

AI-AUGMENTED SIX SIGMA FOR QUALITY 4.0: A DATA-DRIVEN WORKFLOW FOR DEFECT PREDICTION AND PROCESS IMPROVEMENT

Ali Zulqarnain^{*1}, Muhammad Zaid Khan², Rabia Siddiqui³, Muhammad Affaq Khan⁴

^{*1,2,3,4}Department of Industrial & Manufacturing, NED University of Engineering & Technology, Karachi, Pakistan

¹alizul@neduet.edu.pk, ²muhammadzaid@neduet.edu.pk, ³rabias@neduet.edu.pk, ⁴mak96b@gmail.com

DOI: <https://doi.org/10.5281/zenodo.22843632>

Keywords

Quality 4.0; Six Sigma; Define-Measure-Analyze-Improve-Control; Integration of Six Sigma with AI, Quality 4.0 and Six Sigma, Predictive Maintenance; Gradient Boosting.

Article History

Received: 12 June 2026

Accepted: 03 September 2026

Published: 19 September 2026

Copyright @Author

Corresponding Author: *

Ali Zulqarnain

Abstract

Purpose: One of the most important management breakthroughs of the late 20th century is Six Sigma. It is a holistic mindset for attaining process and organizational excellence rather than just a statistical methodology. Intelligent, data-driven quality improvement is made possible by Quality 4.0's integration with Industry 4.0 technology and artificial intelligence. This study develops and empirically evaluates an artificial intelligence-augmented Lean Six Sigma framework for Quality 4.0 by integrating Define-Measure-Analyze-Improve-Control with statistical quality analysis, leakage-controlled machine learning, explainable artificial intelligence, probability calibration and decision-threshold analysis.

Design/methodology/approach: The full University of California, Irvine Artificial Intelligence for Industry 2020 Predictive Maintenance dataset (10,000 observations; 339 failures) was analyzed. Identifier and failure-mode columns were excluded from predictive inputs. Welch tests with Hedges' g and bootstrap 95% confidence intervals compared process variables between failure and normal operation. A multivariable logistic model estimated adjusted odds ratios. Logistic regression, random forest and Extreme Gradient Boosting were evaluated by repeated stratified 5-fold cross-validation repeated three times. area under the receiver operating characteristic curve, area under the precision-recall curve, precision, recall, harmonic mean of precision and recall, specificity and balanced accuracy were reported. Out-of-fold probabilities supported calibration and threshold analysis, and Tree-based Shapley additive explanations were used for model interpretation.

Findings: Failures represented 3.39% of observations. Extreme Gradient Boosting provided the strongest discrimination (area under the receiver operating characteristic curve 0.977; area under the precision-recall curve 0.788) and recall (0.826), while random forest produced the highest harmonic mean of precision and recall (0.707) and better Brier performance. Torque, tool wear and rotational speed were the leading Shapley additive explanation drivers. Threshold analysis demonstrated an explicit trade-off between failure capture and false alarms.

Practical implications: Define-Measure-Analyze-Improve-Control is strengthened by linking Measure to data quality and baselines, Analyze to inferential statistics and explainable artificial intelligence, Improve to

threshold/action policies, and Control to calibration, statistical process control and drift monitoring. Originality/value – The paper reports original full-dataset, leakage-controlled results rather than relying on externally published benchmark scores.

1. Introduction

Artificial intelligence-augmented quality management means that artificial intelligence does not replace Six Sigma, engineering judgment or human decision-making. Instead, it strengthens them by using data-driven models to detect complex patterns, estimate the risk of machine failure, explain which operating conditions influence a prediction, and support more timely process decisions. In this study, “artificial intelligence-augmented” therefore refers to a human-governed system in which statistical quality methods and the Define-Measure-Analyze-Improve-Control cycle remain the management structure, while artificial intelligence adds predictive and explanatory capability. The purpose is augmentation: people retain responsibility for defining the quality problem, validating evidence, choosing interventions and maintaining control, while artificial intelligence helps them use large process datasets more consistently and proactively.

Quality management is moving from periodic inspection toward connected, predictive and increasingly real-time decision support. Quality 4.0 extends established quality principles through industrial Internet of Things, industrial data and artificial intelligence [1-3, 31]. Lean Six Sigma remains important because Define-Measure-Analyze-Improve-Control supplies problem governance, measurement discipline, causal investigation and sustained control, whereas machine learning contributes nonlinear pattern recognition and scalable risk estimation [4,5].

Implementation is not purely technical. A Quality 4.0 implementation and maturity framework for developing-country industries identified scalability, culture and application development as areas requiring particular attention [6]. This is directly relevant to an artificial intelligence-augmented Lean Six Sigma architecture: prediction must be embedded in organizational capability, quality ownership and

operational action rather than treated as an isolated data-science exercise.

Manufacturing failures are also rare events. High overall accuracy can therefore coexist with poor failure detection, making area under the precision-recall curve, recall, precision, harmonic mean of precision and recall, calibration and operating thresholds central to quality-oriented evaluation [7,8]. This study addresses these issues using the Artificial Intelligence for Industry 2020 benchmark [9,10] and translates the empirical evidence into Define-Measure-Analyze-Improve-Control decisions.

2. Literature Review and Conceptual Background

2.1 Quality 4.0 and implementation maturity

Quality 4.0 combines established quality-management principles with digital connectivity, smart sensors, analytics and artificial intelligence. Reviews emphasize predictive quality, process monitoring and data-driven decision support while also warning that deployment depends on data quality, organizational readiness and operational integration [1-3]. The maturity perspective developed for developing-country settings is especially useful because it broadens implementation beyond algorithms to organizational and digital capabilities [6]. Recent mapping work likewise shows that data-science frameworks and quality-management systems are complementary but structurally different [11].

2.2 Lean Six Sigma, Industry 4.0 and Define-Measure-Analyze-Improve-Control

Industry 4.0 technologies increase the speed and granularity of measurement, while Lean Six Sigma provides a disciplined route from problem definition to sustained control. artificial intelligence and machine learning, big-data analytics, smart sensors, automation and simulation have been identified as high-value technologies for Lean Six Sigma integration [4,5,12]. Quality 4.0 has been described as an

evolution of Six Sigma Define-Measure-Analyze-Improve-Control, supporting an architecture in which artificial intelligence is governed by quality-management logic rather than replacing it [13].

2.3 Predictive maintenance, imbalance and explainability

Predictive maintenance is closely linked to quality because equipment condition contributes to variation, downtime and nonconformance. Artificial Intelligence for Industry 2020 was introduced as an explainable predictive-maintenance benchmark [10]. Random forest and Extreme Gradient Boosting can represent nonlinear relationships [14,15], while Shapley additive explanations provides local and global attribution for complex models [16]. For imbalanced outcomes, precision-recall analysis is particularly informative [7,8]. Calibration is necessary when model outputs are used as risk probabilities [17,18], and resampling methods such as synthetic minority over-sampling technique must be restricted to training data to prevent leakage [19].

