American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 5 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 5 Issue 2, Year 2026
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v5i2.8505
https://journals.e-palli.com/home/index.php/ajsts

Cost Sensitive Machine Learning for Rare Failure Prediction: Balancing Missed Failures and False Maintenance Alarms

Kaoutar Bizmour[1*]

Article Information

Received: June 28, 2026

Accepted: September 03, 2026

Published: September 21, 2026

Keywords

Class Imbalance, Cost Sensitive Learning, False Negatives, False Positives, SHAP

ABSTRACT

Rare machine failures pose a difficult predictive-maintenance challenge because extreme class imbalance can result in deceptively high classification accuracy while disguising missed failures and the operational overhead of false alarms. The AI4I 2020 Predictive Maintenance Dataset is applied in this study to develop a leakage-controlled, maintenance-oriented computational learning framework that includes classifier comparison, imbalance correction, cost-sensitive threshold selection, and explainable artificial intelligence. Logistic Regression, Random Forest, XGBoost, and LightGBM were all tested on a stratified independent test set. Random Forest got the highest baseline precision (97.44%) and F1-score (84.44%), but XGBoost and LightGBM had higher failure recall (80.39%). LightGBM was employed as the probabilistic model in following analyzes. Random oversampling outperformed all other imbalance treatments, raising recall to 88.24%, lowering missed failures to six and false positives to five, and earning an F1-score of 89.11%. Cost-sensitive threshold optimization indicated a clear maintenance trade off. FN-to-FP cost ratios of 1:1 and 2:1 chose a threshold of 0.22, resulting in eight missed failures and nine false alarms, whereas a 10:1 ratio dropped the threshold to 0.01 and missed failures to five while increasing false alarms to 59. SHAP analysis revealed that tool wear, torque, and rotational speed were the most influential predictors. The data indicates that balancing class representation is not the same as balancing maintenance risk. As a result, the proposed framework moves rare-failure evaluation away from accuracy centered model selection and toward interpretable FN-FP operating points that explicitly link probabilistic predictions with maintenance priorities and provide a transparent basis for selecting deployment thresholds in industrial predictive maintenance applications with asymmetric operational consequences.

INTRODUCTION

Predictive maintenance (PdM) has emerged as an important component of intelligent manufacturing, as unexpected machinery breakdowns can result in production disruption, equipment damage, increased maintenance costs, and decreased operational reliability. Recent predictive-maintenance applications have also demonstrated the use of Random Forest for equipment health assessment and remaining useful life prediction, illustrating the broader role of data driven models in supporting proactive maintenance planning (Owusu & Baidoo, 2026). With the growing availability of operational and condition-monitoring data, computational intelligence algorithms have been extensively researched for detecting aberrant machine states and assisting with data-driven maintenance decisions. Recent work has similarly demonstrated the use of current and temperature signals with artificial intelligence for detecting thermal faults in induction motors, highlighting the value of condition-monitoring data for intelligent fault assessment (Owusu *et al.*, 2026). Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) algorithms are particularly well-suited to structured industrial data because they can capture nonlinear correlations between several operating variables. The growing application of artificial intelligence to intelligent control and monitoring in mechanical engineering further demonstrates the potential of data driven methods for supporting fault-related analysis and maintenance decision-making (Al Dosari & Abouellail, 2023). Class imbalance is a significant challenge for machine-failure prediction. For the majority of its service life, industrial equipment functions normally; genuine malfunctions happen quite infrequently.

As a result, there may be significantly more normal observations in a dataset than failure observations. In these circumstances, a model may accurately identify the majority of normal data while missing a significant fraction of the comparatively few failures, making high classification accuracy deceptive. In order to enhance minority-class recognition, previous research has investigated class weighting, random oversampling, undersampling, and synthetic minority oversampling, while comparative studies have further demonstrated the importance of selecting appropriate resampling strategies for imbalanced classification problems (de Giorgio *et al.*, 2023; Qawqzeha, 2025). However, as a higher failure recall may also result in a higher number of false alarms, greater class balance does not always translate into better maintenance decisions. When anticipated failure probability are translated into maintenance alerts, this problem becomes especially crucial. While a false positive (FP) indicates normal operation mistakenly categorized as failure and may result in needless inspection or

[1] School of Mechatronic Engineering, China University of Mining and Technology, Xuzhou, 221116, China
[*] Corresponding author's e-mail: 27230007@cumt.edu.cn

maintenance, a false negative (FN) shows a true failure classified as normal and so constitutes a missed warning. These unequal outcomes are not specifically taken into consideration by the traditional probability criterion of 0.50. Therefore, the operational threshold at which the balance between missed failures and false maintenance alerts becomes acceptable should be taken into account in addition to model discrimination in rare-failure prediction. For assessing highly imbalanced maintenance issues, metrics including precision, recall, PR-AUC, and Matthews correlation coefficient (MCC) are therefore more relevant than accuracy alone. Recent applications of artificial intelligence in factory monitoring further demonstrate the growing role of data driven systems in supporting industrial equipment supervision and operational decision making (Nicola, 2026). Explainability is also vital since maintenance people require more information than a simple failure prediction. They must understand which operational conditions contribute to an increased expected risk. Explainable artificial intelligence methods, such as SHapley Additive ExPlanations (SHAP), may quantify global and observation specific feature contributions and are increasingly used in predictive maintenance research (Cummins *et al.*, 2024). Recent research has also incorporated imbalance therapy, cost awareness, and explainability in machinery failure classification (Obeten & Jimoh, 2026). As a result, the individual use of SHAP, oversampling, class weighting, or cost-sensitive learning cannot be considered adequate methodological innovation.

