

A Calibrated Machine Learning Framework for Predictive Maintenance in Industry 4.0: Balancing Reliability and Economic Optimization

Md. Efatuazzaman Efat

Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
efatuazzaman@gmail.com

Sayed Md. Fazle Rabbi

Department of Electrical and Electronic Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
sayedfazlerabbi45@gmail.com

Md. Saiful Islam

Professor
Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
mdsaifulislam@iict.buet.ac.bd

Abstract

The integration of predictive maintenance (PdM) into industrial systems is a crucial aspect of achieving the operational goals of Industry 4.0. In the present study, we propose the development of a calibrated and interpretable machine learning model to issue multi-label predictions of key machine failure modes, with a strong emphasis on economic feasibility and reliability. We apply a forward-moving pipeline using the AI4I 2020 dataset, which includes grid search selection, cost-sensitive thresholding, and probabilistic calibration using both Brier and Expected Calibration Error (ECE) scores. We also quantify uncertainty through conformal prediction. In addition to classical classification measures, our scheme evaluates models based on life-cycle cost (LCC), payback period, and net present value (NPV), modeling the consequences of decisions based on realistic operational policies (e.g., preventive maintenance, inspection, or deferral). By comparing and contrasting a variety of algorithms, we demonstrate that prediction fidelity and economic payoff vary significantly. Notably, our methodology ensures interpretability and robustness by combining sensor noise stress testing with machine-type subgroup analysis. The results show not only higher effectiveness in terms of failure predictions, but also higher cost savings in the long term than reactive policies, making the strategic relevance of machine learning-based PdM in modern manufacturing plants.

Keywords

Predictive Maintenance, Industry 4.0, Cost-Optimized Machine Learning, Maintenance economics and Reliability engineering.

1. Introduction

Predictive maintenance (PdM) takes advantage of operational and sensor data and predicts the risk of failure early enough to act on it before it occurs. Compared to reactive maintenance, PdM creates less exposure to unplanned maintenance and the ensuing exposure to production and safety risks—expenses that can prevail in high-throughput industrial environments. High-frequency sensing and decision-making, with the help of cyber-physical systems and Industrial Internet of Things (IIoT) infrastructures, provide the chance to transform raw signals into actionable maintenance policies in Industry 4.0 (Zheng et al. 2018; Liao et al. 2017). However, predictive accuracy alone is not enough to provide economic benefit, as maintenance choices do not have symmetric costs (unplanned failures, preventive measures, and false alarms); this requires precise probability estimates and cost-conscious decision rules. Digital twin principles also promote such integration, given the connection between physical equipment and analytical models that support economically informed action across the asset lifecycle (Kritzinger et al. 2018; Tao et al. 2019).

1.1 Objectives

- (i) Train a multi-label PdM model to model five failure modes using sensor features and to obtain total machine-failure risk.
- (ii) Develop decision logic that is cost-sensitive in design to covert predicted probabilities into the do nothing / inspect / preventive maintenance actions at different asymmetric costs.
- (iii) Evaluate predictive performance using ROC curves/ROC-AUC, test F1-score, and per-mode PR-AUC (average precision) under class imbalance.
- (iv) Test probability reliability based on the model of economic preference of the Brier score, ECE and reliability diagnostics.
- (v) Assess economic impact based on annual cost, five-year discounted LCC/NPV savings and payback, and strength through subgroup analysis and response to cost parameters.

2. Literature Review

Predictive maintenance (PdM) has transformed traditional condition monitoring into data-oriented analytical solutions built upon existing sensing and industrial data infrastructures. Initial research focused primarily on controlling machining processes and tool life to observe deviant behavior and avoid unintentional downtime (Byrne et al. 1995; Dimla 2000; Teti et al. 2010). Subsequent efforts aimed toward standardizing condition-based maintenance strategies, which use monitoring information to anticipate failures and take preemptive actions (Jardine et al. 2006). These developments led to the predictive maintenance paradigm in contemporary manufacturing systems. Industry 4.0 has accelerated the rate at which PdM is being integrated with cyber-physical systems, Industrial Internet of Things (IIoT) infrastructures, and digital twins. In this regard, sensor data are now being collected on a grand scale and are constantly processed to assist with maintenance decisions (Zheng et al. 2018; Liao et al. 2017). Digital twin architectures extend this functionality by connecting real equipment with virtual analytical models used to monitor the well-being of a system over the lifecycle of industrial equipment (Kritzinger et al. 2018; Tao et al. 2019). Consequently, PdM has become a major concept for achieving higher reliability, operational efficiency, and cost management in intelligent manufacturing settings. Recent works have focused on the growing application of machine learning in predictive maintenance. Industrial PdM surveys demonstrate that supervised learning methods dominate and remain the preferred approach due to the existence of labeled operational information and the possibility of predicting failure issues (Theissler et al. 2021). Meanwhile, deep learning architectures have been well-studied for extracting patterns from complex sensor signals and time-series data. The application of architectures such as convolutional neural networks, recurrent neural networks, and autoencoder models has also been noted as applicable to PdM (Serradilla et al. 2022; Li et al. 2024). Despite the potential of these methods to improve predictive performance, their practical application can be hindered by issues such as interpretability, computational cost, and model selection. Another area of research is the link between predictive models and maintenance decision policies. Reinforcement learning techniques have been explored as a way of learning optimal maintenance actions in the face of uncertainty and cost limitations (Siraskar et al. 2023). Even though these approaches provide a conceptually useful framework for sequential decision processes, they often require the entire modeling of system states and reward architectures, which can limit their practical use in many industrial environments. Therefore, more basic decision mechanisms that can operationalize probabilistic predictions remain a research concern.

