

Predictive Maintenance Optimization Using Machine Learning for Industrial Equipment Reliability

Yassine HADDIOUI^{1*}, Irwandi Rizki Putra²

¹ University Mohamed V in Rabat, Morocco

² Universitas Riau Indonesia, Indonesia

email: yassine_haddioui@um5.ac.ma¹

Article Info :

Received:

02-09-2026

Revised:

08-10-2026

Accepted:

14-10-2026

Abstract

Predictive maintenance models are often evaluated primarily as classifiers, although practical maintenance decisions depend on the relative consequences of preventive interventions and missed failures. This study develops a reliability-oriented predictive maintenance framework that integrates machine-learning failure probabilities with cost-sensitive maintenance thresholds. A reproducible simulation of 80,000 industrial-equipment cycles was constructed from the documented process envelopes and failure mechanisms of the AI4I 2020 benchmark, while the original UCI dataset was used only as a calibration reference. Six raw operating variables were supplemented with temperature difference, mechanical power, and a type-adjusted wear-torque index. Four classifiers—logistic regression, random forest, extra trees, and XGBoost—were evaluated using stratified training, validation, and independent test partitions. Random forest achieved the strongest test PR-AUC of 0.967, ROC-AUC of 0.989, F1 of 0.965, and MCC of 0.964. Incorporating physics-derived variables increased PR-AUC from 0.943 to 0.966 and reduced the Brier score from 0.0088 to 0.0035. Cost-sensitive thresholds reduced normalized maintenance costs by 46.3–88.9% relative to reactive maintenance across failure-to-preventive-maintenance cost ratios of 2:1 to 20:1. Moderate sensor perturbations preserved predictive performance, whereas 10% standard-deviation noise substantially reduced PR-AUC and recall. These findings demonstrate that predictive performance, engineering-informed feature design, sensor robustness, and maintenance economics should be evaluated jointly before industrial deployment.

Keywords: Cost-Sensitive Optimization, Failure Prediction, Industrial Reliability, Machine Learning, Predictive Maintenance.



©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.
(<https://creativecommons.org/licenses/by-nc/4.0/>)

INTRODUCTION

Industrial maintenance has shifted from fixed schedules toward condition-dependent decisions as sensors, edge devices, and industrial data infrastructures have become increasingly accessible across modern manufacturing environments. Predictive maintenance uses observed operating conditions to estimate failure risk, degradation trajectories, or remaining useful life before functional loss occurs, allowing maintenance actions to be positioned closer to the actual condition of an asset. Recent developments have expanded the field from conventional condition monitoring toward integrated data-driven prognostics in which sensing, machine learning, failure prediction, and maintenance decision-making are treated as interconnected components of an industrial reliability system (Achouch et al., 2022). The technical value of this transition is not limited to higher diagnostic accuracy because a prediction becomes operationally meaningful only when it supports timely intervention, limits unplanned downtime, preserves asset availability, and allocates maintenance resources under practical constraints. Systematic evidence indicates that machine-learning-based predictive maintenance has increasingly incorporated heterogeneous sensor information, prognostic models, and industrial decision processes rather than relying exclusively on isolated fault-classification tasks (Fernandes et al., 2022). This evolution places equipment reliability at the intersection of predictive analytics and maintenance optimization, creating a research problem in which the quality of the prediction and the quality of the resulting intervention must be considered jointly.

Machine-learning models are now applied to rotating machinery, production lines, induction motors, and sensor-rich manufacturing systems because nonlinear relationships among temperature, load, speed, vibration-related indicators, and degradation processes are difficult to represent through

fixed alarm thresholds. Studies of induction motors demonstrate that artificial neural networks and other learning architectures can identify condition patterns that are not readily captured by conventional rule-based monitoring (Drakaki et al., 2022). Industrial applications further show that the integration of IoT-enabled sensing with machine learning can support continuous condition assessment and earlier identification of potentially abnormal operating states (Elkateb et al., 2024). The resulting predictive capability, however, depends strongly on the representativeness of the training data, the availability and quality of condition variables, class imbalance, feature construction, and the extent to which experimental observations reflect actual operating regimes. Research on sensor-based predictive maintenance has shown that neural models can provide useful predictive performance when sufficiently informative operational data are available, but model effectiveness remains closely connected to the structure and quality of the underlying measurements (Fordal et al., 2023). Public benchmarks therefore remain valuable for controlled methodological comparison, particularly because industrial failure data are often sparse, proprietary, and difficult to disclose, yet their statistical and synthetic characteristics need to be considered when findings are interpreted beyond the benchmark environment. This distinction is particularly important for reliability-oriented research because a model that performs well under controlled data conditions does not automatically provide an economically or technically defensible maintenance policy in production.

The AI4I 2020 predictive-maintenance dataset provides a widely used benchmark for investigating these methodological questions because it combines product type, air temperature, process temperature, rotational speed, torque, tool wear, machine-failure status, and several documented failure modes within a structured dataset of 10,000 observations (UCI Machine Learning Repository, 2020). Its design has supported research on explainable predictive maintenance and has provided a reproducible basis for evaluating how machine-learning algorithms translate operating variables into failure predictions (Matzka, 2020). Yet, the controlled structure that makes AI4I attractive for benchmarking also limits the extent to which conventional classification results can be interpreted as evidence of field reliability. Recent work on industrial predictive-maintenance datasets has emphasized the importance of irregularities, heterogeneous operating conditions, and more complex industrial settings when developing models intended for realistic deployment (Autran et al., 2024). A further concern is that synthetic failure mechanisms may create statistical relationships that are easier for algorithms to identify than equivalent relationships would be in naturally occurring equipment degradation. The literature therefore contains a conceptual tension between the need for standardized datasets that enable reproducible model comparison and the need for realistic operational variability that permits stronger conclusions about reliability. Resolving this tension requires experimental designs that preserve benchmark reproducibility while explicitly testing feature engineering, robustness, and decision performance rather than treating classification accuracy as the final research endpoint.

A second limitation concerns the separation between predictive performance and maintenance optimization, because a high discrimination score does not by itself determine whether an alert should trigger inspection, repair, component replacement, or continued operation. False-positive alerts can consume maintenance labor, spare parts, and production capacity, whereas false-negative predictions may expose equipment to failure and generate substantially larger operational consequences. Recent predictive-maintenance research has therefore moved toward decision-oriented frameworks that combine prognostic uncertainty with maintenance actions rather than evaluating prediction models independently from operational consequences (Wen et al., 2022). Bayesian decision models, dynamic scheduling, reinforcement learning, and multi-objective optimization have increasingly been used to connect estimated degradation or failure probability with intervention timing and resource constraints (Lee & Mitici, 2023). This development suggests that maintenance optimization should be formulated as a threshold-selection problem in which the acceptable probability of failure is determined not solely by statistical discrimination but also by the relative consequences of preventive intervention and equipment failure. The practical challenge becomes more pronounced when the same prediction model may generate different optimal intervention thresholds under different cost structures, production conditions, or reliability requirements. A maintenance strategy that ignores these decision parameters can therefore produce technically accurate predictions while remaining operationally inefficient, making the translation from machine-learning output to maintenance policy a central unresolved issue for industrial reliability research.