The current study addresses a more particular question: is improved rare-failure detection still useful when the false-maintenance alarm load is explicitly considered? Matzka's AI4I 2020 Predictive Maintenance Dataset serves as an appropriate baseline for this inquiry. It contains 10,000 observations depicting industrial operating circumstances, of which only 339 are related to machine failure, yielding a failure prevalence of 3.39% and a normal-to-failure ratio of around 28.5:1. This pronounced imbalance makes an ideal context for studying how imbalance treatment and probability-threshold selection effect maintenance-oriented classification.

As a result, this study creates a framework for predicting infrequent failures and determining operating points that are cost-effective. Representative machine learning classifiers are initially compared using a leakage-controlled predictor set. Alternative imbalance remedies are then assessed for their impact on minority failure recognition. Rather than assuming that the standard 0.50 probability threshold is the best maintenance option, an asymmetric-cost criterion is used to validate forecasts and evaluate alternative operating points. Finally, SHAP analysis is utilized to determine which operating factors are accountable for the estimated failure risk.

The principal contributions of this study are threefold: (1) a controlled comparison of imbalance-handling strategies that explicitly considers both missed failures and false alarms rather than overall accuracy alone; (2) a validation-based cost-sensitive operating-point approach for quantifying how different penalties for missed failures alter the FN-FP trade-off; and (3) an explainable maintenance risk assessment linking model predictions with the operational variables responsible for elevated failure risk. The resulting framework therefore treats rare-failure prediction as both a classification problem and a maintenance decision problem.

The remaining sections of this paper are organized as follows. Section 2 describes the dataset and approach. Section 3 presents the experimental findings, Section 4 examines their consequences, Section 5 addresses limitations, and Section 6 ends the study.

MATERIALS AND METHODS

Dataset and Preprocessing

The studies were carried out using the AI4I 2020 Predictive Maintenance Dataset, which contains 10,000 observations detailing the operational state of industrial machinery. The dataset contains five failure-mode indicators: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failure (RNF). Table 1 summarizes the key factors and their functions in the present analysis.

Table 1: Variables used in the rare-failure prediction framework

Variable	Unit/type	Role
Product type	L, M, H	Predictor
Air temperature	K	Predictor
Process temperature	K	Predictor
Rotational speed	rpm	Predictor
Torque	Nm	Predictor
Tool wear	min	Predictor
Machine failure	Binary	Prediction target
TWF, HDF, PWF, OSF, RNF	Binary	Failure-mode interpretation

There were none missing values or duplicate observations found in the initial data quality evaluation. Out of the 10,000 data, 339 constituted machine failures and 9,661 represented normal operation, with failure and normal proportions of 3.39% and 96.61%, respectively. The prediction function is a highly unbalanced classification problem, as evidenced by the ensuing normal-to-failure imbalance ratio of roughly 28.5:1.

The IDs UDI and Product ID were eliminated because they do not describe physical operating conditions. More crucially, TWF, HDF, PWF, OSF, and RNF were omitted from the predictor matrix since they explicitly characterize failure processes and may induce target leakage when predicting the aggregate Machine failure variable. They were kept only for post-model failure mode analysis. The final base predictor set included product type, air temperature, process temperature, rotational speed, torque, and tool wear. Two engineering-derived variables were created to describe thermal and mechanical operating conditions. The temperature differential and mechanical power were computed as:

$$\Delta T = T_p - T_a, \qquad P_m = \tau\left(\frac{2\pi n}{60}\right) \qquad (1)$$

Where T_p and T_a represent process and air temperatures, τ represents torque, and n is rotational speed. These changes were used as physically interpretable predictors rather than being presented as unique feature-engineering techniques. Prior to modeling, the product type was one-hot encoded. Numerical variables were scaled for Logistic Regression, although scaling was not required for tree-based models. To preserve the original rare-event distribution during evaluation, the dataset was divided into 70% training, 15% validation, and 15% test sets. The training set had 7,000 observations, including 237 failures, whereas the validation and test sets each had 1,500 observations and 51 failures. The entire experimental technique is shown in Figure 1. All pre-processing parameters and imbalance treatments were selected solely from the training data set.

