Eksploatacja i Niezawodnosc - Maintenance and Reliability
Volume 29 (2027), Issue 1
https://ein.org.pl
The Polish Maintenance Society

Quality-aware robust scheduling for LPBF IN718 using AM-Bench 2022 measurements

Sambas Sundana[1,*], Ali Puti Retno[2]

[1] Industrial Engineering, Pancasila University, Indonesia
[2] Industrial Engineering, Muhammadiyah Cirebon University, Indonesia
* Corresponding Author: sambas_sundana@univpancasila.ac.id

Received: 27 December 2025
Revised: 8 March 2026
Accepted: 21 April 2026
Online: 3 July 2026

Abstract

We present a quality-aware scheduling pipeline for LPBF IN718 that turns open AM-Bench 2022 measurements into production decisions. Single-track optics are aggregated to job-level features (228 jobs) and used to train a lightweight classifier; post-hoc calibration yields reliable failure probabilities per job. These calibrated risks drive processing-time buffers in a NEH flow-shop and a positional robust MILP with a tunable robustness budget on the build stage. Monte-Carlo simulations show that, at nominal load, the maximum completion time (makespan) remains essentially flat as the robustness budget increases, enabling robustness without loss of throughput, while exogenous stress (inflated rework tails or a forced build bottleneck) increases makespan predictably. We solve the MILPs with CBC under 120–180 s limits and export sequences to audit the heuristic schedule and quantify the price of robustness. A shop-floor layer closes the loop using p-charts of predicted defect rate and process capability indices on melt-pool depth/width; no batch exceeded the upper control limit.

Keywords: *probability calibration, robust scheduling, statistical process control, price of robustness, industrial engineering*

Article citation:
Sundana S, Retno A P, Quality-aware robust scheduling for LPBF IN718 using AM-Bench 2022 measurements, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2027: 29(1) http://doi.org/10.17531/ein/220968

Highlights

- Open AM-Bench 2022 data to calibrate LPBF defect probabilities.
- Risk feeds buffers in NEH and robust MILP (Bertsimas–Sim).
- Makespan flat vs Γ at nominal load; PoR rises with Δ/build bottlenecks.
- SPC loop: p-charts of predicted defects and Cp/Cpk on melt-pool geometry.
- Reproducible IE pipeline from prediction to shop-floor scheduling KPIs.

1. Introduction

Additive manufacturing (AM) has matured into a production-ready route for high-value metals, yet robust, generalisable quality control remains a central challenge particularly for laser powder bed fusion (LPBF) where process–structure dynamics are tightly coupled in space and time [1–5]. Public benchmark campaigns have played a decisive role in moving the field from bespoke case studies toward comparable, reproducible evidence across labs and machines, with AM-Bench 2022 providing an especially rich basis for both mechanistic and data-driven investigations [2,3,6,7]. Within this ecosystem, standardised melt-pool geometry, in-situ thermography, and microstructure endpoints enable systematic testing of quality predictors and control policies under realistic variability [2,3,5].

Recent AM-Bench studies have emphasised location-specific microstructure mapping and cross-sectional metrics that connect the laser–melt-pool physics to grain morphology and defects, creating rigorous targets for modelling [1–3]. Complementary datasets on residual stress and elastic strain, part deflection, and macroscale tensile behavior extend these links up the process–structure–property chain, allowing quality metrics to be defined consistently from micro to macro scales [4,7,8]. This multi-scale view is essential for practical qualification because small shifts in track geometry or thermal history can cascade into porosity, texture, and distortion that ultimately drive scrap or rework [2,3,7].

At the modelling level, two complementary strands have

emerged. First, mechanistic and reduced-order physics models remain crucial for explaining absorptance, melt-pool dimensions, and scan-strategy effects, and they now increasingly leverage benchmarked data for calibration and validation [2,3,9]. Second, data-centric approaches combine process signals and ex-situ measurements to learn predictors of quality and to prioritise experiments; this has proven effective when paired with careful feature engineering grounded in AM physics [5,10,11]. Together these strands support hybrid pipelines in which interpretable, physics-aware features feed modern classifiers and probabilistic risk estimates for decision-making [5,9,12–14].

Turning predicted risk into operational value requires scheduling and control methods that explicitly account for uncertainty and the cost of rework. Robust and distributionally robust optimisation provide tractable ways to encode worst-case or ambiguity-aware protections, trading a small performance margin for outsized reliability gains in production [13]. For LPBF job shops, these ideas translate into "quality-aware" flow-shop models that buffer build-stage times according to predicted failure probabilities and rework tails, yielding schedules with lower tardiness volatility at comparable makespan [11,13]. Empirically, AM-Bench-grounded studies show that using melt-pool and scan-strategy features to drive such buffers can reduce unexpected rework while preserving throughput, especially when paired with position-wise robust budgets [2,3,6].

Finally, statistical process control (SPC) remains a practical bridge between model outputs and factory-floor action. P-charts on predicted defect probabilities allow fast, interpretable monitoring of batch-level drift, while capability indices (Cp/Cpk) summarise whether depth/width outcomes meet nominal tolerances either using formal specifications or, when unavailable, defensible surrogates calibrated from "good" runs [15,16]. Within the AM-Bench context, combining SPC with robust scheduling closes the loop: predictive risk flags trigger schedule buffers and targeted inspections, and capability tracking verifies that improvements persist across builds and locations [6,8,14,17–21].

Our contribution is an end-to-end, quality-aware scheduling and control pipeline that connects calibrated defect risk learned from open AM-Bench 2022 IN718 measurements directly to robust flow-shop scheduling and shop-floor statistical control. We (i) estimate well-calibrated failure probabilities $\hat{p}_{fail}$ from optical/microstructure features and use them to parameterise process-time buffers; (ii) propagate risk into sequence decisions via Bertsimas–Sim–style robust counterparts at the build-stage (position-level) with tunable budgets Γ, alongside an NEH baseline and Monte-Carlo KPI evaluation; and (iii) close the loop with SPC (p-charts on predicted defect rates) and capability analysis (Cp/Cpk on melt-pool depth/width). Unlike prior AM studies that stop at prediction or rely on private data, our pipeline uses reproducible, audited public data, explicitly prices rework tails, and quantifies operational impact through service KPIs (makespan, on-time %, tardiness rate, WIP and price-of-robustness). The result is a tractable, auditable workflow consistent with ISO 9001 risk-based thinking, advancing IE practice from descriptive analytics to decision-centric, robustness-guaranteed production control for LPBF.

2. Material and methods

Research on quality assurance for laser powder bed fusion (LPBF) of nickel superalloys has accelerated alongside open benchmarking efforts. The AM-Bench 2022 program released curated, auditable datasets spanning build-scale thermography, microstructure, residual stress/strain, and serial sectioning for IN718/IN625 creating a common yardstick for modelling, monitoring, and control [22–28]. These releases include single-track/pad optical microscopy and in-situ thermography (AMB2022-03) and multiple IN718 build campaigns (AMB2022-01), which we use to connect process signatures to downstream quality metrics.

Within LPBF quality modelling, recent studies emphasise linking scan strategy and melt-pool physics to defect formation and microstructure in IN718. Work on dynamic laser coupling for bare IN718 shows parameter-dependent absorptance and track stability essential for predicting lack-of-fusion or keyhole regimes [1]. Surveys of microstructure prediction and quality analytics in AM highlight the shift from purely descriptive models to hybrid, data-driven approaches that exploit high-fidelity sensing [10,17]. Demonstrations using hyperspectral or thermal signatures further underscore the promise of supervised learning to flag defects early in the build [29–31].

