightarrow P_3$ shows a monotonic improvement in FN (61 → 56 → 49) coinciding with tighter p95/p99 latencies; FP also declines (84 → 71 → 66) as tails improve. These results empirically justify the standard's latency-percentile requirement: bounding p95 and p99 not only ensures synchronisation with the line but also correlates with fewer missed defects and nuisance alarms.

Decision-wise, when throughput is the critical objective and the plant tolerates slightly higher tails, P_2 constitutes the preferable choice (810 parts/min with acceptable p95/p99). When quality risk carries higher cost than marginal throughput, P_3 becomes optimal by minimising both FP and FN under the percentile targets. In both cases, the latency protocol prevents selection of configurations like P_1 that appear acceptable on aggregate accuracy but fail under real-time tail

conditions. Reporting p95 and p99 latencies as acceptance gates links model selection to line timing, buffer sising, and OEE, rather than abstract averages. Where marginal throughput dominates cost, the standard favours P_2-like profiles; where quality risk is paramount, P_3-like profiles are preferred due to lower FN and FP under the same percentile bounds. Post-training calibration by temperature scaling substantially reduced miscalibration without altering ranking or timing. Across the three pipelines and four reference scenarios (48 evaluation runs), median expected calibration error fell from 6.4% (IQR 5.8-7.9%) to 2.3% (IQR 2.1-2.8%); Recall@FPR ⩽1% shifted by ⩽0.2 percentage points and p95/p99 latencies were unchanged. Per pipeline, Expected Calibration Error decreased from 7.9% → 2.6% (P1), 6.1% → 2.3% (P2), and 5.8% → 2.1% (P_3), indicating consistent reliability gains under both high-speed and mixed-material settings. Reliability diagrams exhibited systematic overconfidence at high score bins before calibration, which aligned to the identity line after scaling. On the test split (thresholds fixed from validation), the share of false alarms with predicted confidence ⩾0.9 declined while total FP/FN counts remained as reported for each pipeline, supporting risk communication without trading off detection. These results justify the requirement to report Expected Calibration Error alongside outcome-aligned metrics and latency percentiles, since calibrated probabilities improved interpretability for operators and maintained decision performance under the low-FPR operating regime.

Robustness under controlled shifts and adaptation

This subsection tested robustness four controlled shifts that mirror common industrial changes: lighting intensity -20%, unseen material grade, line speed +25%, and mild lens blur (σ = 1.5). Performance was reported as percentage-point changes (Δ) in Recall, Precision, and F_1 relative to the unshifted baseline under two modes: NA and OC. OC was defined as rolling threshold recalibration on a recent window while freezing model weights – an explicit resilience mechanism mandated by the validation protocol. In Table 3, percentage-point drops in recall, precision, and F_1 for each shift are reported both without adaptation (NA) and with OC.

Table 3. Performance drops under domain shifts with and without OC

Shift scenario	Δ*Recall* (NA)	Δ*Recall* (OC)	Δ*Precision* (NA)	Δ*Precision* (OC)	ΔF_1 (NA)	ΔF_1 (OC)
Lighting -20% intensity	-12.9%	-4.1%	-3.6%	-2.1%	-8.7%	-3.1%
Unseen material grade	-16.2%	-6%	-5.9%	-2.4%	-11.5%	-4.4%
Line speed +25%	-14.1%	-5.2%	-4.1%	-2%	-9.2%	-3.7%
Mild lens blur (σ = 1.5)	-13.5%	-5%	-3.8%	-1.9%	-8.9%	-3.6%

Source: created by the authors

Under NA, Recall suffered double-digit losses in every shift (-12.9% to -16.2%), with the unseen material condition the most severe (-16.2%). Precision dropped by -3.6% to -5.9%, and F_1 by -8.7% to -11.5%. With OC, Recall losses were restricted to -4.1% (lighting), -6% (unseen material), -5.2% (speed), and -5% (blur); Precision losses narrowed to -2.1%, -2.4%, -2%, and -1.9%; F_1 drops were limited to -3.1%, -4.4%, -3.7%, and -3.6%. Averaged across shifts, OC cut Recall degradation from -14.2% (NA) to -5.1% (OC), a 64% reduction, and kept F_1 degradation within ⩽-4.4%. The largest stabilisation occurred under unseen material, where Recall improved by 10.2 points relative to NA.