2.4 Literature gap and positioning

This paper addresses that gap empirically. Rather than treating model performance as the endpoint, it links each analytical output to a Define-Measure-Analyze-Improve-Control decision: baseline and measurement integrity to Measure; effect sizes, adjusted associations and Shapley additive explanations to Analyze; threshold trade-offs to Improve; and calibration, statistical process control and drift monitoring to Control. This positioning extends prior Quality 4.0 and Lean Six Sigma 4.0 frameworks by making the transition from prediction to operational quality control explicit and quantitatively testable.

However, these streams are often evaluated separately. Predictive-maintenance studies commonly emphasize discrimination metrics, whereas Lean Six Sigma studies emphasize improvement governance and Quality 4.0 maturity studies emphasize organizational readiness. Recent work comparing data-science and quality-management frameworks reinforces

this separation: data-science lifecycles are primarily model- and data-centric, whereas Define-Measure-Analyze-Improve-Control is process-centric and explicitly concerned with sustained control [11]. The unresolved issue is therefore not whether machine learning can predict failures, but how statistical evidence, nonlinear prediction, explanation, calibration and operating thresholds can be assembled into a quality-improvement architecture with explicit managerial responsibilities.

Existing Quality 4.0 scholarship establishes three complementary foundations for this study. First, reviews of manufacturing quality analytics show that digital quality increasingly depends on the integration of process knowledge, sensor data and predictive methods rather than on automated inspection alone [2,3]. Second, the Lean Six Sigma 4.0 literature identifies artificial intelligence and machine learning, smart sensing, simulation and big-data analytics as important enablers, while retaining Define-Measure-Analyze-Improve-Control as the process-improvement logic that converts analytical evidence into controlled action [4,12,13]. Third, implementation research shows that organizational maturity, culture, scalability and application-development capability condition whether Quality 4.0 tools become operational routines, particularly in developing-country settings [6].

3. Materials and Methods

3.1 Dataset and leakage control

The University of California, Irvine Artificial Intelligence for Industry 2020 Predictive Maintenance Dataset contains 10,000 complete observations [9]. Predictors were product type, air temperature, process temperature, rotational speed, torque and tool wear. Machine failure was the binary critical-to-quality outcome. unique data identifier and product identifier were excluded as identifiers. tool-wear failure, heat-dissipation failure, power failure, overstrain failure and random failure were used only for descriptive Pareto analysis because using them to predict the aggregate failure label would introduce target leakage.

Table I. Analytical variables and roles.

Variable	Role	Unit/coding
Type	Product category	low / medium / high
Air temperature	Process input	kelvin
Process temperature	Process input	kelvin
Rotational speed	Process input	revolutions per minute
Torque	Process input	newton-metres
Tool wear	Degradation input	minutes
Machine failure	Primary critical-to-quality	zero = normal; one = failure
tool-wear failure / heat-dissipation failure / power failure / overstrain failure / random failure	Failure descriptors	Descriptive only

3.2 Statistical analysis

Failure prevalence was summarized with a 95% confidence interval for a binomial proportion [20]. Continuous variables were compared between failure and normal observations using Welch tests. Hedges' g quantified standardized mean differences and bootstrap resampling (1,000 replicates) provided 95% intervals. A multivariable logistic regression estimated adjusted odds ratios after standardizing continuous predictors. process capability indices were not calculated because customer specification limits are unavailable.

A multicollinearity sensitivity check was also conducted for the five standardized continuous predictors. Variance inflation factors ranged from 1.00 to 4.31, remaining below the commonly used threshold of 5; the largest values reflected the expected temperature and torque-speed relationships. The logistic coefficients are therefore reported as conditional associations, while nonlinear model results are used as the primary predictive evidence.

3.3 Machine-learning validation

Product type was one-hot encoded and continuous predictors standardized inside each model pipeline. Logistic regression and random forest used balanced class weights. Extreme Gradient Boosting used 400 trees, maximum depth 4, learning rate 0.05, subsampling 0.9 and a positive-class weight equal to the observed class ratio. Repeated stratified 5-fold cross-validation with three repeats (15 held-out folds; seed 42) evaluated area under the receiver operating characteristic curve, area under the precision-

recall curve, precision, recall, harmonic mean of precision and recall, specificity and balanced accuracy. Approximate intervals are mean plus or minus 1.96 standard errors across folds and are descriptive because cross-validation folds are correlated [21].

Table IV summarizes mean performance across repeated stratified 5-fold cross-validation (three repeats). Figures 3-7 and Table V use predictions from a separate single stratified 5-fold out-of-fold procedure to obtain one probability per observation for curves, calibration and threshold analysis. Minor differences between the repeated-CV means and the single-OOF metrics are therefore expected and do not represent numerical inconsistency.

3.4 Explainability, calibration and decision analysis

Tree-based Shapley additive explanations were applied to the fitted Extreme Gradient Boosting model. Five-fold out-of-fold probabilities were used for receiver operating characteristic and precision-recall curves, Brier score, calibration intercept/slope and threshold analysis. Thresholds from 0.10 to 0.80 were reported rather than declaring a universal optimum, because the appropriate operating point depends on plant-specific failure costs, safety consequences and maintenance capacity.

4. Results

4.1 Baseline and failure structure

All 10,000 observations were complete. Machine failure occurred in 339 observations (3.39%); 9,661 were normal. A naive always-normal

classifier would therefore obtain 96.61% accuracy while detecting no failures.