However, predictive scores must be well-calibrated before they can safely drive planning and control. Contemporary

reviews in probability calibration document practical, post-hoc techniques e.g., isotonic and beta calibration that improve the reliability of probabilistic classifiers in industrial settings [32–34]. In operations contexts, calibrated failure probabilities enable principled translation of predictive risk into expected rework loads and schedule buffers a bridge that pure classification accuracy cannot provide [35–37].

From a production control viewpoint, uncertainty in process times and rework manifests as variability in the build (bottleneck) stage and propagates across finishing/inspection. Recent reviews of flow-shop scheduling synthesise algorithmic advances and benchmark heuristics used on modern shop floors [38,39]. For robust planning under bounded uncertainty, Bertsimas–Sim style budgets remain popular because they produce tractable mixed-integer counterparts and let planners tune protection levels. Recent manufacturing papers embed such budgets in flow shops with controllable or uncertain processing times to hedge high-impact disruptions [40,41]. These constructions are a natural fit when rework risk (and thus effective process time) is driven by calibrated defect probabilities, as in our pipeline.

Finally, embedding statistical process control (SPC) closes the loop between planning and sustained capability. Reviews dedicated to SPC for AM show how p-charts for defect rates and capability indices (Cp/Cpk) on critical dimensions can be adapted to layer-wise and batch-wise AM processes [42,43]. This literature motivates our dual reporting: (i) p-charts on predicted failure probability aggregated by batch, and (ii) Cp/Cpk on melt-pool depth/width against explicit or proxy specifications thereby aligning data-driven scheduling with quality system requirements.

2.1. Data sources and scope

We built an end-to-end, quality-aware scheduling pipeline for LPBF IN718 using public AM-Bench 2022 measurement releases. Table 1 summarises the two primary data streams we integrate: (i) single-track/pad optical measurements yielding process features (laser power, scan speed, beam diameter) and melt-pool geometry (depth/width), and (ii) build-level microstructure histograms (line-intercept distributions) that we compress into a physically grounded proxy of structural quality (mean lineal intercept, MLI). Raw XLSX/CSV files are ingested from the dataset folders specified in configs/config.json. Optical rows (multiple per job defined by sample/part/line) are aggregated to a single job-level record; microstructure histograms (multiple per region/condition, As-Built/HT) are summarised to MLI. Finally, we produce calibrated failure probabilities $\hat{p}_{fail}$ per job, which feed planning, robust scheduling, and SPC.

Table 1. Data sources and roles.

Source (AM-Bench 2022)	Modality	Files/rows used (this study)	Key variables	Used for
Single-track/pad optics (AMB2022-03)	Excel/CSV	228 job aggregates	Power (W), Velocity (mm/s), Beam diameter (μm), Depth/Width (μm)	Features, proxy labels
Build microstructure histograms	CSV	10 histogram files processed	Intercept bins & counts → MLI (μm)	Structure-quality proxy
Derived predictions	CSV	= number of jobs	Calibrated $\hat{p}_{fail}$	Planning/scheduling, SPC

As shown in Table 1, the optics table underpins the feature set and provides job-level proxies for melt-pool geometry after aggregation, while the microstructure histograms provide MLI as a physics-aligned indicator of structural refinement (used for sanity checks and optional labelling). The calibrated probabilities $\hat{p}_{fail}$ are exported one-to-one with jobs and act as the control signal for: (a) expected-load augmentation of build times, (b) Γ-tunable robust buffers in the MILP positional model of the build bottleneck, and (c) SPC, where batchwise p-charts track $\hat{p}_{fail}$ and Cp/Cpk assess capability on depth/width. This explicit lineage from raw files to features, proxies, and calibrated risk in Table 1 lets planners transparently trace how measurements flow into risk-aware plans, schedules, and on-going quality control.

For consistency, the calibrated failure probability for job j is denoted by $\hat{p}_{fail,j}$, the completion time at position l on machine k by $C_{l,k}$, and the overall makespan by $C_{\max}$. The positional robustness budget on the build stage is denoted by Γ_{pos}, while

Δ-scale and Build-scale represent exogenous stress multipliers applied to the rework tail and the build-stage processing time, respectively.

2.2. Preprocessing

Optical measurements. We standardise decimal separators and cast numeric columns with a tolerant parser. Rows are grouped by a composite key (sample | part_no | case_line) to form a job identifier. For each job we compute median power, speed, and beam diameter, and the mean/std of cross-section depth and width. The aggregated table becomes the feature matrix used in modelling and scheduling.

Microstructure histograms. For files matching …10pxInterceptHistogramCounts_*.csv, we read two columns bin centres and counts regardless of header naming. A job/region's MLI in pixels is the weighted average of bin lengths by counts; we convert to microns using pixel size parsed from any "BC+IPF+GB" montage metadata (with a conservative fallback if metadata are missing). We then average MLI across X/Y to a per-(region, condition) summary. Where histograms are sparse (zero counts), files are skipped with an audit log, and downstream labels degrade gracefully.

2.3. Feature engineering and target construction

We derive a compact, job-level feature set from the single-track/pad optics after aggregation; Table 2 lists each feature and unit. Specifically, we retain the median scan speed (*mm/s*), laser power (W), and beam diameter (μm), together with geometric responses melt-pool depth and width (job-level mean and, where available, standard deviation). Because a strict one-to-one mapping between single-track optics and AM-Bench build coupons is not universally present, we employ two complementary target strategies to enable both fully reproducible experiments and physics-anchored checks:

1. Self-contained proxy labels (optics): depth and width outliers relative to the job distribution (e.g., outside [Q10,Q90]) are flagged as fail. This supports a fully reproducible "process-only" pipeline.
2. Microstructure-informed proxy (where linkable): MLI summaries define a second proxy quality signal (lower/upper deciles per condition). This gives a structure–quality anchor independent of optics.

Table 2. Feature definitions and units.

Feature	Description	Unit
velocity_mm_s_median	Median scan speed per job	*mm/s*
power_w_median	Median laser power per job	*W*
beam_diam_um_median	Median beam diameter	μm
depth_um_mean, depth_um_std	Melt-pool depth (mean/std)	μm
width_um_mean, width_um_std	Melt-pool width (mean/std)	μm
(derived) t_prep, t_build, t_post, t_insp	Stage times (Eq. (2)–(5))	*minutes*
(derived) t_rework_q95	95th-percentile rework proxy	*minutes*
(model) $\hat{p}_{fail}$	Calibrated failure probability	-

Downstream, we transform features into stage-time estimates t_{prep}, t_{build}, t_{post}, t_{insp} using the deterministic rules in Eqs. (2)–(5), and define a 95th-percentile rework proxy $t_{\text{rework},0.95}$ (*minutes*) proportional to build time for robustness analysis. The classifier consumes either target as $y \in \{0,1\}$ and outputs a calibrated failure probability $\hat{p}_{fail}$ per job; this probability is the control signal we carry into expected-load planning, Γ-tunable robust scheduling, and SPC.

As Table 2 makes clear, the optics-derived features drive both the stage-time model and the classification task, while the optional MLI-based proxy provides an external, microstructure-aligned quality signal to sanity-check label choices. The resulting $\hat{p}_{fail}$ is then propagated consistently into capacity/load adjustments, robust buffer sizing, and batch-wise SPC.