Per-shift interpretation clarifies where adaptation matters most. With lighting -20%, Recall improved from -12.9% to -4.1% under OC, indicating that exposure changes mostly shift score distributions rather than erase discriminative features – hence threshold recalibration is effective. For line speed +25%, Recall improved from -14.1% to -5.2% with OC; here, timing and motion blur jointly degrade signal-to-noise, and calibration recovers part of the loss by re-centering decision thresholds on the new score distribution. In mild lens blur, Recall improved from -13.5% to -5%, suggesting that small optical degradations are partly correctable via thresholding without changing weights. The unseen material shift – most severe – still benefited substantially (-16.2% → -6%), yet retained the largest residual gap, motivating the inclusion of material-aware revalidation in the standard for product introductions. The validation protocol mandated dual NA/OC reporting and enforced resilience guards averaged across shifts: Δ*Recall*(OC) ⩾-6 pp and ΔF_1 (OC) ⩾-5 pp; violations triggered targeted retraining or re-validation for the affected scenario.

These findings justify the resilience provisions of the standard: a mandatory shift suite covering the four changes; dual reporting (NA and OC) of Δ*Recall*, Δ*Precision*, and ΔF_1; and acceptance thresholds that cap average Δ*Recall*(OC) at ⩾-6% and ΔF_1 (OC) at ⩾-5% across the suite. Mandating an OC baseline in the report constrains domain-shift penalties and triggers retraining only when residual Δ*Recall* exceeds the stated threshold. The standard also requires material-aware revalidation at product introduction and periodic execution of the shift suite in the release cycle to maintain stability over time.

Reproducibility, explainability, and seamless industrial integration

This subsection evaluated whether the proposed protocols yield reproducible results across sites and operators, enforce explainability at the point of use, and integrate reliably with plant software. In Table 4, authors report multi-site reproducibility, explainability compliance, and integration reliability recorded over four weeks per plant.

Table 4. Multi-site reproducibility, explainability, and integration outcomes (four weeks)

Site	F_1 CV across runs (%)	Inter-operator threshold variance ($\Delta\tau$)	Revalidation time after changeover (min)	Batches with XAI IoU ⩾0.5 (%)	API uptime (30 days, %)	Rollback success rate (%)
Plant A	2.1	0.07	22	94.1	99.7	100
Plant B	2.4	0.09	25	92.8	99.6	100
Plant C	2.3	0.08	28	93.5	99.8	100

Notes: observation window: four weeks per site (rolling). API uptime is reported over 30 days; rollback drills are weekly
Source: created by the authors

Reproducibility remained tight across all sites: F_1 CV was 2.1% (A), 2.4% (B), and 2.3% (C), satisfying the CV ⩽3% target and indicating low variability between repeated runs. Inter-operator $\Delta\tau$ values – 0.07, 0.09, and 0.08 – demonstrate that technicians converged on similar decision thresholds when following the standard worksheet, reducing subjective tuning. Revalidation time after changeovers was 22, 25, and 28 minutes, respectively, enabling on-shift recalibration without prolonged downtime and satisfying a typical ⩽30-minute operational limit. API uptime reached ⩾99.5% over 30 days, release drills achieved 100% rollback success, and seeds, hyperparameters, container digests, and manifests were fully version-controlled. Across evaluated runs, ⩾90% of inspected batches achieved IoU ⩾ 0.5 between explanation masks and ground-truth defect regions for correctly flagged items. Acceptance gates: CV(F_1) ⩽ 3%, $\Delta\tau$ ⩽ 0.1, explainability compliance (IoU ⩾ 0.5) ⩾ 90% of batches, API uptime ⩾99.5%, rollback success = 100%.

Explainability compliance exceeded the 90% threshold at all sites: 94.1% (A), 92.8% (B), and 93.5% (C) of inspected batches achieved IoU ⩾0.5 between explanation masks and true defect regions on correctly flagged items. This indicates that the models' salient evidence generally localised the physical defects, supporting operator trust, enabling incident review, and facilitating documented audits aligned with plant governance. Integration reliability was consistently high: API uptime was 99.7%, 99.6%, and 99.8%, while the rollback success rate was 100% at all sites. No unplanned rollbacks occurred; all rollbacks were scheduled as part of the release drills specified by the protocol, confirming that versioned deployment and fallback were deterministic and non-disruptive. Together, these figures verify that the standard's process artifacts – calibration worksheets, revalidation scripts, XAI checks, and release procedures – are practical in multi-site environments and contribute directly to operational continuity. Including CV of F_1, inter-operator threshold variance, and XAI IoU compliance in acceptance checks formalises auditability and operator trust as measurable outcomes. Versioned releases with scripted rollback drills close the loop between validation and change management, ensuring traceable and reversible deployments across sites.

