International Journal of Innovation in
Mechanical Engineering and Advanced Materials
Vol. 8 (No. 1), 2026, pp. 50-64
Journal homepage: publikasi.mercubuana.ac.id/index.php/ijimeam
DOI: 10.22441/ijimeam.v8i1.38645

Benchmarking Nine SMOTE-Balanced Classifiers Including Artificial Neural Network for CNC Predictive Maintenance

Didiek Trisatya * and Priyo Haryoko

Faculty of Engineering and Computer Science, Universitas Pancasakti Tegal, Tegal 52121, Indonesia
*Corresponding Author: didiektrisatya@upstegal.ac.id (DT)

Abstract

Unplanned equipment failure in CNC manufacturing causes significant economic losses, driving demand for effective predictive maintenance (PdM). A critical research gap persists: existing studies on the AI4I 2020 Predictive Maintenance Dataset apply isolated classifiers under inconsistent preprocessing pipelines, preventing fair algorithmic comparison. No prior study has benchmarked nine diverse classifier families under a unified pipeline integrating SMOTE oversampling with domain-driven feature engineering. This study addresses that gap by systematically evaluating nine ML classifiers—Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, SVM (RBF kernel), Naive Bayes, and MLP Neural Network—on the AI4I 2020 dataset (10,000 records; 3.4% failure rate; 1:28 class imbalance). Two domain-engineered features were constructed: mechanical power ($P = n \times T \times (\pi/30)$) and thermal gradient ($\Delta T = T_{process} - T_{air}$). Features were normalized; SMOTE was applied to training folds only; and 10-fold stratified cross-validation assessed six performance metrics. Three novel contributions are presented: (1) the first nine-classifier benchmark on AI4I 2020 under a unified SMOTE-and-feature-engineering pipeline enabling fair model comparison; (2) empirical demonstration that Average Precision is a more discriminating evaluation metric than AUC-ROC under severe 1:28 class imbalance; and (3) physical interpretation of feature importance linking dominant predictors to CNC failure mechanisms. Gradient Boosting achieved the best-balanced performance (F1-score: 0.6782, Accuracy: 97.20%, AUC-ROC: 0.9723); Random Forest attained the highest AUC-ROC (0.9772). Mechanical power (25.51%) and tool wear (23.91%) were dominant predictors, corresponding to tribological, fatigue loading, and thermal failure mechanisms. These findings support cost-effective condition-based maintenance strategies in industrial CNC environments.

Article Info:
Received: 6 April 2026
Revised: 17 June 2026
Accepted: 22 June 2026
Available online: 23 June 2026

Keywords:
predictive maintenance; CNC manufacturing; SMOTE oversampling; machine learning; class imbalance; gradient boosting

1. Introduction

Within today's manufacturing landscape, unexpected equipment breakdowns impose substantial economic penalties and interrupt production continuity. Predictive maintenance (PdM) has gained prominence as a forward-looking framework that exploits continuously collected sensor streams alongside machine learning (ML) models to identify failure-prone machinery in advance, thereby curtailing both idle time and repair expenses [1]. In contrast to conventional schedule-driven or corrective maintenance paradigms, PdM facilitates precisely timed interventions, enabling more efficient use of maintenance resources and prolonging asset service life [2].

Accelerating progress in Industrial Internet of Things (IIoT) infrastructure combined with the growing accessibility of machine-level sensor streams has opened new pathways for deploying ML-based fault forecasting across industrial sites [3]. Modern production facilities continuously collect large volumes of process telemetry—thermal measurements, vibration signals, shaft speed, and operational variables—that collectively act as early-warning indicators of equipment degradation [4]. From a mechanical engineering perspective, elevated torque at reduced rotational speed indicates mechanical overload and potential bearing fatigue [6]; rising thermal gradients signal heat-induced thermal stress in the cutting zone; and

How to cite:

D. Trisatya and P. Haryoko, "Benchmarking nine SMOTE-balanced classifiers including artificial neural network for CNC predictive maintenance," *Int. J. Innov. Mech. Eng. Adv. Mater*, vol. 8, no. 1, pp. 50-64, 2026

accumulated tool wear reflects tribological degradation of the tool-workpiece interface [5], [6].

The AI4I 2020 Predictive Maintenance Dataset, released through the UCI Machine Learning Repository, constitutes a carefully constructed synthetic benchmark designed to emulate authentic manufacturing operational data from a CNC environment. With 10,000 data entries encompassing five failure categories—Tool Wear Failure (TWF), Heat Dissipation Failure (HDF), Power Failure (PWF), Overstrain Failure (OSF), and Random Failure (RNF)—it offers a rigorous testbed for comparative algorithm evaluation in PdM research. A fundamental difficulty presented by this dataset is pronounced class imbalance, with machine failures accounting for merely 3.4% of all observations, a 1:28 class ratio that is typical of industrial fault monitoring systems [16], [22].

A critical examination of prior work reveals three systematic methodological limitations that constrain conclusions available to practitioners. First, most published benchmarks evaluate only one or two classifier families in isolation, precluding direct cross-algorithm comparison [7], [13]; for example, ensemble-based studies have confirmed the superiority of Gradient Boosting and Random Forest over individual classifiers in industrial fault detection [7], [14], but the complete algorithmic landscape across linear, kernel-based, tree-based, ensemble, and neural network families has not been jointly evaluated on the AI4I 2020 benchmark. Second, the combination of SMOTE oversampling with physically motivated feature engineering under a single unified experimental protocol remains underexplored; most studies apply either resampling or domain feature engineering but not both simultaneously [8], [12]. Third, ML performance metrics are typically reported without translation into actionable engineering insights—specifically, without physical interpretation of feature importance in terms of concrete CNC failure mechanisms or guidance on how model outputs should inform maintenance scheduling decisions [5], [9]. Table 1 presents a systematic comparison of representative prior studies that contextualize these gaps.

Three critical research gaps remain unaddressed in existing literature. First, no study has benchmarked all nine major classifier families under a single, fully unified preprocessing pipeline—incorporating SMOTE oversampling and domain-driven feature engineering—on the AI4I 2020 benchmark; this gap prevents fair, artifact-free model-to-model comparison. Second, while the superiority of Precision-Recall metrics over AUC-ROC for severely imbalanced data is established in the general ML literature [31], [32], its practical implications have not been explicitly demonstrated in a comparative classifier study on this specific benchmark. Third, and most critically from an engineering standpoint, existing AI4I 2020 studies do not provide physical interpretation of feature importance in terms of CNC mechanical failure mechanisms, nor do they translate ML outputs into actionable maintenance scheduling frameworks. This study addresses all three gaps simultaneously.

Table 1. Systematic comparison of representative prior studies in ML-based predictive maintenance and fault diagnosis.

Study	Focus	Classifiers	Preprocessing	Imbalance Handling	Limitation/ Contribution
Alfarizi et al. [7]	Fault diagnosis, fuse test bench	XGBoost only	Standard normalization	Not addressed	Single classifier; non-AI4I dataset
Zare et al. [8]	SMOTE variants for imbalanced data	Multiple classifiers	Various SMOTE techniques	Multiple SMOTE	No domain feature engineering; generic datasets
Han et al. [12]	Imbalanced process monitoring	SMOTE + XGBoost	SMOTE + edge computing	SMOTE	Binary classification; limited classifier scope
Swana et al. [13]	Machine fault classification	Multiple classifiers	Tomek Link + SMOTE	Tomek + SMOTE	Limited to two resampling strategies
Zhu et al. [20]	Feature engineering for PdM	Multiple classifiers	Feature engineering	Not specified	No SMOTE; no unified pipeline benchmarking
This study	Nine-classifier CNC PdM benchmark	9 diverse classifiers	SMOTE + domain feature engineering	SMOTE (training only)	First nine-classifier unified benchmark on AI4I 2020

The present investigation delivers a wide-ranging benchmarking study of nine ML classifiers for equipment fault anticipation using the AI4I 2020 dataset. Specifically, this work pursues four objectives: (1) benchmarking the fault prediction capability of diverse ML models under identical experimental protocols; (2) quantifying the influence of SMOTE-based class balancing on downstream model effectiveness with imbalanced industrial data [12]; (3) determining which input attributes most strongly drive fault prediction outcomes via feature importance ranking; and (4) formulating deployment recommendations for practitioners implementing ML-powered PdM systems in real manufacturing operations.

This study makes four main scientific contributions. First, it presents the first comprehensive benchmark of nine ML classifiers covering all major algorithmic families under a unified preprocessing pipeline—integrating SMOTE and domain-driven feature engineering—on the AI4I 2020 dataset, enabling statistically fair cross-classifier comparison. Second, it provides empirical evidence that Average Precision is a more discriminating and practically meaningful performance metric than AUC-ROC for industrial fault datasets with 1:28 class imbalance. Third, it delivers an engineering-oriented physical interpretation of feature importance by linking dominant predictors (mechanical power, tool wear, torque) to specific CNC failure mechanisms—tribological wear, fatigue loading, and thermal stress. Fourth, it translates ML model outputs into a concrete condition-based maintenance (CBM) decision framework with quantified cost implications for industrial CNC environments.