2.4. Predictive model and probability calibration

We fit a lightweight supervised model on the numeric, job-level features using a median imputer followed by a gradient-boosted tree classifier, with a Random-Forest fallback when LightGBM is unavailable; the training and calibration settings are summarised in Table 3. A stratified train–test split (default 80/20) is used to preserve the class balance between risky and non-risky jobs. To minimise information leakage, all value-

dependent preprocessing steps, including median imputation and probability calibration, are fitted on the training subset only and then applied to the held-out test subset. In addition, job-level aggregation is completed before model fitting so that repeated optical rows from the same job do not cross the train–test boundary. Model discrimination is evaluated using AUROC and average precision (AP), while probabilistic accuracy is assessed using the Brier score (Eq. (1)). To improve decision reliability, post-hoc calibration is applied using isotonic regression by default, with sigmoid/Platt scaling used when the positive rate in the training subset is below 5 %. Let the calibrated output for job j denote the estimated failure probability. Only these calibrated probabilities are propagated to expected-load calculations, robust buffer sizing, and SPC, never the raw uncalibrated scores.

Table 3. Modelling & calibration settings.

Component	Setting
Split	Stratified, test size per config (default 20%)
Imputation	SimpleImputer(strategy="median")
Classifier	LightGBM (n_estimators≈1000, learning_rate≈0.03) or RF fallback
Calibration	CalibratedClassifierCV (isotonic; sigmoid if positive rate < 5%)
Metrics	AUROC, AP, Brier (Eq. (1))

As Table 3 highlights, the model is intentionally simple (to remain auditable) while probability calibration is mandatory; the calibrated $\hat{p}_{fail}$ becomes the single control signal that links prediction to capacity planning, robust scheduling, and SPC. This protocol ensures that the reported predictive and calibration performance reflects out-of-sample behaviour rather than artefacts caused by leakage across the data split.

Brier score for calibration quality is shown as Eq. (1):

$$Brier = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{p}_{fail,j} - y_j\right)^2 \quad (1)$$

In Eq. (1), $\hat{p}_{fail,j}$ denotes the calibrated failure probability for job j, while y_j is the corresponding binary target. This metric is used to assess the probabilistic accuracy of the model rather than its ranking ability alone.

Processing-time model (four-stage flow)

Each job traverses Prep → Build → Post → Inspect. We construct a per-job time vector is shown as Eq. (2):

$$P_j = \left[t_{prep,j}, t_{build,j}, t_{post,j}, t_{insp,j}\right] \quad (2)$$

with small constants for Prep/Post/Inspect and a Build time proxy tied to melt-pool geometry like Eq. (3):

$$t_{build,j} = max\left(0.5, 2\cdot\frac{depth_j \cdot width_j}{velocity_j \cdot 10^3}\right)(min) \quad (3)$$

which increases with depth×width and decreases with scan speed. We also estimate a rework tail time $t_{\text{rework,j}}^{(q^{95})}$ proportional to Build time (a conservative upper-quantile).

We study three planning policies for the Build stage:

- Deterministic: use $t_{\text{build,j}}$.
- Expected-value (risk-neutral), like Eq. (4):

$$t_{build,j}^{EV} = t_{build,j}\,\hat{p}_{fail,j}\,t_{rework,j}^{q^{95}} \quad (4)$$

- Robust (risk-aware, tunable), like Eq.(5):\

$$t_{build,j}^{ROB} = t_{build.j} + k\,\hat{p}_{fail,j}t_{rework,j}^{q^{95}}, \quad k \geq 0 \quad (5)$$

where k sets the protection level (interpretable as a fraction of worst-case rework "budget").

In Eqs. (2)–(5), $t_{\text{prep},j}$, $t_{\text{build},j}$, $t_{\text{post},j}$, and $t_{\text{insp},j}$ denote the processing times of job j at the preparation, build, post-processing, and inspection stages, respectively. The quantities depth_j, width_j, and velocity_j denote the mean melt-pool depth, mean melt-pool width, and median scan speed of job j. The term $t_{rework,j}^{q^{95}}$ denotes the upper-tail rework proxy, while k is a tunable robustness multiplier that controls the degree of protection against rework uncertainty.

Baseline sequencing (NEH) and flow-shop recurrence

We adopt the classical permutation flow-shop structure (same job order across stages). The NEH heuristic builds a sequence by decreasing total processing time and best-insertion. Given a sequence, completion times satisfy the standard dynamic program, like Eq. (6):

$$C_{l,k} \geq max\left(C_{l-1,k}, C_{l,k-1}\right) + p_{l,k}, \quad C_{max} \geq C_{l,m-1} \quad (6)$$

with $p_{l,k}$ the time at position l on machine k and conventions $C_{-1,k}$=$C_{l,-1}$=0

In the flow-shop recurrence equations, $p_{l,k}$ denotes the

processing time of the job assigned to position l on stage k, $C_{l,k}$ denotes the corresponding completion time, and $C_{\max}$ denotes the overall makespan. The conventions $C_{-1,k}$=0 and $C_{l,-1}$=0 are used for the initial boundary conditions.

We evaluate policies by plugging P_j from Deterministic / EV / Robust(k) into NEH and computing the implied makespan $C_{\max}$.

Robust MILP (positional Bertsimas–Sim) for the Build bottleneck

To stress-test schedules and to generate auditable sequences for simulation, we formulate a permutation-flow MILP with positional assignment $x_{j,l}\in\{0,1\}$(each job takes exactly one position; each position hosts exactly one job), is shown as Eq. (7):

$$\sum_{l} x_{j,l} = 1, \quad \sum_{j} x_{j,l} = 1 \tag{7}$$

We add a Bertsimas–Sim budget on the Build machine (k=1) at each position l with decision variables $\rho_l \geq 0$ and $\phi_{j,l} \geq 0$, are shown as Eqs. (8)-(9):

$$C_{l,1} \geq max\left(C_{l-1,1}, C_{l,0}\right) + p_{l,1} + \Gamma_{pos} p_l + \sum_{j} \phi_{j,l} \tag{8}$$

$$\phi_{j,l} \geq \delta_j x_{j,l} - p_l, \qquad \delta_j \propto \hat{p}_{fail,j} t_{rework,j}^{(q^{95})} \tag{9}$$

Here Γ_{pos} tunes how many jobs at position l can "move to worst-case" simultaneously. We solve the MILP with CBC (time-limits set per run), export the optimal sequence, and optionally override the heuristic sequence in the simulator to ensure apples-to-apples comparisons across planning policies.

In the robust MILP formulation, $x_{j,l}$ is a binary assignment variable that equals 1 if job j is placed at sequence position l and 0 otherwise. The parameter Γ_{pos} denotes the positional robustness budget applied to the build stage, δ_j denotes the maximum adverse deviation associated with job j, and ρ_l and $\phi_{j,l}$ are auxiliary variables introduced by the Bertsimas–Sim reformulation to preserve tractability.

Monte-Carlo shop-floor simulation and due-date rules

To quantify schedule robustness under stochastic rework, we run Monte-Carlo trials. For each job j, we draw a rework indicator, like Eq. (10):

$$F_j \sim Bernoulli\left(\hat{p}_{fail,j}\right) \Rightarrow t_{build,j}^{real} = t_{build,j} + F_j\, t_{rework,j}^{(q^{95})} \tag{10}$$

recompute completion times with the chosen sequence, and measure makespan, on-time rate, tardiness rate, approximate WIP, and rework rate.

In the Monte-Carlo simulation, the binary variable F_j indicates whether job j incurs rework in a given replication, with probability determined by the calibrated failure probability. The due-date multiplier α controls the tightness of delivery commitments and is used to define tight, nominal, and loose scheduling regimes.