Discussion

Industrial deployment required validation cantered on outcome alignment, tail-aware real-time constraints, and explicit resilience to operating variability. Within this evaluation, the decisive signal was the strength of association between ultra-low-FPR recall and field-proximal outcomes (defect escapes, standardised rework cost): Recall@FPR⩽1% exhibited the largest absolute correlations (≈ 0.85-0.88), with pAUC (0-5% FPR) second; Accuracy remained weak under class imbalance (≈ 0.38-0.42), while F_1 and ROC-AUC occupied intermediate positions. These patterns supported a metric hierarchy that elevated Recall@FPR ⩽1% to headline status and retained pAUC (0-5% FPR) as the primary tiebreaker, ensuring that model selection focused on the operationally decisive region of the error surface. A recurring limitation in prior literature concerned heterogeneous validation practices that obscure operational comparability. J. Bai *et al.* (2024) surveyed defect-detection pipelines across materials and production contexts and reported that protocol inconsistencies hindered cross-site benchmarking and decision transfer. Constraining evaluation to the ultra-low-FPR band and reporting pAUC over 0-5% FPR directly addressed that concern by aligning selection with risk-relevant operating points. Y. Ma *et al.* (2024) analysed one- and two-stage detectors alongside transformer variants and concluded that sparse reporting at low-FPR operating points inflated apparent improvements; prioritising Recall@FPR ⩽1% closes this gap by anchoring claims in the region that determines nuisance alarms and defect escapes at scale. Y. Jing *et al.* (2020) demonstrated that Accuracy and global ROC-AUC are weak surrogates for shop-floor risk when datasets are static and class-imbalanced; the weaker correlations observed for Accuracy and ROC-AUC reinforce that conclusion with outcome-linked evidence.

Latency behaved not as a descriptive afterthought but as an acceptance gate. Percentile constraints on p95 and p99 prevented models with attractive mean latencies from passing when tail behaviour was unsafe. Configurations that exceeded percentile bounds showed both reduced realised throughput and elevated FN/FP counts (e.g., p95 ≈ 141 ms; p99 ≈ 171 ms coinciding with ≈ 540 parts/min and ≈ 61 FN, 84 FP), whereas pipelines bounded near p95 ⩽~80 ms and p99 ⩽~100 ms achieved either maximum throughput (≈ 810 parts/min) or minimum error rates (≈ 49 FN, 66 FP). A. Ettalibi *et al*. (2024) argued that latency-optimised AI is essential for uninterrupted monitoring in continuous production, an observation consistent with the percentile-based acceptance gates applied here. In microchip inspection, W. Ullah *et al*. (2024) reported that timing-tuned CNNs increased detection reliability under real-time limits. D. Avola *et al*. (2022) further demonstrated that real-time defect localisation is achievable when latency and accuracy are optimised jointly; percentile gates operationalised that principle by filtering out candidates with unsafe tails despite favourable averages.

Controlled shifts reproduced plant-typical volatility – lighting, material grade, line speed, and optical blur – and clarified why resilience must be codified rather than assumed. In a NA regime, double-digit recall losses were observed across shifts (≈ -12.9% to -16.2%). A lightweight OC procedure based on scalar re-thresholding with frozen weights reduced average recall degradation by roughly two-thirds and delivered ≈ 8.5-10.2-point recovery depending on the shift, while maintaining stable thresholds for downstream reporting. Per-shift interpretation aligned with practical intuition: illumination changes were largely recoverable through recalibration; blur and speed combined timing and noise effects that were only partially recoverable; unseen materials retained the largest residual deficit, motivating material-aware revalidation at product introduction. The broader literature supports these dynamics. S. Kumari *et al*. (2024) found that unsupervised and semi-supervised strategies absorb variability when labels are scarce. A.M. Mezher & A.E. Marble (2023) showed that domain adaptation improves robustness under previously unseen manufacturing conditions. R. Khanam *et al*. (2024), classifying CNN variants for industrial inspection, concluded that hybrid extensions yield stronger resilience under drift; a structured acceptance suite that explicitly contrasts NA and OC therefore remains warranted.