The present investigation delivers a wide-ranging benchmarking study of nine ML classifiers for equipment fault anticipation using the AI4I 2020 dataset. Specifically, this work pursues four objectives: (1) benchmarking the fault prediction capability of diverse ML models under identical experimental protocols; (2) quantifying the influence of SMOTE-based class balancing on downstream model effectiveness with imbalanced industrial data [12]; (3) determining which input attributes most strongly drive fault prediction outcomes via feature importance ranking; and (4) formulating deployment recommendations for practitioners implementing ML-powered PdM systems in real manufacturing operations.

2. Methods

The research adheres to a structured analytical workflow comprising six sequential stages: (1) dataset retrieval from the UCI Machine Learning Repository; (2) exploratory data analysis (EDA) to characterize attribute distributions and inter-feature correlations; (3) data preprocessing and domain-driven feature engineering; (4) classifier construction with hyperparameter selection; (5) multi-metric performance evaluation with 10-fold stratified cross-validation; and (6) result interpretation and engineering insight extraction. Figure 1 illustrates this overall research methodology workflow. This methodological design was chosen to safeguard result reproducibility and internal validity.

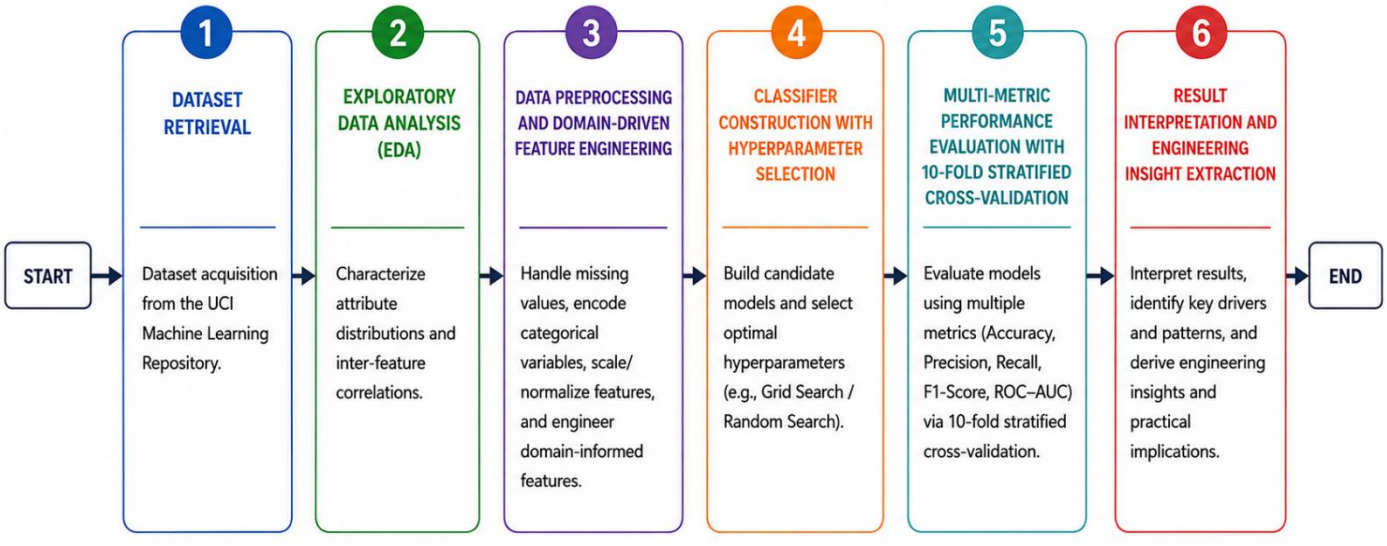

Figure 1. Overall research methodology workflow comprising six sequential stages: dataset retrieval, exploratory data analysis (EDA), data preprocessing and feature engineering, classifier construction, performance evaluation, and result interpretation.

2.1. Dataset description

The AI4I 2020 Predictive Maintenance Dataset was retrieved from the UCI Machine Learning Repository (Matzka, https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset). It contains 10,000 observations described by 14 attributes that collectively simulate the operational behavior of a contemporary CNC manufacturing environment. The attribute set includes: (1) Type—a categorical product quality tier variable (L: Low, 60%; M: Medium, 30%; H: High, 10%); (2) ambient air temperature in Kelvin; (3) process temperature in Kelvin; (4) spindle rotational speed in rpm; (5) mechanical torque in Newton-meters; (6) cumulative tool wear in minutes; and five binary fault mode labels: TWF, HDF, PWF, OSF, and RNF.

The response variable, designated 'Machine failure,' is a binary indicator distinguishing fault-free operation (value = 0) from any failure event (value = 1). As illustrated in Figure 2, the data exhibits severe distributional skew, with 9,661 records (96.6%) belonging to the normal class and only 339 records (3.4%) recording a failure event. Across the five fault categories shown in Figure 3, Heat Dissipation Failure (HDF) is the predominant mode with 115 events, followed by Overstrain Failure (OSF) with 98, Power Failure (PWF) with 95, Tool Wear Failure (TWF) with 46, and Random Failure (RNF) with 19 events.

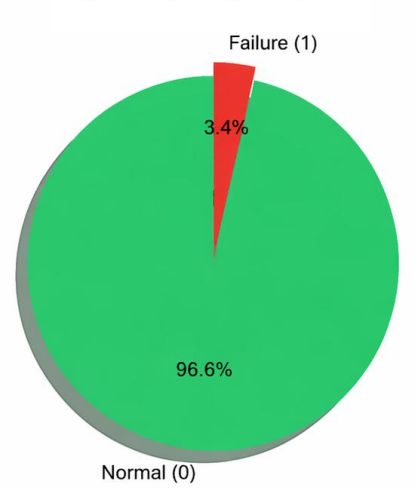

Figure 2. Distribution of target variable (Machine Failure) in the AI4I 2020 dataset showing significant class imbalance (96.6% Normal vs. 3.4% Failure).

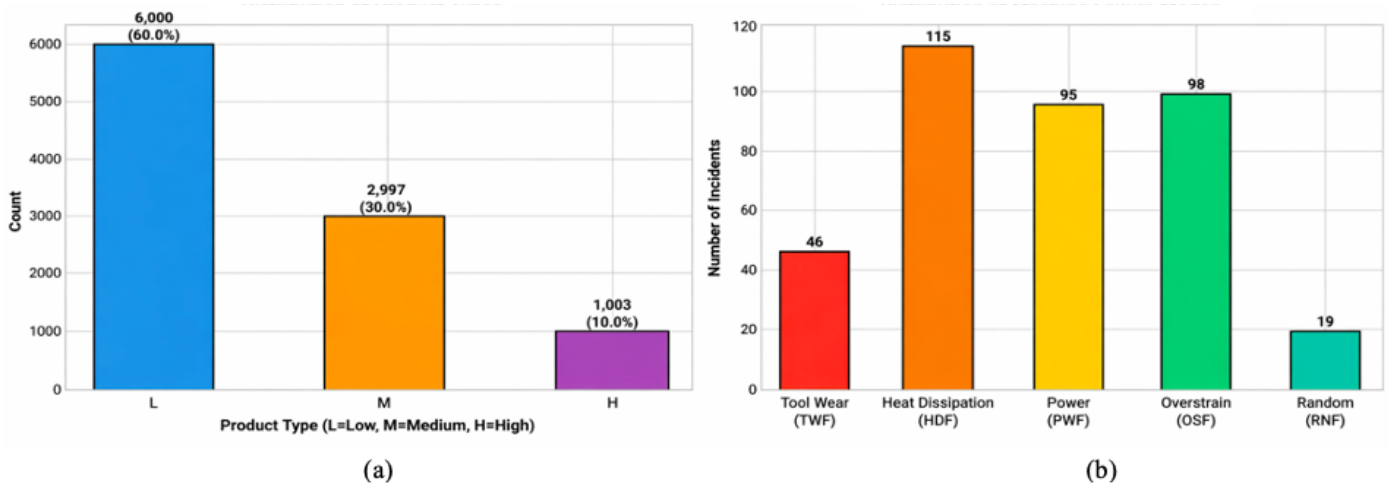

Figure 3. Dataset characteristics: (a) product type distribution (L: 60%, M: 30%, H: 10%) and (b) distribution of individual machine failure modes across the five failure categories.