Model interpretation and physical consistency provide another dimension of reliability because maintenance engineers require signals that can be connected to observable degradation mechanisms rather than probability estimates that are difficult to audit or operationalize. Reviews of condition-based maintenance emphasize that interpretability can strengthen confidence in machine-learning predictions by allowing practitioners to examine which operational variables contribute to predicted equipment conditions (Sharma et al., 2024). Digital-twin research similarly positions physically meaningful state information as an important bridge between monitoring, prognostics, and maintenance planning, particularly when machine-learning models are embedded within broader representations of industrial assets (Chen et al., 2023). Derived variables such as thermal difference, mechanical power, and wear–load interaction can consequently provide a useful intermediate layer between raw sensor measurements and statistical predictions because they encode engineering relationships while remaining calculable from ordinary operational channels. Such variables may also improve the stability and interpretability of failure classification by representing interactions that individual sensor variables cannot fully express. The methodological issue is not simply whether engineered features increase predictive accuracy, but whether their contribution remains meaningful when the model is exposed to measurement noise and when the resulting predictions are converted into maintenance decisions. A reliability-oriented framework therefore needs to evaluate predictive discrimination, feature contribution, threshold behavior, and robustness within a common experimental structure.

This research addresses these gaps through a controlled reliability experiment that extends predictive-maintenance analysis beyond accuracy-only model comparison. A large operational-cycle simulation is calibrated to the documented AI4I process envelopes and failure mechanisms and is used to compare four classification models under a leakage-resistant train–validation–test design. The analysis evaluates whether physics-derived variables provide incremental predictive information and examines whether their contribution remains stable when sensor measurements are perturbed by controlled noise. Maintenance thresholds are selected exclusively from validation data under alternative failure-to-preventive-maintenance cost ratios, allowing the predictive output to be translated into explicit intervention policies rather than treated as an end-point classification score. The resulting policies are then evaluated on an untouched test partition to distinguish model selection effects from genuine out-of-sample performance. Permutation-based feature importance is incorporated to provide a transparent account of the operational variables underlying the predictions and to examine whether engineering-derived features contribute consistently across models. The study consequently aims to establish a reproducible methodological link between machine-learning-based failure probability, engineering-informed feature representation, cost-sensitive maintenance thresholds, and equipment reliability proxies, while explicitly defining the limits of inference associated with synthetic benchmark-based experimentation.

RESEARCH METHODS

The methodological framework was designed as a controlled simulation-based predictive-maintenance experiment in which the AI4I 2020 dataset served as a calibration reference rather than as the direct modelling sample, consistent with its role as a synthetic benchmark for industrial predictive-maintenance research (UCI Machine Learning Repository, 2020). The framework integrated six raw operational variables—product type, air temperature, process temperature, rotational speed, torque, and tool wear—with three physics-derived predictors comprising temperature difference, mechanical power, and a type-adjusted wear–torque index, so that the classification task retained explicit links to thermal, mechanical, and load–wear failure mechanisms. This physics-informed formulation was adopted to improve the engineering interpretability of machine-learning predictions rather than relying on unrestricted feature expansion, an approach consistent with recent work emphasizing trustworthy and interpretable predictive-maintenance systems (Obeten & Jimoh, 2026). Four supervised classifiers were specified: logistic regression with class balancing and standardized numerical predictors, random forest and extra trees with 350 trees and a minimum leaf size of two, and XGBoost with 450 boosting rounds, depth four, a learning rate of 0.05, 0.85 row and column subsampling, and positive-class weighting based on training prevalence. The simulated population was stratified into 70% training, 15% validation, and 15% test partitions, while all model hyperparameters were fixed before final test evaluation to prevent information from the independent test partition from influencing model development, following the importance of systematic model comparison and evaluation in predictive-

maintenance applications (Li et al., 2025). The conceptual framework therefore connected operational sensing, physics-informed representation, machine-failure classification, probabilistic prediction, and subsequent maintenance decision optimization within a single analytical pipeline.

The technical experiment generated 80,000 equipment-cycle observations using Monte Carlo simulation with the fixed random seed 20260916 to provide a sufficiently large and reproducible basis for threshold optimization and robustness analysis. Product types followed an approximate 50:30:20 low–medium–high mixture, air temperature was generated around 300 K with stochastic variation, process temperature was coupled around 310 K, torque was centered near 40 Nm, rotational speed was inversely associated with torque around 1,538 rpm, and tool wear accumulated sequentially before stochastic replacement. Five failure mechanisms were simulated, namely heat-dissipation failure when the temperature difference was below 8.6 K with speed below 1,380 rpm, power failure when mechanical power fell outside 3.5–9.0 kW, overstrain failure when wear–torque exceeded a product-specific threshold, probabilistic tool-wear failure near replacement, and random failure at a 0.1% probability, with machine failure coded as one when at least one mechanism occurred. Model performance was evaluated using ROC-AUC, PR-AUC, precision, recall, F1, Matthews correlation coefficient, balanced accuracy, confusion matrices, and Brier score, while 250 bootstrap replicates were used to estimate 95% intervals for the ROC-AUC and PR-AUC of the model with the highest validation PR-AUC, consistent with the use of probabilistic and decision-oriented evaluation in predictive-maintenance research (Wang et al., 2024). Maintenance thresholds were selected exclusively from validation predictions by minimizing $C(\theta) = C_{PM}(TP + FP) + C_{MF}FN$ under cost ratios $C_{MF}/C_{PM} = 2, 5, 10, \text{ and } 20$, and each selected threshold was then applied unchanged to the independent test partition to estimate the resulting maintenance performance (Xu & Zhang, 2026). Robustness was assessed by independently adding Gaussian sensor noise at 0%, 1%, 2%, 5%, and 10% of each signal's standard deviation, recalculating the derived features after perturbation, and reporting the resulting changes in predictive performance and the reliability proxy $R^* = 1 - FN/N$, which was interpreted only as an intervention-based reliability proxy rather than a lifetime reliability or survival function (Sharma et al., 2024).

RESULTS AND DISCUSSION

Failure Structure and Operating-State Separation

The 80,000-cycle simulation generated 3,627 cycles with at least one failure mechanism, corresponding to a machine-failure prevalence of 4.534%. Heat-dissipation failure represented the largest individual failure component at 1.903%, followed by power failure at 1.195% and overstrain failure at 1.180%. Tool-wear failure accounted for 0.321% of cycles, whereas random failure occurred in only 0.085% of observations. As shown in Table 1, the sum of the individual failure-mode frequencies exceeds neither nor exactly reproduces the composite failure prevalence because a single equipment cycle can activate more than one mechanism. The resulting positive-class proportion creates a distinctly imbalanced classification problem in which the operational population is dominated by non-failure cycles, making accuracy an inadequate standalone indicator of predictive quality. Such imbalance is characteristic of industrial fault-diagnosis settings, where meaningful evaluation requires attention to minority-class discrimination, precision–recall behavior, and the operational consequences of missed failures (Fernandes et al., 2022).