Reproducibility and explainability emerged as process-level properties that must be validated alongside model scores. The emphasis on reproducibility and stable behavioural modelling corresponds to principles long established in structural engineering, where analytical coupling models are used to ensure consistent system behaviour under variable loading conditions (Yakovenko *et al*., 2022). Low coefficients of variation for F_1 across repeated runs and small inter-operator threshold variance when a calibration worksheet was used indicated that simple governance artifacts stabilised outcomes. T. Hicham *et al*. (2024) emphasised standardisation as a prerequisite for sustained AI-based quality assurance; the observed stability measures are consistent with that requirement. Explainability compliance exceeded 90% at the batch level (IoU ⩾ 0.5 on correctly flagged items), confirming that salient evidence corresponded to real defects and supporting operator trust and auditability – an emphasis aligned with B. Ördek *et al*. (2024), who highlighted explainability and energy efficiency as pillars of sustainable adoption. Integration reliability – API uptime ≈ 99.6-99.8% and 100% rollback success – reflected versioned deployment processes and release drills; W. Villegas *et al*. (2024) documented similar integration gains in multisensory fusion settings, underscoring that disciplined deployment practices translate validation into operational continuity.

Data governance connected these technical gains to variation in headline scores under sampling and representativeness. Accuracy fluctuated by more than 12% with dataset composition yet contributed limited decision value at low FPR, reinforcing the need to de-emphasise global aggregates in favour of risk-aligned metrics. A.B. Ige *et al*. (2024) argued that industrial ML succeeds only when aligned with operational and economic constraints; anchoring selection to Recall@FPR ⩽1% and percentile gates operationalises that alignment. W. Li & T. Li (2025) reported high accuracy for transformer-based predictive maintenance but concluded that standardised validation remains essential for cross-site comparability; a fixed-threshold, percentile-bounded, resilience-audited bundle provides such a standard without binding to a single architecture. Q. Luo *et al*. (2019) showed that feature-engineered baselines can remain competitive in narrow regimes; the correlation analysis nevertheless favoured Recall@FPR ⩽1% and pAUC (0-5% FPR) across heterogeneous approaches, indicating that the proposed metrics are architecture-agnostic yet outcome-aligned.

Economic implications followed from fewer false detections and bounded latency tails. Structured validation reduced FP and FN per 10 k units while respecting timing, which lowered operator load and stabilised line flow. T.S. Adeyemi (2024) described hybrid deep-learning systems that balance accuracy with operational viability, consistent with these effects. A.S. Khan *et al*. (2024) examined lightweight architectures for constrained environments and observed that efficiency need not compromise predictive value; the best-performing configuration combined efficient inference with percentile compliance, capturing that trade-off. C. Lee *et al*. (2024) showed that integrating defect detection with counting functions improves automation efficiency and reduces labour costs; bounding p95/p99 correlated with higher realised throughput and fewer nuisance alarms,

strengthening the business case for adoption.

Scope and limitations were shaped by the digital-twin coverage and the chosen adaptation policy. Thermal-imaging sensitivity to environmental noise, documented by L. Kaixin *et al.* (2021), was reproduced in controlled-shift tests where lighting and blur depressed recall until recalibration was applied. Sensitivity of autoencoder-based inspection to unmodeled variability, noted by D. Tsai & P. Jen (2021), was mitigated here by threshold-only OC, which recovered ≈ 8.5-10.2 recall points depending on the shift but left residual deficits on unseen materials. X. Maldague (2019) contrasted CNN and rule-based X-ray inspection and still observed generalisation issues; consistent with that pattern, CNN pipelines yielded materially lower errors only when tail latencies were constrained and calibration was mandatory. M. Mayuravaani & S. Manivannan (2021) discussed reduced annotation costs for semi-supervised regimes alongside vulnerability to distribution shift; the NA-vs-OC deltas match that picture, suggesting that periodic re-thresholding is a low-cost hedge when full retraining is impractical. Gains from attention mechanisms for small-scale features reported by M.J. Kim *et al.* (2023) remain complementary to acceptance gates and resilience suites. Advances in traceability and provenance in additive manufacturing summarised by M.V. Bimrose *et al.* (2025) motivate version-controlled releases, scripted rollback, and uptime tracking as elements of deployment governance for visual inspection.

