Robust Baselines and Probability Calibration for TPM-Oriented Predictive Maintenance

José Roberto Dale Luche[1], Blaha Gregory Correia dos Santos Goussain[2] and Claudia Regina de Freitas[3]

[1,2,3]*São Paulo State University (UNESP), School of Engineering and Sciences, Guaratinguetá, 12516-410, São Paulo, Brazil*

dale.luche@unesp.br
blaha.goussain@unesp.br
claudia.freitas@unesp.br

ABSTRACT

Predictive maintenance (PdM) under severe class imbalance challenges model evaluation and deployment, especially when probabilities inform maintenance decisions. Using the AI4I 2020 dataset, this study establishes a reproducible baseline for failure detection with emphasis on rigorous validation and probability calibration. Models such as Random Forest (RF), Multilayer Perceptron (MLP), and classical baselines were evaluated via nested cross-validation with strict leakage control. Metrics included Average Precision, recall, precision, Brier score, and Expected Calibration Error (ECE), while the impact of SMOTE on class imbalance was analyzed. RF achieved the most robust balance between discrimination and calibration reliability, whereas MLP with SMOTE improved sensitivity but incurred calibration and false-positive trade-offs. Torque and tool wear emerged as dominant predictors, aligning with physical degradation mechanisms. By explicitly linking predictive performance to probability calibration and operational cost considerations, this work provides an actionable, cost-aware reference baseline for PdM within Total Productive Maintenance frameworks.

Keywords: Predictive Maintenance, Total Productive Maintenance, Probability Calibration, Class Imbalance, Machine Learning.

1. INTRODUCTION

Increased competitive pressure has driven manufacturers to adopt integrated, data-centric maintenance strategies that raise equipment availability and stabilize throughput. With the diffusion of Industry 4.0, operational data and machine learning (ML) models are reshaping how incipient faults are detected and acted upon on the shop floor, complementing established maintenance programs and accelerating decision cycles (Tortorella et al., 2021; Zonta et al., 2020). Predictive maintenance (PdM) specifically aims to anticipate failures from condition monitoring data, reducing unplanned downtime and maintenance costs while enhancing production stability. Recent reviews document PdM's maturation from proof-of-concept pilots to plant-level deployments, but also emphasize persistent challenges such as severe class imbalance, data heterogeneity, and validation under real operating constraints (Zonta et al., 2020).

In parallel, Total Productive Maintenance (TPM) remains a foundational program for asset management, linking operations and maintenance to systematically eliminate losses and raise Overall Equipment Effectiveness (OEE). Classical TPM studies showed that autonomous maintenance, focused improvement, and planned maintenance sustain reliability gains (Ahuja & Khamba, 2008), while more recent perspectives on TPM 4.0 highlight the integration of the Industrial Internet of Things (IIoT), Artificial Intelligence (AI), and predictive analytics to support real-time decision-making (Gomaa, 2025). The integration of ML-driven PdM within TPM frameworks thus offers the potential to enhance equipment reliability further by embedding data-driven insights into preventive actions and resource allocation (Muchiri & Pintelon, 2008).

A significant challenge in applying ML to PdM is the inherent class imbalance in industrial failure data, where failure events represent only a small fraction of observations. This imbalance motivates the adoption of evaluation strategies that emphasize minority-class detection and operational relevance, rather than relying on accuracy-driven assessment.

Recent literature on PdM has explored advanced architectures such as attention-based temporal networks (Liu & Su, 2024), ordinal classification frameworks (Yürek & Birant, 2024), and continual learning under non-stationary conditions (Benatia et al., 2025). Nevertheless, classical ML models, including tree-based ensembles and multilayer perceptrons (MLPs), remain highly relevant in industrial settings due to their interpretability, computational efficiency, and ease of integration into existing TPM workflows. Moreover, their performance serves as a vital baseline for evaluating more complex approaches.

However, two methodological issues remain underexplored in PdM studies using tabular industrial data. First, many

https://doi.org/10.36001/IJPHM.2026.v17i1.4659

published works rely on single train–test splits or non-nested cross-validation, which may yield optimistic performance estimates when hyperparameters are tuned on the same data used for evaluation. Second, while discrimination metrics such as ROC-AUC and AP are widely reported, fewer studies assess probability calibration, a critical requirement for risk-based maintenance decisions. Miscalibrated models may appear accurate yet produce unreliable probability estimates, potentially undermining TPM decision-making routines. Addressing these gaps requires the combined use of nested cross-validation and calibration metrics such as the Brier score and ECE, allowing fair model comparison and trustworthy operational deployment.