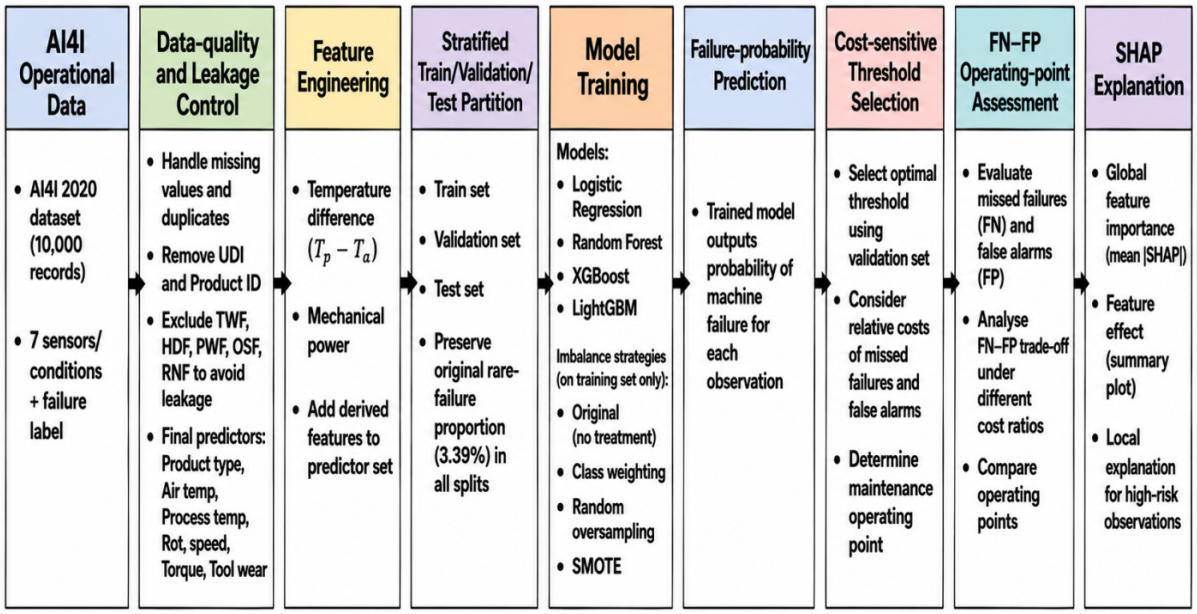

Figure 1: Proposed cost sensitive rare failure prediction and maintenance operating point assessment structure

Rare Failure Prediction and Cost Sensitive Structure

To compare various learning techniques, four supervised classifiers Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting) Light Gradient Boosting Machine (XGBoost), and (LightGBM) were assessed. XGBoost and LightGBM represented gradient-boosting techniques that could describe nonlinear interactions among operating variables, RF represented bagging based ensemble learning, and LR served as a linear baseline. Conclusions about rare failure behavior are also prevented from being reliant on a single classifier by using multiple well-established models. Four imbalance conditions no imbalance treatment, class weighting, random oversampling (ROS), and SMOTE were examined because only 3.39% of the observations matched machine failure. After the original dataset was separated into training, validation, and test sets, resampling was only done on the training partition. As a result, the test and validation distributions did not change. The various treatments were compared based on their impacts on failure recall, precision, PR-AUC, MCC, and especially the quantities of false negatives and false positives, rather than assuming that balancing inevitably improves prediction.

During model fitting, failure observations were given more weight for the class weighted models. The training distribution, for which the normal-to-failure ratio was roughly 28.54:1, was used to determine the weighting. Without creating fake data, this tactic raises the penalty for improperly classified minority observations. While SMOTE creates artificial minority observations from nearby failure samples, ROS replicates training failure

observations to boost minority representation. In order to ascertain whether increases in failure sensitivity are accompanied by an intolerable rise in false alarms, these options were included. The approach primary maintenance-oriented component deals with turning projected failure probabilities into decisions. When a traditional classifier's projected probability above a threshold typically 0.50 it issues a failure label. Nevertheless, the distinctions between the outcomes of a false maintenance alert and a missed failure are tacitly ignored by this fixed operating point. Thus, the threshold-dependent decision cost is defined in the current study as:

$$C(\theta) = C_{FN}FN(\theta) + C_{FP}FP(\theta) \qquad (2)$$

Where the numbers of false-negative and false-positive predictions produced at threshold θ are denoted by FN(θ) and FP (θ), respectively. While C_{FN} denotes the supposed result of giving an unjustified failure notice, C_{FP} indicates the assumed penalty of dismissing a genuine machine failure. Afterward, the cost-sensitive operational threshold was established as:

$$\theta^* = \frac{\arg\min C(\theta)}{\theta} \qquad (3)$$

The AI4I dataset missing information on downtime, inspection, production losses, and component replacement costs, hence C_{FN} and FP(θ) are interpreted as scenario-based relative costs, not monetary estimates. A cost-ratio sensitivity analysis is performed to investigate how increasing the projected consequence of a missed failure affects the desired threshold and the resulting FN-FP balance. The independent test set is not used to determine thresholds. Table 2 summarizes the predictive models and imbalance techniques employed during the experiments.

Table 2: Predictive models and imbalance handling configurations

Component	Configuration
Logistic Regression	Linear baseline; standardized numerical predictors
Random Forest	500 decision trees
XGBoost	400 estimators; maximum depth 4; learning rate 0.05
LightGBM	400 estimators; 15 leaves; learning rate 0.05
Class weighting	Minority class weighted from training distribution
Random oversampling	Minority observations replicated in training set only
SMOTE	Synthetic minority observations generated in training set only
Decision threshold	Selected using validation-set asymmetric cost
Final evaluation	Independent stratified test set

SHapley Additive exPlanations (SHAP) were then used to obtain model explanations. The additive explanation for a prediction f(x) is shown as:

$$f(x) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \qquad (4)$$

The variables most influential across predictions were identified using global SHAP importance, and the operating conditions responsible for a subset of high-risk observations were examined using local explanations, with $\emptyset_0$ representing the baseline model output and $\emptyset_j$ representing the contribution of predictor j. This makes it possible to add an engineering interpretation of the anticipated failure risk to the statistical FN-FP analysis.

Experimental Evaluation

On the independent test set, model performance was assessed using precision, recall, F1-score, specificity, balanced accuracy, Matthews correlation coefficient (MCC), ROC-AUC, and PR-AUC. Because machine faults account for just 3.39% of the dataset, overall accuracy was considered supplementary rather than the primary evaluation criterion. Failure recall and PR-AUC, as well as the absolute numbers of false negatives (FN) and false positives (FP), were used to evaluate each model's capacity to detect unusual failures without producing too many false maintenance alarms.

The experiments were carried out in two stages. To begin, LR, RF, XGBoost, and LightGBM were assessed using a common probability threshold of 0.50 to determine their baseline predictive behavior. The selected core classifier was then assessed using the original unbalanced distribution, class weighting, ROS, and SMOTE. The validation data was used to determine the model and imbalance approach, while the test partition was set aside for ultimate performance evaluation.