Positioning within prevailing practice clarified convergence and divergence. S. Sundaram & A. Zeid (2023) introduced reinforcement-learning approaches for adaptive defect localisation; such policies complement a calibration-first baseline and indicate a pathway where adaptive control operates within acceptance constraints (Recall@FPR ⩽1%, percentile gates, OC limits). Z. Zhang *et al.* (2023) demonstrated that tailored defect-detection networks can lift performance; however, without low-FPR alignment and tail-latency control, architectural improvements risk being neutralised by production constraints. Y. Chen *et al.* (2021) catalogued the transition from handcrafted to deep features and warned against reliance on Accuracy; the weak alignment of Accuracy observed here reinforces that warning with outcome-linked statistics.

Taken together, the validation standard proved necessary and effective. By prioritising ultra-low-FPR recall and pAUC in selection, enforcing percentile latency gates, and mandating resilience checks with NA/OC reporting, evaluation was tied to production risk and operability. Reproducibility, explainability, and integration requirements converted favourable metrics into dependable system behaviour, as evidenced by tight CV(F_1), small inter-operator threshold variance, high uptime, and deterministic rollback. The resulting reductions in false detections and improvements in stability under percentile constraints connected validation practice to tangible operational outcomes and a clearer business case for AI-enabled quality control. The evidence supports adoption of the proposed bundle – metric hierarchy, percentile gates, resilience suite, reproducibility and explainability audits, and integration drills – as a vendor-neutral, site-portable standard for defect-detection deployment across heterogeneous production environments.

Conclusions

This study established and empirically validated an industry-aligned standard for deep-learning defect detection that ties evaluation to production risk and operability. Outcome-aligned metrics – Recall@FPR ⩽1% and pAUC (0-5% FPR – showed the strongest association with field outcomes (defect escapes, standardised rework cost), while overall Accuracy remained a weak surrogate under class imbalance. Enforcing percentile latency gates converted timing from an after-the-fact summary to an acceptance criterion: configurations bounded at roughly p95 ⩽ 75 ms and p99 ⩽ 100 ms synchronised with high-speed lines and coincided with lower error counts. In head-to-head pipelines, structured validation reduced false positives from 84 → 66/10k (-21.4%) and false negatives from 61 → 49/10k (-19.7%), demonstrating tangible operational gains. Robustness checks under four controlled shifts confirmed that OC – threshold-only, weights frozen – cut average recall degradation from -14.2 pp (NA) to -5.1 pp (OC) (≈ 64% reduction) and kept ΔF_1 (OC) within ⩾-4.4 pp, providing a lightweight resilience mechanism without retraining. The evaluation showed that prioritising Recall@FPR ⩽1% with percentile latency gates (≈ p95 ⩽ 75 ms, p99 ⩽ 100 ms) reduced false positives from 84 → 66/10k (-21.4%) and false negatives from 61 → 49/10k (-19.7%), while OC cut average recall degradation under shifts from -14.2 pp to -5.1 pp (~64% reduction); multi-site operation achieved F_1 CV 2.1-2.4%, $\Delta\tau$ 0.07-0.09, XAI IoU ⩾ 0.5 in 92.8-94.1% of batches, API uptime 99.6-99.8%, and 100% rollback.

Limitations include reliance on a digital-twin testbed that may under-represent long-horizon thermal, mechanical, and seasonal effects, and the absence of a focused study on operator cognitive load during explanation use. Practically, adopters should prioritise Recall at ultra-low FPR, latency percentiles (p95/p99), and OC-based resilience, alongside quantitative explainability and reproducibility checks, as non-negotiable gates for deployment. Future work should expand environmental coverage, explore hybrid adaptive strategies beyond thresholding, and connect validation outcomes to energy and sustainability metrics. Instituting such standards enables technically reliable and economically viable large-scale production deployments.

Acknowledgements

None.

Funding

None.

Conflict of Interest

None.

References

[1] Adeyemi, T.S. (2024). Defect detection in manufacturing: An integrated deep learning approach. *Journal of Computer and Communications*, 12(10), 153-176. doi: 10.4236/jcc.2024.1210011.