This study addresses the following research question: How effectively can well-established ML classification models predict equipment failures in a manufacturing context while remaining compatible with TPM practices and OEE-oriented decision making? Accordingly, our objectives are to: (i) benchmark strong tabular baselines (e.g., tree-based and margin-based classifiers) and MLPs using leakage-controlled validation, including nested cross-validation; (ii) quantify performance trade-offs that matter operationally, especially recall versus precision, given the high cost of missing failures; and (iii) discuss how these models and metrics, including calibration indicators, can be integrated into TPM routines to enable actionable maintenance planning. We deliberately scope this work to classically effective methods that are widely accessible and interpretable.

The main contributions of this study can be summarized as follows:

- A leakage-controlled nested cross-validation benchmark for AI4I under severe class imbalance;
- A joint evaluation of discrimination and probability calibration (Brier + ECE + reliability diagrams) in a PdM context;
- An operational interpretation of recall–precision–calibration trade-offs within TPM/OEE decision frameworks;
- A fully reproducible experimental pipeline with documented hyperparameter search and implementation details.

The remainder of this paper is organized as follows. Section 2 reviews related work and recent advances on the AI4I dataset, situating our contribution within the broader PdM research landscape. Section 3 details the materials and methods, including dataset characteristics, preprocessing, and validation protocols. Section 4 reports the experimental results, followed by Section 5, which discusses their implications for TPM and outlines limitations. Finally, Section 6 concludes with key takeaways and future research directions.

2. RELATED WORK

PdM has progressed rapidly with the diffusion of Industry 4.0 technologies. Recent reviews highlight its transition from proof-of-concept pilots to large-scale deployments, while emphasizing enduring challenges such as data imbalance, concept drift, data heterogeneity, and the need for explainability (Zonta et al., 2020; Nunes et al., 2023). These challenges remain critical for ensuring that predictive models are not only accurate but also actionable within established maintenance programs such as TPM.

Within this landscape, the AI4I 2020 PdM Dataset has become a benchmark for evaluating supervised learning approaches, a crucial step for developing interpretable models in line with the principles of Explainable AI (XAI) (Arrieta et al., 2020) and widely recognized in PdM research (Iqbal et al., 2023). Its synthetic design mitigates the scarcity of public failure data while preserving key challenges of practical deployments, including severe class imbalance and multivariate operational signals. As a result, it has been widely used to train and test models for failure forecasting and maintenance scheduling.

Early research with AI4I emphasized established supervised classifiers and straightforward process-improvement integrations. For example, Shivaramu (2025) embedded Random Forests (RF), Support Vector Machines (SVM), and simple neural networks within a Lean Six Sigma (DMAIC) framework, reporting reduced simulated failure rates by guiding targeted interventions. Similarly, Çiftpınar et al. (2025) compared ensemble strategies such as bagging and majority voting under explicit imbalance handling, finding that ensembles generally outperformed single learners on AI4I in terms of failure-prediction accuracy. These studies provide solid baselines for classical, well-understood models on this dataset.

Moving beyond binary nominal classification, ordinal formulations have been explored to reflect graded health states. Yürek and Birant (2024) proposed OPMEB (Ordinal PdM with Ensemble Binary Decomposition), which decomposes ordered multi-class targets into binary subproblems that respect severity order. Evaluated on AI4I and other datasets, OPMEB achieved superior results relative to nominal multi-class baselines, indicating that leveraging the inherent order of degradation can improve sensitivity across health stages.

Recognizing that industrial data streams are non-stationary, another line of research investigates continual and online learning under concept drift. Esteban et al. (2025) introduced an online ensemble of Hoeffding Adaptive Trees for multi-label streaming fault prediction simulated with AI4I, reporting marked gains over static batch models under evolving conditions. In complementary work, Prashanth et al. (2025) simulated drift on AI4I to compare static versus incrementally updated models, showing that while static

RFs perform well on stationary data, incremental approaches better sustain accuracy under distributional shifts.

Given the safety- and cost-critical nature of maintenance, XAI methods have also been applied to AI4I-trained models. Presciuttini et al. (2024) contrasted SHAP, LIME, and counterfactual explanations for model outputs on AI4I, documenting how global and local attributions highlight drivers such as tool wear and rotational speed for specific failure modes, thereby linking predictions to actionable maintenance cues.

Finally, the constraints of data privacy and cross-site collaboration have motivated federated and privacy-preserving approaches using AI4I as a proxy. Alshkeili et al. (2025) proposed an explainable federated learning framework that integrates XAI into decentralized training, reporting high predictive performance while avoiding centralization of sensitive data and enabling local interpretability.

To further situate our work within the current state of the art and explicitly highlight its positioning, we conducted a directed mapping of studies employing the AI4I 2020 dataset in PdM scenarios. This landscape, synthesized in Table 1, illustrates methodological trends, gaps, and extension opportunities that directly inform the choices made in this manuscript.