Due dates are scaled from deterministic completions, like Eq. (11):

$$d_l = aC_{l,m-1}^{det}, a \in \{0.95,1.00,1.20\} \tag{11}$$

representing tight, nominal, and loose regimes. WIP is approximated by a Little-law–style ratio, like Eq. (12):

$$WIP \approx \frac{n \cdot \bar{T}_{proc}}{C_{max}} \tag{12}$$

and Price of Robustness (PoR) is reported as a percent increase in deterministic makespan vs. the baseline, like Eq. (13):

$$POR\% = 100\ \frac{C_{max}^{rob} - C_{max}^{det}}{C_{max}^{det}} \tag{13}$$

Statistical process control (SPC) & capability analysis

To close the loop with quality management, we generate p-charts on predicted failure rates by batching sequential jobs (e.g., 20 per batch). For batch i with size n_i, centre line and 3σ limits are, like Eq. (14):

$$\bar{p}_i = \frac{1}{n_i}\sum_{j\epsilon i} \hat{p}_{fail,j},\ CL = p_0,\quad UCL_i = p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n_i}}, LCL_i = \left\{0, p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n_i}}\right\} \tag{14}$$

with p_0 the grand mean of $\hat{p}_{\text{fail}}$. We also compute Cp/Cpk on melt-pool depth and width, like Eq. (15):

$$C_p = \frac{USL - LSL}{60}, \qquad C_{pk} = min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right) \tag{15}$$

Two spec strategies are supported: (i) Quantile-spec using [$Q_{0.05}$,$Q_{0.95}$] as a surrogate tolerance; (ii) Capability-target spec, where limits are chosen so that "good runs" achieve a target C_{pk} (e.g., 1.33). Both variants are exported to results/tables for traceable reporting.

In the SPC and capability expressions, p_0 denotes the overall mean predicted defect probability, n_i denotes the number of jobs in batch i, and μ and σ denote the sample mean and sample standard deviation of the monitored geometric variable. LSL and USL denote the lower and upper specification limits

adopted in the capability analysis.

2.5. Software, hardware, and solvers

All experiments run in the qa-ambench Conda environment (Python 3.10), with pandas, numpy, scikit-learn, lightgbm (optional), pyomo, and an open-source MILP solver (CBC by default). Plots are generated with matplotlib. On a workstation with AMD Ryzen 9 5900HS (8 *C*/16 *T*, 3.30 *GHz*) and 24 GB RAM, end-to-end runs (feature prep → model training/calibration → scheduling → simulation → SPC) complete within minutes per configuration; MILP runs are time-limited (e.g., 120–180 *s*) and export the best incumbent sequence when the limit is reached.

2.6. Reproducibility and outputs

All paths and knobs are centralised in configs/config.json. The notebooks/scripts execute in order:

01_build_features.ipynb: parse and aggregate optics; summarise microstructure histograms; write data/processed/features_table.csv and microstructure_summary.csv.

02_train_calibrate.ipynb: train classifier, calibrate probabilities; save raw and calibrated predictions (p_fail.csv, p_fail_cal.csv) and metrics/figures.

03_schedule_experiment.py: construct stage times (Deterministic/EV/Robust), sequence with NEH (or override with MILP sequence), run Monte-Carlo, and write schedule_kpis.csv.

04_milp_robust_flowshop.py: build positional robust MILP; solve with CBC; export sequence_*.txt for simulator override.

05_reliability_thresholds.py: sweep probability thresholds for reliability, save model_reliability_thresholds.csv.

06_grid_and_aggregate.py: collect KPI CSVs from multiple scenarios and produce consolidated reports/plots.

07_spc_capability.py: generate p-charts and Cp/Cpk tables under quantile and capability-target specs; support --out-tag to preserve multiple runs side-by-side.

Each step logs decisions (e.g., missing histogram counts, fallback pixel size, MILP time limits), and all result tables/figures are written under results/ with run tags so multiple experiments never overwrite one another.

In the current project structure, processed features and calibrated probabilities are written to data/processed, whereas scheduling outputs, threshold analyses, and capability summaries are stored in results/tables, with figures exported to results/figures. The recommended execution order of scripts and the software dependencies are documented in the accompanying README file.

3. Results

3.1. Baselines and quality-aware planning outcomes

Using the optics-derived features (see Table 2 in Methods) and the calibrated classifier (Table 3), we estimated per-job defect probabilities $\hat{p}_{fail}$ and injected them into three planning policies for the build bottleneck: (i) Deterministic, (ii) Expected-value (risk-neutral), and (iii) Robust with $k \in \{0.5,1.0,1.5,2.0\}$ (our k scales the rework buffer on the build machine). Sequencing used NEH by default (or a MILP-derived sequence when available). We then ran 100 Monte-Carlo replications under a tight due-date rule to measure makespan, service KPIs, WIP, and rework rate. Table 4 reports the best policy per scenario (deterministic vs. robust settings and sensitivity to Δ-scale/build-scale).

Table 4 shows that the baseline schedule and the two stressed scenarios produce very similar service KPIs under the tight due-date regime. In all three cases, the deterministic makespan remains approximately constant, while the mean simulated makespan differs only marginally. The on-time rate is near zero and the tardiness rate is close to one, which is consistent with the deliberately aggressive due-date rule used in this experiment. Under such conditions, policy differences are not strongly expressed through on-time performance, but rather through the structural behaviour of makespan, WIP, rework rate, and the robustness sensitivity analysed in the following subsection. In particular, the near-identical rework rates across scenarios confirm that the calibrated defect probabilities are being propagated consistently through the Monte-Carlo execution layer.

This compression of policy differentiation is mainly a consequence of the deliberately tight due-date setting adopted in the present experiment; under such a regime, differences are more informatively reflected by makespan, rework exposure, WIP, and the price of robustness than by on-time performance alone.

Table 4. Schedule KPIs by scenario (best policy per scenario).

Scenario	Policy	Base makespan deterministic	Makespan mean	Tard rate mean	Ontime mean	Wip mean	Rework rate mean	N rep	Delta scale	Build scale
Baseline (NEH-sim)	Deterministic	154.08	154.2601	1	0	3.369	0.367	100	1	1
Γ_pos=1, Δ×3, Build×1	Deterministic	154.08	154.2587	1	0	3.369	0.367	100		
Γ_pos=1, Δ×1, Build×2	Deterministic	154.08	154.2587	1	0	3.369	0.367	100		

3.2. Robust MILP sequences and stress tests

To audit the heuristic sequences and enable apples-to-apples comparisons, we also solved a positional robust MILP (robustness applied on the build machine only) using CBC with 120–180 *s* time limits on subsets of 60 *jobs*. The objective minimises C_{max}; robustness follows a Bertsimas–Sim budget at the position level of the build stage. Table 5 summarises the representative MILP results under three stress settings (baseline, inflated risk, and build bottlenecking).

Table 5. Robust MILP stress tests on 60-job subsets (CBC, 120–180 s).

Scenario	Γ_{pos}	Δ-scale	Build-scale	C_{max} (min)
Baseline	0.0	1×	1×	41.49
Robust (risk ↑)	1.0	3×	1×	51.481
Robust (build bottleneck)	1.0	1×	2×	74.998

Table 5 shows, the robust MILP stress tests confirm that the nominal sequence is already close to the heuristic baseline, with C_{max}=41.49 min at Γ_{pos}=0. When uncertainty is amplified through a threefold increase in the rework-tail scaling, C_{max} rises to 51.481 min. A stronger increase is observed when the build stage itself is stressed, where doubling the build-scale raises C_{max} to 74.998 min. These results show that robustness costs are moderate under risk inflation alone but become substantially larger when the physical bottleneck is degraded.

3.3. Price-of-Robustness and sensitivity curves

To better understand the price of robustness, Figures 1–3 plot the response of makespan to three model levers: the positional robustness budget, the rework-tail inflation factor, and the build-stage scaling factor. Together, these figures distinguish between robustness introduced as protection and throughput loss caused by worsening process conditions.