Economic analysis has become a recognized and important aspect of predictive maintenance systems. Most studies report enhancements in classification accuracy or ROC-AUC but do not quantify the impact on the cost of predictive decisions. Recent research highlights that cost-benefit factors and profitability of investment are important for the implementation of PdM technologies (Meng et al. 2022). Equally, economic decision models have shown that the algorithm with the most accurate predictive performance does not always provide the most operational value when taking into account, for example, maintenance expenses and the consequences of errors (Fabri et al. 2025). Systematic reviews on predictive maintenance adoption also show that the absence of standardized economic assessment models is one of the biggest barriers to industrial adoption (Benhanifia et al. 2025).

Another problem in PdM study is predictive probability reliability and uncertainty quantification. Poorly calibrated probability estimates are common in machine learning models and can lead to poor maintenance decisions if the prediction is used directly as an operational plan. Calibration techniques are used in probabilistic forecasts to make projections more accurate (Guo et al. 2017). Conformal prediction has also recently become an interesting technique for uncertainty quantification in industrial-based fault detection systems, as it provides statistically valid prediction sets and false alarm control (Diallo et al. 2025). Similarly, uncertainty intervals have been estimated based on similar methods for remaining useful life (RUL) prediction in IoT-enabled predictive maintenance systems (Moccardi et al. 2025). These developments represent a growing interest in the integration of uncertainty-conscious prediction mechanisms into maintenance analytics.

Data availability and realism are further issues in PdM research. To benchmark machine learning models, public datasets like the AI4I 2020 dataset are frequently used. Recent research, however, suggests that these datasets might not model the irregularities and noise present in real-world industrial setups. In particular, Autran et al. (2024) presented a modified AI4I dataset with irregular sampling patterns and missing values, showing that model performance can be affected dramatically in more realistic settings (Autran et al. 2024). This points to the need for testing predictive maintenance models outside the benchmark context of ideal conditions.

More recent works have also addressed methods dedicated to multi-failure-mode prediction. For example, the model proposed by Ghasemkhani et al. (2025) is a balanced framework of Hoeffding tree forests covering multi-label characteristics and the high skewness of predictive maintenance data (Ghasemkhani et al. 2025). Their strategy achieved good classification results on the AI4I 2020 dataset, demonstrating the usefulness of specialized algorithms in multi-failure-mode diagnostics. However, the vast majority of these studies focus on predictive accuracy and show less interest in probability calibration, uncertainty-sensitive decision processes, and their economic implications. Generally speaking, uncertainty-conscious predictive maintenance is an increasingly important research path. Recent attempts involving hybrid learning frameworks that rely on uncertainty quantification algorithms—including Monte Carlo dropout and conformal prediction—have been more robust for remaining useful life estimation (Yang et al. 2026). These advances confirm the growing importance of predictive maintenance systems that are not only accurate but also provide credible estimates of confidence for making operational decisions.

Research Gap and Contribution: Despite remarkable progress in the field of predictive maintenance, three major limitations remain in the literature. First, most studies are based on classification accuracy measures and do not study the calibration quality of predicted probabilities. Second, uncertainty regarding model predictions is rarely incorporated into operational decision policies to trigger maintenance actions. Third, predictive performance is not often associated directly with financial performance measures such as Net Present Value (NPV) or life-cycle cost savings.

3. Methodology

3.1 Dataset Description (AI4I 2020)

The UCI Machine Learning Repository is used in this research, where the AI4I 2020 Predictive Maintenance Dataset (10,000 instances) has been used. Based on the studies of Marco Matzka of Hochschule Furtwangen University (HFU) (Table 1).

Table 1. Statistical Overview of Dataset

Feature	Unit	Mean	Std	Min–Max
Air temperature	K	300.00	2.00	295.3–304.5
Process temperature	K	310.01	1.48	305.7–313.8
Rotational speed	rpm	1538.78	179.28	1168–2886
Torque	Nm	39.99	9.97	3.8–76.6
Tool wear	min	107.95	63.65	0–253

Germany, this dataset is a realistic synthetic proxy applied to an industrial PdM context and overcomes the lack of publicly available real-world data. It incorporates six operational and sensory characteristics, including Type (categorical), air temperature, process temperature, rotational speed, torque, and tool wear. It also includes a single generic failure mode, machine failure, as well as five specific failure modes, such as Excessive tool wear (TWF), Overheating-related breakdown (HDF), Energy supply interruption (PWF), Structural strain failure (OSF), and Unpredictable/Random error (RNF). Matzka (2020) presented the data and released it together with a predictive maintenance workflow framework that can be explained (Matzka 2020).

3.2 Dataset preprocessing and model training pipeline

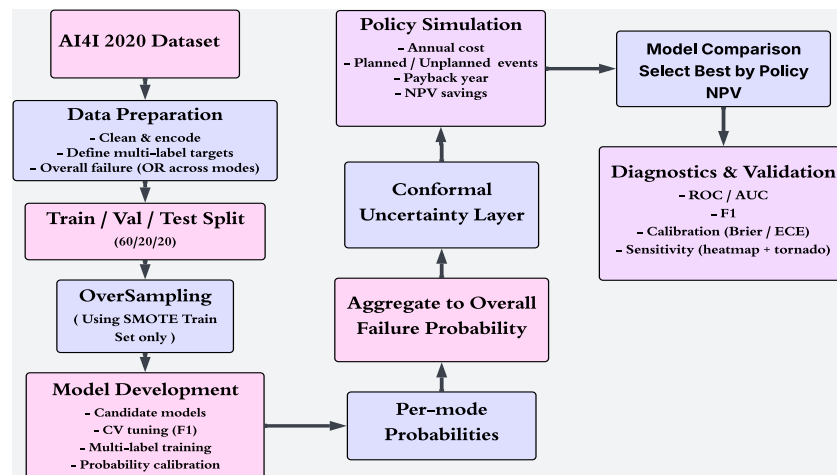


Figure 1. A Flowchart Representing End to End Pipeline of Predictive Maintenance Framework