Table 1. Failure Composition in the 80,000-Cycle Simulation

Failure Mode	Cycles	Share of All Cycles (%)
Tool wear failure	257	0.321
Heat dissipation failure	1,522	1.903
Power failure	956	1.195
Overstrain failure	944	1.180
Random failure	68	0.085
Any machine failure	3,627	4.534

The dominance of heat-dissipation failure indicates that thermal operating conditions constitute the most frequent route to failure under the simulated equipment logic, while power and overstrain mechanisms contribute substantial secondary sources of risk. The relatively small contribution of random failure is also analytically important because it prevents stochastic events from dominating the classification boundary and keeps most detectable failures connected to observable operating states. The frequency structure therefore reflects a system in which failure probability is primarily generated by interactions among measurable physical variables rather than by an undifferentiated random process. This design is compatible with predictive-maintenance frameworks that distinguish condition indicators from failure outcomes and seek to transform operational measurements into actionable prognostic information (Achouch et al., 2022). The distribution also explains why the model evaluation cannot rely exclusively on ROC-based discrimination, because a classifier can maintain high ranking performance while producing an operationally unsuitable alert burden when positive events are sparse. The use of PR-AUC, MCC, recall, and precision is consequently appropriate for assessing whether the learned boundary preserves meaningful discrimination for the minority failure class.

The contrast between failed and normal operating states further demonstrates that the simulated failures were associated with measurable changes in the equipment's physical condition. As reported in Table 2, the mean temperature difference declined from 10.034 K during normal operation to 9.275 K among failed cycles, while mean rotational speed decreased from 1,540.454 to 1,477.821 rpm. Failed cycles also exhibited higher mean torque, increasing from 39.865 to 43.273 Nm, and substantially greater tool wear, rising from 110.882 to 142.663 min. Mechanical power showed a smaller difference between states, with a mean of 6.023 kW in failed cycles compared with 6.266 kW in normal cycles, which is expected because the simulated power-failure mechanism can occur at either the lower or upper boundary of the allowable power envelope. The mean therefore conceals the two-sided nature of this mechanism and illustrates why aggregate descriptive statistics alone cannot fully characterize failure conditions.

Table 2. Operating-Variable Distribution and Failure-State Contrast

Variable	Mean	SD	Failure Mean	Normal Mean
Air temperature (K)	300.000	2.000	300.679	299.968
Process temperature (K)	310.000	1.442	309.954	310.002
Temperature difference (K)	10.000	1.131	9.275	10.034
Rotational speed (rpm)	1537.614	181.138	1477.821	1540.454
Torque (Nm)	40.020	10.017	43.273	39.865

The wear–torque index provides an even clearer separation between operating states, increasing from 0.377 under normal conditions to 0.562 among failed cycles. This difference is consistent with the underlying overstrain mechanism because the index explicitly combines accumulated wear with instantaneous torque relative to the product-type-specific failure threshold. Temperature difference and rotational speed similarly retain direct mechanistic relevance because their joint configuration determines the simulated heat-dissipation failure condition. The stronger separation of these derived variables suggests that engineering relationships can concentrate failure-relevant information that is distributed across individual sensor measurements. Knowledge-guided feature construction has been identified as an important direction for predictive-maintenance systems because physically meaningful representations can improve both model transparency and the connection between predictions and equipment behavior (Ucar et al., 2024). The result also provides a substantive rationale for evaluating the physics-derived feature set separately from the raw sensor variables in the subsequent model analysis.

The observed operating-state separation should not, however, be interpreted as evidence that identical numerical boundaries apply to a particular industrial machine. The simulated thresholds were intentionally specified to create transparent and reproducible relationships between operating conditions and failure labels, whereas real manufacturing environments contain changing production regimes, sensor degradation, irregular observations, maintenance interventions, and failure mechanisms that may not be explicitly measured. Benchmark-oriented research has highlighted the limitations of clean predictive-maintenance datasets when irregularities encountered in industrial settings are absent

or insufficiently represented (Autran et al., 2024). The controlled simulation remains valuable because it isolates the effects of feature construction, classification, decision thresholds, and sensor perturbation without confounding these components with uncontrolled field conditions. Its principal analytical contribution is therefore not direct estimation of plant-level failure probabilities but examination of how a mechanism-informed predictive-maintenance pipeline behaves under a known failure-generating structure. This distinction is important when interpreting subsequent performance metrics and prevents the simulated results from being presented as direct evidence of general industrial reliability.

The combined failure composition and operating-state contrasts establish a clear analytical basis for the predictive modelling stage. Failure events were sufficiently sparse to create a meaningful imbalanced-learning problem, yet the principal mechanisms generated identifiable changes in thermal, mechanical, and wear-related operating states. The coexistence of multiple failure mechanisms also creates interactions that are unlikely to be represented adequately by a single linear boundary, particularly where failure depends on conditional combinations such as low temperature difference with low rotational speed or excessive wear under elevated torque. Predictive-maintenance research similarly emphasizes that machine-learning performance depends on the structure and quality of condition-monitoring data as well as on the selected modelling strategy (Ward et al., 2024). The present results therefore support the subsequent comparison of ensemble and linear classifiers while providing an engineering interpretation for the inclusion of temperature difference, mechanical power, and wear–torque index. The next analysis evaluates whether these physically structured representations translate into measurable gains in predictive discrimination and probability quality.

Predictive Performance and Value of Physics-Derived Features

The independent test results demonstrate substantial differences in predictive behavior across the four evaluated classifiers. Random forest achieved the highest PR-AUC at 0.967, narrowly exceeding XGBoost at 0.965 and extra trees at 0.954, while logistic regression reached only 0.458. XGBoost obtained the highest ROC-AUC at 0.991, but this marginal advantage did not translate into comparable precision at the default decision threshold. As presented in Table 3, random forest produced precision of 0.998, recall of 0.934, F1 of 0.965, and MCC of 0.964, accompanied by the lowest Brier score of 0.0035. These results indicate that the selected ensemble model was able to distinguish failure and non-failure cycles while maintaining a substantially lower false-alert burden than the alternative boosting model. The contrast between ROC-AUC and PR-AUC also demonstrates why model selection for sparse industrial failures should consider the operational distribution of the positive class rather than relying on ranking discrimination alone (Li et al., 2025).

Table 3. Comparative Predictive Performance of Machine-Learning Models on the Independent Test Set

Model	ROC-AUC	PR-AUC	Prec.	Recall	F1	MCC	Brier
Random forest	0.989	0.967	0.998	0.934	0.965	0.964	0.0035
XGBoost	0.991	0.965	0.837	0.943	0.887	0.883	0.0100
Extra trees	0.988	0.954	0.897	0.928	0.912	0.908	0.0100
Logistic regression	0.938	0.458	0.254	0.869	0.393	0.429	0.0967

Random forest detected 508 of the 544 failures in the independent test partition at the 0.50 threshold while generating only one false alert. XGBoost detected a slightly larger number of failures, with recall of 0.943, but produced 100 false alerts and consequently reduced precision to 0.837. Extra trees occupied an intermediate position, whereas logistic regression combined relatively high recall with substantially weaker precision and probability quality. This pattern shows that maximizing sensitivity alone does not guarantee a practically favorable maintenance classifier because each false positive can represent an unnecessary inspection, intervention, or production interruption. The very low Brier score of random forest further indicates that its predicted probabilities were more closely aligned with observed outcomes than those of the competing models. Probability quality becomes particularly relevant when classification outputs are subsequently transformed into maintenance thresholds rather than treated as fixed binary labels (Wang, Zhu, & Zhao, 2024). The model comparison therefore

supports random forest as the analytical basis for the subsequent cost-sensitive decision analysis, while retaining the distinction between statistical performance and universal model superiority.