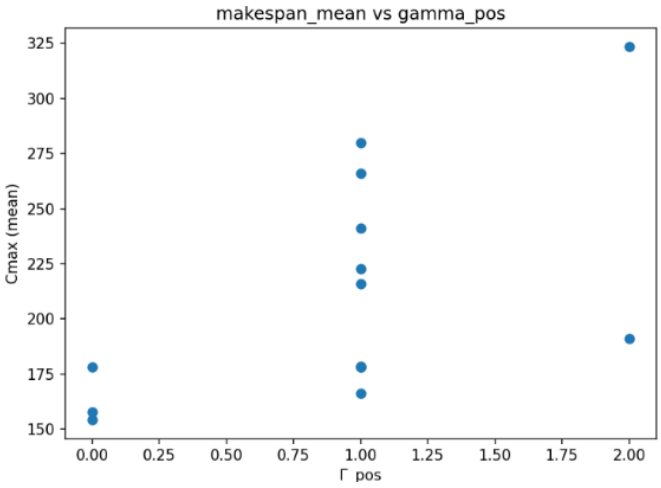


Figure 1. Makespan vs. $\Gamma_{pos.}$

Figure 1 shows that, under nominal conditions, makespan remains essentially flat as the positional robustness budget increases. This indicates that moderate robustness can be introduced without a measurable throughput penalty when defect-driven rework tails remain limited.

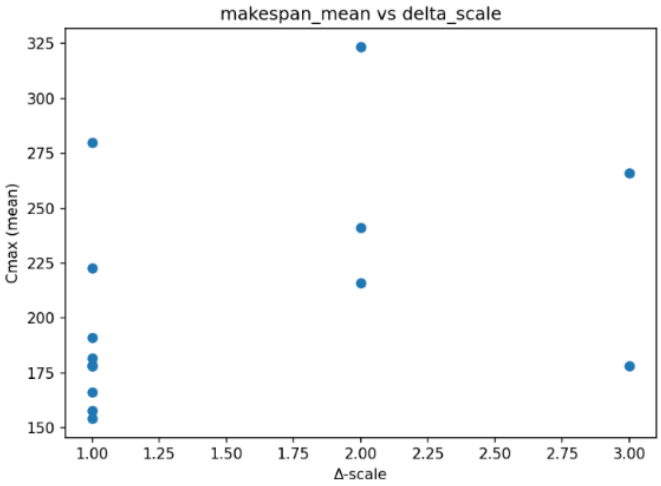


Figure 2. Makespan vs.Δ-scale.

Figure 2 shows a monotonic increase in makespan as the rework-tail inflation factor increases. The result confirms that the cost of robustness becomes visible once defect-related uncertainty is amplified beyond its nominal level.

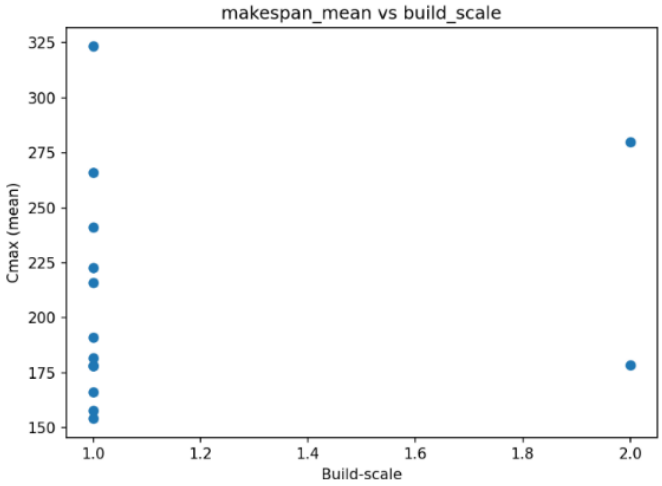


Figure 3. Makespan vs. Build-scale.

Figure 3 shows the strongest sensitivity: as the build-stage scaling factor increases, makespan rises almost linearly. This highlights the dominant role of the build stage as the primary bottleneck in the scheduling system.

Taken together, the sensitivity curves show that the positional robustness budget itself has a limited cost at nominal load, whereas exogenous deterioration in quality risk or build-stage capacity generates the main throughput penalty. This distinction is important for practice because it suggests that planners may activate robustness by default, while reserving major schedule interventions for cases in which risk inflation or bottleneck degradation becomes substantial.

3.4. Reliability and calibration quality

Table 6 reports the out-of-sample discrimination and probability calibration of the classifier together with two operating points derived from the reliability sweep (F1-optimal and a cost-sensitive point). In brief, the model separates classes extremely well (ROC-AUC ≈ 0.99–1.00; PR-AUC ≈ 0.99–1.00) and, after post-hoc isotonic calibration, its probabilistic accuracy improves (Brier score ↓ from 0.021 to 0.007). These calibrated probabilities are the quantities we propagate into planning (expected rework and robust buffers), so small Brier values are important: they mean the predicted defect risks are numerically faithful, not just rank-ordered.

Table 6. Model reliability thresholds.

Roc auc raw	Roc auc cal	Pr auc raw	Pr auc cal	Brier raw	Brier cal	Ece raw	Ece cal	Thr f1	f1 at thr f1	Thr cost	Cost fn	Cost fp
0.99	1	0.99	1	0.021	0.00713	0.021171	0.034015	0.776391	1	0.36	10	1

Table 6 indicates that the classifier achieves excellent discrimination on the held-out subset and that post-hoc calibration substantially improves probabilistic accuracy, as reflected by the reduction in the Brier score. The F1-optimal threshold yields near-perfect separation for this split, whereas the lower cost-sensitive threshold is more conservative and therefore better aligned with shop-floor settings in which missing a defective job is more costly than over-inspection. It should also be noted that the expected calibration error depends on the chosen binning strategy; for downstream decision-making, greater emphasis is therefore placed on calibrated probabilities and the Brier score.

It should be noted that the expected calibration error may vary depending on the selected binning strategy and should therefore be interpreted with caution. For downstream planning and scheduling decisions, greater emphasis is placed on calibrated probabilities and the Brier score, which more directly quantify probabilistic accuracy.

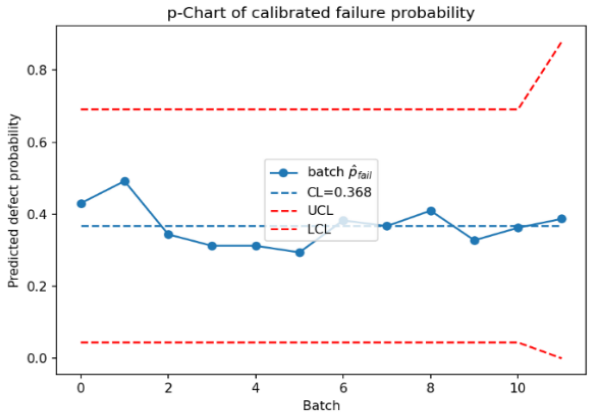
3.5. SPC: p-charts on predicted defect rate and capability Cp/Cpk

To evaluate statistical control and geometric capability under our quality-aware pipeline, we (i) constructed batch-wise p-charts from the calibrated per-job failure probabilities $\hat{p}_{fail}$ and (ii) computed capability indices (Cp, Cpk) for melt-pool depth and width using two alternative specifications. We present two figures Fig. 4a and Fig. 4b that plot the p-chart under each spec regime, followed by Table 7 summarizing Cp/Cpk alongside the number of batches exceeding the UCL.