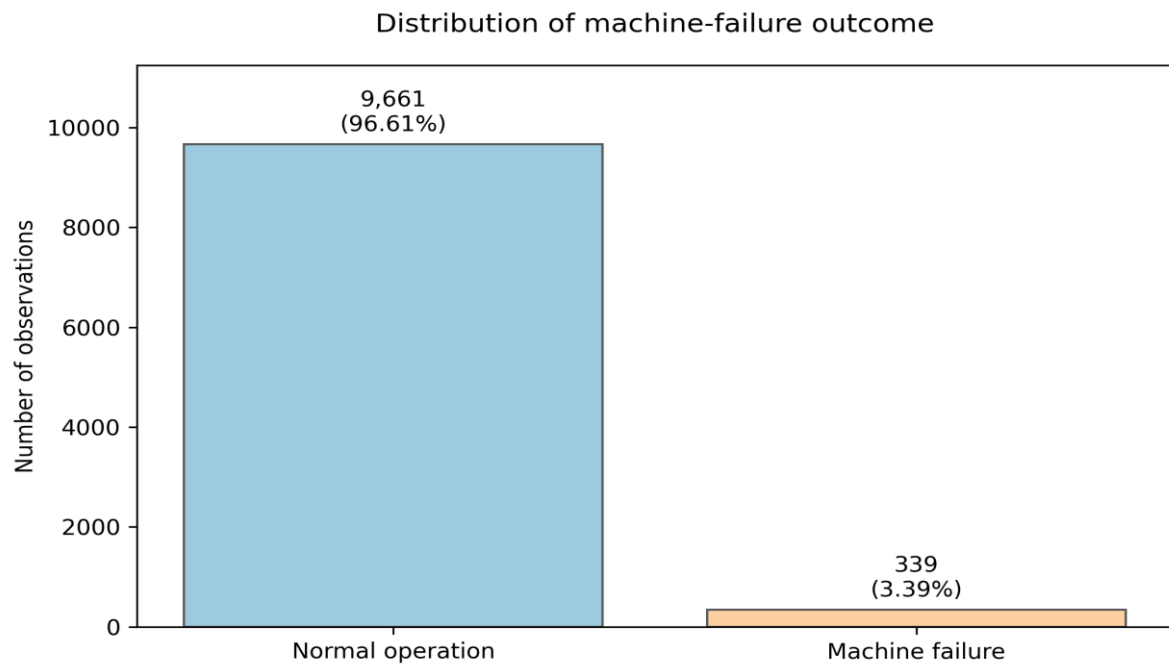


Figure 1. Distribution of the primary critical-to-quality outcome.

Failure-mode flags were not mutually exclusive. Heat-dissipation failure was most frequent (115),

followed by overstrain (98), power failure (95), tool-wear failure (46) and random failure (19).

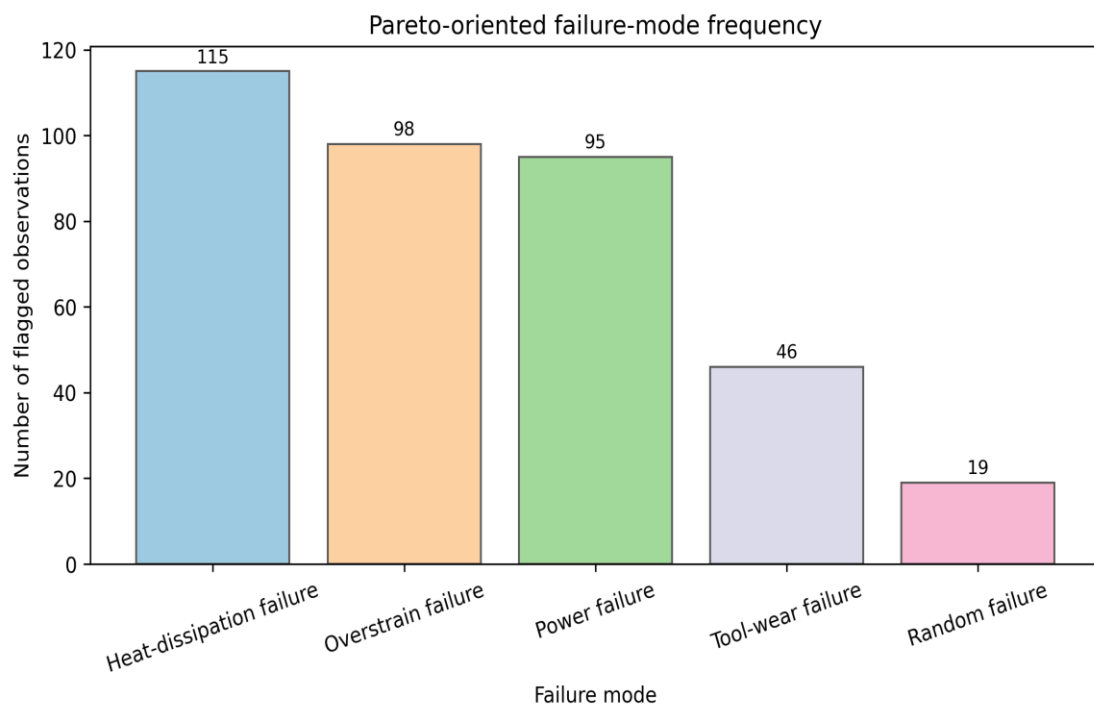


Figure 2. Pareto-oriented failure-mode frequencies

4.2 Inferential statistical results

Table II. Failure/non-failure comparisons and Hedges' g effect sizes.

Variable	Failure mean	Normal mean	Welch t	p	Hedges g	confidence interval lower bound	confidence interval upper bound
Air temperature (kelvin)	300.886	299.974	7.982	<0.001	0.458	0.343	0.570
Process temperature (kelvin)	310.290	309.996	3.898	<0.001	0.199	0.096	0.298
Rotational speed [revolutions per minute]	1496.487	1540.260	-2.087	0.038	-0.244	-0.478	-0.001
Torque [newton-metres]	50.168	39.630	11.781	<0.001	1.077	0.882	1.275
Tool wear (minutes)	143.782	106.694	9.264	<0.001	0.586	0.461	0.710

Torque showed the largest univariate standardized difference ($g = 1.077$), followed by tool wear ($g = 0.586$) and air temperature ($g =$

0.458). Failures occurred at lower mean rotational speed, although this difference was smaller ($g = -0.244$).

Table III. Adjusted logistic regression; continuous predictors are standardized.

Term	Odds Ratio	Confidence Interval (Lower bound)	Confidence Interval Upper bound	p
Air temperature (kelvin)	4.66	3.51	6.20	<0.001
Process temperature (kelvin)	0.33	0.25	0.44	<0.001
Rotational speed [revolutions per minute]	7.90	6.55	9.52	<0.001
Torque [newton-metres]	16.40	13.10	20.53	<0.001
Tool wear (minutes)	2.33	2.02	2.68	<0.001
Low-quality product type versus high-quality product type	1.81	1.09	3.03	0.023
Medium-quality product type versus high-quality product type	1.32	0.76	2.28	0.321

In adjusted analysis, torque had the largest standardized association with failure (odds ratio 16.40, 95% confidence interval 13.10–20.53), followed by rotational speed (odds ratio 7.90, 6.55–9.52), air temperature (odds ratio 4.66,

3.51–6.20) and tool wear (odds ratio 2.33, 2.02–2.68). Process temperature had an inverse conditional association (odds ratio 0.33, 0.25–0.44). These are associations in a synthetic benchmark and are not causal effects.