The end to end pipeline of the predictive maintenance framework is shown in Fig. 1. Preprocessing starts by removing non-informative identifiers (UDI and Product ID) to avoid spurious correlations. The "Type" feature is a categorical feature which is coded numerically (L, M, H as 0, 1, 2) as required to run the algorithm. To prevent data leakage during the training of the model, both the overall machine failure label and all individual failure mode labels are excluded from the feature matrix. A 60/20/20 data split is used to separate the data into a training dataset, validation dataset, and test dataset, with stratification by overall machine failure so that the proportion of classes remains consistent between data subsets. Due to the low instances of failure (approximately 3.39 percent), the Synthetic Minority Over-sampling Technique (SMOTE) is used to address the problem of class imbalance (Elreedy and Atiya 2019); this is implemented strictly in the training pipeline (Chawla et al. 2002). All candidate models are combined into a pipeline involving feature scaling (StandardScaler), resampling (SMOTE), and training the classifier. Decision trees, random forests, gradient boosting, logistic regression, support vector machines, XGBoost, and LightGBM are evaluated. Five-fold stratified cross-validation is performed on the training data, and hyperparameters are selected using the F1 score. The optimum configuration is then fixed and expanded to a multi-output system consisting of a single calibrated classifier for each type of failure. Sigmoid calibration is used for calibrating probabilities (Guo et al. 2017). Finally, the cost-sensitive thresholds and the conformal uncertainty are calculated on validation data prior to the final test assessment.

3.3 Evaluation Framework Overview

Model performance is assessed from three complementary perspectives: i) predictive accuracy, ii) probabilistic calibration, and iii) economic impact.

First, predictive performance is evaluated in terms of per-mode Precision-Recall AUC (average precision) and overall machine-failure statistics such as the F1-score, ROC curves, and ROC AUC (Fawcett 2006; Davis and Goadrich 2006; Saito and Rehmsmeier 2015). These metrics are necessary measures of discriminatory power in the presence of class imbalance.

Second, calibration quality is assessed using the Brier score and Expected Calibration Error (ECE), along with the reliability diagram of the economically optimal model (Brier 1950; Guo et al. 2017). This is necessary because maintenance decisions are based on probability estimates and not simply binary outputs.

Third, as a measure of economic performance, the annual reactive versus predictive maintenance cost, five-year discounted life-cycle costs, estimated Net Present Value (NPV) savings, and the payback period are included. Both costing based on the confusion matrix and decision policy costing (including inspection and abstention) are adopted, which is in accordance with the requirements of cost-sensitive learning (Elkan 2001).

4. Results

4.1 Class imbalance and resampling

Overall machine failure is highly imbalanced. In the 6,000-sample training set, we observe 5,797 non-failure samples versus 203 failure samples (3.38%). After SMOTE (Synthetic Minority Over-sampling Technique), the training distribution becomes balanced at 5,797 vs. 5,797 (Figure 2). This step is consistent with using synthetic oversampling to increase the minority-class learning signal (Chawla et al. 2002).

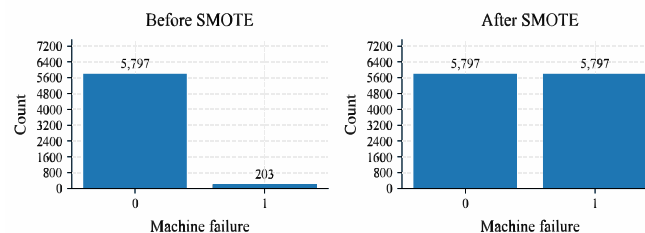


Figure 2. Class distribution before/after SMOTE (train set)

4.2 Predictive performance (overall machine failure)

Across all models, the ROC AUC values are high (0.873-0.983) which suggests that the model was able to rank strongly on the test set (Figure 3). The best ROC AUC is obtained with SVM (AUC=0.983) and Logistic Regression (AUC=0.980) while with Gradient Boosting the AUC is 0.956 (Figure 3). This illustrates an important point from evaluation theory, which is that high AUC does not necessarily mean that the fixed threshold F1 or downstream decision value will be the best (Fawcett 2006). F1-scores (test) are in the range of 0.611 to 0.794 except Decision Tree having score 0.716 (Figure 4). It is interesting to note that Gradient Boosting performs the best in terms of F1 (approx0.79), which indicates a balance in precision-recall according to the learned thresholds on the validation dataset.

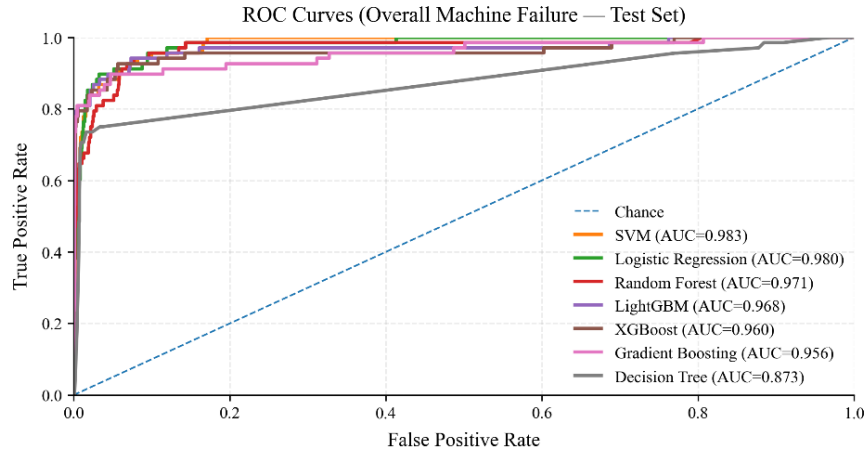


Figure 3. ROC curves (overall machine failure, test set)