The comparatively weak performance of logistic regression provides evidence that the simulated failure structure contains interactions that are not efficiently represented by a single linear decision boundary. Its ROC-AUC of 0.938 was lower than the values obtained by all three ensemble approaches, while its PR-AUC of 0.458 indicates a much greater difficulty in maintaining discrimination within the sparse failure class. The result is consistent with the presence of conditional mechanisms in which failure depends on combinations of temperature difference, speed, torque, wear, and derived mechanical quantities rather than on isolated additive effects. Tree-based ensembles can partition such multidimensional operating spaces into localized regions without requiring the analyst to specify the interaction structure in advance. Similar predictive-maintenance research has emphasized the relevance of nonlinear learning approaches for representing complex equipment conditions and degradation patterns (Taşcı et al., 2023). The finding also clarifies why the present experiment did not treat model complexity as an objective in itself, since the relevant criterion is whether the algorithm captures the failure structure with adequate discrimination and manageable alert consequences.

Feature ablation provides direct evidence that the physics-derived variables contributed information beyond the six raw sensor variables. As shown in Table 4, the random-forest model using only raw variables achieved a ROC-AUC of 0.986 and PR-AUC of 0.943, whereas adding temperature difference, mechanical power, and the type-adjusted wear–torque index increased these values to 0.989 and 0.966, respectively. The Brier score also declined from 0.0088 to 0.0035, indicating an improvement in probability quality alongside the increase in discrimination. Because the derived variables were calculated from measurements already available in the simulated operating state, their contribution represents a transformation of existing information rather than an expansion of sensor requirements. Temperature difference concentrates the thermal relationship relevant to heat-dissipation failure, mechanical power expresses the combined torque–speed relationship underlying the power envelope, and the wear–torque index expresses load relative to accumulated wear and product-specific overstrain capacity. Knowledge-guided representation of this type is consistent with predictive-maintenance approaches that integrate engineering knowledge with data-driven modelling to strengthen the physical interpretability of prognostic outputs (Li, Chen, Hu, & Zhang, 2023). The ablation result therefore supports the proposition that feature engineering based on failure mechanisms can provide greater analytical value than indiscriminate expansion of the predictor space.

Table 4. Predictive Performance of Random Forest With and Without Physics-Derived Features

Feature set	ROC-AUC	PR-AUC	Brier score
Raw sensor variables	0.986	0.943	0.0088
Raw + physics-derived variables	0.989	0.966	0.0035

The improvement in PR-AUC is particularly relevant because the positive class represents only 4.534% of the simulated population, making precision–recall behavior more informative for evaluating separation of failure cycles than conventional accuracy alone. An increase from 0.943 to 0.966 represents a 0.023 absolute gain in PR-AUC, while the reduction in Brier score from 0.0088 to 0.0035 indicates a substantial improvement in probabilistic prediction. These changes suggest that the derived features did not merely improve ranking at the extreme ends of the score distribution but also contributed to more reliable probability estimates. The result aligns with the broader requirement for trustworthy industrial AI systems in which predictive performance should remain connected to observable equipment conditions and engineering meaning (Ucar et al., 2024). It also provides an analytical explanation for the strong performance of random forest: the algorithm operates on a feature space in which key failure mechanisms have already been expressed through physically meaningful relationships. The contribution of these variables is consequently methodological as well as predictive, because they establish a transparent bridge between sensor observations and the later maintenance decision rule.

The selected random-forest model achieved a ROC-AUC of 0.989 and a PR-AUC of 0.967 on the independent test partition, with bootstrap 95% intervals of 0.982–0.994 and 0.955–0.978, respectively. The relatively narrow intervals indicate that the observed discrimination was stable across

the 250 bootstrap resamples of the test predictions. The precision of 0.998 at the default threshold further shows that the model generated very few unnecessary positive classifications under the controlled simulation conditions. The PR-AUC curve in **Figure 1** provides a complementary threshold-independent view of the model's ability to maintain precision while increasing failure detection. Its placement before the final synthesis emphasizes that predictive discrimination should be considered separately from the operational threshold later evaluated through cost-sensitive analysis. Predictive-maintenance research increasingly distinguishes model evaluation from maintenance-policy optimization because a high-quality probability estimate still requires an appropriate decision rule before it can support intervention planning (Wang, Li, & Zhao, 2024).

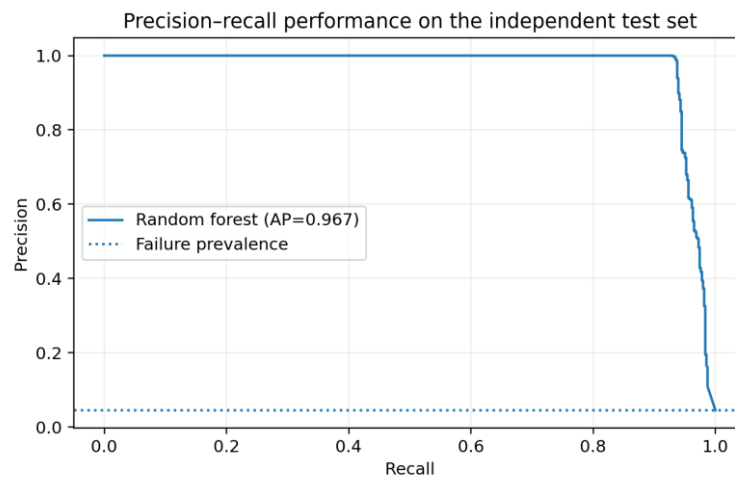


Figure 1. Precision–Recall Curve for the Selected Random-Forest Classifier on the Independent Test Set

The strong baseline performance establishes a quantitative foundation for examining whether the classifier remains useful when its probabilities are converted into cost-sensitive maintenance decisions. In particular, the combination of high PR-AUC, high MCC, and low Brier score indicates that random forest provides both class discrimination and probability estimates suitable for downstream policy analysis. The feature-ablation results further show that this performance is not solely attributable to algorithm selection but is supported by representations that encode thermal, mechanical, and wear-related relationships. Such integration of engineering knowledge with machine learning has become increasingly relevant in predictive-maintenance systems where interpretability and operational trust are required alongside predictive accuracy (Sharma, Mittal, & Soni, 2024). The next analysis therefore treats the random-forest probability output as a decision variable whose threshold can be varied according to the relative consequences of missed failures and preventive interventions. This distinction preserves the analytical separation between predictive performance and maintenance decision optimization, which is essential when evaluating machine-learning systems for industrial reliability.