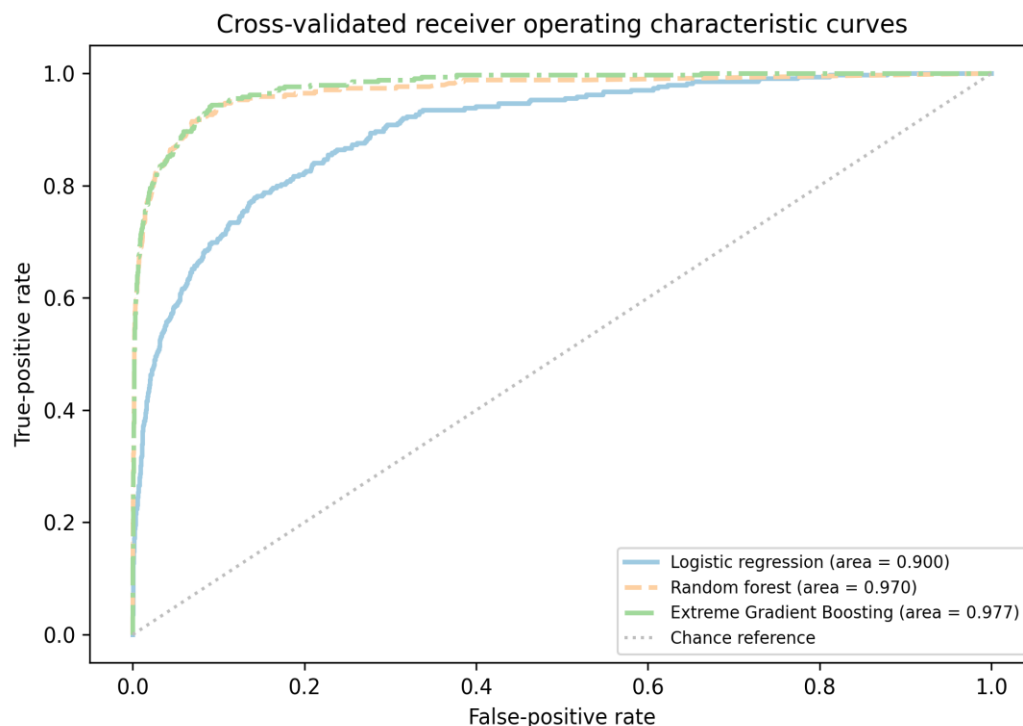
4.3 Original repeated cross-validated model performance

Table IV. Repeated stratified 5-fold cross-validation: mean (approximate 95% interval).

Model	area under the receiver operating characteristic curve	area under the precision-recall curve	Precision	Recall	harmonic mean of precision and recall	Specificity	Balanced acc.
Logistic Regression	0.900 (0.891–0.908)	0.429 (0.405–0.452)	0.137 (0.135–0.139)	0.812 (0.796–0.829)	0.234 (0.231–0.238)	0.820 (0.817–0.824)	0.816 (0.809–0.824)
Random forest	0.970 (0.964–0.976)	0.776 (0.759–0.794)	0.837 (0.815–0.859)	0.615 (0.584–0.647)	0.707 (0.685–0.729)	0.996 (0.995–0.996)	0.806 (0.790–0.821)
Extreme Gradient Boosting	0.977 (0.973–0.981)	0.788 (0.767–0.808)	0.557 (0.540–0.574)	0.826 (0.796–0.856)	0.664 (0.648–0.681)	0.977 (0.975–0.979)	0.901 (0.887–0.916)

Extreme Gradient Boosting achieved the strongest discrimination (area under the receiver operating characteristic curve 0.977; area under the precision-recall curve 0.788), failure recall (0.826) and balanced accuracy (0.901). Random

forest achieved the highest precision (0.837) and harmonic mean of precision and recall (0.707). Balanced logistic regression retained high recall (0.812) but low precision (0.137), implying a much larger false-alarm burden.



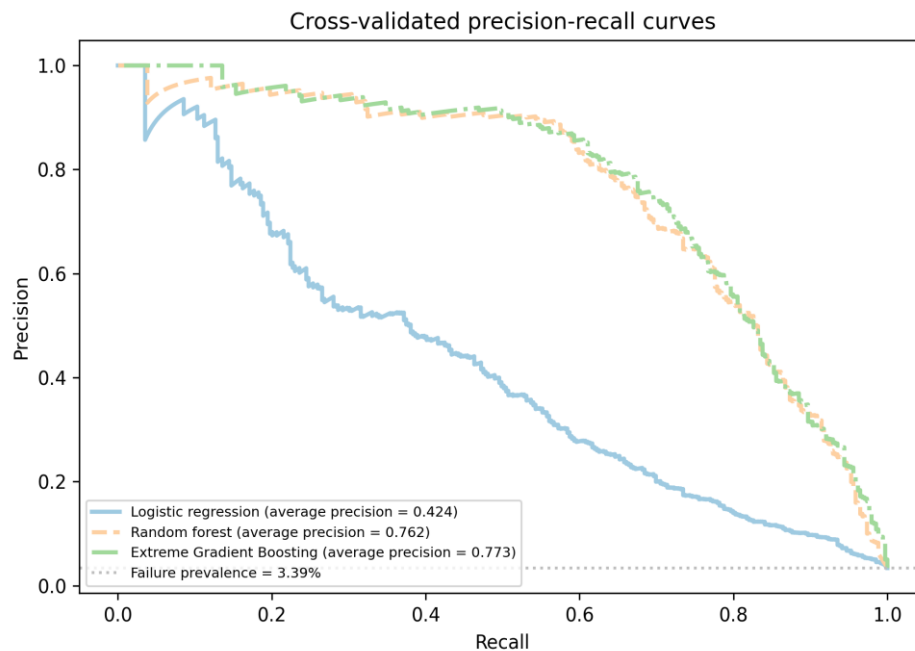


Figure 4. Five-fold out-of-fold precision-recall curves.

4.4 Explainable Artificial Intelligence

Tree Shapley additive explanations ranked torque as the most influential global predictor, followed closely by tool wear, rotational speed, air temperature and process temperature. Product-type indicators contributed

substantially less. The Shapley additive explanations ranking complements the statistical analysis by revealing the features most influential to the nonlinear model without treating those contributions as causal effects.

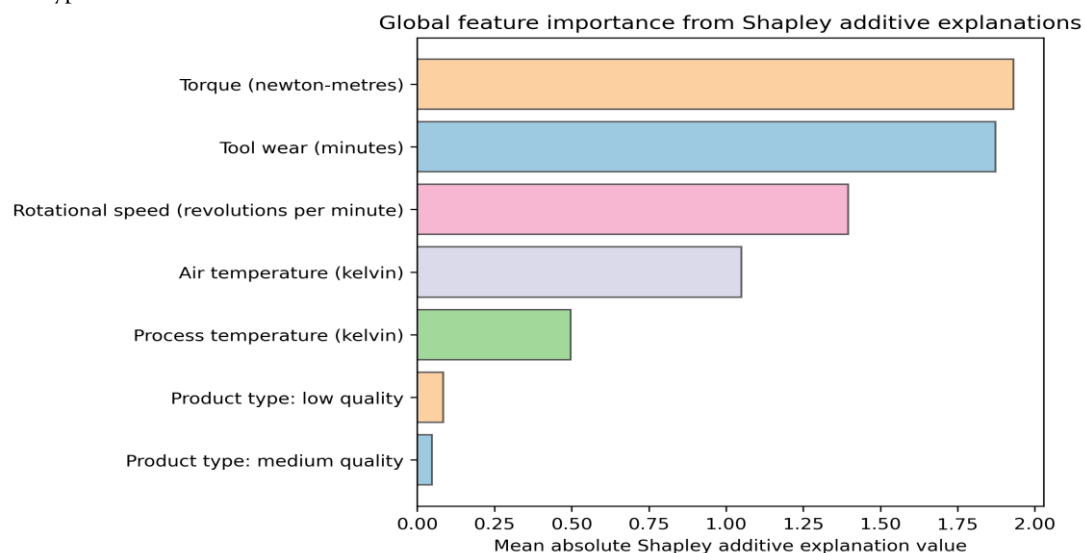


Figure 5. Shapley additive explanations summary plot for the fitted Extreme Gradient Boosting model.