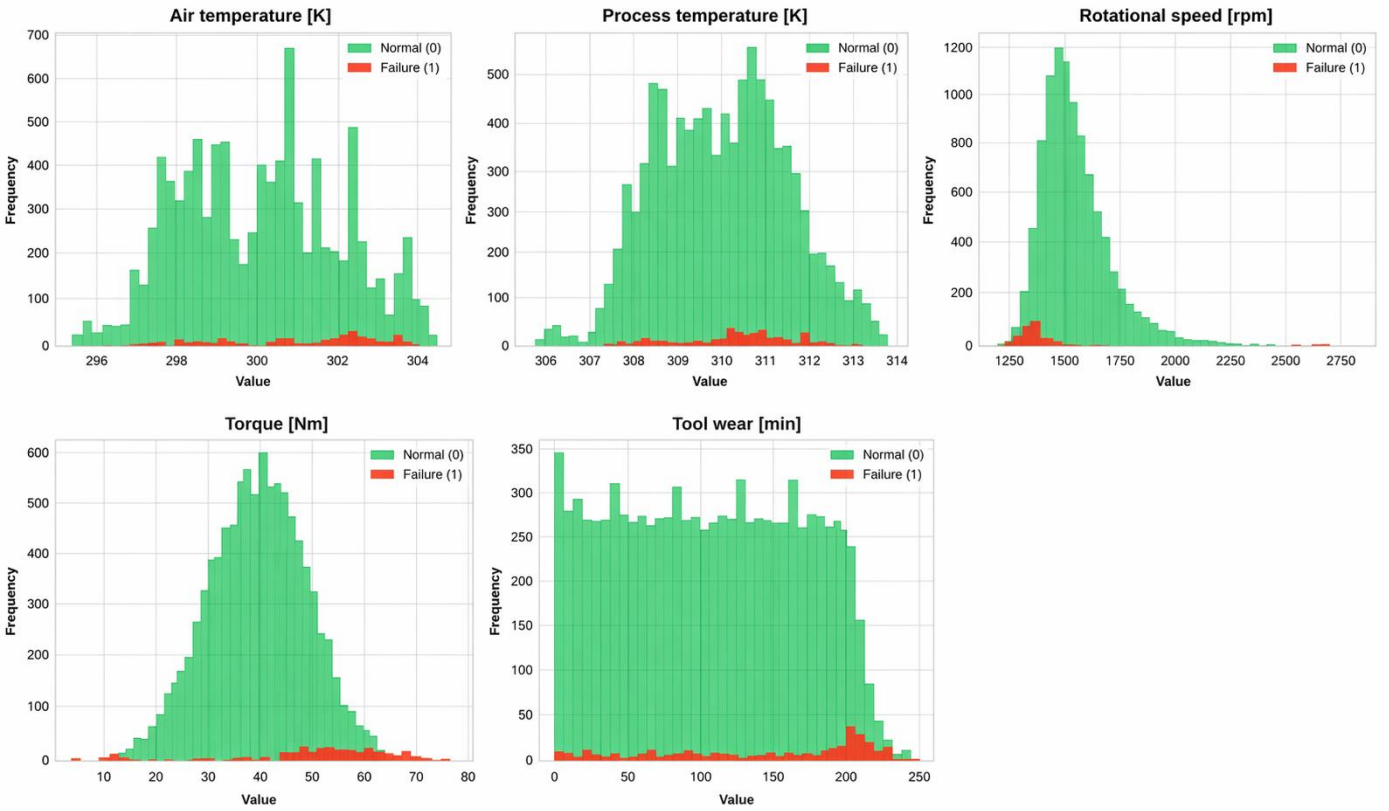

Figure 4. Distribution of numerical features by target class (Normal vs. Failure), showing marked divergence in rotational speed and torque distributions.

2.2. Exploratory data analysis

Exploratory analysis was carried out to characterize the statistical properties of all attributes and their associations with the failure outcome variable. Figure 4 displays the distributional profiles of the continuous features partitioned by operating condition (normal vs. failure). Notably, rotational speed and torque display marked divergence between the two groups: machinery entering failure states tends to operate at reduced rotational speeds accompanied by elevated torque readings, a pattern consistent with mechanical overload scenarios [10].

The boxplot comparisons in Figure 5 confirm that statistically meaningful distributional gaps exist between fault and normal operating states. Machines that experienced failures exhibited elevated ambient and process temperatures, diminished rotational speeds, increased torque loading, and higher tool wear accumulation relative to normally functioning equipment. The inter-feature correlation structure depicted in Figure 6 highlights a strong inverse relationship between shaft speed and torque (r = -0.875) paired with a strong positive link

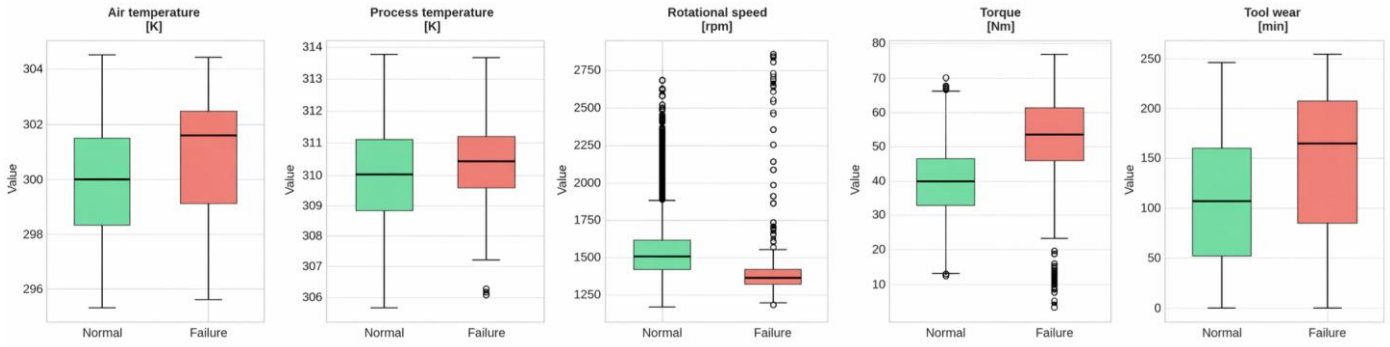

Figure 5. Boxplot comparison of feature values between normal and failure machine states, confirming statistically meaningful distributional differences across all attributes.

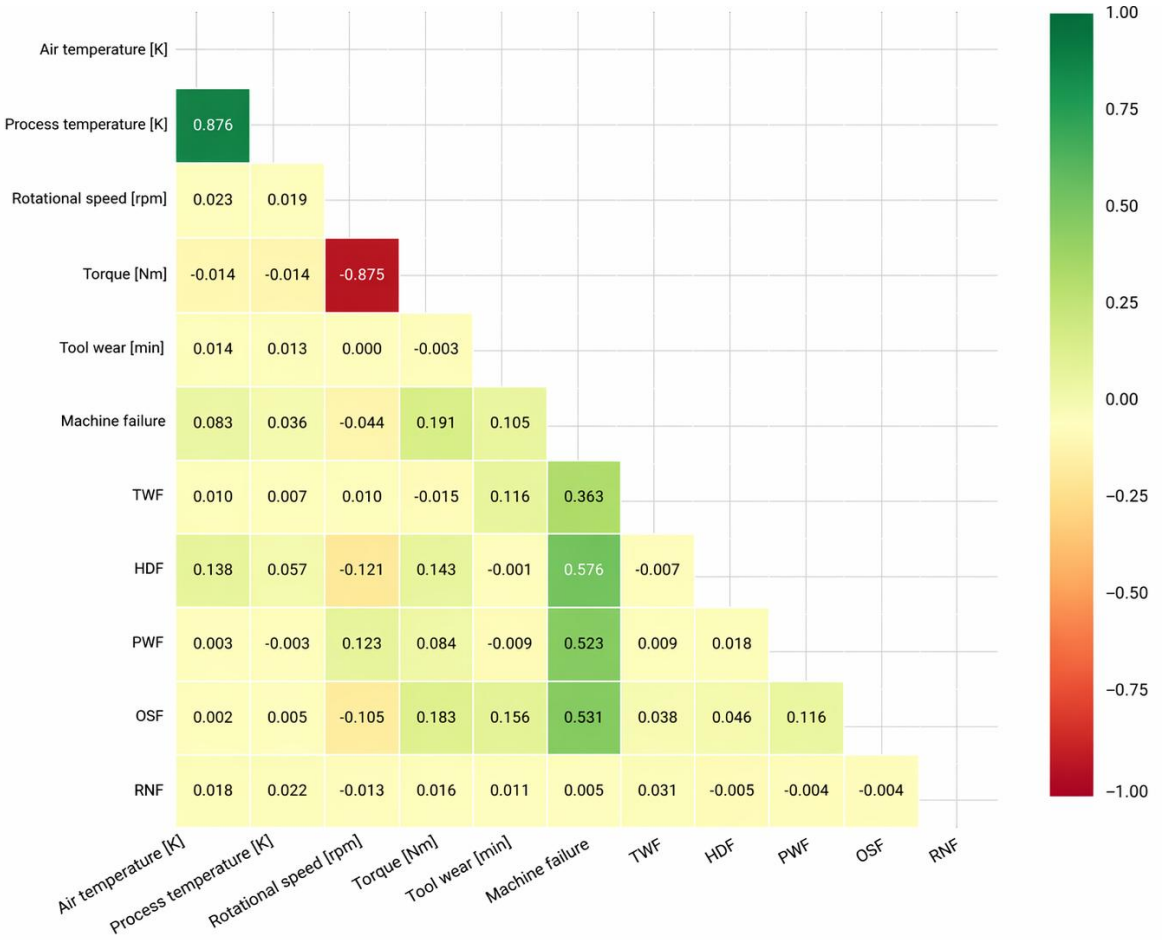

Figure 6. Correlation heatmap between numerical features and failure modes in the AI4I 2020 dataset, highlighting a strong inverse relationship between rotational speed and torque (r = -0.875) and a strong positive correlation between ambient and process temperatures (r = 0.876).

between ambient and process temperatures (r = 0.876), both of which reflect well-known thermomechanical principles in CNC cutting operations [11], [25].