Collectively, the literature on AI4I spans from classical supervised baselines and ordinal formulations to continual/online adaptation, XAI, and federated learning, mapping a trajectory toward models that are not only accurate but also adaptable, interpretable, and deployable under realistic operational constraints. These contributions confirm the vitality of PdM research but also underscore the difficulty of balancing model complexity, interpretability, and operational relevance. While deep learning and privacy-preserving methods extend the frontier, they often demand high computational resources and remain challenging to integrate into existing TPM workflows, which typically prioritize simplicity, reproducibility, and explainability.

Reference	Dataset	Task	Methods	Key metrics	Contribution
Lyubchyk et al., 2023	AI4I 2020	Failure classification (multiclass) and failure-rate regression	Soft-regularized clustering with kernel regression; SVM and k-NN as comparators	Not reported	Data aggregation via regularized clustering for tabular PdM
Yürek & Birant, 2024	AI4I 2020	Ordinal classification (2/3/4/5 classes)	OPMEB with LightGBM, CatBoost, RF, XGBoost	5-class Acc ≈ 91.8%, Macro-F1 ≈ 0.90; 3-class Acc ≈ 80.1%, F1 ≈ 0.81; 2-class Acc ≈ 83.1%, F1 ≈ 0.59	Ordinal reformulation outperforms OVO/OVA/chain baselines on AI4I
Ronzoni et al., 2022	AI4I 2020; transmission lines	Failure type (multiclass) and status (binary), multi-output	AdaBoost, Gradient Boosting, RF, MLP with Bayesian Optimization	RF best on AI4I (macro metrics)	BO pipeline for imbalanced data; multi-output setup
Presciuttini et al., 2024	AI4I 2020	Failure type classification	RF with XAI (SHAP, LIME, counterfactuals)	Not reported (RF reported as high accuracy)	Turns explanations into operational decisions (thresholds and actions)
Liu & Su, 2024	AI4I; HDFS; AWS CloudWatch	Classification with temporal learning	Transformers with channel-spatial attention; Bi-LSTM; classic ML baselines	High AUROC; gains vs. baselines (per paper's ablations)	Temporal DL with attention; reduction of false alarms
Autran et al., 2024	AI4I-PMDI (extension)	PdM with irregular sampling and missing data	Dataset curation and data-quality analysis	Not applicable	Realistic scenarios with timestamps, missingness, and multiple machines
Benatia et al., 2025	AI4I plus real data	Failure prediction under continual learning	CL with self-supervision and fine-tuning	On AI4I: AUROC ≈ 0.910, F1 ≈ 0.958	Robustness to drift and sequential tasks
Kang et al., 2024	AI4I (benchmark) plus others	Privacy-preserving training and inference	Multi-key homomorphic encryption (MK-CKKS)	Not applicable (focus on feasibility and overhead)	Privacy for collaborative learning across parties
Kamel, 2022	AI4I 2020	Binary status (failure vs. normal)	Feed-forward ANN with threshold tuning	TPR ≈ 99.5%, TNR ≈ 68.7% (validation)	Introductory ANN case on AI4I

Table 1. Landscape of AI4I 2020 studies (2022–2025), highlighting methodological trends and contributions

In this context, there is a persistent need for transparent and reproducible baselines. Classical ML methods such as RF and MLPs remain widely accessible in industry, offering computational efficiency and inherent interpretability. Their

performance provides a crucial reference point for evaluating newer approaches, and their integration into TPM workflows can directly inform OEE-oriented decision making.

Despite substantial progress in PdM research, relatively few studies jointly address leakage-controlled validation protocols and probability calibration. This gap motivates the present study, which focuses on establishing a reproducible and decision-aware baseline for fair model comparison under severe class imbalance.

3. MATERIALS AND METHODS

This section details the dataset, preprocessing procedures, imbalance-handling strategy, model specifications, validation protocol, evaluation metrics, and implementation details adopted in this study. All methodological choices were designed to ensure statistical rigor, prevent information leakage, and provide transparent, reproducible baselines for PdM under severe class imbalance.

3.1. Dataset and Target Definition

The experiments were conducted using the AI4I 2020 PdM Dataset, a well-established benchmark in PHM research due to its realistic operational variability and pronounced class imbalance between failure and non-failure observations. The dataset comprises 10,000 observations described by fourteen attributes, including continuous variables related to mechanical and thermal behavior, air temperature [K], process temperature [K], rotational speed [rpm], torque [Nm], and tool wear [min], as well as a categorical indicator of motor type (L, M, or H).