Second, the cost-sensitive operating point specified in Eq. (2) and (3) was examined using the validation predictions. To find out how increasingly harsher penalties for missed failures impacted the chosen threshold, failure recall, and false alarm load, several relative FN-to-FP cost scenarios were examined. Each threshold was applied to the independent test set without modification after it was chosen. To ascertain whether decreases in missed failures outweighed the related rise in false maintenance alarms, the resulting operating points were compared with the traditional 0.50 threshold. The chosen model was then

subjected to SHAP analysis in order to determine which operating variables were in charge of both individual and global failure-risk forecasts.

RESULTS AND DISCUSSION

Failure Characteristics and Model Performance

In order to ascertain whether the minority failure class showed discernible patterns, the operating attributes linked to normal and failure data were analyzed prior to assessing the prediction models. The significant class imbalance mentioned in Section 2.1 was confirmed by the fact that 9,661 (96.61%) of the 10,000 observations showed normal functioning and 339 (3.39%) showed machine failure. Several operating variables demonstrated distinct variations between the two machine modes, despite the very limited number of failure observations.

The most noticeable variations were found in mechanical power, torque, and tool wear. The average tool wear rose by around 34.76%, from 106.69 minutes during regular operation to 143.78 minutes after failure observations. Similarly, the mean torque increased by 26.59%, from 39.63 Nm to 50.17 Nm. The mechanical power derived by engineering grew by 16.63%, from roughly 6244.55 W to 7282.82 W. On the other hand, the mean rotational speed dropped by about 2.84%, from 1540.26 rpm during normal operation to 1496.49 rpm during failure. The results showed less variations in the thermal variables. The average process temperature rose from 310.00 K to 310.29 K, and the average air temperature rose from 299.97 K to 300.89 K. From roughly 10.02 K to 9.40 K, the derived temperature difference dropped. These descriptive findings imply that mechanical loads and cumulative tool wear are more powerful differentiators of failure data than straightforward variations in mean temperature. Figure 2 shows the main distinctions between normal and failed observations, and Table 3 summarizes the relative predictive performance of LR, RF, XGBoost, and LightGBM on the independent test set at the traditional probability threshold of 0.50. These group-level statistics, however, do not determine feature importance within the predictive models; SHAP analysis is used later to assess the impact of individual variables.

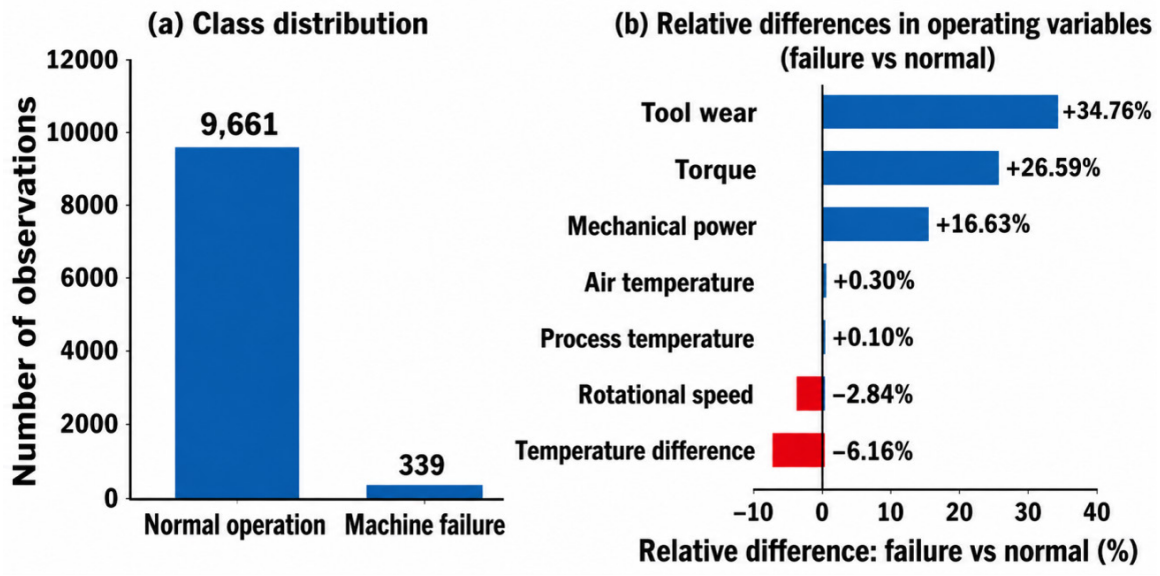

Figure 2: Distribution and operating characteristics of normal and machine-failure observations: (a) class distribution and (b) relative differences in selected operating variables between normal and failure states.

Table 3: Comparative test performance of machine-failure prediction models at a probability threshold of 0.50

Model	Accuracy	Precision	Recall	F1-score	FN	FP
Logistic Regression	0.9687	0.6429	0.1765	0.2769	42	5
Random Forest	0.9907	0.9744	0.7451	0.8444	13	1
XGBoost	0.9893	0.8723	0.8039	0.8367	10	6
LightGBM	0.9893	0.8723	0.8039	0.8367	10	6

The findings show that evaluation of a rare-failure prediction requires more than just precision. Although the overall accuracy of Logistic Regression was 96.87%, its failure recall was just 17.65%, meaning that 42 of the 51 failure observations in the test set were missed. Therefore, rather than accurately identifying the minority failure class, the comparatively high accuracy was primarily due to accurate categorization of the dominating normal class. The minority-class recognition was much improved by the tree-based models. Random Forest produced just one false positive and the highest precision of 97.44%, but its recall of 74.51% led to 13 missed failures. At the traditional probability threshold of 0.50, XGBoost and LightGBM achieved the same classification results:

98.93% accuracy, 87.23% precision, 80.39% recall, and an F1-score of 83.67%. Ten of the fifty-one failure observations were overlooked by both models, and they produced six false-positive maintenance alerts.