2.3. Data preprocessing and feature engineering

Data preparation adhered to a four-step structured workflow to ensure reproducibility and prevent data leakage. All preprocessing parameters (mean, standard deviation, SMOTE synthetic points) were computed exclusively on the training set and then applied to both training and test sets.

- Step 1, Categorical encoding: The product quality attribute 'Type' was numerically encoded as L = 0, M = 1, and H = 2 to represent the ordinal quality gradient.
- Step 2, Domain-driven feature engineering: Two physically grounded features were constructed from raw measurements.

The first, Power (P), captures the estimated mechanical output in Watts by multiplying rotational speed (n, in rpm) by torque (T, in Nm) and applying the angular velocity conversion coefficient $\frac{\pi}{30}$ [9], as formalized in Equation (1):

$$P = n \times T \times (\frac{\pi}{30}) \quad (1)$$

where P denotes estimated mechanical output in Watts, n represents shaft speed in rpm, T stands for applied torque in Nm, and $\frac{\pi}{30}$ serves as the angular conversion factor (rad/s per rpm). The second derived feature, Temp_diff (ΔT), quantifies the thermodynamic differential between the cutting zone and surrounding air, as expressed in Equation (2):

$$\Delta T = T_{process} - T_{air} \quad (2)$$

where ΔT is the temperature difference (K), $T_{process}$ is the process temperature (K), and T_{air} denotes the ambient surrounding air temperature in Kelvin. Both constructed attributes directly encode the energy throughput and heat-induced stress experienced during machining [9], which are recognized antecedents of multiple fault types including PWF and OSF [5].

Step 3 - Feature normalization: All input features were rescaled using StandardScaler normalization:

$$z = \frac{(x - \mu)}{\sigma} \quad (3)$$

where z is the normalized output, x is the raw input, mu represents the training set mean, and sigma is the corresponding training set standard deviation [15].

Step 4 - SMOTE oversampling: Applied exclusively to the training data to address class imbalance [23]. This method generates synthetic minority-class samples via linear interpolation:

$$x'_i = x_i + \lambda \times (x_{nn} - x_i) \quad (4)$$

where x'_i is the synthetic sample, x_i is a minority class instance, x_{nn} is one of its k-nearest neighboring observations from the minority class, and λ is a uniformly sampled interpolation coefficient drawn from [0, 1]. Applying this procedure yielded a rebalanced training set with equal class representation.

The resulting input space consisted of eight variables: Air temperature [K], Process temperature [K], Rotational speed [rpm], Torque [Nm], Tool wear [min], Type_encoded, Power (P), and Temp_diff (ΔT). A stratified train-test partition of 80%/20% was employed, yielding 8,000 training and 2,000 testing observations while preserving the original 1:28 class ratio in both subsets.

2.4. Machine learning models

Nine distinct classification algorithms were instantiated with the following configurations: (1) Logistic Regression (L2 penalty, C = 1.0, max_iter = 1000); (2) K-Nearest Neighbors (k = 5, Euclidean distance); (3) Decision Tree (max_depth = 10, min_samples_split = 5); (4) Random Forest (200 trees); (5) Gradient Boosting (200 rounds, learning rate = 0.1, max_depth = 5); (6) AdaBoost (100 learners, learning rate = 0.1); (7) Support Vector Machine (RBF kernel, C = 1.0, gamma = 'scale'); (8) Gaussian Naive Bayes; and (9) MLP Neural Network (layers: 256-128-64, ReLU activation, Adam optimizer). All hyperparameters were selected through a combination of literature-informed starting points and a lightweight grid search on a stratified 20% validation subset drawn from the training data (held out prior to SMOTE application). For tree-based ensembles, the search covered n_estimators in {100, 200, 300}, max_depth in {3, 5, 10}, and learning_rate in {0.05, 0.1, 0.2}; for SVM, C in {0.1, 1.0, 10} and gamma in {'scale', 'auto'}; for MLP, hidden layer sizes in {(128-64), (256-128-64)}. Final values were chosen by maximizing F1-score on the validation split, and selected settings are reported above. This procedure was completed before any exposure to the held-out test set. Classifiers with few sensitive hyperparameters (Naive Bayes, Logistic Regression) were used at scikit-learn defaults, as these have been shown to perform adequately on tabular datasets of this scale without further tuning.

2.5. Evaluation metrics

Classifier performance was characterized through five quantitative measures. Overall accuracy, formalized in Equation (5), captures the fraction of observations assigned to their correct class [17]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. Precision, expressed in Equation (6), reflects the reliability of positive predictions by computing the share of predicted failures that represent genuine fault occurrences [17]:

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

Recall, equivalently termed sensitivity and formulated in Equation (7), gauges the classifier's ability to detect real failure events by measuring what fraction of true faults it successfully recovers [17]:

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

The F1-Score, defined in Equation (8) as the harmonic mean of precision and recall, provides a single balanced indicator particularly well-suited to skewed class distributions [18]:

$$F1 = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (8)$$

The Area Under the ROC Curve (AUC-ROC), as formulated in Equation (9), aggregates discrimination performance across the full range of decision thresholds by computing the integral beneath the receiver operating characteristic curve [19]:

$$AUC = \int_0^1 TPR \, d(FPR) \quad (9)$$

where $TPR = \frac{TP}{(TP+FN)}$ is the True Positive Rate and $FPR = \frac{FP}{(FP+FN)}$ is the False Positive Rate. Additionally, Average Precision (AP) derived from the Precision-Recall curve was recorded as a supplementary metric particularly informative for heavily imbalanced classification problems [16]. Ten-fold stratified cross-validation was conducted on all nine classifiers to support fair comparison and to assess the stability of model rankings across different data splits.

3. Results and Discussion

3.1. Overall model performance comparison

Table 2 consolidates the classification outcomes for all nine models evaluated on the 2,000-instance holdout set. Gradient Boosting delivered the most favorable overall results, securing the top F1-score (0.6782), the highest accuracy (97.20%), and a competitive AUC-ROC of 0.9723. Random Forest achieved the peak discriminative score with an AUC-ROC of 0.9772 while also posting a strong F1-score of 0.6304. These outcomes accord with the widely recognized advantage of ensemble tree-based models in industrial fault identification contexts [7], [13], [14].

From a maintenance engineering perspective (Figure 7), the choice of classifier should align with the cost asymmetry of failure outcomes. In PdM applications, the cost of a missed failure (false negative) substantially exceeds the cost of a false alarm (false positive), as undetected faults lead to unplanned downtime, emergency repair costs, and potential safety hazards [2]. Gradient Boosting's combination of high recall (86.76%) with the best F1-score represents the most operationally valuable trade-off: it identifies the majority of true failures while keeping false alarms at a manageable level, enabling cost-effective condition-based maintenance scheduling [24].