The principal prediction target in this study is Machine failure, defined as a binary variable indicating whether a failure occurred. Although the dataset provides five auxiliary failure-type indicators, Tool Wear Failure (TWF), Heat Dissipation Failure (HDF), Power Failure (PWF), Overstrain Failure (OSF), and Random Failure (RNF), all experiments in this work strictly address a binary classification task (failure vs. no failure). The individual failure-type indicators are used exclusively for descriptive analysis of class imbalance and failure mechanisms, with aggregation into a single positive class during model training and evaluation.

Table 2 summarizes the class distribution before balancing. Only 348 instances (3.48%) correspond to failure events, resulting in a highly imbalanced learning problem with important implications for model validation and metric selection.

Failure Type	Occurrences
No failure	**9,652**
Heat dissipation failure (HDF)	112
Power failure (PWF)	95
Overstrain failure (OSF)	78
Tool wear failure (TWF)	45
Random failures (RNF)	18
Total failures	**348 (3.48%)**

Table 2. Class distribution before balancing

Figure 1 illustrates a bivariate scatter plot of torque versus rotational speed; two variables directly associated with mechanical load and operational stress. Despite a clear physical relationship between these variables, substantial overlap between normal and failure conditions is observed across the operating range, indicating that simple univariate or low-dimensional decision rules are insufficient for reliable failure detection.

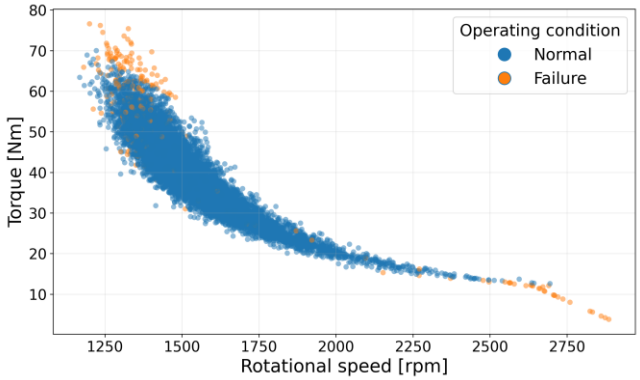

Figure 1. Bivariate distribution of torque and rotational speed under normal and failure conditions

3.2. Exploratory Analysis and Preprocessing

All preprocessing steps were designed to preserve the integrity of the nested cross-validation protocol and to prevent information leakage across folds. The dataset was first examined for duplicate entries; none were found. Exploratory data analysis revealed substantial natural variability in the continuous operating variables, particularly torque, temperatures, and tool wear.

Outlier analysis was conducted using boxplots for the continuous variables (Figure 2). Although extreme values were observed, these were retained in the dataset, as they correspond to plausible industrial operating conditions and reflect genuine physical behavior rather than data artifacts. Removing such observations could artificially simplify the classification task and bias performance estimates.

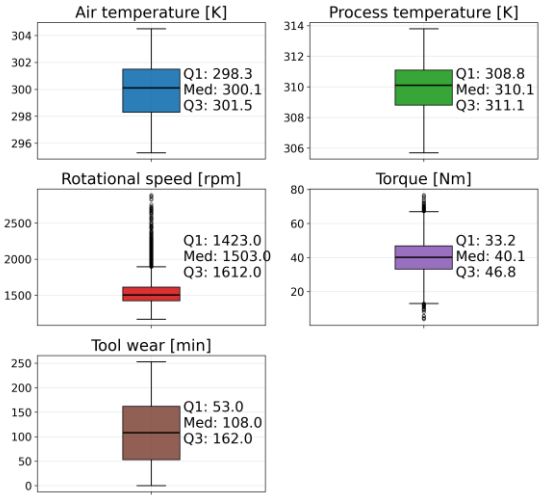

Figure 2. Boxplots of continuous variables in AI4I 2020 (individual scales)

No manual feature engineering was performed. Instead, the study focuses on evaluating the effectiveness of well-established models under realistic data conditions, emphasizing transparency and reproducibility.

3.3. Class Imbalance Handling

The severe imbalance between failure and non-failure observations motivates the use of imbalance-aware training strategies. In this study, SMOTE was applied to generate synthetic minority-class samples when explicitly indicated. To avoid information leakage, SMOTE was applied exclusively to training data within the cross-validation procedure.

Specifically, SMOTE was applied to the aggregated binary failure class and never to validation or test folds, following best practices for leakage control in imbalanced learning as discussed by Santos et al. (2018). The technique was not applied to individual failure mechanisms, and the learning task remained strictly binary throughout all experiments. This design ensures that class rebalancing improves minority-class representation during training without altering the evaluation protocol or introducing distributional bias.

3.4. Models and Hyperparameters Configuration

This study evaluates a set of well-established classification models that are frequently adopted in industrial PdM and condition monitoring applications. The emphasis is deliberately placed on models that combine solid predictive performance with transparency, robustness, and ease of deployment in TPM contexts, rather than on highly complex or data-hungry architectures.