4.5 Calibration and threshold trade-offs

Model	Brier	Calibration intercept	Calibration slope	area under the receiver operating characteristic curve	area under the precision-recall curve
Logistic regression	0.125	-3.353	1.043	0.900	0.424
Random forest	0.016	-0.053	1.249	0.970	0.762
Extreme Gradient Boosting	0.022	-1.671	0.861	0.977	0.773

Table V. Out-of-fold calibration and discrimination diagnostics.

Random forest had the lowest Brier score (0.0155) and a calibration intercept near zero. Extreme Gradient Boosting had the strongest

discrimination but showed probability miscalibration (intercept -1.671; slope 0.861), demonstrating that ranking performance and probability reliability are distinct properties.

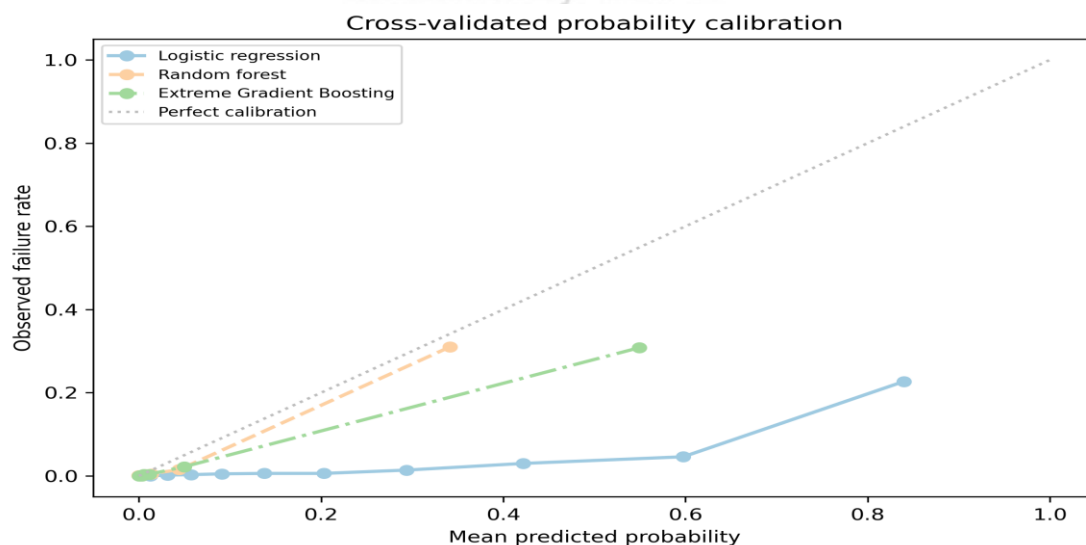


Figure 6. Cross-validated calibration curves.

Threshold	Recall	Precision	False-alarm rate	Specificity	harmonic mean of precision and recall	true positives	false positives	false negatives
0.100	0.920	0.287	0.080	0.920	0.438	312	775	27
0.200	0.885	0.369	0.053	0.947	0.520	300	514	39
0.300	0.844	0.431	0.039	0.961	0.571	286	377	53
0.400	0.829	0.491	0.030	0.970	0.617	281	291	58
0.500	0.805	0.538	0.024	0.976	0.645	273	234	66
0.600	0.782	0.600	0.018	0.982	0.679	265	177	74
0.700	0.746	0.666	0.013	0.987	0.704	253	127	86
0.800	0.696	0.744	0.008	0.992	0.720	236	81	103

Table VI. Extreme Gradient Boosting operating-threshold trade-offs.

At threshold 0.10, recall was 0.920 with an 8.0% false-alarm rate. At 0.50, recall was 0.805 with a 2.4% false-alarm rate. At 0.80, recall decreased

to 0.696 while the false-alarm rate fell below 1%. These scenarios make the Improve-phase decision transparent without inventing a universal cost ratio.

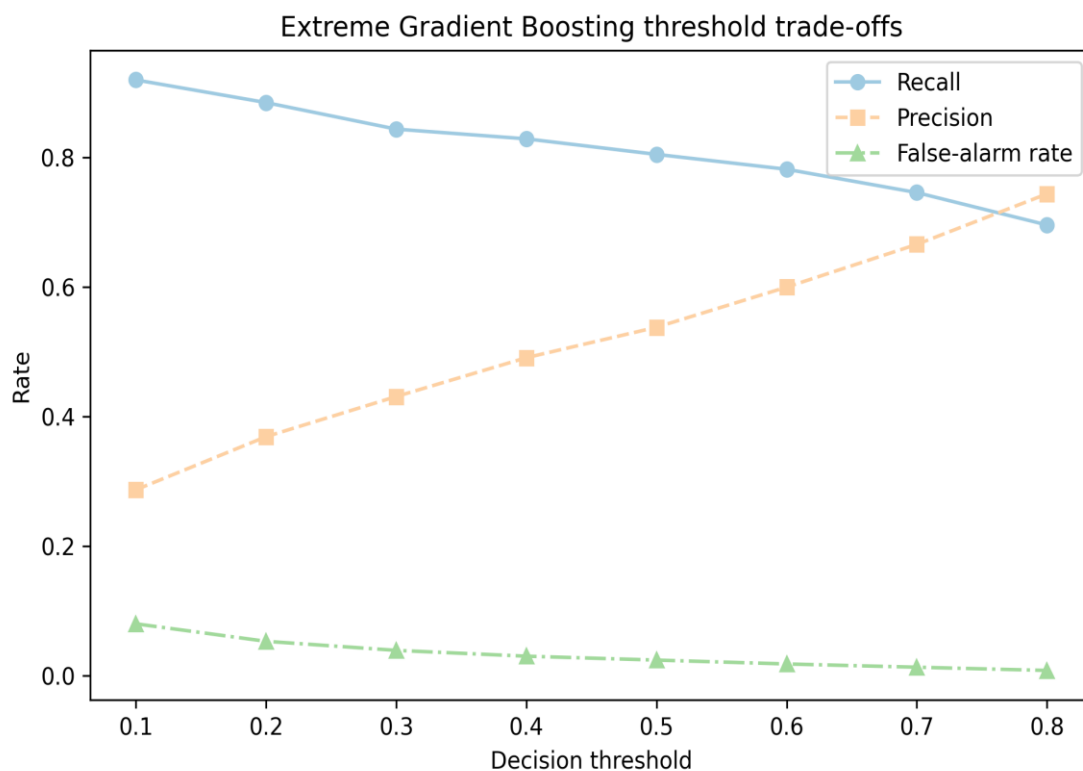


Figure 7. Extreme Gradient Boosting threshold trade-offs

5. Empirical Artificial Intelligence-Augmented Define-Measure-Analyze-Improve-Control Framework

Table VII. Empirically grounded artificial intelligence-augmented Define-Measure-Analyze-Improve-Control architecture.