Cost-Sensitive Maintenance Thresholds and Reliability Implications

The maintenance decision analysis demonstrates that the probability threshold should be treated as an operational parameter rather than a fixed statistical constant. When the relative cost of an undetected failure was set at 2:1, the validation procedure selected a threshold of 0.43, and the same threshold remained optimal at a 5:1 ratio. Increasing the failure-to-preventive-maintenance ratio to 10:1 reduced the selected threshold to 0.18, while the 20:1 scenario produced a threshold of 0.17. This downward movement reflects the increasing consequence assigned to missed failures, making earlier intervention acceptable despite a larger number of preventive actions. Cost-aware predictive-maintenance research similarly emphasizes that prognostic outputs should be translated into maintenance actions according to operational consequences rather than interpreted only through conventional classification metrics (Khazaelpour & Hashemkhani Zolfani, 2024). The resulting validation-selected thresholds and corresponding test-set outcomes are presented in Table 5.

Table 5. Validation-Selected Thresholds and Independent Test-Set Maintenance Outcomes

Failure:PM cost	Threshold	Interventions	Caught	Missed	Precision	Recall
2:1	0.43	514	509	35	0.990	0.936
5:1	0.43	514	509	35	0.990	0.936
10:1	0.18	567	511	33	0.901	0.939
20:1	0.17	573	512	32	0.894	0.941

The outcomes in Table 5 show that progressively lower thresholds increased the number of maintenance interventions while also increasing the number of detected failures. At the 2:1 and 5:1 ratios, 514 interventions captured 509 failures and left 35 failures undetected, with precision remaining at 0.990. At 10:1, the policy increased interventions to 567 and captured 511 of the 544 failures, reducing missed failures to 33 while precision declined to 0.901. The 20:1 scenario generated the largest intervention count of 573 and captured 512 failures, producing the lowest number of missed cases at 32 but also reducing precision to 0.894. This pattern illustrates the trade-off between preventive intervention frequency and failure avoidance that arises when maintenance decisions incorporate asymmetric failure and intervention costs (Wang, Li, & Zhao, 2024). The threshold movement is therefore a consequence of the decision objective rather than a modification of the random-forest model itself. Such separation between prognostic estimation and maintenance-policy selection is also consistent with dynamic predictive-maintenance frameworks that use model outputs as inputs to subsequent intervention decisions (Zhuang et al., 2023).

The cost-sensitive analysis becomes more consequential as the penalty associated with a missed failure increases. Under the 10:1 scenario, the threshold falls to 0.18 because the decision rule places greater emphasis on identifying potentially failing cycles before failure occurs. **Figure 2** illustrates the corresponding validation-cost relationship and shows why threshold selection should be based on the minimum of the specified cost function rather than on a default classification cutoff. The threshold was determined using the validation partition and was then transferred unchanged to the independent test partition, preserving the separation between model selection and final performance evaluation. This procedure is important because threshold optimization directly on the test data could introduce optimistic bias into the reported maintenance outcomes. Maintenance optimization studies that integrate predictive information with intervention scheduling similarly emphasize the need to align decisions with explicit cost and operational constraints (Wang, Li, & Zhao, 2024). The resulting framework consequently evaluates the classifier not only as a predictor of failure but also as a probabilistic input to a maintenance decision system.

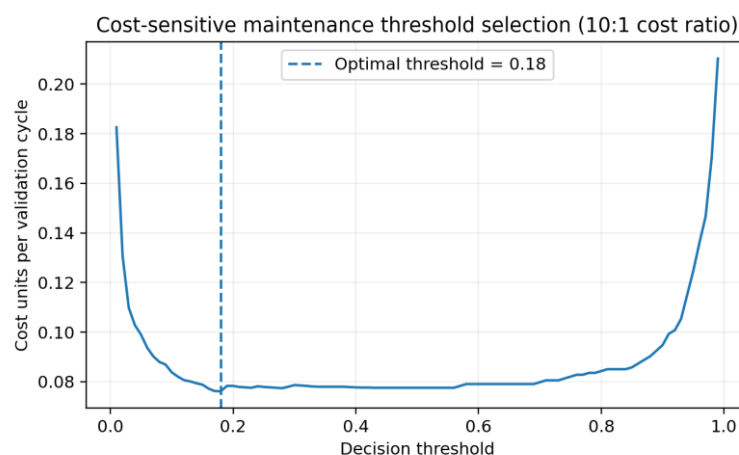


Figure 2. Validation Cost as a Function of the Random-Forest Decision Threshold for a 10:1 Failure-to-Preventive-Maintenance Cost Ratio.

At the 10:1 ratio, the independent test-set policy required 567 interventions, caught 511 of the 544 failures, and missed 33 cases. Under the normalized cost function, this produced a policy cost of

897 compared with 5,440 for reactive maintenance, corresponding to an 83.5% reduction. The residual failure proportion was 0.275%, yielding a reliability proxy of 99.725% under the defined cycle-level formulation. Table 6 compares this scenario with the other cost structures and demonstrates that increasingly high failure penalties produce larger normalized cost reductions. The 20:1 scenario generated the largest reduction at 88.9%, but it also required 573 interventions and produced a policy cost of 1,213, indicating that greater failure avoidance was accompanied by greater preventive-maintenance expenditure. Similar trade-offs have been identified in predictive-maintenance optimization, where maintenance timing and intervention decisions depend jointly on failure risk, maintenance costs, and available operational resources (Feng et al., 2023).

Table 6. Cost and Reliability Proxy Across Failure-Cost Scenarios

Failure:PM Cost	Policy Cost	Reactive Cost	Reduction (%)	Residual Failure (%)	Reliability Proxy (%)
2:1	584	1088	46.3	0.292	99.708
5:1	689	2720	74.7	0.292	99.708
10:1	897	5440	83.5	0.275	99.725
20:1	1213	10880	88.9	0.267	99.733

The values in Table 6 indicate that the normalized cost reduction increased from 46.3% at a 2:1 ratio to 88.9% at a 20:1 ratio, while the residual failure proportion declined from 0.292% to 0.267%. The relatively small change in residual failures compared with the larger change in intervention requirements suggests that increasingly aggressive thresholding produces diminishing changes in failure avoidance within this simulated environment. This relationship is important for maintenance planning because an increase in preventive actions does not automatically imply a proportionate improvement in reliability-related outcomes. Risk-aware predictive-maintenance approaches similarly emphasize balancing failure consequences against intervention burden when converting prognostic information into operational policies (Xu & Zhang, 2026). The 10:1 scenario illustrates this balance clearly, with 567 interventions, 33 missed failures, and an 83.5% normalized cost reduction relative to the reactive benchmark. The numerical results should nevertheless be interpreted as conditional on the specified cost ratios, simulated failure mechanisms, and assumption that an intervention can be executed when an alert is generated. In an industrial setting, maintenance duration, spare-parts availability, production schedules, and intervention capacity could alter the cost surface and consequently shift the selected threshold (Feng et al., 2023).