As baseline references, simple probabilistic and tree-based models were included to provide interpretable performance anchors. Logistic Regression was considered as a linear probabilistic baseline, offering a transparent decision function and well-calibrated probability estimates under appropriate regularization. Random Forests were employed as a strong non-linear ensemble baseline, capable of capturing complex interactions among process variables while retaining a degree of interpretability through global feature importance analysis.

In addition to classical baselines, a MLP classifier was included to represent a lightweight neural approach that remains feasible in industrial environments. The MLP architecture was intentionally kept shallow, avoiding deep or highly overparameterized networks, in order to maintain comparability with classical models and to reduce sensitivity to overfitting under limited minority-class samples. The MLP therefore serves not as a state-of-the-art deep learning solution, but as a reference point for assessing how modest neural models behave under strict validation and calibration requirements.

All models were implemented using standardized preprocessing pipelines, including feature scaling when required. For the MLP, input features were standardized using statistics computed exclusively from the training data within each cross-validation fold. When applicable, oversampling via SMOTE was integrated into the pipeline and applied only to training subsets, following best practices for leakage control.

Hyperparameter tuning was performed within the inner loop of the nested cross-validation procedure. Rather than exploring exhaustive or excessively large search spaces, the tuning strategy focused on a compact set of hyperparameters known to exert the strongest influence on model capacity and regularization. This choice reflects a practical engineering perspective, prioritizing reproducibility and interpretability over marginal performance gains.

Table 3 summarizes the hyperparameters tuned for each model, their corresponding search ranges, and the fixed settings adopted across all experiments to ensure deterministic and reproducible behavior.

In addition to the primary models discussed, a small set of additional classical classifiers was included to provide broader baseline coverage and facilitate comparative analysis. Specifically, Linear Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (k-NN) were evaluated using standard configurations and compact hyperparameter search spaces. These models were not the primary focus of the study, but serve as complementary references to contextualize the performance of the main baselines under the same validation protocol.

Model / Condition	Tuned hyperparameters (inner CV)	Search range / values	Fixed settings (for reproducibility)
Dummy (Stratified)	–	–	strategy = "stratified"
Logistic Regression	C	{0.01, 0.1, 1.0, 10.0}	solver = "liblinear", max_iter = 500
Linear SVM	C	{0.01, 0.1, 1.0, 10.0}	kernel = "linear", max_iter = 1000
Decision Tree	max_depth, min_samples_leaf	max_depth ∈ {None, 5, 10}; min_samples_leaf ∈ {1, 5, 10}	random_state = 42
Random Forest	max_depth, min_samples_leaf	max_depth ∈ {None, 5, 10}; min_samples_leaf ∈ {1, 5, 10}	n_estimators = 200, random_state = 42, n_jobs = 1
k-NN	n_neighbors	{3, 5, 7}	metric = "minkowski", weights = "uniform"
MLP (no SMOTE)	hidden_layer_sizes, alpha, learning_rate_init	hidden_layer_sizes ∈ {(50,), (100,), (100, 50)}; alpha ∈ {1e−4, 1e−3, 1e−2}; learning_rate_init ∈ {1e−3, 1e−2}	StandardScaler, max_iter = 400, early_stopping = True, n_iter_no_change = 10, validation_fraction = 0.1, random_state = 42
MLP (SMOTE 1:1)	Same as MLP (no SMOTE)	Same as MLP (no SMOTE)	Same as MLP (no SMOTE) + SMOTE(random_state = 42) applied inside the pipeline (training folds only; default sampling_strategy="auto" → 1:1)

Table 3. Hyperparameter search space used in nested cross-validation (inner loop)

3.5. Validation Protocol

A rigorous nested cross-validation (CV) protocol was employed to obtain statistically unbiased performance estimates and to prevent hyperparameter-induced optimism, a common pitfall in predictive maintenance (PdM) studies involving highly imbalanced datasets. The complete end-to-end workflow of the experimental methodology is formally described in Algorithm 1 (Figure 3), which provides an algorithmic overview (pseudocode) of the nested cross-validation procedure with leakage-controlled preprocessing and oversampling.

The outer loop consisted of 10 stratified folds, ensuring that the proportion of failure events was preserved across all splits. In each outer iteration, one fold was held out as an untouched test set and used exclusively for final evaluation, while the remaining folds formed the corresponding training set. This design guarantees that model assessment is performed on data that played no role in preprocessing, resampling, or hyperparameter tuning, thereby providing an unbiased estimate of generalization performance.

Within each outer iteration, an inner loop with 3 stratified folds was used for hyperparameter optimization. Candidate hyperparameter configurations were evaluated using a standard training–validation procedure, and model selection was based on the mean AP computed across the inner validation folds. This strict separation between model selection and evaluation ensures that performance estimates are not inflated by repeated exposure to test data.