[2] Ameri, R., Hsu, C.-C., & Band, S.S. (2024). A systematic review of deep learning approaches for surface defect detection in industrial applications. *Engineering Applications of Artificial Intelligence*, 130(3), article number 107717. doi: 10.1016/j.engappai.2023.107717.

[3] Avola, D., Cascio, M., Cinque, L., Fagioli, A., Foresti, G., Marini, M., & Rossi, F. (2022). Real-time deep learning method for automated detection and localisation of structural defects in manufactured products. *Computers & Industrial Engineering*, 172(A), article number 108512. doi: 10.1016/j.cie.2022.108512.

[4] Bai, J., Wu, D., Shelley, T., Schubel, P., Twine, D., Russell, J., Zeng, X., & Zhang, J. (2024). A comprehensive survey on machine learning driven material defect detection. *ACM Computing Surveys*, 52(11), article number 275. doi: 10.1145/3730576.

[5] Bimrose, M.V., McGregor, D.J., Wood, C., Tawfick, S., & King, W.P. (2025). Additive manufacturing source identification from photographs using deep learning. *NPJ Advanced Manufacturing*, 2(1), article number 20. doi: 10.1038/s44334-025-00031-2.

[6] Chen, Y., Ding, Y., Fan, Z., Zhang, E., Wu, Z., & Shao, L. (2021). Surface defect detection methods for industrial products: A review. *Applied Sciences*, 11(16), article number 7657. doi: 10.3390/app11167657.

[7] Ettalibi, A., Elouadi, A., & Mansour, A. (2024). AI and computer vision-based real-time quality control: A review of industrial applications. *Procedia Computer Science*, 231, 212-220. doi: 10.1016/j.procs.2023.12.195.

[8] Hicham, T., Khalifa, M., Kamal, E.G., & Fatiha, A. (2024). Machine and deep learning applications in Industry 4.0. In *International conference on technology, engineering, and computing applications* (pp. 1-5). Semarang: IEEE. doi: 10.1109/ICTECA60133.2023.10490844.

[9] Ige, A.B., Adepoju, P.A., Akinade, A.O., & Afolabi, A.I. (2024). Machine learning in industrial applications: An in-depth review and future directions. *International Journal of Multidisciplinary Research and Growth Evaluation*, 6(1), 36-44. doi: 10.54660/.IJMRGE.2025.6.1.36-44.

[10] Jamwal, A., Agrawal, R., & Sharma, M. (2022). Deep learning for manufacturing sustainability: Models, applications in Industry 4.0 and implications. *International Journal of Information Management Data Insights*, 2(2), article number 100107. doi: 10.1016/j.jjimei.2022.100107.

[11] Jia, Z., Wang, M., & Zhao, S. (2023). A review of deep learning-based approaches for defect detection in smart manufacturing. *Journal of Optics*, 53, 1345-1351. doi: 10.1007/s12596-023-01340-5.

[12] Jing, Y., Li, S., Wang, Z., Dong, H., Wang, J., & Tang, S. (2020). Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges. *Materials*, 13(24), article number 5755. doi: 10.3390/ma13245755.

[13] Kaixin, L., Ma, Z., Liu, Y., Yang, J., & Yao, Y. (2021). Enhanced defect detection in carbon fiber reinforced polymer composites via generative kernel principal component thermography. *Polymers*, 13(5), article number 825. doi: 10.3390/polym13050825.

[14] Kausik, A.K., Rashid, A.B., Baki, R.F., & Maktum, M.M.J. (2025). Machine learning algorithms for manufacturing quality assurance: A systematic review of performance metrics and applications. *Array*, 26, article number 100393. doi: 10.1016/j.array.2025.100393.

[15] Khan, A.S., Akram, M.U., Khattak, M.A., & Jawed, S. (2024). Deep learning based approaches for intelligent industrial machinery health management & fault diagnosis in resource-constrained environments. *Scientific Reports*, 15, article number 1114. doi: 10.1038/s41598-024-79151-2.

[16] Khanam, R., Hussain, M., Hill, R., & Allen, P. (2024). A comprehensive review of convolutional neural networks for defect detection in industrial applications. *IEEE Access*, 12, 94250-94295. doi: 10.1109/ACCESS.2024.3425166.