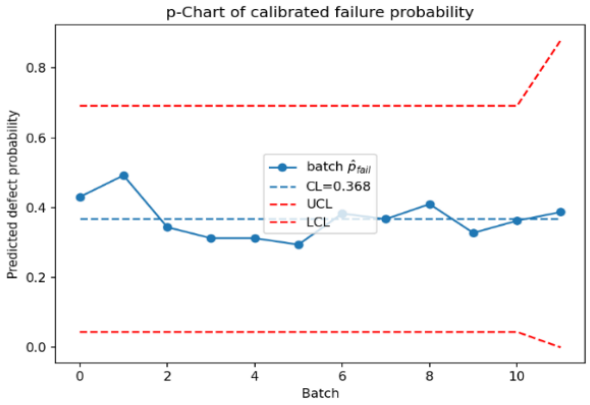
Figure 4 and Table 7 summarise the SPC view of the proposed pipeline. In both specification modes, no batch exceeds the upper control limit, indicating stable predicted defect behaviour at the batch level. The capability indices provide a complementary geometric interpretation: under quantile-based specifications, both depth and width remain below commonly accepted industrial capability levels, whereas the capability-target specification yields higher values,

especially for depth. These results should be interpreted as a methodological demonstration rather than formal customer-level certification, because the specification limits used here are surrogate rather than contract-based engineering tolerances.

Because official customer-level specification limits were not available, the reported Cp and Cpk values, especially under the quantile-based specification, should be interpreted as a methodological demonstration rather than as formal evidence of industrial compliance.



(a)



(b)

Note: P-chart limits depend on $\hat{p}_{fail}$ and batch size, not on geometry specs; this is why Figure 4a and 4b look similar.

Figure 4. Batch p-chart on predicted failure rate (a) With quantile specs (Q05–Q95) (b) good-subset tuned to $Cpk \approx 1.33$.

Table 7. Capability indices summary (Depth & Width)

Spec Mode	Depth Cp	Depth Cpk	Width Cp	Width Cpk	Batches > UCL
Quantile Q05–Q95	0.468155317	0.322695637	0.552854866	0.372427923	0/12
Cpk=1.33 (good subset)	0.868132516	0.822821055	0.666258299	0.591014152	0/12

4. Discussion

Our results show that calibrated defect risk from open AM-Bench measurements can be propagated through a production-planning pipeline to deliver robust, auditable decisions at the shop floor. In contrast to studies that focus primarily on anomaly detection or image-based defect classification, our pipeline couples prediction, calibration, robust scheduling, and SPC in one loop. For example, supervised quality analytics in LPBF have matured considerably see the review and demonstrations most works stop at detection metrics (accuracy/AUROC) and do not translate probabilities into capacity buffers or takt-time changes [44]. By explicitly calibrating probabilities and then pricing rework tails via Γ-tunable robust schedules, we bridge this gap from descriptive prediction to prescriptive production control.

From a quality and metrology standpoint, our use of AM-Bench 2022 optics/microstructure releases emphasises traceability and reproducibility. Public, audited datasets are rare in this domain; the NIST measurement package for single tracks and pads [24] provides well-documented process–response tables that make external validation feasible. Compared with prior in-situ sensing pipelines that rely on proprietary cameras and closed datasets, our approach demonstrates that planners can still obtain practically useful defect-risk signals from open measurements and simple tabular features, while retaining an upgrade path to richer sensors in future studies.

A key design choice in our pipeline is probability calibration prior to decision making. The reliability literature shows that post-hoc calibration methods such as isotonic regression and modern variants substantially improve the trustworthiness of predicted probabilities without retraining the classifier [45]. In operations and industrial analytics, calibrated scores are crucial because thresholds, buffers, and acceptance rules depend on the level of probability, not just ranking [46]. Empirically, we observed large Brier reductions after isotonic calibration and stable expected-cost optima across thresholds, which aligns with these findings and supports our decision to carry only calibrated p^_fail forward.

On the scheduling side, our findings are consistent with the

broader flow-shop literature: baseline NEH sequences provide strong starts, and positional MILP models yield reference optima at moderate problem sizes [47]. Where we extend the state of the art for AM is the risk-aware robustification targeted at the build bottleneck: we embed a Bertsimas–Sim-style budget at the position level to hedge a limited number of high-impact rework realisations. This follows robust-optimisation best practice tunable, tractable mixed-integer counterparts while remaining interpretable for planners [47]. Our sensitivity curves show that increasing Γ at nominal load has negligible throughput penalty, whereas inflation of risk magnitudes (Δ-scale) or physical slow-downs (Build-scale) raise Cmax roughly linearly intuitive, planner-friendly "price-of-robustness" trade-offs that complement the algorithmic insights reported in the scheduling literature.

Finally, we close the loop with SPC artefacts. Recent reviews argue that SPC in metal AM should combine part-quality capability (Cp/Cpk) and process-monitoring charts tuned to the AM workflow [48–50]. Our paired outputs batch-level p-charts on predicted defect probability and Cp/Cpk on melt-pool depth/width mirror that guidance: in our runs we saw no batch-level out-of-control points under either spec mode, while capability indices made tolerance assumptions explicit (proxy quantiles vs. capability-targeted specs). This operational pairing makes it straightforward for quality engineers to track both risk drift (p-charts) and geometric capability (Cp/Cpk) under ISO-style, risk-based thinking. In practice, planners can first set Γpos to hedge sporadic rework with near-zero PoR, then tune acceptance thresholds via the reliability sweep to match inspection capacity, and finally monitor Δ (risk inflation) through SPC triggering rescheduling only when PoR curves steepen.

We note several limitations and directions for future work. First, official lower and upper specification limits (LSL/USL) for melt-pool depth and width were unavailable; we therefore reported two surrogate specification schemes, namely quantile-based and capability-targeted limits, and recommend replacing them with customer-approved specifications as soon as they become accessible. Second, the microstructure-informed labels serve as proxies because a one-to-one provenance link between single-track optics and build coupons was not universally available; tighter data lineage would strengthen causal interpretation. Third, our robust MILP experiments were time-boxed with the open-source CBC solver and run on 60-job subsets to keep wall-clock time practical; larger instances could be addressed with commercial MIP solvers, decomposition, or matheuristics while preserving the same Bertsimas–Sim structure. Fourth, because the present study intentionally adopted a tight due-date regime to stress-test schedule robustness, policy differentiation in service-level metrics such as on-time rate was partially compressed; future work may extend the comparison to medium or loose due-date settings. Finally, although we intentionally used simple and replicable features, incorporating richer in-situ signatures, such as thermal or optical fields, may further improve calibration, especially at low false-positive rates, thereby supporting more precise and auditable production control in advanced manufacturing.

5. Conclusion

This study demonstrated an end-to-end, quality-aware scheduling pipeline for LPBF IN718 that starts from open AM-Bench 2022 measurements, learns calibrated failure probabilities p^_fail, and propagates them into production decisions from takt-time adjustments and robust buffers to shop-floor sequencing and SPC. On our dataset, simple numeric features from single-track/pad optics, combined with post-hoc probability calibration, produced well-behaved risk scores that could be "spent" in planning. Under nominal load, Γ-tunable robust schedules matched deterministic throughput (flat Cmax vs. Γ), while exogenous stressors (inflated rework tails or build bottlenecking) increased Cmax in a predictable way, giving planners an explicit price-of-robustness curve. Closing the loop, p-charts on predicted defect rate showed no batch-level drift across our runs, and capability indices (Cp/Cpk) on melt-pool depth/width provided a complementary geometric view of process capability.