Although Random Forest had the best precision and F1-score of the models tested, XGBoost and LightGBM had superior failure sensitivity, detecting 41 of the 51 failure observations compared to 38 recognized by Random Forest. LightGBM was used as the core probabilistic classifier in the subsequent imbalance-treatment, cost sensitive, and SHAP analyzes because the goal of this study is to investigate how the operating threshold of a failure-probability model can be adjusted in response to asymmetric maintenance consequences. The 10 missed failures at the customary threshold serve as a baseline for determining whether enhanced failure sensitivity can be obtained without imposing an excessive false-alarm burden.

Effect of Imbalance Treatment

After the baseline comparison, the potential for explicit class-imbalance treatment to enhance rare-failure detection was studied using LightGBM. Four training settings were compared: SMOTE, random oversampling (ROS), class weighting, and the initial unbalanced data. To ensure that the comparison represented performance under the real imbalance level of the dataset, the validation and test sets maintained their original class distributions. Table 4 presents the results achieved on the independent test set.

Table 4: Effect of imbalance handling strategy on LightGBM test performance

Strategy	Precision	Recall	F1-score	PR-AUC	FN	FP
Original	0.8723	0.8039	0.8367	0.8892	10	6
Class weighting	0.7963	0.8431	0.8190	0.8801	8	11
Random oversampling	0.9000	0.8824	0.8911	0.9091	6	5
SMOTE	0.6406	0.8039	0.7130	0.8497	10	23

The effects of the imbalance handling techniques on failure sensitivity and false-alarm burden varied significantly. Class weighting reduced missed failures from 10 to 8 and raised recall from 80.39% to 84.31%, but it also raised false positives from 6 to 11. Random oversampling, on the other hand, had the best overall performance, raising recall to 88.24% while lowering false positives to 5 and missed failures to 6. Among the four assessed conditions, it also attained the highest precision (90.00%), F1-score (89.11%), and PR-AUC (0.9091). The similar advantage was not offered by SMOTE. While precision dropped significantly to 64.06% and false positives rose from 6 to 23, recall stayed at 80.39%, the same as the previous unbalanced model. Additionally, its PR-AUC and F1-score decreased to 0.8497 and 0.7130, respectively.

Therefore, in the current experimental setup, simulated oversampling raised the false-alarm load without enhancing failure detection. The results presented show that whether minority representation increases during training should not be the only criterion used to evaluate imbalanced treatment. While SMOTE reduced overall decision quality and class weighting traded fewer missed failures for more false alarms, random oversampling increased failure sensitivity and the FN-FP balance. The results therefore serve as motivation for the cost sensitive analysis in Section 3.3, where maintenance operating points are examined independently of the training distribution adjustment via threshold selection.

Cost Sensitive FN-FP Operating Point Analysis

LightGBM was selected as the basic probabilistic classifier for this investigation because it provided relatively high failure sensitivity at the standard threshold and allowed the effect of asymmetric FN-FP costs to be investigated through threshold change. The validation set was used to find cost-sensitive operational points, which were then tested independently. Table 5 summarizes the corresponding outcomes. Cost-sensitive threshold selection resulted in distinct maintenance operating points as the estimated relative significance of a missed failure rose. The validation criterion was 0.22 when FN and FP costs were equal (1:1). Applying these criteria to the independent test set resulted in 8 missed failures and 9 false maintenance alerts, for an 84.31% recall. The 2:1 cost scenario used the same threshold, resulting in the same test-set operating point. Increasing the relative FN cost to 5:1 dropped the specified threshold to 0.18, resulting in 8 missed failures and 10 false positives.

Table 5: Validation selected cost-sensitive maintenance operating points and corresponding test performance

FN:FP cost ratio	Selected threshold	Precision	Recall	F1-score	Specificity	Balanced accuracy	MCC	FN	FP
1:1	0.22	0.8269	0.8431	0.8350	0.9938	0.9185	0.8291	8	9
2:1	0.22	0.8269	0.8431	0.8350	0.9938	0.9185	0.8291	8	9
5:1	0.18	0.8113	0.8431	0.8269	0.9931	0.9181	0.8209	8	10
10:1	0.01	0.4381	0.9020	0.5897	0.9593	0.9306	0.6117	5	59

These findings highlight the main trade-off explored in this study. Lowering the judgment threshold averted three further missed failures, but this improvement necessitated 36 more false alarms. As a result, the increased failure sensitivity came at a cost. Precision fell from 93.18% to 53.01%, while F1-score dropped from 86.32% to 65.67%. In contrast, balanced accuracy rose because the lower threshold boosted minority-class sensitivity while preserving relatively high specificity.

Under the 10:1 situation, a much different operating point appeared, and the chosen threshold dropped to 0.01. False positives rose to 59 as a result, although recall rose to 90.20% and missed failures decreased to 5. As a result, precision dropped to 43.81% and the F1-score to 58.97%. In contrast to the 1:1 operating point, 50 more false maintenance alarms were needed to stop three more missed failures. The main trade-off addressed in this study is revealed by these findings. The chosen operating point did not always alter much when the estimated penalty of missed failures was increased moderately. The 5:1 scenario only resulted in a slight decrease from 0.22 to 0.18, although the 1:1 and 2:1 scenario chose the same threshold. By contrast, the classifier was moved toward a highly failure sensitive operational point and the false-alarm load was significantly increased when missed failures were given a 10-fold relative consequence.