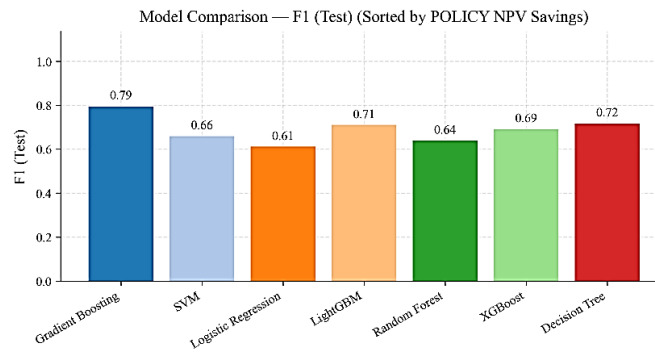


Figure 4. Model comparison by F1 (test)

4.3 Performance comparison between models

Table 2. Performance comparison across all models

Model	F1 (Test overall)	TP	FP	FN	TN	Brier (Overall)	ECE (Overall)	Annual Reactive Cost (CM) [\$]	Annual PdM Cost (CM) [\$]	NPV Savings (CM) [\$]	NPV Savings (Policy) [\$]
Decision Tree	0.716418	48	18	20	1,914	0.017506	0.005014	1,700,000	668,000	3,892,092	-943,057
Random Forest	0.639053	54	47	14	1,885	0.016130	0.005463	1,700,000	605,500	4,129,016	3,353,800
Gradient Boosting	0.794118	54	14	14	1,918	0.009797	0.008112	1,700,000	523,000	4,441,756	4,133,755
Logistic Regression	0.611399	59	66	9	1,866	0.015184	0.008815	1,700,000	540,500	4,375,417	4,043,723
SVM	0.658537	54	42	14	1,890	0.013458	0.006327	1,700,000	593,000	4,176,401	4,060,782
XGBoost	0.690476	58	42	10	1,890	0.011012	0.002597	1,700,000	503,000	4,517,572	2,915,964
LightGBM	0.709677	55	32	13	1,900	0.010353	0.006157	1,700,000	545,500	4,356,463	3,562,293

4.4 Multi-label failure-mode performance (PR-AUC)

From Table 2, there are large differences in the prevalence of failure modes (e.g., RNF is extremely rare), and PR-AUC (Precision-Recall Area Under the Curve) is used to ensure quality sensitivity (Davis and Goadrich 2006; Saito and Rehmsmeier 2015). We consistently find high PR-AUC for OSF (around 0.87–0.97) and good PR-AUC for HDF/PWF for most ensemble models. On the other hand, RNF PR-AUC hovers around the base-rate range, which is expected for extremely rare labels. This supports our design decision to treat the problem as multi-label and report per-mode metrics instead of simply an "all labels" metric.

4.5 Calibration results (Brier, ECE)

From Table 2, among all models, the lowest Brier score is obtained with Gradient Boosting (0.009797), which indicates high probability accuracy averaged over outcomes. The lowest ECE is obtained with XGBoost (0.002597), indicating that its average predicted confidence across frequency bins closely matches the relative frequency distribution (Figure 5). Calibration is important because we use probabilities (versus labels) to set the economic policy, which chooses between "do nothing," "PM," and "inspect" (Brier 1950; Guo et al. 2017).

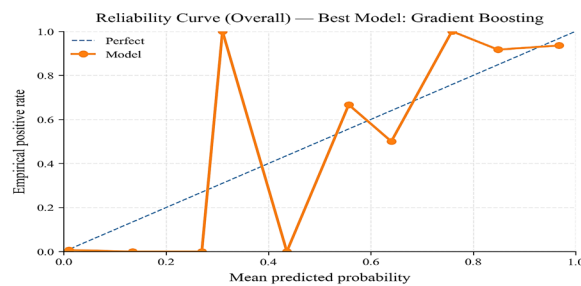


Figure 5. Reliability curve for the best policy-value model (Gradient Boosting)

4.6 Economic results: annual costing, NPV and payback

We model a five-year horizon with a 10% discount rate, an annual volume of 10,000 units, and the following unit costs: unplanned failure = \$5,000; false alarm = \$500; planned maintenance = \$500; annual operations = \$3,000; capital expenditure = \$20,000; inspection = \$200 with 70% conversion efficiency; and a conformal alpha = 0.10 (90% coverage). Under these assumptions, the reactive cost is \$1.70 million per year (340 failures per year $\times$ \$5,000). In confusion-matrix-based economics, XGBoost provides the highest Net Present Value savings ($\approx$ \$4.52 million), reflecting the lowest confusion-matrix-based annual predictive maintenance cost ($\approx$ \$503,000 per year).

However, policy-based economics re-ranks the models: Gradient Boosting produces the best policy Net Present Value savings ($\approx$ \$4.13 million) with a payback period starting in year one, while the Decision Tree yields a negative policy Net Present Value ($\approx$ -\$0.94 million). This negative outcome is largely driven by an extreme abstention and inspection rate (71.55%). These results are clearly visible in Figure 6 – Figure 8.