Implications for industrial engineering are the pipeline operationalises predictive quality in a tractable, auditable way: (i) calibrated p^_fail converts model output into expected rework minutes and Γ-budgeted buffers; (ii) robust MILP/heuristics translate those buffers into sequences with clear throughput trade-offs; and (iii) SPC artefacts (p-charts, Cp/Cpk) provide a governance layer that quality and operations teams can review together. Practically, this means IE groups can

move from “predict-and-inspect” to “predict-to-plan,” using calibrated risk to set acceptance thresholds, rebalance capacity, and justify buffer placement before parts are built.

Our study limitations are first, our microstructure linkage used a small number of histogram files and surrogate labels; while defensible for method development, broader mapping between single-track optics and coupon-level builds would strengthen external validity. Second, the lack of official LSL/USL for depth/width required proxy specifications (quantiles or capability-targeted surrogates), which should be replaced by customer-approved tolerances when available. Third, the scheduling scope focused on the build bottleneck with four aggregated stages; real cells may require multi-machine, parallel-server, or re-entrant constraints and richer changeover logic. Fourth, calibration quality was excellent on our split, but drift across machines/material lots, scanners, or parameter windows warrants periodic recalibration. Finally, solver runs were time-limited on a workstation-class CPU; different hardware or commercial solvers could tighten optimality gaps on larger instances.

For the future work on the data side, (i) expand and formalise the mapping between process logs, in-situ sensing, and ex-situ microstructure to replace surrogates with measured acceptance criteria; (ii) incorporate additional predictors (path strategy, hatch spacing, layer time, thermography) to reduce residual risk. On modelling, (iii) extend the robust counterpart to multi-machine/parallel builds with sequence-dependent setups, and explore distributionally robust or chance-constrained forms that price not only mean rework but tail exceedances; (iv) integrate explicit cost models (scrap, reprint, lateness penalties) so Γ and thresholds can be tuned economically; (v) evaluate human-in-the-loop policies where engineers can override or “pin” critical jobs, with audit trails. On control, (vi) connect the SPC layer to feedback actions e.g., adjust scan speed or beam diameter when predicted risk or Cp/Cpk trends degrade and (vii) deploy drift monitors to trigger automatic recalibration or re-scheduling.

The key contribution is not a single algorithm but a reproducible workflow that links calibrated quality signals to capacity, sequence, and control artefacts that factories already use. By making risk explicit, tunable, and visible in both planning (Γ, PoR) and quality (p-charts, Cp/Cpk), the approach provides a practical path for industrial engineering teams to move from descriptive analytics to decision-centric, robustness-guaranteed production control in metal AM.

References

1. Levine L E, Williams M E, Creuziger A, Stoudt M R, Young S A, Moon K W, Lane B M. Location-Specific Microstructure Characterization Within AM Bench 2022 Nickel Alloy 718 3D Builds. *Integrating Materials and Manufacturing Innovation* 2024; 13: 585–97. https://doi.org/10.1007/s40192-024-00371-5.
2. Weaver J S, Deisenroth D, Mekhontsev S, Lane B M, Levine L E, Yeung H. Cross-Sectional Melt Pool Geometry of Laser Scanned Tracks and Pads on Nickel Alloy 718 for the 2022 Additive Manufacturing Benchmark Challenges. *Integrating Materials and Manufacturing Innovation* 2024; 13: 363–79. https://doi.org/10.1007/s40192-024-00355-5.
3. Simonds B J, Tanner J, Artusio-Glimpse A, Parab N, Zhao C, Sun T, Williams P A. Ability to Simulate Absorption and Melt Pool Dynamics for Laser Melting of Bare Aluminum Plate: Results and Insights from the 2022 Asynchronous AM-Bench Challenge. *Integrating Materials and Manufacturing Innovation* 2024; 13: 175–84. https://doi.org/10.1007/s40192-023-00336-0.
4. Moser N, Benzing J, Kafka O L, Weaver J, Derimow N, Rentz R, Hrabe N. AM Bench 2022 Macroscale Tensile Challenge at Different Orientations (CHAL-AMB2022-04-MaTTO) and Summary of Predictions. *Integrating Materials and Manufacturing Innovation* 2024; 13: 155–74. https://doi.org/10.1007/s40192-023-00333-3.
5. Wang Z, Yang W, Liu Q, Zhao Y, Liu P, Wu D, Banu M, Chen L. Data-driven modeling of process, structure and property in additive manufacturing: A review and future directions. *Journal of Manufacturing Processes* 2022; 77: 13–31. https://doi.org/10.1016/j.jmapro.2022.02.053.
6. Levine L E, Lane B M, Seppala J, Simonds B J, Stoudt M R, Weaver J, Yeung H, Zhang F. Outcomes and Conclusions from the 2022 AM Bench Measurements, Challenge Problems, Modeling Submissions, and Conference. *Integrating Materials and Manufacturing Innovation* 2024; 13: 598–621. https://doi.org/10.1007/s40192-024-00372-4.
7. Praniewicz M, Fox J C, Tarr J. Part Deflection Measurements of AM Bench IN718 3D Build Artifacts. *Integrating Materials and*

Manufacturing Innovation 2023; 12: 386–96. https://doi.org/10.1007/s40192-023-00310-w.

8. Phan T, Șeren H, Das A, Ko P, Nygren K, Levine L. Elastic Residual Strain Measurements of 3D Additively Manufactured Builds of Nickel Alloy 718 AM Bench 2022 Artifacts Using Energy Dispersive Synchrotron X-ray Diffraction. *Integrating Materials and Manufacturing Innovation* 2025; 14: 14–24. https://doi.org/10.1007/s40192-024-00388-w.
9. H.L. Wei, T. Mukherjee, W. Zhang, J.S. Zuback, G.L. Knapp, A. De, T. DebRoy. Mechanistic models for additive manufacturing of metallic components. *Progress in Materials Science* 2021; 116: 100703. https://doi.org/10.1016/j.pmatsci.2020.100703.
10. Levine L E, Williams M E, Stoudt M R, Weaver J S, Young S A, Deisenroth D, Lane B M. Location-Specific Microstructure Characterization Within AM Bench 2022 Laser Tracks on Bare Nickel Alloy 718 Plates. *Integrating Materials and Manufacturing Innovation* 2024; 13: 380–95. https://doi.org/10.1007/s40192-024-00361-7.
11. Liu F, Dai Y. Product quality prediction method in small sample data environment. *Advanced Engineering Informatics* 2023; 56: 101975. https://doi.org/10.1016/j.aei.2023.101975.
12. Liu J, Ye J, Silva Izquierdo D, Vinel A, Shamsaei N, Shao S. A review of machine learning techniques for process and performance optimization in laser beam powder bed fusion additive manufacturing. *Journal of Intelligent Manufacturing* 2023; 34: 3249–75. https://doi.org/10.1007/s10845-022-02012-0.
13. Rahimian H, Mehrotra S. Frameworks and Results in Distributionally Robust Optimization. *Open Journal of Mathematical Optimization* 2022; 3: 4. https://doi.org/10.5802/ojmo.15.
14. Inayathullah S, Buddala R. Review of machine learning applications in additive manufacturing. *Results in Engineering* 2025; 25: 103676. https://doi.org/10.1016/j.rineng.2024.103676.
15. Keller D S, Reif de Paula T, Yu G, Zhang H, Al-Mazrou A, Kiran R P. Statistical Process Control (SPC) to drive improvement in length of stay after colorectal surgery. *The American Journal of Surgery* 2020; 219: 1006–11. https://doi.org/10.1016/j.amjsurg.2019.08.029.
16. da Silva G J, Borges A C. Statistical Process Control in the Environmental Monitoring of Water Quality and Wastewaters: A Review. *Water* 2025; 17: 1281. https://doi.org/10.3390/w17091281.
17. Zhang F, Johnston-Peck A C, Levine L E, Katz M B, Moon K-W, Williams M E, Young S A, Allen A J, Borkiewicz O, Ilavsky J. Phase Composition and Phase Transformation of Additively Manufactured Nickel Alloy 718 AM Bench Artifacts. *Integrating Materials and Manufacturing Innovation* 2024; 13: 185–200. https://doi.org/10.1007/s40192-023-00338-y.
18. Liu Z, Zhao D, Wang P, Yan M, Yang C, Chen Z, Lu J, Lu Z. Additive manufacturing of metals: Microstructure evolution and multistage control. *Journal of Materials Science and Technology* 2022; 100: 224–36. https://doi.org/10.1016/j.jmst.2021.06.011.
19. Simonds B J, Tanner J, Artusio-Glimpse A, Williams P A, Parab N, Zhao C, Sun T. The causal relationship between melt pool geometry and energy absorption measured in real time during laser-based manufacturing. *Applied Materials Today* 2021; 23: 101049. https://doi.org/10.1016/j.apmt.2021.101049.
20. Seenivasan R, Pachiyappan J K, Reddy Murthannagari V, Krishnan Ganesh G N. Optimizing Metformin HCl manufacturing: A Six Sigma approach to assess process capability. *Journal of Applied Pharmaceutical Science* 2024. https://doi.org/10.7324/JAPS.2024.177701.
21. Marín Díaz G. Comparative Analysis of Explainable AI Methods for Manufacturing Defect Prediction: A Mathematical Perspective. *Mathematics* 2025; 13: 2436. https://doi.org/10.3390/math13152436.
22. Team NIST AM-Bench. AM Bench 2022 Microstructure Measurements for IN718 3D Builds (AMB2022-01). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2692.
23. Team NIST AM-Bench. In-situ Thermography and Scan Strategy for Single Tracks/Pads (AMB2022-03). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2716.
24. Team NIST AM-Bench. Measurement Results: Optical Microscopy of Laser-scanned Single Tracks and Pads (AMB2022-03). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2718.
25. Team NIST AM-Bench. Serial Sectioning & X-ray CT Measurement Data (IN718). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2767.
26. Team NIST AM-Bench. 3D Builds In-situ Thermography & Scripts (AMB2022-01). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2715.
27. Team NIST AM-Bench. 3D Build with Custom Laser Scan Strategies (AMB2022-02). *National Institute of Standards and Technology*