Crucially, these operational locations do not correspond to distinct prediction models. The threshold used to translate anticipated probabilities into maintenance decisions was adjusted, while the underlying LightGBM probability model stayed the same. As a result, without knowledge of the real-world repercussions of overlooked failures and needless maintenance, no operating point can be deemed universally ideal. Figure 3 illustrates this tendency by showing how different FN-to-FP cost assumptions result in different maintenance operating points and how changes in the classification threshold affect the competing FN and FP results.

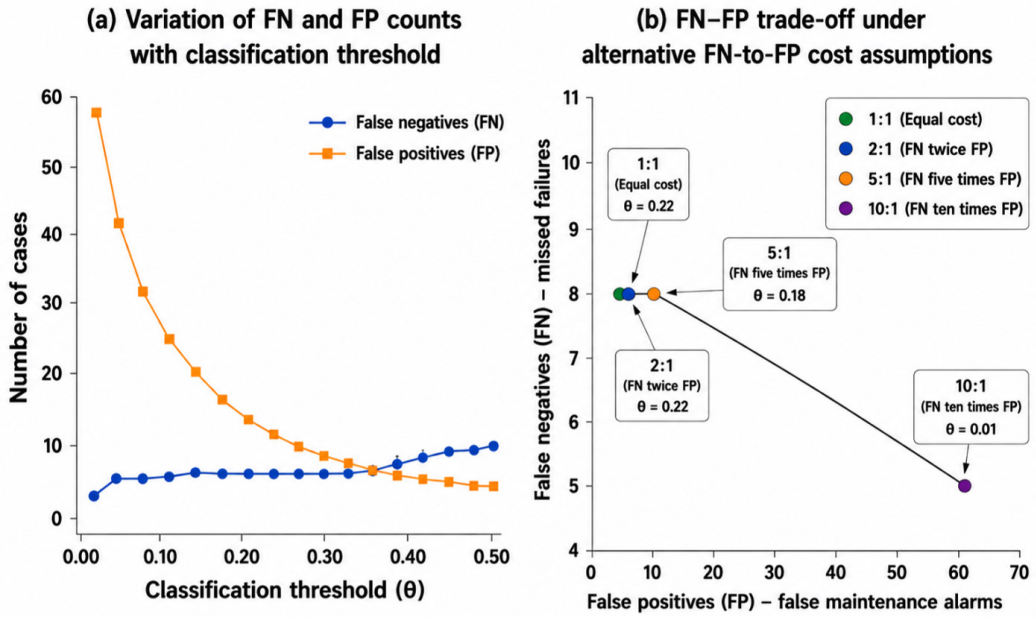

Figure 3: Cost sensitive maintenance operating-point analysis showing (a) variation of false-negative and false-positive counts with classification threshold and (b) the resulting trade-off between missed failures and false maintenance alarms under alternative FN-to-FP cost assumptions.

The findings thus validate the main idea of the suggested framework: failure recall should not be the only criterion used to assess rare-failure prediction. Improvements in failure detection must be weighed against the additional false-alarm burden they induce. The operational threshold establishes how probabilistic forecasts are converted into maintenance activities.

Explainable Failure Risk Analysis

In order to supplement the cost-sensitive and predictive analyzes, SHAP was used to the chosen LightGBM model to investigate the contribution of various operating variables to machine-failure predictions. In order to avoid using the five failure-mode indicators as explanatory inputs, the analysis was carried out using the leakage-controlled predictor set outlined in Section 2.1. The predictors were classified via global SHAP analysis based on their average absolute contribution to the model's output. With a mean absolute SHAP value of 0.939, tool wear was the most significant predictor, as seen in Figure 4(a). Torque (0.675), rotational speed (0.634), temperature difference (0.545), air temperature (0.431), mechanical power (0.392), process temperature (0.305), and product type (0.175) were next. The magnitude and direction of feature contributions are further displayed in Figure 4(b) SHAP summary distribution. The model output is shifted toward the failure class by positive SHAP values and away from failure by negative SHAP values. These findings show that when separating failure from typical observations, the model most heavily depended

on cumulative tool wear and mechanical operating circumstances.

Additionally, a representative high-risk failure observation was subjected to local SHAP analysis. Tool wear contributed the most positively (+1.256), as seen in Figure 4(c), followed by torque (+0.782), rotational speed (+0.468), and temperature differential (+0.332). On the other hand, the model output was changed in the opposite direction by air temperature (-0.174), mechanical power (-0.142), process temperature (-0.076), and product type (-0.034). This local explanation supplements the FN-FP analysis by revealing the operating conditions that underlie individual maintenance alerts and shows how several operating factors collectively influenced a single high-risk forecast.