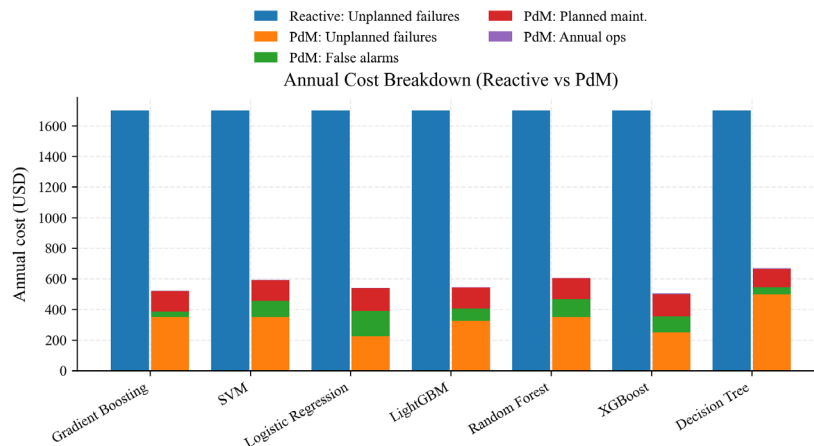


Figure 6. Annual cost breakdown (reactive vs PdM) by model

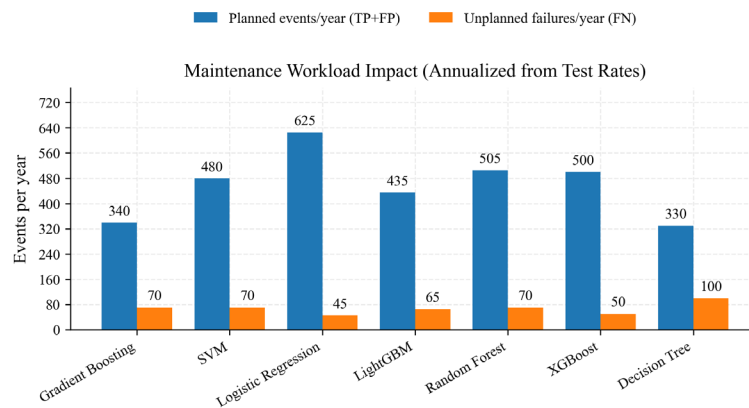


Figure 7. Maintenance workload impact (annualized planned vs unplanned events)

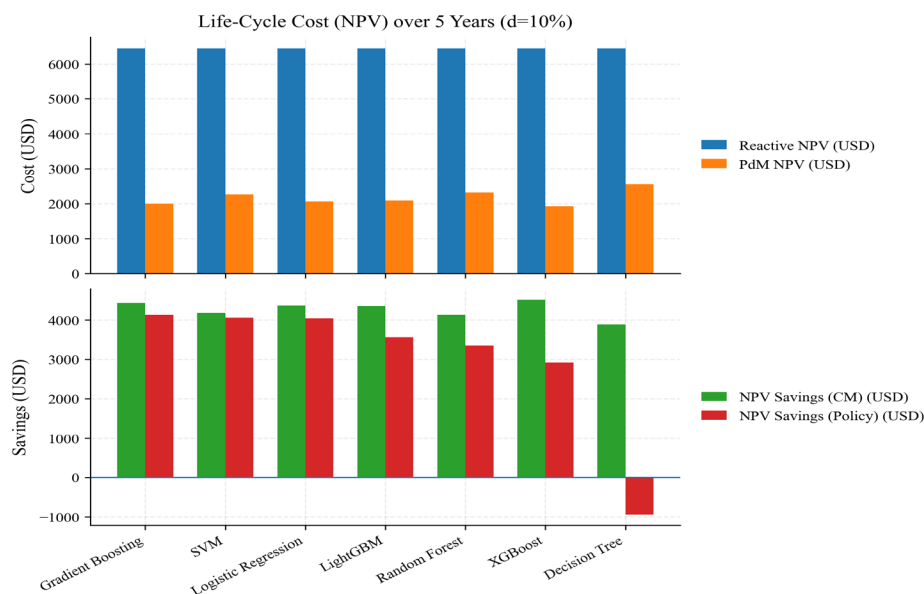


Figure 8. Life-cycle cost (NPV) and NPV savings (CM vs policy)

4.6 Sensitivity analysis and robustness

For the best policy model (Gradient Boosting), the scenario heatmap shows that the Net Present Value (NPV) savings grow rapidly with the unplanned failure cost multiplier (C_{UF}); however, the sensitivity to the false alarm cost multiplier (C_{FA}) is of less importance (Figure 9). The results of the tornado analysis ($\pm 30\%$) show the primary driver is C_{UF} , followed by the discount rate and the planned maintenance cost (C_{PM}) (Figure 10). These findings are consistent with the intuitive economics behind predictive maintenance: unplanned failures mainly drive the financial upside (Busse et al. 2019). Probability drift robustness tests for the best model under a 2% sensor noise stress test (mean $|\Delta p|$ approx. 0.003826) were found to be small. Subgroup results by machine type provide modest variation in failure rate and calibration; an interesting detail is that Type 2 has the lowest failure rate and lowest average policy cost per case in our split.

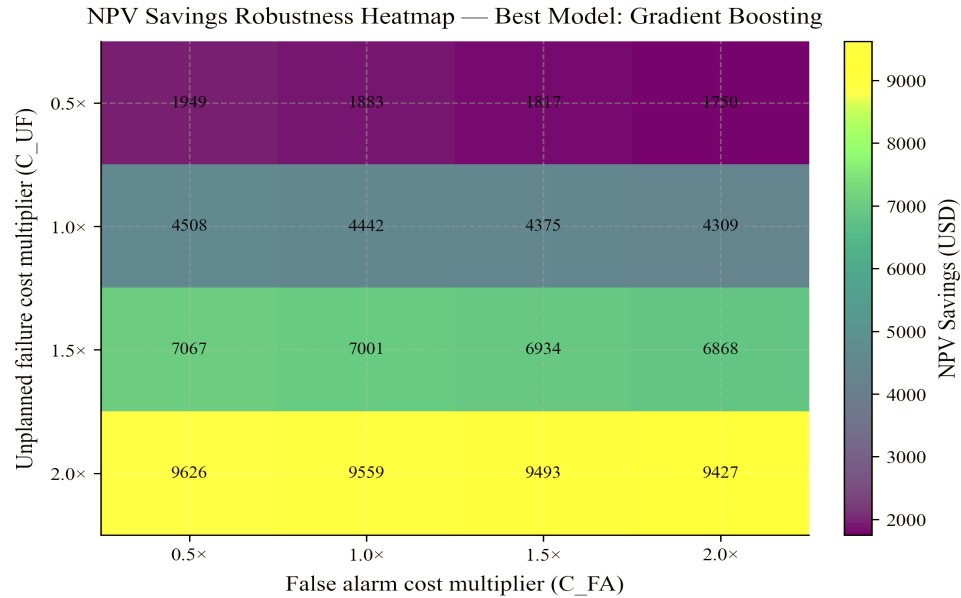


Figure 9. NPV robustness heatmap for best model (varying C_UF and C_FA)