The reliability interpretation requires particular caution because the reported reliability proxy is defined as $R^* = 1 - FN/N$, where FN denotes missed failures and N denotes the evaluation population. The value of 99.725% under the 10:1 scenario therefore represents the proportion of simulated cycles without a missed-failure classification under the specified policy, rather than the probability that a machine will remain operational for a particular period. It should not be interpreted as mean time between failures, survival probability, or component lifetime reliability because the experiment does not model time-to-event dependence or progressive degradation trajectories. Remaining-useful-life research distinguishes classification-based failure detection from temporal prognostics because lifetime estimation requires information about degradation evolution and event timing (Ferreira & Gonçalves, 2022). This distinction is also relevant to maintenance approaches that integrate probabilistic remaining-useful-life estimates with maintenance scheduling and decision optimization (Lee & Mitici, 2023). The cycle-level reliability proxy remains useful within the present experiment because it provides a transparent quantitative connection between missed classifications and the simulated maintenance objective. Its translation into plant-level reliability indicators would require longitudinal equipment data, observed failure times, actual maintenance histories, and explicit modelling of intervention effects.

The cost-sensitive results establish that predictive-maintenance optimization involves a decision layer beyond the predictive model itself. The random-forest probabilities provide the basis for identifying potentially failing cycles, whereas the cost function determines how aggressively those probabilities should trigger maintenance actions. This distinction becomes particularly important when failure consequences vary across equipment types, production stages, or operational environments,

because the same classifier can produce different economically justified thresholds under different cost structures. Contemporary predictive-maintenance research increasingly combines machine-learning prognostics with decision frameworks to account for uncertainty, maintenance timing, and system-level consequences (Zeng & Liang, 2023). The present findings extend that logic by demonstrating quantitatively that lower thresholds reduce missed failures but increase intervention frequency, while the magnitude of the trade-off depends on the assumed failure-to-maintenance cost ratio. The results therefore provide a decision-sensitive interpretation of predictive performance, while the reliability proxy remains explicitly limited to the cycle-level simulation framework used in this study.

Robustness, Feature Importance, and Deployment Boundaries

The robustness analysis indicates that the selected random-forest model retained strong discrimination under modest perturbations of the sensor inputs but progressively deteriorated as measurement noise increased. With no additional noise, the model achieved a ROC-AUC of 0.989 and PR-AUC of 0.967, while the corresponding values at 1% noise remained almost unchanged at 0.989 and 0.960. At 2% noise, ROC-AUC decreased slightly to 0.987 and PR-AUC to 0.941, accompanied by a decline in recall from 0.939 to 0.910. The deterioration became more pronounced at 5% noise, where PR-AUC fell to 0.894 and recall to 0.860, while 10% noise reduced these metrics further to 0.796 and 0.805. The pattern indicates that the classifier can tolerate limited measurement perturbation but becomes increasingly sensitive when uncertainty obscures the relationships separating normal and failure states. Robustness to imperfect sensor information is particularly relevant for industrial predictive-maintenance systems because practical deployments commonly encounter measurement variability, sensor degradation, and heterogeneous data quality (Elkateb et al., 2024).

Table 7. Predictive Performance Under Increasing Sensor-Measurement Noise

Noise (% SD)	ROC-AUC	PR-AUC	Precision	Recall	Missed	False Alarms
0	0.989	0.967	0.901	0.939	33	56
1	0.989	0.960	0.892	0.930	38	61
2	0.987	0.941	0.881	0.910	49	67
5	0.982	0.894	0.814	0.860	76	107
10	0.960	0.796	0.722	0.805	106	169

The values in Table 7 reveal that degradation affects both sides of the maintenance decision, increasing missed failures as well as false alarms. At 1% noise, missed failures increased from 33 to 38 and false alarms from 56 to 61, while at 5% noise these values increased to 76 and 107, respectively. At 10% noise, the model missed 106 failures and generated 169 false alarms, demonstrating that severe measurement perturbation can simultaneously weaken failure detection and increase unnecessary interventions. This behavior is consistent with the role of sensor quality in data-driven condition monitoring, where noisy or unreliable measurements can propagate uncertainty into downstream predictive decisions (Fordal et al., 2023). The deterioration also indicates that the strong baseline metrics should not be interpreted as evidence of unconditional robustness across all data-quality conditions. Practical deployment would therefore require sensor-quality monitoring, calibration procedures, and explicit validation under the measurement uncertainty expected at the target facility.

The permutation analysis provides additional insight into which variables sustain the model's predictive structure. Temperature difference produced the largest mean decrease in PR-AUC at 0.2770, followed by rotational speed at 0.2229, mechanical power at 0.1036, and the wear-torque index at 0.0801. The relatively small decreases associated with tool wear, torque, type L, and process temperature indicate that their incremental contribution was limited after the dominant physical relationships had already been represented. These results are consistent with the mechanism-oriented feature design, because temperature difference directly represents the thermal relationship associated with heat-dissipation behavior, while speed and mechanical power capture important aspects of the operating envelope. Explainable predictive-maintenance research emphasizes that feature importance can help connect algorithmic predictions with equipment conditions and provide a basis for engineering interpretation (Dereci & Tuzkaya, 2024). The ranking therefore provides evidence that the model's

predictive behavior is concentrated in physically interpretable variables rather than being distributed uniformly across the available feature space.

Table 8. Permutation Feature Importance of the Selected Random-Forest Model

Feature	Mean Decrease in PR-AUC	SD
Temperature difference K	0.2770	0.0161
Rotational speed rpm	0.2229	0.0286
Mechanical power kW	0.1036	0.0034
Wear torque index	0.0801	0.0042
Tool wear min	0.0172	0.0013
Torque Nm	0.0065	0.0013
Type L	0.0016	0.0007
Process temperature K	0.0015	0.0020

The importance structure in Table 8 is further visualized in Figure 3, where the magnitude of the mean decrease in PR-AUC differentiates the dominant predictors from variables with limited incremental contribution. Temperature difference and rotational speed form the most influential pair, followed by mechanical power and the wear–torque index. The comparatively small importance values for tool wear, torque, type L, and process temperature indicate that much of their predictive information is either weaker individually or represented through the derived variables. This pattern strengthens the interpretation that mechanism-based feature construction can concentrate relevant information into variables that are more directly interpretable from an engineering perspective. Explainable artificial intelligence research in predictive maintenance similarly emphasizes the value of identifying influential variables to support model transparency and facilitate communication between data-driven systems and maintenance personnel (Matzka, 2020). The visualization therefore complements the numerical permutation results without implying that feature importance represents causal influence.