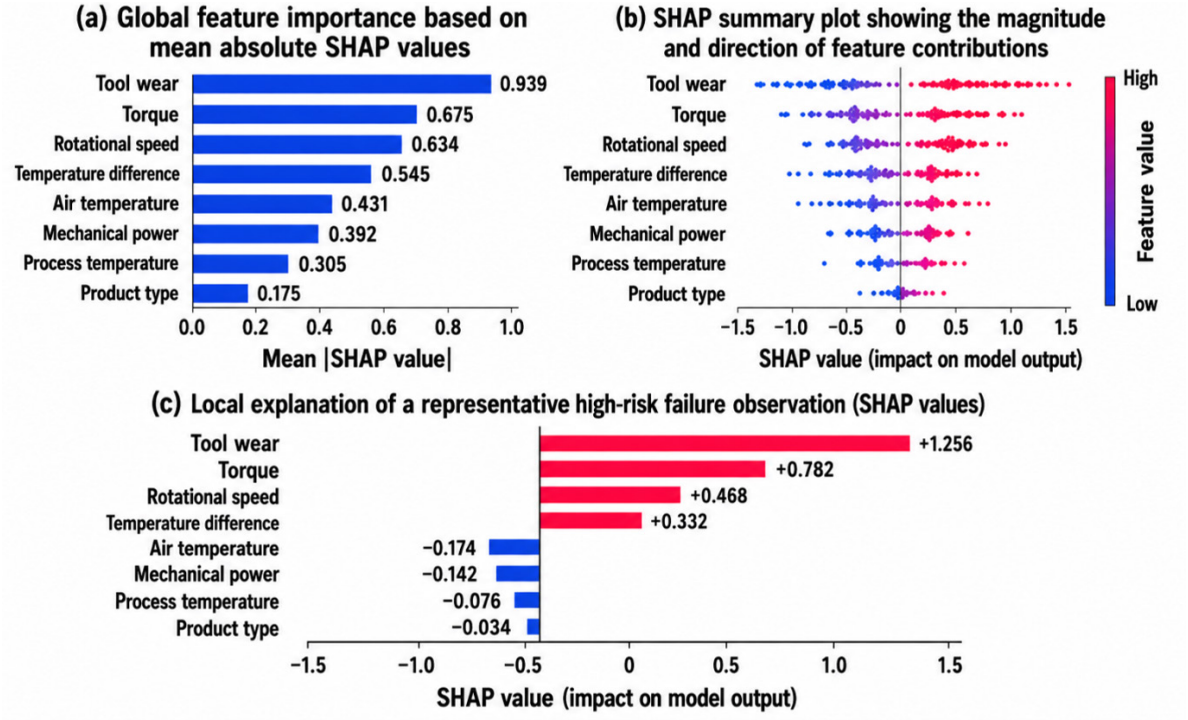

Figure 4: SHAP-based explanation of machine-failure predictions: (a) global feature importance based on mean absolute SHAP values, (b) SHAP summary plot showing the magnitude and direction of feature contributions, and (c) local explanation of a representative high-risk failure observation

The SHAP results offer an engineering understanding of the specified predictive model and enable a comparison between the operating-condition variations found in Section 3.1 and the factors associated to enhanced failure risk. In example, the very considerable positive differences found for these variables between failure and normal observations are generally consistent with the high impact of tool wear and torque. Crucially, SHAP values do not indicate a causal relationship between an operating variable and machine failure; rather, they characterize the predictive behavior of the model.

Discussion

The results obtained indicate that in broad terms classification accuracy alone is insufficient for evaluating rare machine-failure prediction. Due to the most of normal data, all assessed models obtained rather good accuracy; nonetheless, there were significant differences in their capacity to identify the minority failure class. This was especially noticeable for Logistic Regression, which had an accuracy of 96.87% yet only identified 17.65% of the failures, meaning that 42 failures were misidentified.

The tree-based models, on the other hand, produced noticeably better minority-class recognition. Random Forest had the best F1-score (84.44%) and precision (97.44%), but 13 missed failures were caused by its recall of 74.51%. Although their F1-score was marginally lower at 83.67%, XGBoost and LightGBM both achieved higher recall of 80.39% and reduced the number of missed failures to 10. These results highlight the significance of assessing rare-failure classifiers using a variety of minority-sensitive metrics as opposed to just overall accuracy (de Giorgio *et al.*, 2023). The imbalance treatment studies also demonstrated that the technique chosen has a significant impact on the outcome of minority class treatment. Class weighting decreased missed failures from 10 to 8 and raised failure recall from 80.39% to 84.31%; nevertheless, false positives rose from 6 to 11. The best overall outcome was obtained by random oversampling, which increased recall to 88.24% while concurrently lowering false positives to 5 and missed failures to 6. Among the assessed imbalance circumstances, it also had the greatest PR-AUC (0.9091) and F1-score (89.11%).

SMOTE, on the other hand, increased false positives from 6 to 23 and decreased precision to 64.06% without improving recall beyond the original model. These findings

show that improving minority representation during training does not ensure better maintenance-decision quality, and that strategies for handling imbalances should be evaluated based on their combined effects on failure detection and false-alarm burden (Mahale *et al.*, 2025).

More significantly, the cost-sensitive analysis showed that maintenance decision making and model training are connected but separate phases. Modifying the classification threshold affected how projected probabilities were converted into maintenance warnings, but it had no effect on the underlying LightGBM model. The validation process used the same threshold of 0.22 under FN-to-FP cost ratios of 1:1 and 2:1, resulting in 8 missed failures and 9 false alarms on the test set. Even though the 5:1 scenario had a marginally lower threshold of 0.18, it still resulted in 8 missed failures and 10 false alarms.

At 10:1, there was a more noticeable trade-off: false alarms rose to 59 while the threshold dropped to 0.01 and missed failures dropped to 5. As a result, there isn't a threshold that is always ideal regardless of maintenance priorities. While a system where inspections or needless shutdowns are costly would necessitate a more conservative operating point, an operating environment where an undetected breakdown carries serious implications might justify accepting more false alarms. Rather than only reporting the highest possible recall, this FN-FP approach offers a more practical interpretation of rare-event classification.