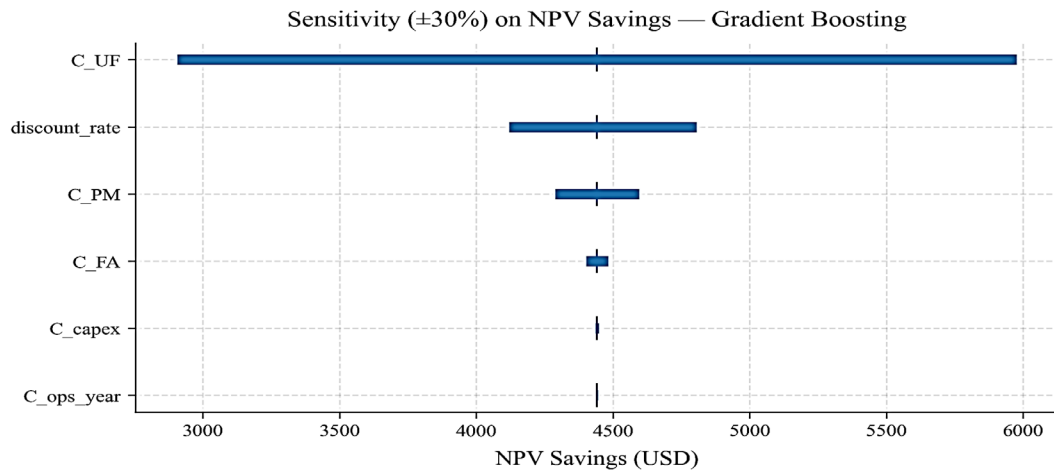


Figure 10. Tornado sensitivity ($\pm 30\%$) on NPV savings (best model)

5. Discussion

Predictive maintenance studies often report accuracy, F1-score, and ROC-AUC, but these metrics alone do not describe how a model behaves when its outputs are converted into real maintenance actions (Table 3).

Table 3. Comparison of our work with State-of-the-art works

Study	Year	Dataset used	Key metrics reported	Economic analysis present	Comparison with our study
Theissler et al.	2021	Survey across multiple industrial PdM datasets	Taxonomy of ML methods, performance trends	No	Provides overview of ML applications in PdM; our work adds calibration and economic decision modeling.
Serradilla et al.	2022	Public PdM datasets (survey comparison)	Model architecture comparisons and predictive accuracy	No	Focuses on deep learning architecture design; our study evaluates models using calibration and financial impact.
Meng et al.	2022	Industrial case study (battery systems)	Cost-benefit analysis of PdM adoption	Yes	Highlights economic feasibility of PdM adoption; our work links predictive probabilities directly to NPV and lifecycle cost.

Siraskar et al.	2023	Survey across industrial PdM applications	Reinforcement learning decision policies	Limited	Explores RL-based maintenance policies; our framework uses calibrated probabilities with a simpler deployable action policy.
Li et al.	2024	Survey of multiple PdM datasets	Deep learning architecture comparisons	No	Summarizes DL pipelines for PdM; our study focuses on reliability of predictions and economic evaluation.
Autran et al.	2024	AI4I-PMDI (modified AI4I 2020 dataset)	Classification performance comparison	No	Shows benchmark datasets may be overly clean; our study uses AI4I while focusing on decision-oriented evaluation.
Benhanifia et al.	2025	Survey of manufacturing PdM studies	Implementation challenges and adoption analysis	Yes	Identifies ROI evaluation gaps; our framework explicitly models NPV and payback.
Fabri et al.	2025	Industrial sensor dataset (large-scale case study)	Prediction error vs economic cost analysis	Yes	Shows algorithm selection should consider economic cost; our work integrates calibration and uncertainty-aware decisions.
Diallo et al.	2025	Tennessee Eastman process datasets	Fault detection accuracy and false-alarm control	No	Applies conformal prediction to fault detection; our framework links uncertainty to inspection policy and economics.
Moccardi et al.	2025	IoT predictive maintenance datasets	RUL prediction with uncertainty intervals	No	Focuses on uncertainty intervals for RUL; our work applies uncertainty to classification-based PdM decisions.
Ghasemkhani et al.	2025	AI4I 2020 predictive maintenance dataset	Multi-label classification accuracy	No	Develops multi-label prediction models; our study adds calibration and economic decision evaluation.
Yang et al.	2026	NASA C-MAPSS datasets	RUL prediction accuracy and uncertainty metrics	No	Focuses on transformer-based RUL prediction with uncertainty; our work studies classification-based PdM with economic decision modeling.

Table 3 aligns this paper with some of the most recent developments in PdM: surveys and architecture reviews summarize modeling choices (e.g., Theissler et al.; Serradilla et al.; Li et al.), other contributions target best practices or economics (e.g., Meng et al.; Benhanifia et al.; Fabri et al.), and the issues of uncertainty-aware approaches are increasingly moving into the orbit of conformal prediction, especially for the risk control of industrial monitoring (e.g., Diallo et al.; Moccardi et al.; Yang et al.). Our results demonstrate the logic behind the simultaneous examination of prediction, calibration, uncertainty, and economics.