(NIST); 2022. https://doi.org/10.18434/MDS2-2617.
28. Team NIST AM-Bench. 3D Build Modeling Challenge Description (AMB2022-01). *National Institute of Standards and Technology* (NIST); 2022. https://doi.org/10.18434/MDS2-2607.
29. Feng S, Chen Z, Bircher B, Ji Z, Nyborg L, Bigot S. Predicting laser powder bed fusion defects through in-process monitoring data and machine learning. *Materials and Design* 2022; 222: 111115. https://doi.org/10.1016/j.matdes.2022.111115.
30. Oster S, Breese P P, Ulbricht A, Mohr G, Altenburg S J. A deep learning framework for defect prediction based on thermographic in-situ monitoring in laser powder bed fusion. *Journal of Intelligent Manufacturing* 2024; 35: 1687–706. https://doi.org/10.1007/s10845-023-02117-0.
31. Guillen D, Wahlquist S, Ali A. Critical Review of LPBF Metal Print Defects Detection: Roles of Selective Sensing Technology. *Applied Sciences* 2024; 14: 6718. https://doi.org/10.3390/app14156718.
32. Nixon J, Dusenberry MW, Jerfel G, Nguyen T, Liu J, Zhang L, Tran D. Measuring calibration in deep learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019. p. 38–41.
33. Kull M, Silva Filho TM, Flach P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics* 2017; 11. https://doi.org/10.1214/17-EJS1338SI.
34. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*. 2017. p. 1321–30.
35. Bertsimas D, Kallus N. From Predictive to Prescriptive Analytics. *Management Science* 2020; 66: 1025–44. https://doi.org/10.1287/mnsc.2018.3253.
36. Li JY-M. Inverse Optimization of Convex Risk Functions. *Management Science* 2021; 67: 7113–41. https://doi.org/10.1287/mnsc.2020.3851.
37. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal Risk Control. *The Twelfth International Conference on Learning Representations* (ICLR); 2024. Available from: https://openreview.net/forum?id=33XGfHLtZg.
38. Sun Q, Dou J, Zhang C. Robust optimization of flow shop scheduling with uncertain processing time. In: *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*. 2020. p. 512–7. https://doi.org/10.1109/ICMA49215.2020.9233523.
39. Rahmativala S, Ghahremani J. A robust optimization method for hybrid flow shop scheduling with uncertain setup times. *Decision Analytics Journal* 2025; 16: 100609. https://doi.org/10.1016/j.dajour.2025.100609.
40. Hegyháti M, Bakon K A, Holczinger T. Optimization with uncertainties: a scheduling example. *Central European Journal of Operations Research* 2023; 31: 1239–63. https://doi.org/10.1007/s10100-023-00854-4.
41. Tliba K, Diallo T M L, Penas O, Ben Khalifa R, Ben Yahia N, Choley J-Y. Digital twin-driven dynamic scheduling of a hybrid flow shop. *Journal of Intelligent Manufacturing* 2023; 34: 2281–306. https://doi.org/10.1007/s10845-022-01922-3.
42. Mattera G, Mattera R, Otto P. Hybrid Statistical Process Monitoring of Wire Arc Additive Manufacturing With Frequency‐Informed Deep Learning. *Quality and Reliability Engineering International* 2025. https://doi.org/10.1002/qre.70041.
43. Montgomery DC. Introduction to Statistical Quality Control. *Wiley*; 2020.
44. Witherell P, Lane B, Yeung H, Yang Z, Law K. Anomaly detection of laser powder bed fusion melt pool images using combined unsupervised and supervised learning methods. In: *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. St Louis, MO, US; 2022.
45. Kull M, Perello-Nieto M, Kängsepp M, Filho TS, Song H, Flach P. Beyond temperature scaling: obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019.
46. Xiong R, Shi Y, Jing H, Liang W, Nakahira Y, Tang P. Calibrating subjective data biases and model predictive uncertainties in machine learning-based thermal perception predictions. *Building and Environment* 2024; 247: 111053. https://doi.org/10.1016/j.buildenv.2023.111053.
47. Neufeld J S, Schulz S, Buscher U. A systematic review of multi-objective hybrid flow shop scheduling. *European Journal of Operational Research* 2023; 309: 1–23. https://doi.org/10.1016/j.ejor.2022.08.009.
48. Sahin A, Rey P, Panoutsos G. Self-tuning multi-model statistical process control for process monitoring in additive manufacturing. In:

2022 8th International Conference on Control, Decision and Information Technologies (CoDIT). 2022. p. 1049–54. https://doi.org/10.1109/CoDIT55151.2022.9803964.

49. Hou Z-J, Wang Q, Zhao C-G, Zheng J, Tian J-M, Ge X-H, Liu Y-G. Online Monitoring Technology of Metal Powder Bed Fusion Processes: A Review. *Materials* 2022; 15: 7598. https://doi.org/10.3390/ma15217598.
50. Tebianian M, Aghaie S, Razavi Jafari N S, Elmi Hosseini S R, Pereira A B, Fernandes F A O, Farbakhti M, Chen C, Huo Y. A Review of the Metal Additive Manufacturing Processes. *Materials* 2023; 16: 7514. https://doi.org/10.3390/ma16247514.