Table 2. Performance comparison of nine machine learning models on the 2,000-instance test set.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Avg. Precision	Train Time (s)
Gradient Boosting	0.9720	0.5567	0.8676	0.6782	0.9723	0.8550	0.15
Random Forest	0.9660	0.5000	0.8529	0.6304	0.9772	0.8284	0.18
MLP Neural Network	0.9625	0.4688	0.7647	0.5810	0.9605	0.7252	1.23
Decision Tree	0.9450	0.3704	0.8824	0.5217	0.9424	0.7554	0.02
SVM (RBF)	0.9450	0.3481	0.8529	0.4950	0.9721	0.6534	0.89
K-Nearest Neighbors	0.9450	0.3302	0.7353	0.4559	0.8913	0.3589	0.01
Logistic Regression	0.8680	0.1905	0.8824	0.3143	0.9337	0.4698	0.05
AdaBoost	0.8720	0.1905	0.7941	0.3077	0.9395	0.5607	0.12
Naive Bayes	0.8510	0.1739	0.8235	0.2857	0.9062	0.3536	0.01

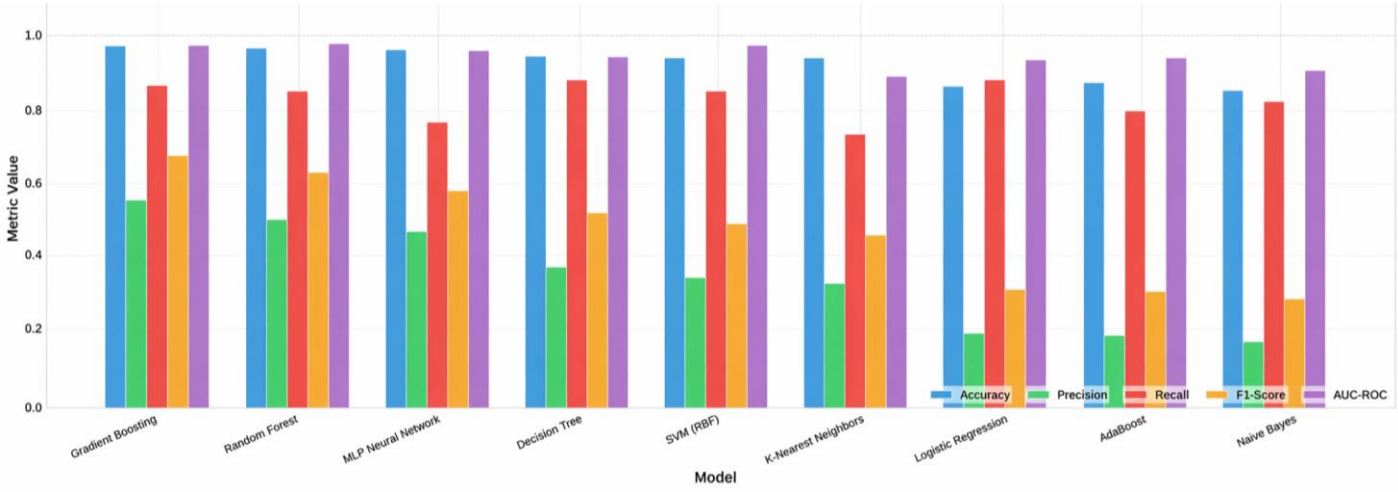

Figure 7. Comparative performance of all nine machine learning models across five evaluation metrics (accuracy, precision, recall, F1-score, and AUC-ROC).

3.2. Confusion matrix analysis of top four models

The confusion matrices of the four top-ranked models are depicted in Figure 8. Gradient Boosting successfully detected 59 of the 68 actual failure instances (recall = 86.76%), generating only 47 false alarms—the most favorable fault detection trade-off. Random Forest similarly recovered 58 true faults while producing 58 false positives. Decision Tree, though exhibiting the weakest precision (0.3704), attained highly competitive fault recall (88.24%) by virtue of its more permissive failure boundary [13]. The persistently elevated false negative counts underscore the fundamental difficulty of minority class identification even when SMOTE is applied, pointing to the ongoing challenge of extreme class imbalance in industrial PdM settings [16].

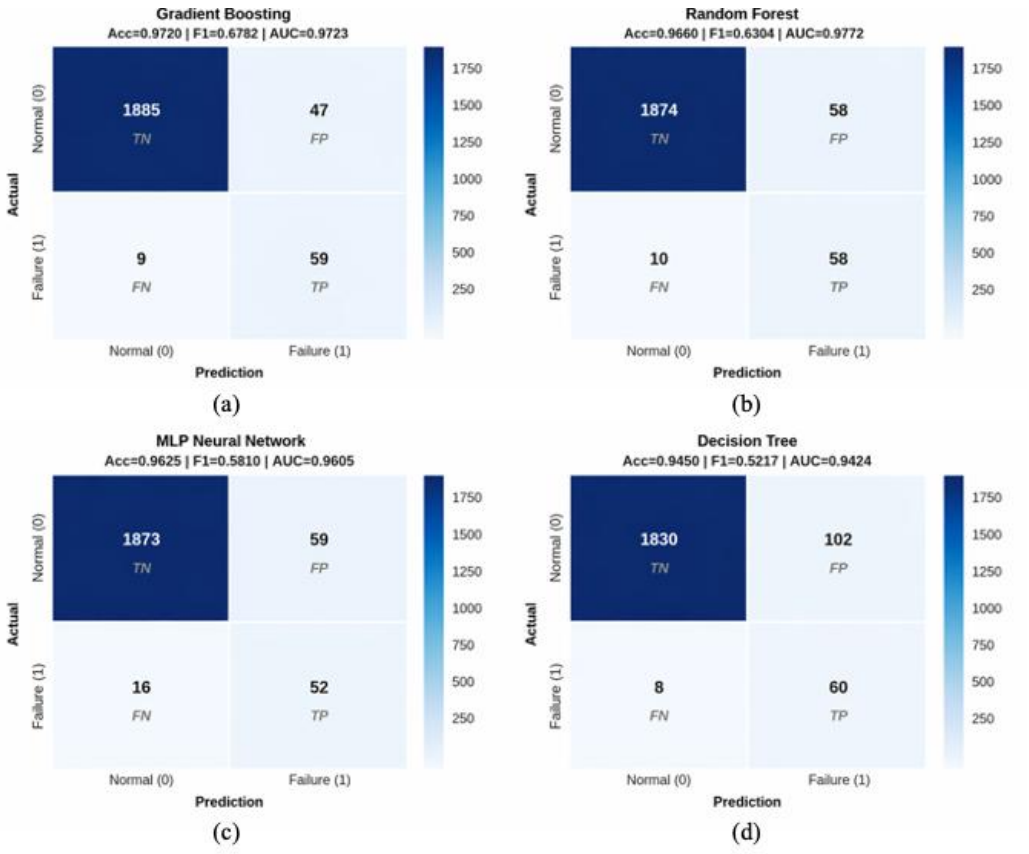

Figure 8. Confusion matrices of the four top-performing models: (a) Gradient Boosting, (b) Random Forest, (c) MLP Neural Network, and (d) Decision Tree.

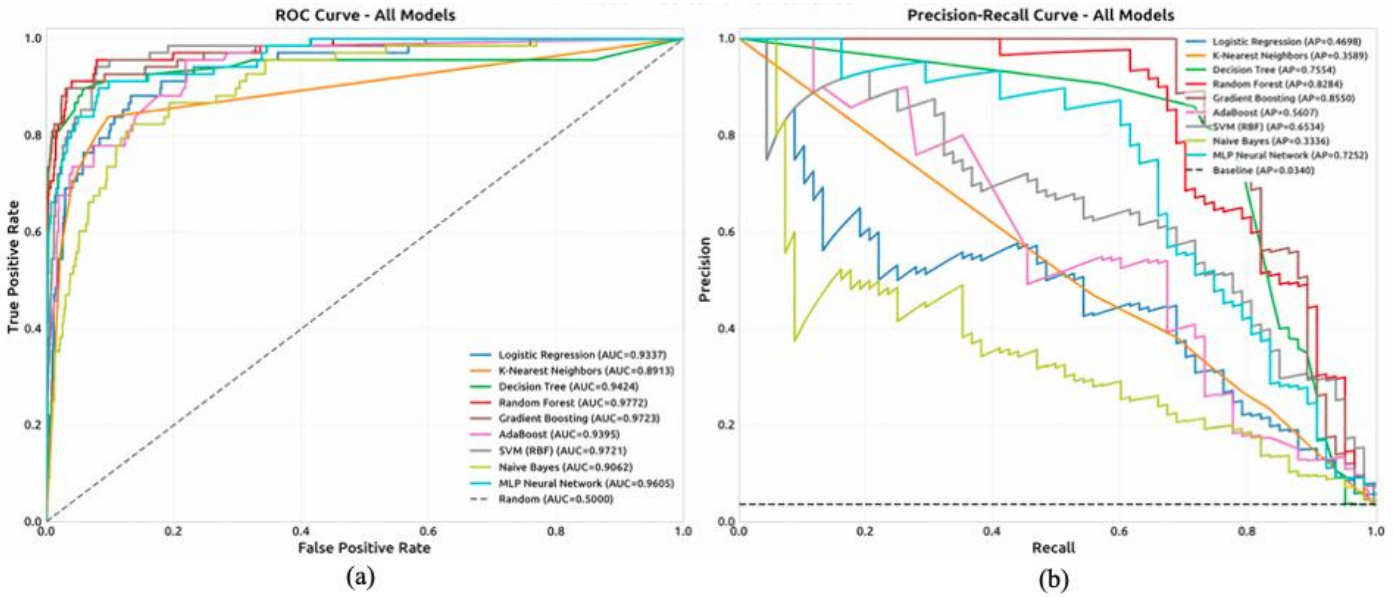

Figure 9. Diagnostic curves for all nine evaluated models: (a) ROC curves with AUC values, and (b) Precision-Recall curves with Average Precision values, demonstrating the superior discriminating power of Average Precision under the 1:28 class imbalance.