An additional connection between statistical prediction and engineering interpretation is made possible by the explainability analysis. Tool wear was shown to be the most significant predictor (mean |SHAP| = 0.939) by global SHAP analysis, followed by torque (0.675) and rotational speed (0.634). This suggests that the LightGBM model heavily depended on variables related to degradation and operating conditions when assessing failure risk. High predictive performance by itself does not prove that a model is reacting to significant operating variables, which makes this interpretation crucial. Explainability is becoming more widely acknowledged as a critical prerequisite for reliable predictive-maintenance systems (Shadi *et al.*, 2025). However, rather than being causal proof, the SHAP associations should be seen as explanations of model behavior; a high SHAP contribution suggests that a variable affected the prediction, but it does not establish that altering that variable would independently cause or avoid failure.

When considered together, the results indicate that explainability, operating-threshold selection, imbalance handling, and discrimination performance must be integrated for successful rare-failure prediction. The key practical implication is that the model configuration that yields the best accuracy or recall should not be automatically adopted by maintenance systems. Rather, the operational burden caused by false alarms should be taken into account along with the allowable number of missed failures. With this method, different maintenance policies can be supported by the same probabilistic prediction model based on the relative consequences attributed to the two categories of errors.

Limitations

When considering the results, a number of limitations should be taken into account. First off, although being developed to represent actual industrial predictive-maintenance situations, the AI4I 2020 dataset is artificial. As a result, it is not appropriate to interpret the predictive performance found in this study as proof of comparable performance in an operational factory. The benchmark dataset may not accurately reflect sensor noise, missing measurements, shifting operating regimes, maintenance interventions, aging equipment, and environmental disruptions found in real industrial systems. Therefore, prior to practical deployment, external validation using actual machinery data is required.

Second, instead of full run-to-failure temporal trajectories, the dataset offers machine-state observations. As a result, rather of anticipating degradation or remaining useful life, the current paradigm focuses on rare-failure classification and maintenance operating-point selection. Furthermore, because the dataset lacks information on downtime losses, inspection costs, component replacement, and production interruption, the FN-to-FP cost ratios used in the threshold analysis represent hypothetical relative decision scenarios rather than empirically estimated or monetary maintenance costs. Therefore, in order to ascertain whether the resulting operating thresholds offer quantifiable economic and reliability benefits, future research should test the framework using longitudinal industrial data and organization-specific maintenance costs.

CONCLUSION

The present study used the significantly unbalanced AI4I 2020 Predictive Maintenance Dataset to assess uncommon machine-failure prediction from classification and maintenance-decision perspectives. LightGBM outperformed the other baseline classifiers in terms of failure detection, and it was kept for study focused on maintenance. With an F1-score of 89.11%, a reduction in missed failures to six and false positives to five, and an increase in recall to 88.24%, random oversampling yielded the best imbalance-treatment result. A cost-sensitive threshold study showed a definite trade-off between false maintenance alarms and undetected failures. The selected threshold of 0.22 resulted in eight missed failures and nine false alarms for FN-to-FP cost ratios of 1:1 and 2:1, whereas a 10:1 ratio decreased missed failures to five but increased false alarms to 59. Tool wear, torque, and rotational speed were found to be the most significant predictors by SHAP analysis. Therefore, the results show that better failure sensitivity does not always translate into better maintenance choices. Predictive performance can be taken into account along with maintenance effects thanks to the suggested FN-FP operating-point strategy. Future studies should include experimentally established

maintenance costs and longitudinal industrial data to validate the approach.

REFERENCES

Al Dosari, F. H. M., & Abouellail, S. I. A. D. (2023). Artificial intelligence (AI) techniques for intelligent control systems in mechanical engineering. *American Journal of Smart Technology and Solutions, 2*(2), 55–64. https://doi.org/10.54536/ajsts.v2i2.2188

Cummins, L., Sommers, A., Ramezani, S. B., Mittal, S., Jabour, J., Seale, M., & Rahimi, S. (2024). Explainable predictive maintenance: A survey of current methods, challenges and opportunities. *IEEE Access, 12*, 57574–57602. https://doi.org/10.1109/ACCESS.2024.3391130

de Giorgio, A., Cola, G., & Wang, L. (2023). Systematic review of class imbalance problems in manufacturing. *Journal of Manufacturing Systems, 71*, 620–644. https://doi.org/10.1016/j.jmsy.2023.10.014

Mahale, Y., Kolhar, S., & More, A. S. (2025). Enhancing predictive maintenance in automotive industry: Addressing class imbalance using advanced machine learning techniques. *Discover Applied Sciences, 7*, 340. https://doi.org/10.1007/s42452-025-06827-3

Nicola, A. A. (2026). AI-driven factory robot remote control monitoring equipment using machine learning and cloud-based system. *American Journal of Smart Technology and Solutions, 5*(2), 18–28. https://doi.org/10.54536/ajsts.v5i2.8114

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

Owusu, D. K., Amuzuvi, C. K., & Attachie, J. C. (2026). Thermal fault analysis and optimisation of induction motors using current and temperature signals with artificial intelligence. *American Journal of Data Science and Artificial Intelligence, 2*(1), 33–43. https://doi.org/10.54536/ajdsai.v2i1.5278

Owusu, D. K., & Baidoo, C. E. (2026). AI-enabled prediction of power transformer remaining useful life using dissolved gas analysis and Random Forest regression. *American Journal of Innovation in Science and Engineering, 5*(1), 56–69. https://doi.org/10.54536/ajise.v5i1.6829

Qawqzeha, Y. (2025). Addressing class imbalance in IoT: A comparative analysis of resampling techniques. *American Journal of Smart Technology and Solutions, 4*(1), 16–24. https://doi.org/10.54536/ajsts.v4i1.2912

Shadi, M. R., Mirshekali, H., & Shaker, H. R. (2025). Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects. *Renewable and Sustainable Energy Reviews, 216*, 115668. https://doi.org/10.1016/j.rser.2025.115668