Define-Measure-Analyze-Improve-Control	Empirical/artificial intelligence activity	Quality output
Define	Machine failure as critical-to-quality; consequences	Project charter and decision objective
Measure	3.39% baseline; data/sensor audit	Trusted baseline
Analyze	Tests, adjusted odds ratios, machine learning, Shapley additive explanations	Drivers and nonlinear risk structure
Improve	Threshold scenarios; validated interventions	Action rules and trade-offs
Control	Calibration, statistical process control, drift, retraining	Sustained early-warning control

Define establishes machine failure as the critical-to-quality and specifies the operational consequences of missed failures and false alarms.

Measure establishes the rare-event baseline and verifies sensor and label integrity. Analyze combines effect sizes, adjusted regression,

nonlinear prediction and Shapley additive explanations. Improve converts risk scores into threshold/action scenarios that can be tested in controlled pilots. Control adds calibration,

statistical process control, drift monitoring, human escalation and evidence-based retraining triggers.

Artificial Intelligence-Augmented Define-Measure-Analyze-Improve-Control Workflow

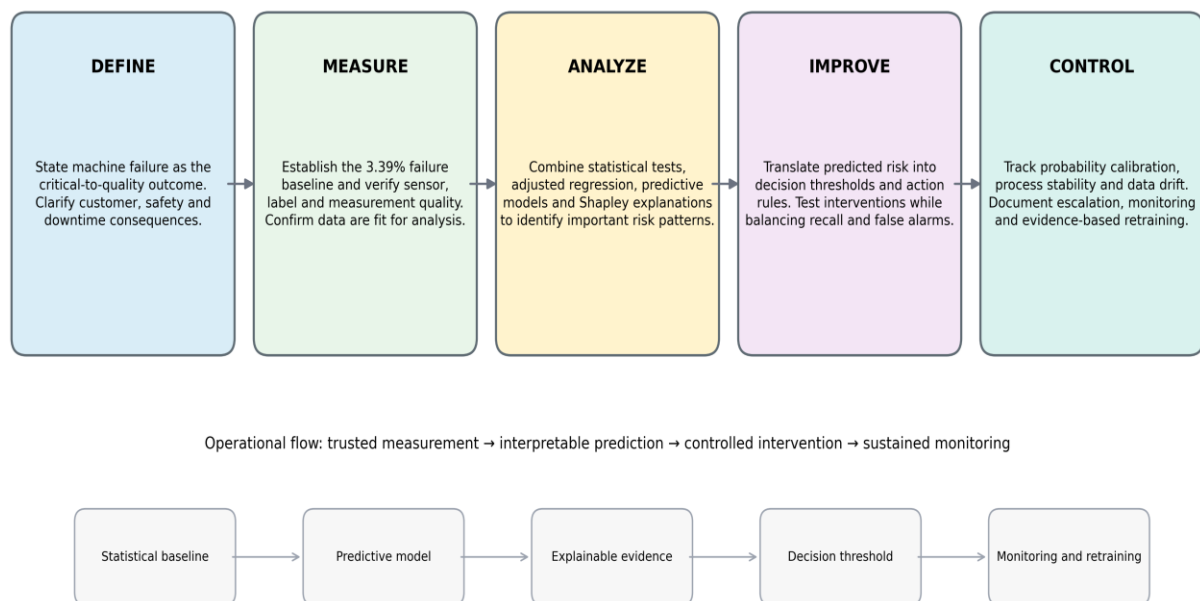


Figure 8. artificial intelligence-augmented Define-Measure-Analyze-Improve-Control workflow linking statistical evidence, explainable prediction, intervention and control.

6. Discussion

6.1 Predictive performance in the context of rare-event quality

The first contribution concerns how predictive performance should be interpreted in a rare-event quality problem. With only 3.39% failures, the 96.61% accuracy of an always-normal rule demonstrates the inadequacy of accuracy as a primary quality metric. This result is consistent with the established argument that precision-recall analysis is more informative when the positive class is uncommon [7,8]. In the present analysis, Extreme Gradient Boosting achieved the strongest area under the receiver operating characteristic curve, area under the precision-recall curve, recall and balanced accuracy, whereas random forest achieved higher precision and harmonic mean of precision and recall. The practical implication is not that one algorithm is universally preferable, but that model choice depends on the quality

loss function. A process in which missed failures carry severe safety or downtime consequences may rationally prioritize recall, while a maintenance system with limited inspection capacity may place greater weight on precision and false-alarm control.

The comparison also clarifies the role of model complexity. Logistic regression retained high recall after class balancing but produced low precision, indicating that a linear decision surface was insufficiently selective for the nonlinear failure structure. The stronger performance of tree ensembles is consistent with the broader predictive-maintenance literature, in which nonlinear classifiers are useful when failure risk arises from interactions among operating conditions rather than from isolated one-variable effects [22]. This empirical result strengthens the Lean Six Sigma 4.0 argument that advanced analytics can extend conventional statistical tools when process relationships are

nonlinear, while Define-Measure-Analyze-Improve-Control remains necessary to govern how those predictions are acted upon [4,5].

6.2 Statistical evidence and explainability as complementary Analyze-phase tools

The second contribution is the deliberate combination of classical statistical evidence with explainable machine learning. Torque showed the largest standardized univariate difference and the largest adjusted odds ratio, while tool wear and temperature variables also remained informative. Tree Shapley additive explanations independently ranked torque, tool wear and rotational speed among the most influential predictive features. The convergence is important because the two analytical layers answer different questions. Effect sizes and adjusted regression quantify average associations and uncertainty under a specified model; Shapley additive explanations decomposes the contribution of features to predictions from a nonlinear ensemble [16]. Their agreement therefore increases interpretive coherence without converting either analysis into causal evidence.

This combined use of statistical inference and explainable artificial intelligence addresses a recurring limitation in digital-quality applications: a high-performing classifier may identify risk without providing an engineering narrative that supports root-cause investigation or corrective action. Prior predictive-maintenance research similarly emphasizes that outputs become more useful to decision makers when explanations accompany binary predictions [23]. Within Define-Measure-Analyze-Improve-Control, explanation therefore has a governance function. Global attribution can prioritize variables for process investigation, while local explanations can support case-level maintenance review. However, because Artificial Intelligence for Industry 2020 is synthetic and its failure mechanisms are engineered, the identified feature hierarchy should be interpreted as benchmark evidence; plant deployment would require measurement-system analysis, engineering validation and, where feasible, designed experiments before interventions are treated as causal.