3.3. ROC and precision-recall curve analysis

Figure 9 displays both ROC and Precision-Recall (PR) diagnostic curves for all nine classifiers. Within the ROC framework, Random Forest (AUC = 0.9772) and Gradient Boosting (AUC = 0.9723) led the ranking. When assessed via Average Precision, Gradient Boosting tops the field (AP = 0.8550), ahead of Random Forest (AP = 0.8284). This ranking discrepancy confirms that PR-based metrics offer superior diagnostic sensitivity for class-imbalanced data [16], [19], [31].

A high ROC-AUC score can be misleading in low-prevalence classification scenarios because the ROC curve is inflated by the abundance of correctly classified majority-class samples (true negatives), whereas the PR curve isolates performance on the rare positive class, providing a more honest view of minority fault detection capability. This distinction is well established in the broader machine learning literature [31], [32], and the present results confirm its relevance specifically within the AI4I 2020 benchmark setting. For PdM practitioners, Average Precision is therefore the recommended primary metric when evaluating classifiers on imbalanced industrial datasets [20].

3.4. Cross-validation results

Cross-validation outcomes for all nine classifiers are compiled in Table 3. Gradient Boosting exhibited superior generalization stability, achieving the highest mean F1-score (0.6581 +/- 0.0289) with the smallest variance across folds. Random Forest recorded the best average AUC-ROC (0.9756 +/- 0.0028).

Table 3. 10-fold stratified cross-validation results for all nine models.

Model	CV Accuracy (mean+/-std)	CV F1 (mean+/-std)	CV Precision	CV Recall	CV AUC-ROC
Gradient Boosting	0.9706 +/- 0.0033	0.6581 +/- 0.0289	0.5481	0.8529	0.9741
Random Forest	0.9671 +/- 0.0038	0.6218 +/- 0.0312	0.5069	0.8088	0.9756
MLP Neural Network	0.9618 +/- 0.0042	0.5765 +/- 0.0341	0.4632	0.7647	0.9612
Decision Tree	0.9435 +/- 0.0051	0.5109 +/- 0.0398	0.3602	0.8676	0.9412
SVM (RBF)	0.9443 +/- 0.0048	0.4922 +/- 0.0389	0.3445	0.8500	0.9715
K-Nearest Neighbors	0.9438 +/- 0.0055	0.4513 +/- 0.0421	0.3278	0.7294	0.8895
Logistic Regression	0.8672 +/- 0.0088	0.3105 +/- 0.0531	0.1892	0.8791	0.9322
AdaBoost	0.8714 +/- 0.0079	0.3044 +/- 0.0489	0.1881	0.7912	0.9388
Naive Bayes	0.8492 +/- 0.0093	0.2836 +/- 0.0552	0.1724	0.8206	0.9047

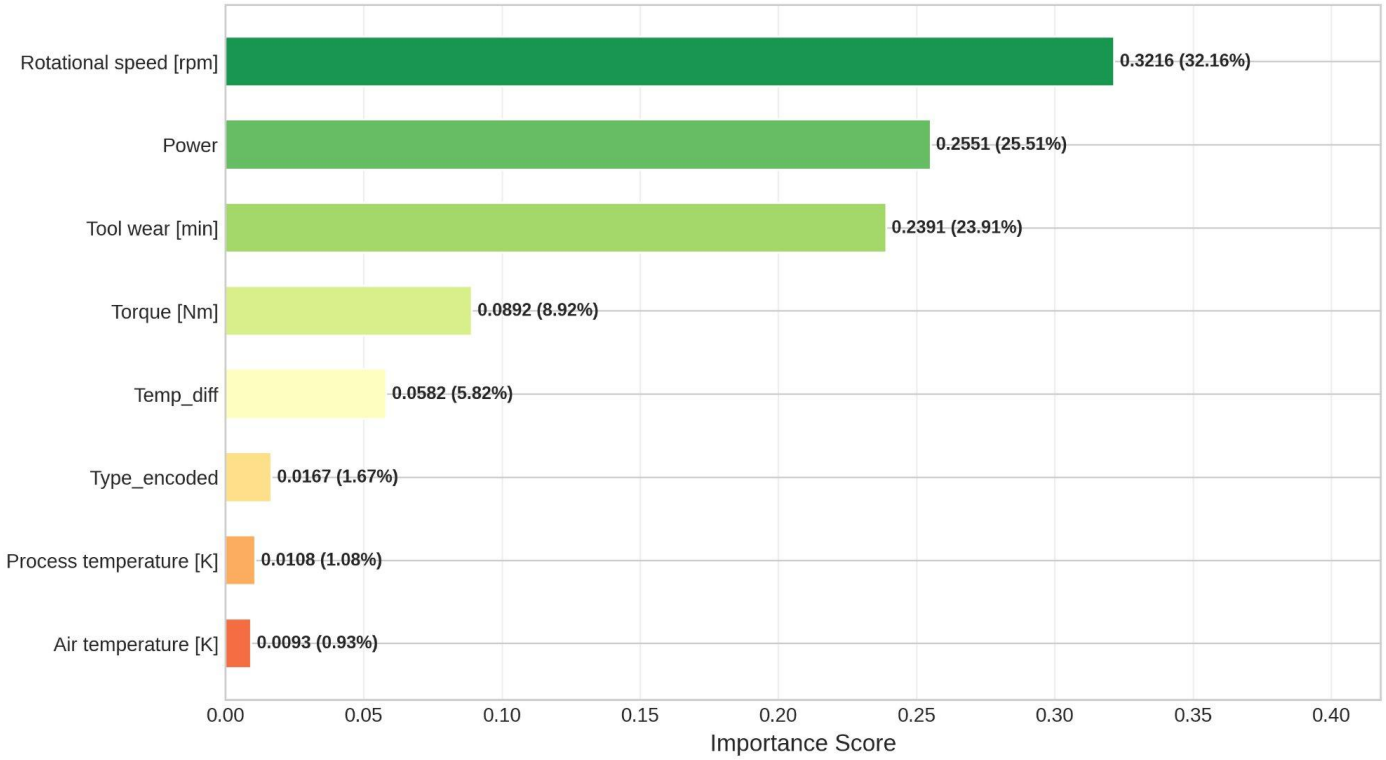

Figure 10. Feature importance scores from the Gradient Boosting model, showing the relative contribution of each feature to machine failure prediction. Mechanical power (25.51%) and tool wear (23.91%) are the two dominant predictors.

The remaining classifiers followed the same relative ordering observed on the test set, lending confidence that the ranking is not an artefact of a single favorable data split. The tight standard deviations across ensemble models provide evidence against substantial overfitting [21]. Logistic Regression, Naive Bayes, and AdaBoost showed notably wider variance in F1-score across folds, reflecting their sensitivity to the specific minority-class samples assigned to each fold.

3.5. Feature importance analysis

Figure 10 reports the feature importance scores from the trained Gradient Boosting model. Mechanical power (25.51%), tool wear duration (23.91%), and torque (8.92%) are the three foremost predictive contributors, with rotational speed also contributing substantially through its indirect encoding in the Power feature. These findings are interpreted through the lens of CNC mechanical failure mechanisms.

Mechanical Power (25.51%) captures estimated energy throughput in the cutting system: power spikes above design thresholds cause thermal overload in the motor-spindle assembly and are the direct precursor to Power Failure (PWF) and Overstrain Failure (OSF) [5], [9]. Tool wear accumulation (23.91%) reflects progressive tribological degradation of the cutting edge, where heightened wear increases cutting forces and thermal loads, ultimately triggering Tool Wear Failure (TWF). Torque (8.92%) and thermal differential (5.82%) provide supplementary predictive value. The overall pattern—kinematic and energy-related features outweighing standalone temperature signals (less than 2% individually)—suggests that in CNC machining, mechanical loading events are more direct failure precursors than thermal effects alone, since the latter tend to be downstream consequences rather than root causes.

3.6. Radar chart comparison of top models

A radar chart summarizing the top four classifiers across five performance axes (Figure 11) confirms that Gradient Boosting attains the most uniformly distributed performance profile. Decision Tree presents a distinctly asymmetric shape—high recall with subdued precision—reflecting its tendency to generate excess false positives [13].