To mitigate the risk of information leakage, all data-dependent transformations strictly followed the training-only principle. Preprocessing steps, including feature scaling for neural-network models, were fitted exclusively on the training portion of each outer fold and subsequently applied to the corresponding validation and test subsets. Likewise, SMOTE was applied only to the inner training splits during hyperparameter tuning and only to the outer training set during final model retraining, while all validation and test folds remained completely untouched, in accordance with established best practices for imbalanced learning.

After identifying the optimal hyperparameter configuration in the inner loop, the model was retrained on the full outer training set using the selected parameters and subsequently evaluated on the held-out outer test fold. This process yielded one unbiased performance estimate per outer iteration. Results are reported as the mean and standard deviation of the evaluation metrics across the ten outer folds, providing a robust and statistically sound assessment of model performance.

For completeness, Algorithm 1 takes as input the dataset $D=(X,y)$; the number of outer folds $K=10$ and inner folds $J=3$ (both stratified); the model family M; the hyperparameter search space Λ; and the set of evaluation

metrics S={AP, Recall, Brier, ECE}. The output consists of the outer-fold performance estimates {s_k} and their aggregated mean ± standard deviation across folds.
The application of SMOTE within the pipeline is model-dependent and conditionally executed only for configurations designed to include class rebalancing.

```
Initialize empty list R
Partition D into stratified outer folds {F_k} for k=1..K
For k = 1..K do
    D_test ← F_k; D_train ← D \ F_k
    (Leakage control) Fit preprocessing pipeline P (e.g.,
    scaling, encoding) on X_train only
    X_train ← P(X_train); X_test ← P(X_test)
    Partition D_train into stratified inner folds {G_j} for
    j=1..J
    For each hyperparameter configuration λ in Λ do
        Initialize empty list V_λ
        For j = 1..J do
            D_val ← G_j; D_in ← D_train \ G_j
            If pipeline includes SMOTE then
                Apply SMOTE only to inner-training split:
                (X_in+, y_in+) ← SMOTE(X_in, y_in)
                Train model M(λ) on (X_in+, y_in+)
            Else
                Train model M(λ) on (X_in, y_in)
            End if
            Compute validation score v_j ← AP on D_val
            Append v_j to V_λ
        End for
        Compute mean inner score v̄_λ ← mean(V_λ)
    End for
    Select best hyperparameters λ* ← argmax_λ v̄_λ
    If pipeline includes SMOTE then
        (X_train+, y_train+) ← SMOTE(X_train, y_train)
        Train final model M(λ*) on (X_train+, y_train+)
    Else
        Train final model M(λ*) on (X_train, y_train)
    End if
    Compute outer test metrics s_k ← {AP, Precision,
    Recall, Brier, ECE} on D_test
    Append s_k to R
End for
Return R and mean ± std across outer folds
```

Figure 3. Nested cross-validation algorithm with leakage-controlled preprocessing and conditional SMOTE

3.6. Evaluation Metrics

Model performance was assessed using evaluation metrics specifically suited to highly imbalanced failure-prediction tasks. The primary evaluation metric adopted in this study was AP, i.e., the area under the Precision–Recall curve (PR-AUC). AP directly reflects a model's ability to rank failure instances ahead of normal operating conditions under severe class imbalance.

To complement AP, additional metrics were computed to provide a more comprehensive characterization of model behavior. Recall (also referred to as sensitivity) quantified the model's ability to correctly identify true failure events, a critical requirement for reducing unplanned downtime in PdM applications. Precision measured the proportion of predicted failures that correspond to actual failures, capturing the operational cost associated with false alarms and unnecessary maintenance actions. Their harmonic mean, the F1-score, summarized the trade-off between detection capability and false-positive control.

Because maintenance decisions are often triggered based on predicted failure probabilities rather than hard class labels, probability calibration was explicitly evaluated. Two complementary calibration metrics were considered. The Brier score measures the mean squared error between predicted probabilities and observed outcomes, providing a global assessment of probabilistic accuracy. ECE quantifies the discrepancy between predicted confidence levels and empirical event frequencies across probability bins, offering a more localized view of probability reliability. Together, these metrics provide insight into whether predicted probabilities can be meaningfully interpreted and used for risk-based decision making in industrial settings.

For completeness, the area under the Receiver Operating Characteristic curve (ROC-AUC) and specificity were also reported, given their widespread use in the PHM literature. However, these metrics played a secondary role in the interpretation of results due to their limited sensitivity under extreme class imbalance. All metrics were computed independently on each outer test fold of the nested cross-validation procedure and are reported as mean ± standard deviation across folds, ensuring statistically unbiased and comparable performance estimates.