Recent empirical work further supports this positioning. One study found Extreme Gradient Boosting competitive for machine-failure prediction on the University of California, Irvine predictive-maintenance benchmark [24], while another showed that realistic industrial irregularities can materially reduce model performance relative to the clean Artificial Intelligence for Industry 2020 benchmark [25]. At the process-improvement level, an applied Define-Measure-Analyze-Improve-Control 4.0 study demonstrated an architecture that integrates data-science methods with measurement and error-prevention activities [26]. Together, these studies reinforce the need to evaluate predictive performance, deployment realism and Define-Measure-Analyze-Improve-Control governance jointly rather than as separate concerns.

6.3 Calibration and threshold governance as the bridge from Analyze to Improve

A third contribution is the separation of discrimination from probability reliability. Extreme Gradient Boosting ranked observations most effectively but displayed poorer calibration than random forest. This distinction is well established in probabilistic prediction: a model may order cases correctly while producing probabilities that are systematically too high or too low [18]. In a Quality 4.0 setting, calibration matters because risk scores may trigger maintenance, inspection or escalation. Treating an uncalibrated score as an absolute probability can therefore distort resource allocation even when area under the receiver operating characteristic curve is high. The Brier and calibration results extend the predictive-maintenance analysis beyond the usual model-ranking exercise and provide a direct Control-phase monitoring target.

The threshold analysis further operationalizes the Improve phase. Moving the Extreme Gradient Boosting threshold from 0.10 to 0.80 reduced the false-alarm rate from 8.0% to below 1%, but recall declined from 0.920 to 0.696. This is a quality-economic trade-off rather than a purely statistical choice. A lower threshold increases preventive workload but reduces escaped failures; a higher threshold reduces unnecessary intervention but accepts more

missed events. The study therefore avoids presenting a universal optimum and instead treats the threshold as a policy parameter to be selected using plant-specific failure costs, maintenance capacity, safety requirements and customer consequences. This is more consistent with Lean Six Sigma than maximizing a generic machine-learning metric because Improve requires a process decision, not merely a better score.

6.4 Contribution to Quality 4.0 and Six Sigma theory

The findings extend existing Quality 4.0 and Lean Six Sigma 4.0 literature in three ways. First, they provide an empirical mechanism for the argument that Quality 4.0 builds on rather than displaces established quality management [1,13]. Statistical baselines, effect sizes and process-control logic remain necessary, while artificial intelligence contributes nonlinear detection and scalable risk estimation. Second, the analysis makes the complementarity between data-science and quality-management lifecycles concrete. Prior framework-mapping research distinguishes the data-centric orientation of data-science frameworks from the process-centric and control-oriented nature of Define-Measure-Analyze-Improve-Control [11]; the present framework connects the two by assigning model outputs to explicit Define-Measure-Analyze-Improve-Control decisions. Third, it adds calibration and threshold governance to the integration literature, two elements that are often technically important but underdeveloped in conceptual Lean Six Sigma 4.0 frameworks. The results also connect technical capability to organizational maturity. Prior implementation research identifies culture, scalability and application development as areas requiring attention in developing-country Quality 4.0 implementation [6]. The present framework shows why those dimensions matter operationally. A calibrated model still requires reliable sensor infrastructure, defined escalation authority, maintenance capacity, retraining ownership and acceptance of data-supported decisions. Thus, Quality 4.0 maturity is not simply the presence of artificial intelligence; it is the institutional capacity to convert validated

analytical signals into controlled, auditable and sustainable improvement routines.

6.5 Boundary conditions and interpretation

The reported thresholds, odds ratios and Shapley additive explanation patterns are benchmark-specific rather than transferable plant constants. The theoretical contribution of this section is therefore the integration logic: statistical evidence, leakage-controlled prediction, explanation and decision analysis are assigned to explicit Define-Measure-Analyze-Improve-Control responsibilities. Detailed external-validity limitations and validation priorities are summarized in Section 8.

7. Managerial and Practical Implications

Manufacturers should implement the framework in stages. First, the Define and Measure phases should establish the critical-to-quality, failure-cost logic, sensor reliability, label integrity and a stable baseline. Only after measurement readiness is demonstrated should teams compare interpretable and nonlinear models under leakage-controlled validation. This sequence prevents an accurate algorithm from being built on an operationally weak measurement system.

Governance responsibilities should be explicit. Quality engineers should own critical-to-quality definitions, process interpretation and control plans; maintenance teams should own escalation and intervention capacity; data-science teams should own model validation, calibration and drift diagnostics; and operations managers should approve threshold policies using safety, downtime and false-alarm costs. Local Shapley additive explanations can support case review, but engineering confirmation should precede irreversible interventions.

A practical control dashboard should combine conventional quality indicators with predictive metrics. The quality layer should track failure prevalence, Pareto structure, process-control signals, intervention completion and recurrence. The predictive layer should track recall, precision, false-alarm burden, calibration/Brier score, feature drift and model performance over time. Threshold changes and retraining should be documented as controlled process changes rather than informal model adjustments.

8. Limitations and Future Research

Artificial Intelligence for Industry 2020 is synthetic, complete and comparatively clean. Real plants contain sensor faults, interventions, temporal dependence, product-mix changes and concept drift; recent Artificial Intelligence for Industry Predictive Maintenance Dataset with Industrial Irregularities evidence shows that such irregularities can materially alter predictive performance [25]. The present results therefore provide a reproducible methodological benchmark rather than evidence that the same coefficients or thresholds will transfer to a factory. Specification limits are unavailable, so process capability indices were not estimated. Cross-validation intervals summarize fold variability but are not exact inferential confidence intervals. The variance inflation factor sensitivity check did not indicate severe multicollinearity (maximum 4.31), but the adjusted logistic model still assumes linear effects on the log-odds scale; the nonlinear ensemble is therefore treated as the primary predictive model. Shapley additive explanations explains the fitted model rather than proving causal mechanisms. External, time-aware and prospective validation should test whether recalibrated, plant-specific model-guided actions reduce defects, downtime and total quality cost.

9. Conclusion

This study provides an original empirical evaluation of an artificial intelligence-augmented Six Sigma workflow using the full Artificial Intelligence for Industry 2020 dataset. Extreme Gradient Boosting provided the strongest discrimination and recall, random forest provided higher precision/harmonic mean of precision and recall and better Brier performance, and statistical/explainable artificial intelligence analyses identified load, wear and speed-related variables as central signals. The resulting Quality 4.0 architecture links Define to critical-to-quality consequences, Measure to baseline/data quality, Analyze to statistics and explainable artificial intelligence, Improve to threshold/action policies, and Control to calibration, statistical process control and drift monitoring. The workflow therefore positions artificial intelligence as a governed analytical extension of Six Sigma rather than a substitute for quality-management discipline.