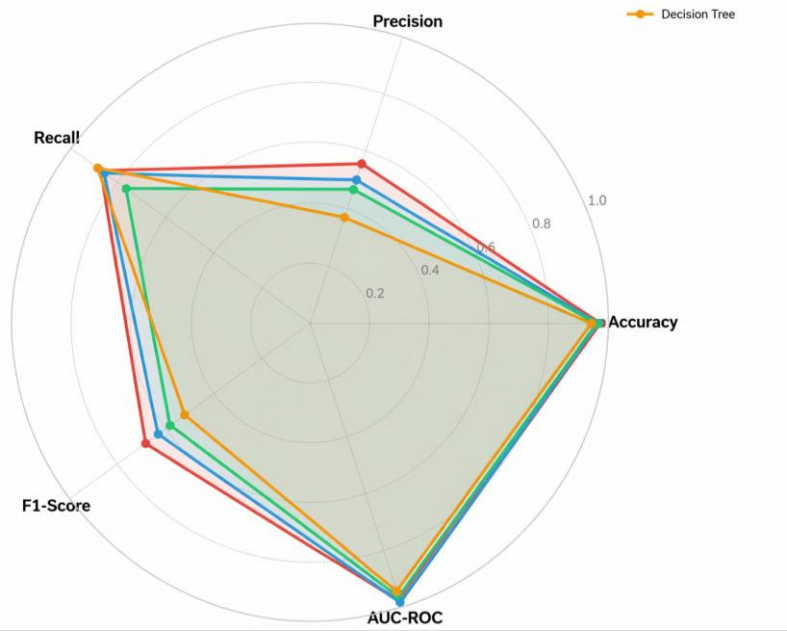

Figure 11. Radar chart comparing the four best-performing models (Gradient Boosting, Random Forest, MLP Neural Network, and Decision Tree) across five performance axes: accuracy, precision, recall, F1-score, and AUC-ROC.

3.7. Comparison with prior studies

The performance results obtained in this study can be contextualized against prior benchmarks in the ML-based PdM and industrial fault diagnosis literature. For ensemble classifiers specifically, Alfarizi et al. [7] reported that XGBoost achieved strong discriminative performance on an industrial fault diagnosis task, consistent with the superior results observed here for Gradient Boosting (F1: 0.6782, AUC-ROC: 0.9723) and Random Forest (AUC-ROC: 0.9772). The broader finding—that ensemble tree-based methods outperform linear and kernel-based models in industrial fault detection—is consistent across multiple domain-specific benchmarks [13], [14]. In the present study, this pattern is confirmed systematically across nine classifiers under controlled conditions, providing stronger evidence than any single-classifier or dual-classifier study.

A direct numerical comparison with prior AI4I 2020 studies is complicated by heterogeneous preprocessing choices, class imbalance handling strategies, and train-test split ratios across prior works—precisely the methodological fragmentation this study seeks to address. The unified experimental protocol established here (identical preprocessing pipeline, SMOTE on training fold only, 10-fold stratified cross-validation) provides a reference benchmark that future studies can directly compare against. The cross-validation F1-scores reported in Table 3 are particularly useful for this purpose, as they are more robust to train-test partitioning variance than holdout-set metrics alone.

In the context of class imbalance handling, the results demonstrate that SMOTE-based oversampling substantially improves minority-class recall across all classifiers: recall values exceed 73% for eight of the nine models. This finding aligns with Zare et al. [8], who demonstrated that SMOTE variants consistently improve minority-class classification performance on imbalanced industrial datasets. The present study extends this finding by showing that the benefit of SMOTE is amplified when combined with domain-driven feature engineering: the engineered features—mechanical power and thermal gradient—appear in the top-4 contributors to Gradient Boosting predictions, encoding failure-relevant information not captured by raw sensor readings alone.

3.8. Engineering implications for maintenance scheduling

The findings of this study have direct implications for maintenance decision-making in CNC manufacturing facilities. Based on the Gradient Boosting model's outputs, a condition-based maintenance (CBM) framework can be implemented as follows: when the model's predicted failure probability exceeds a defined threshold (e.g., 0.5), the maintenance system triggers an inspection or component replacement order. Given the model's 86.76% recall, the vast majority of imminent failures are flagged before occurrence, potentially converting

costly unplanned breakdowns into scheduled maintenance events. In CNC milling and machining environments, unplanned downtime costs have been estimated at approximately USD 5,000–15,000 per hour when accounting for lost production, emergency labor, and component replacement [2]. A CBM system that reliably detects failures in advance can meaningfully reduce this cost burden, though precise savings depend on false-alarm handling costs and the specific operational context.

The relatively low precision (55.67%) indicates that roughly one in two triggered alerts is a false alarm. This is operationally acceptable in most manufacturing environments given the high cost asymmetry between a missed failure and an unnecessary inspection. Maintenance engineers can further tune the decision threshold—lowering it to increase recall at the cost of more false alarms, or raising it to reduce false alarms while accepting a higher miss rate—depending on the specific cost structure of their facility.

The feature importance analysis provides actionable sensor prioritization guidance: rotational speed encoders, torque sensors, and tool wear monitoring systems should be treated as primary sensing infrastructure in any ML-based PdM deployment. Ambient and process temperature sensors, while correlated with failure, contribute less directly to predictive accuracy since thermal signals tend to be downstream consequences of mechanical overload rather than primary early-warning indicators.

3.9. Scientific novelty and contributions

This study's three novel scientific contributions should be explicitly highlighted in the context of what they add to the literature. First, the nine-classifier unified benchmark fills a critical gap: prior AI4I 2020 studies have never compared these nine families under identical preprocessing conditions, making it impossible to determine whether performance differences reflect genuine algorithmic advantages or experimental heterogeneity. By controlling all preprocessing variables, this study enables the first statistically fair comparison of these algorithmic families on this benchmark dataset.

Second, the empirical demonstration that Average Precision provides a more reliable performance gauge than AUC-ROC under the 1:28 class imbalance has direct practical implications for metric selection in future industrial PdM studies. AUC-ROC scores above 0.90 can be misleading under severe class imbalance due to the abundance of correctly classified majority-class samples inflating the true negative rate; Average Precision, by contrast, evaluates only the minority-class detection performance. The ranking discrepancy observed here—Gradient Boosting leads in AP (0.8550) while Random Forest leads in AUC-ROC (0.9772)—illustrates how metric choice can reverse algorithmic rankings, with potentially significant implications for deployment decisions.

Third, the physical interpretation of feature importance—linking mechanical power (25.51%) to PWF and OSF mechanisms, tool wear (23.91%) to TWF, and torque (8.92%) to bearing fatigue—provides actionable physical insight that bridges the gap between statistical ML output and engineering domain knowledge. This interpretation enables practitioners to understand which physical phenomena are driving predictions, building trust in the model and guiding targeted sensor investment decisions.

4. Conclusions

This investigation conducted a systematic nine-classifier benchmark study for CNC predictive maintenance using the AI4I 2020 dataset, yielding five significant and quantitative findings. In terms of model performance, Gradient Boosting demonstrated the most favorable overall profile (F1-score: 0.6782, Accuracy: 97.20%, AUC-ROC: 0.9723, Average Precision: 0.8550), while Random Forest attained the peak discriminative score (AUC-ROC: 0.9772, Average Precision: 0.8284). SMOTE-based class rebalancing elevated minority-class recall beyond 85% for the two best-performing models despite the extreme 1:28 class ratio, confirming the critical role of oversampling in industrial fault detection under severe class imbalance. Feature attribution analysis identified mechanical power (25.51%), tool wear duration (23.91%), and torque (8.92%) as the three highest-impact predictors, jointly accounting for over 58% of Gradient Boosting's predictive contribution. Cross-validation verified generalization stability, with Gradient Boosting sustaining the top mean F1-score (0.6581 +/- 0.0289) across all ten folds. Systematic metric comparison confirmed that Average Precision is a more discriminating and practically meaningful performance gauge than AUC-ROC for severely imbalanced industrial fault datasets.

The scientific contributions of this work are fourfold: (1) it provides the first comprehensive nine-classifier benchmark under a unified SMOTE-and-feature-engineering pipeline

on the AI4I 2020 dataset, enabling fair algorithmic comparison across all major ML families; (2) it demonstrates empirically that Average Precision is the appropriate primary metric for class-imbalanced industrial fault detection; (3) it delivers a physically grounded interpretation of feature importance linking dominant predictors to specific CNC mechanical failure mechanisms (tribological wear, fatigue loading, thermal stress); and (4) it translates ML outputs into a practical CBM decision framework with quantified implications for manufacturing operations.

The practical engineering implication is that Gradient Boosting, deployed alongside the two engineered features (mechanical power and thermal gradient) derived from standard CNC sensor streams, can trigger proactive maintenance alerts with 86.76% fault detection recall—potentially enabling conversion of unplanned breakdowns into scheduled maintenance events and meaningful reduction of unplanned downtime costs (estimated at USD 5,000–15,000 per hour [2]).

This study has three primary limitations that should be acknowledged. First, the AI4I 2020 dataset is synthetic and does not capture real CNC operational dynamics such as sensor drift, multi-modal vibration signatures, and missing data from tool changes; based on related real-data PdM studies, these factors can reduce classification F1-scores by 10–25% relative to synthetic benchmarks [5], [17]. Second, all evaluations are conducted in offline batch mode without addressing real-time inference latency or model drift. Third, the binary classification framework does not distinguish between individual failure modes. Future research directions include: (1) LSTM and TCN architectures for real-time streaming CNC sensor data [26], [27]; (2) federated learning frameworks for privacy-preserving multi-machine training [28]; (3) multi-label classification models for failure-mode-specific diagnostics [29], [30]; and (4) validation on real industrial CNC datasets to assess generalizability beyond the synthetic AI4I 2020 benchmark.