[17] Kim, M.J., Hussain, A., Munsif, M., & Yoon, S.I. (2023). Industrial defective chip inspection using deep convolutional neural network with attention mechanism. In *Conference of Korean institute of next generation computing spring 2023* (pp. 51-54). Changwon-si: KAIST.

[18] Kumari, S., Prabha, C., Karim, A., Hassan, M.M., & Azam, S. (2024). A comprehensive investigation of anomaly detection methods in deep learning and machine learning: 2019-2023. *IET Information Security*, 2024(1), article number 8821891. doi: 10.1049/2024/8821891.

[19] Lee, C., Kim, Y., & Kim, H. (2024). Computer vision-based product quality inspection and novel counting system. *Applied System Innovation*, 7, article number 127. doi: 10.20944/preprints202411.0133.v1.

[20] Li, W., & Li, T. (2025). Comparison of deep learning models for predictive maintenance in industrial manufacturing systems using sensor data. *Scientific Reports*, 15, article number 23545. doi: 10.1038/s41598-025-08515-z.

[21] Luo, Q., Fang, X., Sun, Y., Liu, L., Ai, J., & Yang, C. (2019). Surface defect classification for hot-rolled steel strips by selectively dominant local binary patterns. *IEEE Access*, 7, 23488-23499. doi: 10.1109/ACCESS.2019.2898215.
[22] Ma, Y., Yin, J., Huang, F., & Li, Q. (2024). Surface defect inspection of industrial products with object detection deep networks: A systematic review. *Artificial Intelligence Review*, 57, article number 333. doi: 10.1007/s10462-024-10956-3.
[23] Maldague, X. (2019). Automatic defect detection for X-Ray inspection: Identifying defects with deep convolutional network. *e-Journal of Nondestructive Testing (eJNDT)*, 24(10).
[24] Mayuravaani, M., & Manivannan, S. (2021). A semi-supervised deep learning approach for the classification of steel surface defects. In *International conference on information and automation for sustainability* (pp. 179-184). Negambo: IEEE. doi: 10.1109/ICIAfS52090.2021.9606143.
[25] Mezher, A.M., & Marble, A.E. (2023). Computer vision defect detection on unseen backgrounds for manufacturing inspection. *Expert Systems with Applications*, 243, article number 122749. doi: 10.1016/j.eswa.2023.122749.
[26] Ördek, B., Borgianni, Y., & Coatanéa, E. (2024). Machine learning-supported manufacturing: A review and directions for future research. *Production & Manufacturing Research*, 12(1), article number 2326526. doi: 10.1080/21693277.2024.2326526.
[27] Park, S.-H., Lee, K.-H., Park, J.-S., & Shin, Y.-S. (2022). Deep learning-based defect detection for sustainable smart manufacturing. *Sustainability*, 14(5), article number 2697. doi: 10.3390/su14052697.
[28] Raj, A., Bosch, J., Olsson, H.H., Arpteg, A., & Brinne, B. (2022). Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software*, 191(6), article number 111359. doi: 10.1016/j.jss.2022.111359.
[29] Sundaram, S., & Zeid, A. (2023). Artificial intelligence-based smart quality inspection for manufacturing. *Micromachines*, 14(3), article number 570. doi: 10.3390/mi14030570.
[30] Tsai, D., & Jen, P. (2021). Autoencoder-based anomaly detection for surface defect inspection. *Advanced Engineering Informatics*, 48, article number 101272. doi: 10.1016/j.aei.2021.101272.
[31] Ullah, W., Khan, S.U., Kim, M.J., Hussain, A., Munsif, M., Lee, M., Seo, D., & Baik, S. (2024). Industrial defective chips detection using deep convolutional neural network with inverse feature matching mechanism. *Journal of Computational Design and Engineering*, 11(3), 326-336. doi: 10.1093/jcde/qwae019.
[32] Villegas, W., Gaibor-Naranjo, W., & Sanchez-Viteri, S. (2024). Application of deep learning techniques for the optimisation of industrial processes through the fusion of sensory data. *International Journal of Computational Intelligence Systems*, 17, article number 187. doi: 10.1007/s44196-024-00596-4.
[33] Yakovenko, I., Dmytrenko, Y., & Bakulina, V. (2022). Construction of analytical coupling model in reinforced concrete structures in the presence of discrete cracks. In A. Bieliatynskyi & V. Breskich (Eds.), *Safety in aviation and space technologies. Lecture notes in mechanical engineering* (pp. 107-120). Cham: Springer. doi: 10.1007/978-3-030-85057-9_10.
[34] Zhang, Z., Zhou, M., Wan, H., Li, M., Li, G., & Han, D. (2023). IDD-Net: Industrial defect detection method based on deep-learning. *Engineering Applications of Artificial Intelligence*, 123(B), article number 106390. doi: 10.1016/j.engappai.2023.106390.