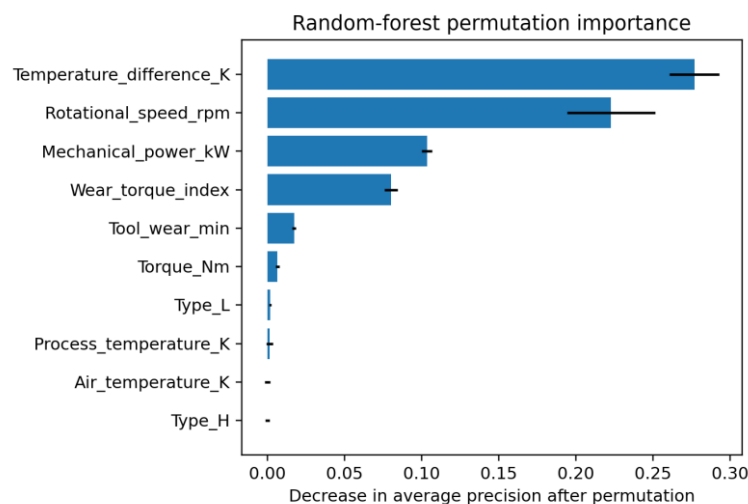


Figure 3. Permutation Importance for the Selected Random-Forest Model; Larger Values Indicate a Larger Decrease in Average Precision After Feature Permutation

The robustness findings also define an important boundary between controlled predictive performance and industrial deployment. The simulation uses mechanism-based failure labels with explicit physical relationships, whereas real equipment operates under changing operators, maintenance histories, environmental conditions, production regimes, and sensor configurations that may not be represented in the training distribution. Research on predictive maintenance identifies domain shift, data availability, irregular operating conditions, and transferability as persistent barriers to moving models from controlled datasets into production environments (Meitz et al., 2025). The present experiment also assumes that every positive alert can trigger an available maintenance intervention and

that the intervention prevents the corresponding failure, conditions that may not hold when maintenance resources, production schedules, or spare parts are constrained. Digital-twin-based predictive-maintenance research highlights the importance of integrating predictive models with continuously evolving equipment states and maintenance processes rather than treating the prediction algorithm as an isolated component (van Dinter, Tekinerdogan, & Catal, 2022). These constraints mean that the reported performance should be interpreted as evidence of methodological feasibility within the defined simulation rather than as a direct estimate of performance at an arbitrary industrial site.

A further limitation concerns the temporal interpretation of reliability and model performance. The experiment evaluates individual cycles and therefore does not represent serial dependence, degradation trajectories, censoring, or the time between successive failures, all of which influence real maintenance decisions. Data-driven prognostic research emphasizes that predictive maintenance increasingly requires temporal degradation modelling and remaining-useful-life estimation when decisions depend on when a failure is expected to occur rather than simply whether a failure is possible (Wen et al., 2022). Bayesian prognostic and risk-aware reinforcement-learning approaches further demonstrate how uncertainty and sequential decision-making can be incorporated when maintenance actions influence subsequent system states (Zhuang et al., 2023; Xu & Zhang, 2026). For practical deployment, the present classifier would therefore require longitudinal validation using site-specific failure histories, probability recalibration, sensor-drift monitoring, and evaluation under realistic maintenance-resource constraints. The noise experiment additionally indicates that data-quality monitoring should form part of the deployment specification because severe measurement perturbation substantially reduces both discrimination and classification precision. The combined robustness, feature-importance, and deployment evidence supports treating the machine-learning model as one component of a broader reliability system that requires continuous validation against physical equipment behavior and operational conditions.

CONCLUSION

The cost-sensitive predictive-maintenance workflow demonstrated that machine learning can provide strong failure discrimination when combined with engineering-informed feature construction and decision-oriented threshold optimization. Across 80,000 reproducible industrial-equipment cycles calibrated to the AI4I process structure, random forest achieved the strongest positive-class performance, with a test PR-AUC of 0.967, F1 of 0.965, and MCC of 0.964. The addition of physics-derived thermal, power, and wear-load variables increased PR-AUC from 0.943 to 0.966 and reduced the Brier score from 0.0088 to 0.0035 without requiring additional sensors. Feature-importance analysis further identified temperature difference, rotational speed, mechanical power, and the wear-torque index as the dominant predictors, providing an interpretable connection between model outputs and the simulated failure mechanisms. Maintenance thresholds varied according to the assumed failure-to-preventive-maintenance cost ratio, ranging from 0.43 under lower failure penalties to 0.17–0.18 when missed failures were substantially more costly. Across the evaluated scenarios, the resulting policies reduced normalized cost by 46.3–88.9% relative to reactive maintenance and reduced residual test failures to approximately 0.27–0.29% of cycles. These findings position predictive maintenance as a decision-oriented framework in which model probabilities, engineering representations, and maintenance consequences must be evaluated together rather than through classification performance alone.

The robustness analysis further showed that the workflow retained substantial predictive capability under modest sensor perturbations but experienced marked deterioration under severe measurement noise, with PR-AUC declining to 0.796 and recall to 0.805 at 10% standard-deviation noise. The results also demonstrate that the reported cost reductions and reliability proxy should not be interpreted as direct estimates of factory-level savings or lifetime reliability. Because the experiment uses a synthetic public benchmark with transparent and partially deterministic failure mechanisms, its predictive performance represents evidence under controlled conditions rather than a direct representation of all industrial operating environments. The cycle-level reliability proxy, defined from missed-failure classifications, likewise does not substitute for time-to-failure, survival, or remaining-useful-life analysis. Practical implementation should therefore reproduce the workflow using temporally separated plant data and incorporate site-specific failure consequences, preventive-maintenance costs, maintenance capacity, intervention effectiveness, and sensor-quality characteristics.

Continuous probability recalibration and drift monitoring should accompany deployment so that changes in equipment conditions and data distributions can be detected before they substantially affect maintenance decisions. Under these conditions, machine learning can serve as an integrated component of industrial reliability management by connecting failure prediction with engineering knowledge, cost-sensitive intervention policies, and operational validation.