Acknowledgements

The authors wish to express their appreciation to Stephan Matzka of HTW Berlin for openly sharing the AI4I 2020 Predictive Maintenance Dataset via the UCI Machine Learning Repository.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors used Claude (Anthropic) to assist in grammar checking and English language editing of this manuscript. All AI-generated content was critically reviewed, revised, and validated by the authors. The authors take full responsibility for the scientific accuracy, originality, and integrity of the manuscript, in accordance with the guidelines of the Committee on Publication Ethics (COPE).

References

[1] I. Hector and R. Panjanathan, "Predictive maintenance in Industry 4.0: A survey of planning models and machine learning techniques," *PeerJ Comput. Sci.,* Art. no. 10:e2016, 2024, doi: 10.7717/peerj-cs.2016.

[2] N. Grabill, S. Wang, H. A. Olayinka, T. P. De Alwis, Y. F. Khalil, and J. Zou, "AI-augmented failure modes, effects, and criticality analysis (AI-FMECA) for industrial applications," *Reliab. Eng. Syst. Saf.,* vol. 250, Art. no. 110308, 2024, doi: 10.1016/j.ress.2024.110308.

[3] M. Soori, B. Arezoo, and R. Dastres, "Internet of Things for smart factories in Industry 4.0: A review," *Internet Things Cyber-Phys. Syst.,* vol. 3, pp. 192-204, 2023, doi: 10.1016/j.iotcps.2023.04.006.

[4] O. Fink, Q. Wang, M. Svensen, P. Dersin, W.-J. Lee, and M. Ducoffe, "Potential, challenges and future directions for deep learning in prognostics and health management applications," *Eng. Appl. Artif. Intell.,* vol. 92, Art. no. 103678, 2021, doi: 10.1016/j.engappai.2020.103678.

[5] L. Polverino, R. Abbate, P. Manco, D. Perfetto, F. Caputo, R. Macchiaroli, and M. Caterino, "Machine learning for prognostics and health management of industrial mechanical systems and equipment: A systematic literature review," *Int. J. Eng. Business Manag.,* vol. 15, 2023, doi: 10.1177/18479790231186848.

[6] M. Soori, B. Arezoo, and R. Dastres, "Machine learning and artificial intelligence in CNC machine tools: A review," *Sustain. Manuf. Serv. Econ.,* vol. 2, Art. no. 100009, 2023, doi: 10.1016/j.smse.2023.100009.

[7] H. A. B. Alfarizi, J. Vatn, and S. Yin, "An extreme gradient boosting aided fault diagnosis approach: A case study of fuse test bench," *IEEE Trans. Artif. Intell.,* vol. 3, no. 6, pp. 974-983, Dec. 2022, doi: 10.1109/TAI.2022.3165599.

[8] M. Zare, P. Kebria, A. Khosravi, and S. Nahavandi, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Trans. Knowl. Data Eng.,* vol. 35, no. 6, pp. 6106-6120, Jun. 2023, doi: 10.1109/TKDE.2022.3179381.

[9] S. N. Tang, J. T. Ma, Z. Q. Yan, Y. Zhu, and B. C. Khoo, "Deep transfer learning strategy in intelligent fault diagnosis of rotating machinery," *Eng. Appl. Artif. Intell.,* vol. 134, Art. no. 108678, 2024, doi: 10.1016/j.engappai.2024.108678.

[10] A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing health monitoring based on Hilbert–Huang transform, support vector machine, and regression," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 52–62, 2015, doi: 10.1109/TIM.2014.2330494.

[11] T. von Hahn, and C. K. Mechefske, "Machine learning in CNC machining: Best practices," *Machines,* vol. 10, no. 12, Art. no. 1233, 2022, doi: 10.3390/machines10121233.
[12] Y. Han, Z. Wei, and G. Huang, "An imbalance data quality monitoring based on SMOTE-XGBOOST supported by edge computing," *Sci. Rep.,* vol. 14, Art. no. 10151, May 2024, doi: 10.1038/s41598-024-60607-w.
[13] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," *Sensors,* vol. 22, no. 9, Art. no. 3246, 2022, doi: 10.3390/s22093246.
[14] Y. Hou, J. Ma, J. Wang, C. Liu, D. Li, and H. Li, "Enhanced generative adversarial networks for bearing imbalanced fault diagnosis of rotating machinery," *Appl. Intell.,* vol. 53, pp. 25201-25215, 2023, doi: 10.1007/s10489-023-04785-8.
[15] A. Windmann, P. Wittenberg, M. Schieseck, and O. Niggemann, "Artificial intelligence in Industry 4.0: A review of integration challenges for industrial systems," *in Proc. IEEE Int. Conf. Ind. Inform. (INDIN)*, Beijing, China, 2024, pp. 1-8, doi: 10.1109/INDIN58382.2024.10774488.
[16] E. Jovic, D. Primorac, M. Cupic, and A. Jovic, "Publicly available datasets for predictive maintenance in the energy sector: A review," *IEEE Access,* vol. 11, pp. 73505-73520, 2023, doi: 10.1109/ACCESS.2023.3294998.
[17] J. Zheng, C. Liu, and L. Zhang, "A comprehensive review of machine learning techniques for condition-based maintenance," *Int. J. Progn. Health Manag.,* vol. 15, no. 2, 2024, doi: 10.36001/ijphm.2024.v15i2.3850.
[18] J. Li, "Area under the ROC curve has the most consistent evaluation for binary classification," *PLOS ONE,* vol. 19, no. 12, Art. no. e0316019, 2024, doi: 10.1371/journal.pone.0316019.
[19] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," *J. Roy. Stat. Soc. C,* vol. 72, no. 5, pp. 1394-1412, Nov. 2023, doi: 10.1093/jrsssc/qlad057.
[20] J. Zhu, B. Chen, and W. Qin, "Feature engineering-based machine learning approaches for predictive maintenance in smart manufacturing," *IEEE Trans. Autom. Sci. Eng.,* vol. 19, no. 4, pp. 2801-2815, Oct. 2022, doi: 10.1109/TASE.2021.3106671.
[21] G. Niu, E. Liu, X. Wang, P. Ziehl, and B. Zhang, "Enhanced discriminate feature learning deep residual CNN for multi-task bearing fault diagnosis with information fusion," *IEEE Trans. Ind. Inform.,* vol. 19, no. 1, pp. 762-770, Jan. 2023, doi: 10.1109/TII.2022.3179011.
[22] C. Lou, M. A. Atoui, and X. Li, "Recent deep learning models for diagnosis and health monitoring: A review of research works and future challenges," *Trans. Inst. Meas. Control,* vol. 46, no. 5, pp. 879-901, Mar. 2024, doi: 10.1177/01423312231197052.
[23] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 34, no. 9, pp. 6390-6404, 2023, doi: 10.1109/TNNLS.2021.3136503.
[24] O. Matania, L. Bechhoefer, and J. Bortman, "Signal processing for the condition-based maintenance of rotating machines via vibration analysis: A tutorial," *Sensors,* vol. 24, no. 2, p. 454, Jan. 2024, doi: 10.3390/s24020454.
[25] M. Tiboni, C. Remino, R. Bussola, and C. Amici, "A review on vibration-based condition monitoring of rotating machinery," *Appl. Sci.,* vol. 12, no. 3, Art. no. 972, Jan. 2022, doi: 10.3390/app12030972.
[26] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of classification methods on unbalanced data sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
[27] X. Thelen, X. Zhang, O. Fink, Y. Lu, S. Ghosh, and B. D. Youn, "A comprehensive review of digital twin—Part 1: Modeling and twinning enabling technologies," *Struct. Multidiscip. Optim.,* vol. 65, p. 354, Dec. 2022, doi: 10.1007/s00158-022-03425-4.
[28] Q. Qian, J. Zhou, and Y. Qin, "Relationship transfer domain generalization network for rotating machinery fault diagnosis under different working conditions," *IEEE Trans. Ind. Inform.,* vol. 19, no. 9, pp. 9898-9908, Sep. 2023, doi: 10.1109/TII.2023.3234970.
[29] B. A. Tama, M. Vania, S. Lee, and S. Lim, "Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals," *Artif. Intell. Rev.,* vol. 56, no. 5, pp. 4667-4709, May 2023, doi: 10.1007/s10462-022-10293-3.
[30] E. Zio, "Prognostics and health management (PHM): Where are we and where do we (need to) go in theory and practice," *Reliab. Eng. Syst. Saf.,* vol. 218, Art. no. 108119, Feb. 2022, doi: 10.1016/j.ress.2021.108119.
[31] M. Owusu-Adjei, J. B. Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digit. Health,* vol. 2, no. 11, Art. no. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290.
[32] M. S. Khaled, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data,* vol. 12, no. 1, Art. no. 268, 2025, doi: 10.1186/s40537-025-01313-4.