REFERENCES

- Escobar, C.A., McGovern, M.E. and Morales-Menendez, R. (2021), "Quality 4.0: a review of big data challenges in manufacturing", *Journal of Intelligent Manufacturing*, Vol. 32, pp. 2319-2334, doi: 10.1007/s10845-021-01765-4.
- Bousdekis, A., Lepenioti, K., Apostolou, D. and Mentzas, G. (2023), "Data analytics in Quality 4.0: literature review and future research directions", *International Journal of Computer Integrated Manufacturing*, Vol. 36, pp. 678-701, doi: 10.1080/0951192X.2022.2128219.
- Markatos, N.G. and Mousavi, A. (2023), "Manufacturing quality assessment in the Industry 4.0 era: a review", *Total Quality Management & Business Excellence*, Vol. 34 Nos 13-14, pp. 1655-1681, doi: 10.1080/14783363.2023.2194524.
- Ibrahim, A. and Kumar, G. (2024), "Selection of Industry 4.0 technologies for Lean Six Sigma integration using fuzzy DEMATEL approach", *International Journal of Lean Six Sigma*, Vol. 15 No. 5, pp. 1025-1042, doi: 10.1108/IJLSS-05-2023-0090.
- Vigneshvaran, R. and Vinodh, S. (2026), "State of the art review on Lean Six Sigma with Industry 4.0 integration", *International Journal of Services and Operations Management*, Vol. 54 No. 2, pp. 229-252, doi: 10.1504/IJSOM.2026.154595.
- Zulqarnain, A., Wasif, M. and Iqbal, S.A. (2022), "Developing a Quality 4.0 implementation framework and evaluating the maturity levels of industries in developing countries", *Sustainability*, Vol. 14 No. 18, 11298, doi: 10.3390/su141811298.
- Davis, J. and Goadrich, M. (2006), "The relationship between Precision-Recall and ROC curves", *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233-240, doi: 10.1145/1143844.1143874.

- Flach, P. and Kull, M. (2015), "Precision-Recall-Gain curves: PR analysis done right", *Advances in Neural Information Processing Systems* 28, pp. 838-846.
- UCI Machine Learning Repository (2020), "AI4I 2020 Predictive Maintenance Dataset", doi: 10.24432/C5HS5C.
- Matzka, S. (2020), "Explainable artificial intelligence for predictive maintenance applications", 2020 International Conference on Artificial Intelligence for Industries (AI4I), pp. 69-74, doi: 10.1109/AI4I49448.2020.00023.
- Bahr, L., Poßner, L., Wewerka, J., Bittencourt, J., Gröger, S. and Daub, R. (2026), "Aligning quality management and data science with AI agents for process optimization in manufacturing: a survey", *Journal of Industrial Information Integration*, Vol. 52, 101129, doi: 10.1016/j.jii.2026.101129.
- Narkhede, G., Naranje, V. and Mahajan, S. (2026), "Tri-dimensional review on integrating Industry 4.0 and Lean Six Sigma for manufacturing excellence", *Discover Sustainability*, Vol. 7, 618, doi: 10.1007/s43621-026-02961-z.
- Escobar, C.A., Macias, D., McGovern, M., Hernández-de-Menéndez, M. and Morales-Menendez, R. (2022), "Quality 4.0 - an evolution of Six Sigma DMAIC", *International Journal of Lean Six Sigma*, Vol. 13 No. 6, pp. 1200-1238, doi: 10.1108/IJLSS-05-2021-0091.
- Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45, pp. 5-32, doi: 10.1023/A:1010933404324.
- Chen, T. and Guestrin, C. (2016), "XGBoost: a scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, doi: 10.1145/2939672.2939785.
- Lundberg, S.M. and Lee, S.I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765-4774.
- Brier, G.W. (1950), "Verification of forecasts expressed in terms of probability", *Monthly Weather Review*, Vol. 78 No. 1, pp. 1-3, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- Niculescu-Mizil, A. and Caruana, R. (2005), "Predicting good probabilities with supervised learning", *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625-632, doi: 10.1145/1102351.1102430.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357, doi: 10.1613/jair.953.
- Wilson, E.B. (1927), "Probable inference, the law of succession, and statistical inference", *Journal of the American Statistical Association*, Vol. 22 No. 158, pp. 209-212.
- Bengio, Y. and Grandvalet, Y. (2004), "No unbiased estimator of the variance of K-fold cross-validation", *Journal of Machine Learning Research*, Vol. 5, pp. 1089-1105.
- Susto, G.A., Schirru, A., Pampuri, S., McLoone, S. and Beghi, A. (2015), "Machine learning for predictive maintenance: a multiple classifier approach", *IEEE Transactions on Industrial Informatics*, Vol. 11 No. 3, pp. 812-820, doi: 10.1109/TII.2014.2349359.
- Dereci, U. and Tuzkaya, G. (2024), "An explainable artificial intelligence model for predictive maintenance and spare parts optimization", *Supply Chain Analytics*, Vol. 8, 100078, doi: 10.1016/j.sca.2024.100078.
- Yadav, D.K., Kaushik, A. and Yadav, N. (2024), "Predicting machine failures using machine learning and deep learning algorithms", *Sustainable Manufacturing and Service Economics*, Vol. 3, 100029, doi: 10.1016/j.smse.2024.100029.

Autran, J.-V., Kuhn, V., Diguët, J.-P., Dubois, M. and Buche, C. (2024), "AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities", *Procedia Computer Science*, Vol. 246, pp. 1201-1209, doi: 10.1016/j.procs.2024.09.546.

Pongboonchai-Empl, T., Antony, J., Neustock, J., Stemann, D., Komkowski, T. and Sony, M. (2025), "DMAIC 4.0 in Action: introducing MSA 4.0 and Poka Yoke 4.0", *International Journal of Computer Integrated Manufacturing*, pp. 1-21, doi: 10.1080/0951192X.2025.2599558.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Kuhn, M. and Johnson, K. (2013), *Applied Predictive Modeling*, Springer, New York, NY, doi: 10.1007/978-1-4614-6849-3.

Hand, D.J. (2009), "Measuring classifier performance: a coherent alternative to the area under the ROC curve", *Machine Learning*, Vol. 77, pp. 103-123, doi: 10.1007/s10994-009-5119-5.

Powers, D.M.W. (2011), "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", *Journal of Machine Learning Technologies*, Vol. 2 No. 1, pp. 37-63.

Zulqarnain, A., Iqbal, S. A., & Khalid, R. (2013). Implementing Six Sigma methodology in a developing country. *International Journal of Process Management and Benchmarking*, 3(3), 314-333.