REFERENCES

- Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in Industry 4.0: Overview, models, and challenges. *Applied Sciences*, 12(16), 8081. <https://doi.org/10.3390/app12168081>
- Alves da Silveira, N. N., Meghoo, A. A., & Tinga, T. (2024). Quantifying the suitability and feasibility of predictive maintenance approaches. *Computers & Industrial Engineering*, 194, 110342. <https://doi.org/10.1016/j.cie.2024.110342>
- Autran, J.-V., Kuhn, V., Diguët, J.-P., Dubois, M., & Buche, C. (2024). AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities. *Procedia Computer Science*, 246, 1201–1209. <https://doi.org/10.1016/j.procs.2024.09.546>
- Chen, C., Fu, H., Zheng, Y., Tao, F., & Liu, Y. (2023). The advance of digital twin for predictive maintenance: The role and function of machine learning. *Journal of Manufacturing Systems*, 71, 581–594. <https://doi.org/10.1016/j.jmsy.2023.10.010>
- Dereci, U., & Tuzkaya, G. (2024). An explainable artificial intelligence model for predictive maintenance and spare parts optimization. *Supply Chain Analytics*, 8, 100078. <https://doi.org/10.1016/j.sca.2024.100078>
- Drakaki, M., Karnavas, Y. L., Tzifafettas, I. A., Linardos, V., & Tzionas, P. (2022). Machine learning and deep learning based methods toward Industry 4.0 predictive maintenance in induction motors: State of the art survey. *Journal of Industrial Engineering and Management*, 15(1). <https://doi.org/10.3926/jiem.3597>
- Elkateb, S., Métwalli, A., Shendy, A., & Abu-Elanien, A. E. B. (2024). Machine learning and IoT-based predictive maintenance approach for industrial applications. *Alexandria Engineering Journal*, 88, 298–309. <https://doi.org/10.1016/j.aej.2023.12.065>
- Fernandes, M., Corchado, J. M., & Marreiros, G. (2022). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: A systematic literature review. *Applied Intelligence*, 52, 14246–14280. <https://doi.org/10.1007/s10489-022-03344-3>
- Ferreira, C., & Gonçalves, G. (2022). Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems*, 63, 550–562. <https://doi.org/10.1016/j.jmsy.2022.05.010>
- Feng, Q., Zhang, Y., Sun, B., Guo, X., Fan, D., Ren, Y., Song, Y., & Wang, Z. (2023). Multi-level predictive maintenance of smart manufacturing systems driven by digital twin: A matheuristics approach. *Journal of Manufacturing Systems*, 68, 443–454. <https://doi.org/10.1016/j.jmsy.2023.05.004>
- Fordal, J. M., Schjølberg, P., Helgetun, H., Skjermo, T. Ø., Wang, Y., & Wang, C. (2023). Application of sensor data based predictive maintenance and artificial neural networks to enable Industry 4.0. *Advances in Manufacturing*, 11, 248–263. <https://doi.org/10.1007/s40436-022-00433-x>
- Jaenal, A., Ruiz-Sarmiento, J. R., & González-Jiménez, J. (2024). MachNet, a general deep learning architecture for predictive maintenance within the Industry 4.0 paradigm. *Engineering Applications of Artificial Intelligence*, 127, 107365. <https://doi.org/10.1016/j.engappai.2023.107365>
- Khazaelpour, P., & Hashemkhani Zolfani, S. (2024). FUCOM-optimization based predictive maintenance strategy using expert elicitation and artificial neural network. *Expert Systems with Applications*, 238, 121322. <https://doi.org/10.1016/j.eswa.2023.121322>
- Lee, J., & Mitici, M. (2023). Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics. *Reliability Engineering & System Safety*, 230, 108908. <https://doi.org/10.1016/j.ress.2022.108908>

- Li, X., Williams, J., Swanson, C., & Berg, T. (2025). A machine learning approach to predictive maintenance: Remaining useful life and motor fault analysis. *Computers & Industrial Engineering*, 206, 111222. <https://doi.org/10.1016/j.cie.2025.111222>
- Li, Y., Chen, Y., Hu, Z., & Zhang, H. (2023). Remaining useful life prediction of aero-engine enabled by fusing knowledge and deep learning models. *Reliability Engineering & System Safety*, 229, 108869. <https://doi.org/10.1016/j.ress.2022.108869>
- Li, Z., Jana, C., Pamucar, D., & Pedrycz, W. (2025). A comprehensive assessment of machine learning models for predictive maintenance using a decision-making framework in the industrial sector. *Alexandria Engineering Journal*, 120, 561–583. <https://doi.org/10.1016/j.aej.2025.02.010>
- Marti-Puig, P., Amar Touhami, I., Colomer Perarnau, R., & Serra-Serra, M. (2024). Industrial AI in condition-based maintenance: A case study in wooden piece manufacturing. *Computers & Industrial Engineering*, 188, 109907. <https://doi.org/10.1016/j.cie.2024.109907>
- Matzka, S. (2020). Explainable artificial intelligence for predictive maintenance applications. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)* (pp. 69–74). IEEE. <https://doi.org/10.1109/AI4I49448.2020.00023>
- Meitz, L., Senge, J., Wagenhals, T., Schöler, T., Hähner, J., Edinger, J., & Krupitzer, C. (2025). A literature review framework and open research challenges for predictive maintenance in Industry 4.0. *Computers & Industrial Engineering*, 206, 111193. <https://doi.org/10.1016/j.cie.2025.111193>
- Obeten, O., & Jimoh, A. (2026). Physics-informed, cost-aware fault classification with comparative explainability for predictive maintenance: A SHAP–LIME agreement analysis. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-64374-2>
- Sharma, J., Mittal, M. L., & Soni, G. (2024). Condition-based maintenance using machine learning and role of interpretability: A review. *International Journal of System Assurance Engineering and Management*, 15(4), 1345–1360. <https://doi.org/10.1007/s13198-022-01843-7>
- Taşcı, B., Omar, A., & Ayvaz, S. (2023). Remaining useful lifetime prediction for predictive maintenance in manufacturing. *Computers & Industrial Engineering*, 184, 109566. <https://doi.org/10.1016/j.cie.2023.109566>
- Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: A systematic review. *Journal of Intelligent Manufacturing*, 33, 1879–1905. <https://doi.org/10.1007/s10845-022-01963-8>
- Ucar, A., Karakose, M., & Kırımca, N. (2024). Artificial intelligence for predictive maintenance applications: Key components, trustworthiness, and future trends. *Applied Sciences*, 14(2), 898. <https://doi.org/10.3390/app14020898>
- UCI Machine Learning Repository. (2020). *AI4I 2020 predictive maintenance dataset* [Data set]. <https://doi.org/10.24432/C5HS5C>
- van Dinter, R., Tekinerdogan, B., & Catal, C. (2022). Predictive maintenance using digital twins: A systematic literature review. *Information and Software Technology*, 151, 107008. <https://doi.org/10.1016/j.infsof.2022.107008>
- Wang, L., Li, B., & Zhao, X. (2024). Multi-objective predictive maintenance scheduling models integrating remaining useful life prediction and maintenance decisions. *Computers & Industrial Engineering*, 197, 110581. <https://doi.org/10.1016/j.cie.2024.110581>
- Wang, L., Zhu, Z., & Zhao, X. (2024). Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method. *Reliability Engineering & System Safety*, 245, 110012. <https://doi.org/10.1016/j.ress.2024.110012>
- Ward, T., Jenab, K., Ortega-Moody, J., & Staub, S. (2024). A comprehensive review of machine learning techniques for condition-based maintenance. *International Journal of Prognostics and Health Management*, 15(2). <https://doi.org/10.36001/ijphm.2024.v15i2.3850>
- Wen, Y., Rahman, M. F., Xu, H., & Tseng, T.-L. B. (2022). Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective. *Measurement*, 187, 110276. <https://doi.org/10.1016/j.measurement.2021.110276>
- Xu, Z., & Zhang, Q. (2026). Predictive maintenance optimization for industrial equipment via reliable prognosis and risk-aware reinforcement learning. *Complex & Intelligent Systems*, 12, 11. <https://doi.org/10.1007/s40747-025-02127-w>

- Zeng, J., & Liang, Z. (2023). A dynamic predictive maintenance approach using probabilistic deep learning for a fleet of multi-component systems. *Reliability Engineering & System Safety*, 238, 109456. <https://doi.org/10.1016/j.ress.2023.109456>
- Zhuang, L., Xu, A., & Wang, X.-L. (2023). A prognostic driven predictive maintenance framework based on Bayesian deep learning. *Reliability Engineering & System Safety*, 234, 109181. <https://doi.org/10.1016/j.ress.2023.109181>