A primary limitation is the extreme disparity in classes. There are only 203 cases of failure out of the 6,000-sample training split (approximately 3.38%). The results in Figure 2 show the balancing of the training distribution through the introduction of SMOTE, which is critical for learning consistent decision boundaries and important in preventing the models from falling back into the majority class. A pure ranking viewpoint of the ROC curves in Figure 3 indicates that all models have a similarly high ROC-AUC (0.873–0.983). This strong discrimination is confirmed by the leading positions of SVM (approx. 0.983) and Logistic Regression (approx. 0.980) in ROC-AUC. However, as shown in Figure 4 and Table 2, the performances at an operating point differ: the Gradient Boosting algorithm has the highest F1-score on the test dataset (approximately 0.794), while the Logistic Regression model performs the worst with the lowest test F1 (approximately 0.611). This disparity supports an important deployment lesson—high ROC-AUC does not imply optimum thresholder performance or optimum downstream value in the context of costs and action threshold sensitivity.

Calibration is a decisive factor because the maintenance policy is probability-based. Table 2 shows that Gradient Boosting scores the lowest on the Brier score (0.009797) and therefore has high probability accuracy (on average), although XGBoost has the lowest ECE (0.002597), which implies a close correspondence between average predicted confidence and frequency. The reliability plot of the optimal policy value model (Gradient Boosting) in Figure 5 is a tangible test to ensure that the predicted probabilities follow the empirical results sufficiently to guide the choice of action.

The question of the "best" model is critical given the economic results. Within the 5-year horizon and considering the assumptions in Section 4.6, the cost of reactive operation reaches approximately USD 1.70M/year. Figures 6–8 and Table 2 suggest that XGBoost generates the highest NPV savings using confusion matrix costing (about USD 4.52M), but this decreases significantly using policy-based costing (about USD 2.92M). On the other hand, Gradient Boosting provides the optimal policy NPV savings (around USD 4.13M), reaching the payback period in year one. A warning sign is the Decision Tree, as it has a relatively good F1 (0.716) but a negative NPV resulting from the policy (-0.94M); this is consistent with the very high abstention/inspection burden reported in the policy evaluation (71.55%), which increases inspection costs and lowers savings.

Lastly, Figures 9–10 show that the primary economic driver is the unplanned failure cost multiplier (CUF): NPV savings grow strongly with the cost of unplanned failures and weaklier with the cost of false alarms (CFA). The noise stress test shows a slight drift in probability (mean $|\Delta p| = -0.003826$), and the subgroup analysis shows little variation between machine types (Type 2 has the lowest failure rate and average policy cost per case). Overall, the results support a decision-oriented PdM evaluation: model selection should rely on predictive metrics, calibrated uncertainty, and policy-level economics—not ROC-AUC alone.

6. Conclusion and Future work

This paper suggests a predictive maintenance framework consisting of multi-label machine learning models, probability calibration, uncertainty-conscious decision policies, and economic analysis. The framework is used to evaluate predictive models applied to the AI4I predictive maintenance dataset in terms of classification performance, as well as the impact of the models on maintenance and financial decisions. These findings suggest that the model with the best predictive power is not necessarily the most cost-effective model for a maintenance decision policy. Gradient Boosting was found to be the ideal model with respect to the proposed policy-based performance, providing a strong trade-off between prediction accuracy, calibrated probability estimates, and cost-value. It is also revealed that the quality of calibration and the management of uncertainty are significant when predictive outputs are used to drive maintenance decisions. These results show the importance of testing models for predictive maintenance in a way that reflects actual operational decision-making.

The proposed framework offers a more practical approach to selecting predictive maintenance solutions by relating the output of the predictors to maintenance policies and economic factors such as Net Present Value and payback period. There are a number of limitations which should be recognized. The AI4I data is a structured benchmark which might fail to reflect the reality of complex industrial scenarios, including sensor noise, data gaps, and changing operating conditions. Furthermore, the economic analysis is based on assumptions of cost parameters and inspection effectiveness, which may vary across different industries. This framework should in the future be extended with more realistic operating conditions and real industrial datasets. The combination of predictive maintenance models and digital twin systems, along with adaptive maintenance scheduling, may further increase the practical implementation of predictive maintenance in Industry 4.0.

Acknowledgements

This research was possible with the help of the Basic Research Grant (June 2025) of BUET (Bangladesh University of Engineering and Technology) awarded to the Principal Investigator, Dr. Md. Saiful Islam. The computational capabilities used in this research were granted to the high-performance computing server, which was hosted on the Institute of Information and Communication Technology (IICT) at the BUET, Dhaka, Bangladesh.

Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

References

- Autran, J.-V., Kuhn, V., Diguët, J.-P., Dubois, M., and Buche, C., AI4I-PMDI: Predictive maintenance datasets with complex industrial settings' irregularities, *Procedia Computer Science*, vol. 246, pp. 1201–1209, 2024.
- Benhanifia, A., Cheikh, Z.B., Oliveira, P.M., Valente, A., and Lima, J., Systematic review of predictive maintenance practices in the manufacturing sector, *Intelligent Systems with Applications*, vol. 26, pp. 200501, 2025.
- Brier, G.W., VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY, *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
- Busse, A., Metternich, J., and Abele, E., Evaluating the Benefits of Predictive Maintenance in Production: A Holistic Approach for Cost-Benefit-Analysis, *Advances in Production Research*, pp. 690–704, 2019.
- Byrne, G., Dornfeld, D., Inasaki, I., Ketteler, G., König, W., and Teti, R., Tool Condition Monitoring (TCM) — The Status of Research and Industrial Application, *CIRP Annals*, vol. 44, no. 2, pp. 541–567, 1995.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- Davis, J. and Goadrich, M., The relationship between Precision-Recall and ROC curves, *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 233–240, 2006.