3.7. Implementation and reproducibility

All experiments were implemented in Python using widely adopted open-source libraries to ensure transparency and reproducibility. Model training, hyperparameter optimization, and evaluation were conducted using scikit-learn, while class imbalance handling via SMOTE was performed using the imbalanced-learn library. The experiments were executed using Python 3.13.7, scikit-learn 1.7.1, and imbalanced-learn 0.14.0.

Stratified data partitions were generated using a fixed random seed (RANDOM_STATE = 42) to ensure consistent fold construction across all experiments. The same seed was used for neural network initialization, SMOTE sample generation, and other stochastic components of the learning algorithms, thereby minimizing variability due to random

processes and enabling fair and reproducible comparisons between classifiers.

Hyperparameter tuning was performed exclusively within the inner cross-validation loop, in accordance with the nested validation protocol described in Section 3.5. All preprocessing operations, including feature scaling, categorical encoding, and oversampling, were encapsulated in reproducible pipelines and fitted solely on training subsets before being applied to validation or test data, ensuring strict control of information leakage.

4. RESULTS

This section reports the empirical results obtained from the benchmark evaluation of predictive models for machine failure detection under severe class imbalance. All results were produced under the nested cross-validation and leakage-control protocol described in Section 3.5 and are summarized using fold-aggregated metrics and pooled out-of-fold predictions.

4.1. Overall Performance Metrics

The fold-averaged performance metrics for all evaluated models are summarized in Table 4 (mean ± standard deviation across the ten outer folds). This comparison differentiates models that are misled by class imbalance from those that effectively detect rare failure events.

Model	AP	Recall (macro)	Precision (macro)	F1 (macro)	ROC-AUC	Accuracy
Dummy (Stratified)	0.034 ± 0.001	0.02 ± 0.02	0.03 ± 0.03	0.03 ± 0.02	0.497 ± 0.012	0.939 ± 0.002
Logistic Regression	0.430 ± 0.080	0.83 ± 0.07	0.14 ± 0.01	0.24 ± 0.02	0.900 ± 0.022	0.820 ± 0.013
Linear SVM	0.428 ± 0.080	0.82 ± 0.07	0.14 ± 0.01	0.24 ± 0.02	0.899 ± 0.022	0.820 ± 0.012
Decision Tree	0.535 ± 0.099	0.88 ± 0.07	0.29 ± 0.03	0.44 ± 0.04	0.940 ± 0.037	0.976 ± 0.004
Random Forest	0.707 ± 0.061	0.85 ± 0.07	0.40 ± 0.05	0.54 ± 0.05	0.973 ± 0.014	0.981 ± 0.003
k-NN	0.561 ± 0.069	0.85 ± 0.06	0.23 ± 0.02	0.37 ± 0.03	0.940 ± 0.034	0.972 ± 0.002
MLP (no SMOTE)	0.747 ± 0.060	0.53 ± 0.15	0.80 ± 0.13	0.61 ± 0.09	0.970 ± 0.012	0.978 ± 0.003
MLP (SMOTE 1:1)	0.730 ± 0.066	0.83 ± 0.05	0.43 ± 0.06	0.56 ± 0.05	0.964 ± 0.012	0.956 ± 0.010

Table 4. Fold-averaged performance (mean ± SD) across 10 folds (nested CV, OOF unified)

As expected, the Dummy (Stratified) classifier achieves near-chance performance in terms of AP, confirming that accuracy-driven baselines are not informative under severe class imbalance. Among classical baselines, RF exhibits a strong and stable performance profile, combining high AP with robust recall and low variance across folds. Its consistently high ROC-AUC further indicates reliable ranking of failure instances while maintaining controlled false-positive behavior.

Linear models (Logistic Regression and Linear SVM) display a recall-oriented operating profile, achieving relatively high sensitivity at the cost of very low precision, which leads to moderate AP values. Decision Tree and k-NN models exhibit intermediate behavior, with improved recall compared to linear models but less stable discrimination and higher variability across folds.

Neural network models highlight the impact of imbalance handling strategies. The unbalanced MLP achieves the highest AP among all evaluated models, but shows substantial variability in recall and precision across folds. Introducing SMOTE (1:1) reshapes the operating profile of the MLP: recall increases markedly, indicating improved sensitivity to rare failures, while precision and probability calibration deteriorate. This confirms that resampling primarily affects the balance between missed failures and false alarms rather than delivering uniform gains across all performance dimensions.

4.2. Precision–Recall Analysis

Under severe class imbalance, PR curves provide a more informative characterization of rare-event detection than ROC curves, as precision directly reflects the cost associated with false positives while recall captures the risk of missed failures. Accordingly, PR behavior is examined at two complementary levels: a global ranking of all models according to AP and a detailed analysis of PR curve shapes for the most competitive classifiers.