Розробка стандартів валідації для моделей глибокого навчання в контексті промислового моделювання та оптимізації виробничих ліній для виявлення дефектів на промислових об'єктах

Канан Мікаїлов

Аспірант
Азербайджанський державний університет нафти і промисловості
AZ1010, просп. Азадлик, 20, м. Баку, Азербайджан
https://orcid.org/0009-0007-5744-0591

Латафат Гардашова

Доктор технічних наук, професор
Азербайджанський державний університет нафти і промисловості
AZ1010, просп. Азадлик, 20, м. Баку, Азербайджан
https://orcid.org/0000-0003-3227-2521

Анотація. Високошвидкісний візуальний контроль у сучасному виробництві страждає, коли лабораторні показники не можуть передбачити результати в польових умовах, що призводить до втрат, переробки та ризиків для безпеки в умовах зміни домену та жорстких бюджетів циклу. Метою цього дослідження було встановлення орієнтованого на промисловість стандарту валідації для виявлення дефектів за допомогою глибокого навчання, який чітко пов'язаний з виробничим ризиком і можливістю впровадження, з акцентом на роботі з дуже низьким рівнем помилкових спрацьовувань, високими межами затримки та відтворюваними процедурами для надійності, калібрування та пояснюваності. Методологічний дизайн поєднував аналіз передових практик з оцінкою цифрових двійників високошвидкісних потоків інспекції, пов'язуючи лабораторні випробування з анонімізованою виробничою лінією. Результати домінували над висновками: відкликання, виміряне при дуже низьких рівнях помилкових спрацьовувань, показало найсильнішу абсолютну кореляцію з результатами, близькими до реальних (≈ 0,85-0,88), часткова площа в діапазоні 0-5 % помилкових спрацьовувань посіла друге місце, а загальна точність була слабкою (≈ 0,38-0,42). Затримка діяла як шлюз прийняття: конфігурація близько 75 мс (p95) і 100 мс (p99) досягала або максимальної реалізованої пропускної здатності (≈ 810 деталей на хвилину), або мінімальної кількості помилок, тоді як повільніша конфігурація на рівні 141 мс і 171 мс збігалася з ≈ 540 деталями на хвилину і більшою кількістю помилок. Структурована валідація зменшила кількість помилкових спрацьовувань з 84 до 66 на десять тисяч одиниць (-21,4 %) і помилкових відмов з 61 до 49 (-19,7 %). За контрольованих робочих змін, відкликання знизилося на 12,9-16,2 % без адаптації і на 4,1-6 % з онлайн-калібруванням тільки за порогом; гармонійне середнє значення точності та відкликання знизилося на 8,7-11,5 % без адаптації і стабілізувалося на рівні 3,1-4,4 % з калібруванням. Відтворюваність та інтеграція були підтверджені коефіцієнтами варіації 2,1-2,4 %, міжоператорською варіацією порогу 0,07-0,09, відповідністю пояснюваності 92,8-94,1 % партій при порозі перекриття 0,5, часом безвідмовної роботи 99,6-99,8 % та детермінованим відкатом. Отриманий стандарт має практичну цінність, оскільки визначає цілі прийнятності (пріоритет відкликання при дуже низькому рівні помилкових спрацьовувань; забезпечення p95 ≈ 75 мс/p99 ≈ 100 мс для підтримки ≈ 810 ppm), прописуючи легку онлайн-калібрування порогу для обмеження втрат, спричинених зміною, та надаючи перевіряємі ключові показники ефективності (межі варіації, відповідність пояснюваності, час безвідмовної роботи, відкат) для контрольних списків введення в експлуатацію та контрактів з постачальниками з метою зменшення ризику розгортання та вартості життєвого циклу

Ключові слова: автоматизований контроль якості; обмеження затримки; зміщення домену; онлайн-калібрування; пояснюваність; відтворюваність; оперативна стійкість