- Diallo, A.R., Homri, L., and Dantan, J.-Y., Reducing false alarms in fault detection: A comparative analysis between conformal prediction and classical methods applied to PCA and autoencoders, *Journal of Process Control*, 2025.
- Dimla, D.E., Sensor signals for tool-wear monitoring in metal cutting operations—a review of methods, *International Journal of Machine Tools and Manufacture*, vol. 40, no. 8, pp. 1073–1098, 2000.
- Elkan, C., The foundations of cost-sensitive learning, *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, pp. 973–978, 2001.
- Elreedy, D. and Atiya, A.F., A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance, *Information Sciences*, vol. 505, pp. 32–64, 2019.
- Fabri, L., Häckel, B., Keller, R., and Oberländer, A.M., How to Select Algorithms for Predictive Maintenance: An Economic Decision Model and Real-World Instantiation, *International Journal of Information Technology & Decision Making*, vol. 24, no. 02, pp. 361–394, 2025.
- Fawcett, T., An introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- Ghasemkhani, B., Kut, R.A., Birant, D., and Yilmaz, R., Balanced Hoeffding Tree Forest (BHTE): A Novel Multi-Label Classification with Oversampling and Undersampling Techniques for Failure Mode Diagnosis in Predictive Maintenance, *Mathematics*, vol. 13, no. 18, pp. 3019, 2025.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q., On Calibration of Modern Neural Networks, 2017.
- Jardine, A.K.S., Lin, D., and Banjevic, D., A review on machinery diagnostics and prognostics implementing condition-based maintenance, *Mechanical Systems and Signal Processing*, vol. 20, no. 7, pp. 1483–1510, 2006.
- Kritzinger, W., Karner, M., Traar, G., Henjes, J., and Sihm, W., Digital Twin in manufacturing: A categorical literature review and classification, *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.
- Li, Z., He, Q., and Li, J., A survey of deep learning-driven architecture for predictive maintenance, *Engineering Applications of Artificial Intelligence*, vol. 133, pp. 108285, 2024.
- Liao, Y., Deschamps, F., Loures, E.D.F.R., and Ramos, L.F.P., Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal, *International Journal of Production Research*, vol. 55, no. 12, pp. 3609–3629, 2017.
- Matzka, S., Explainable Artificial Intelligence for Predictive Maintenance Applications, *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 69–74, 2020.
- Meng, H., Liu, X., Xing, J., and Zio, E., A method for economic evaluation of predictive maintenance technologies by integrating system dynamics and evolutionary game modelling, *Reliability Engineering & System Safety*, vol. 222, pp. 108424, 2022.
- Moccardi, A., Conte, C., Chandra Ghosh, R., and Moscato, F., A Robust Conformal Framework for IoT-Based Predictive Maintenance, *Future Internet*, vol. 17, no. 6, pp. 244, 2025.
- Saito, T. and Rehmsmeier, M., The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, *PLOS ONE*, vol. 10, no. 3, pp. e0118432, 2015.
- Serradilla, O., Zugasti, E., Rodriguez, J., and Zurutuza, U., Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects, *Applied Intelligence*, vol. 52, no. 10, pp. 10934–10964, 2022.
- Siraskar, R., Kumar, S., Patil, S., Bongale, A., and Kotecha, K., Reinforcement learning for predictive maintenance: a systematic technical review, *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12885–12947, 2023.
- Tao, F., Qi, Q., Wang, L., and Nee, A.Y.C., Digital Twins and Cyber-Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison, *Engineering*, vol. 5, no. 4, pp. 653–661, 2019.
- Teti, R., Jemielniak, K., O'Donnell, G., and Dornfeld, D., Advanced monitoring of machining operations, *CIRP Annals*, vol. 59, no. 2, pp. 717–739, 2010.
- Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., and Elger, G., Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry, *Reliability Engineering & System Safety*, vol. 215, pp. 107864, 2021.
- Yang, C.-L., Meles, T.Y., Yilma, A.A., and Teshome, M.M., Uncertainty aware predictive maintenance using a hybrid Transformer with Monte Carlo Dropout and conformal prediction, *Ain Shams Engineering Journal*, vol. 17, no. 2, pp. 103992, 2026.
- Zheng, P., Wang, H., Sang, Z., Zhong, R.Y., Liu, Y., Liu, C., Mubarak, K., Yu, S., and Xu, X., Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives, *Frontiers of Mechanical Engineering*, vol. 13, no. 2, pp. 137–150, 2018.

Biographies

Md. Efatazzaman Efat is a student completed his B.Sc. in Mechanical Engineering at the Khulna University of Engineering and Technology (KUET) and currently pursuing a Post Graduate Diploma (PGD) in Business IT at the Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET). The challenge of solving real-world problems is what motivates him as a result of a combination of smart computational processes and high-tech engineering systems. Nowadays, his work is dedicated to Material Informatics, the application of AI and ML to the work of mechanical systems, and IoT to the industrial market.

Sayed Md Fazle Rabbi is presently undertaking his B.Sc. in Electrical and Electronic engineering at Ahsanullah University of Science and Technology (AUST). He has been active in researching interdisciplinary use of emerging technologies, which he started studying in 2021. His areas of research include machine learning, cybersecurity, optical communication, photonic crystal fiber (PCF) and material informatics. The combination of smart computational processes and the high-tech engineering systems to solve the challenges in the real world is what motivates him in particular.

Md. Saiful Islam is a Professor at the Institute of Information and Communication Technology (IICT) of Bangladesh University of Engineering and Technology (BUET). He obtained B. Sc. in Electrical and Electronic Engineering from BUET in 1989, M. Sc. in Computer Engineering from Shanghai University, China in 1997, and PhD from BUET in 2008. He served as the Director of IICT for more than 6 years and as Director of ICT Cell for 2 years at BUET. He has designed, coordinated and implemented many IT projects at the national level and published many articles in peer-reviewed journals. His research interest includes cybersecurity, artificial intelligence, optical communication and real-world AI applications. In his long academic career, he has supervised about 60 postgraduate students, which resulted in solutions to various real-life ICT problems.