Figure 4 reports the global AP ranking computed from pooled OOF predictions. The ranking confirms that models explicitly designed to capture nonlinear interactions, RF and MLP variants, substantially outperform linear baselines in terms of ranking rare failure events. The Dummy (Stratified) classifier performs at near-chance level, reinforcing that discrimination under extreme imbalance cannot be achieved through prevalence-aware random guessing.

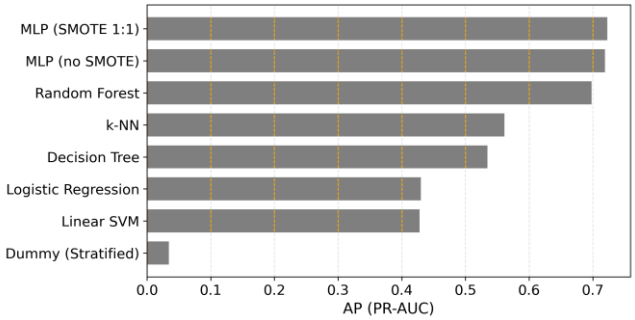

Figure 4. Average Precision ranking.

Minor numerical differences between the AP values reported in Table 4 and those shown in Figure 4 arise from the aggregation procedure. Specifically, Table 4 reports fold-wise averaged AP values (mean ± standard deviation across outer folds), whereas Figure 4 reports AP computed once from pooled OOF predictions. Both representations are complementary and consistent, providing statistically unbiased yet operationally informative views of model discrimination.

While the unbalanced MLP attains the highest AP, the difference relative to the SMOTE-enhanced MLP and RF is marginal, indicating that these models achieve broadly comparable ranking performance. This observation suggests that AP alone is insufficient to fully characterize operational behavior, motivating a closer inspection of how precision degrades as recall increases.

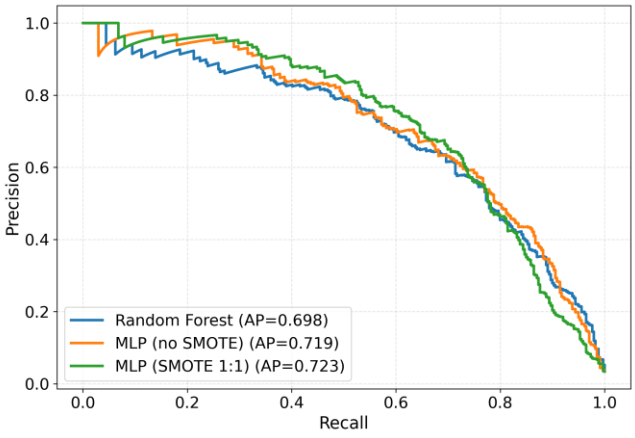

Figure 5. PR curves for the top-performing models based on pooled outer-loop OOF predictions.

Figure 5 presents the full PR curves for the three most competitive classifiers. Random Forest exhibits a conservative yet stable PR profile, maintaining relatively high precision over a wide recall interval. This behavior indicates a gradual trade-off between sensitivity and false alarms, which is desirable in operational settings where abrupt increases in maintenance workload are costly.

The unbalanced MLP displays high precision at low-to-moderate recall levels but experiences a sharper decline as recall approaches higher values, reflecting a rapidly increasing false-positive burden when sensitivity is pushed. In contrast, the SMOTE-enhanced MLP shifts the operating regime toward higher recall, sustaining improved sensitivity to rare failures in the mid-to-high recall range, albeit at the expense of lower precision. This shift confirms that resampling primarily redistributes errors along the PR curve rather than yielding uniform improvements in ranking performance.

Taken together, these results indicate that the three top-performing models occupy distinct regions of the precision–recall trade-off space. RF offers a balanced and predictable operating profile, while the MLP variants enable more aggressive recall-oriented strategies when missed failures are considered more costly than false alarms. This distinction motivates the subsequent threshold-dependent and calibration analyses.

4.3. Error Analysis and Calibration

Beyond global discrimination metrics and PR behavior, error forensics and probability calibration provide deeper insight into the operational risks associated with deploying PdM models in practice. While PR curves offer a threshold-independent assessment of ranking quality, real maintenance decisions are inherently threshold-based and rely on predicted probabilities as proxies for failure risk. In this context, both the type of classification errors and the reliability of probability estimates become critical.

To analyze model behavior under a realistic operating condition, confusion matrices were computed from pooled OOF predictions using a fixed decision threshold of 0.50. Figure 6 presents the aggregated confusion matrices for the three most competitive classifiers identified in the previous analysis: Random Forest, MLP without SMOTE, and MLP with SMOTE (1:1).

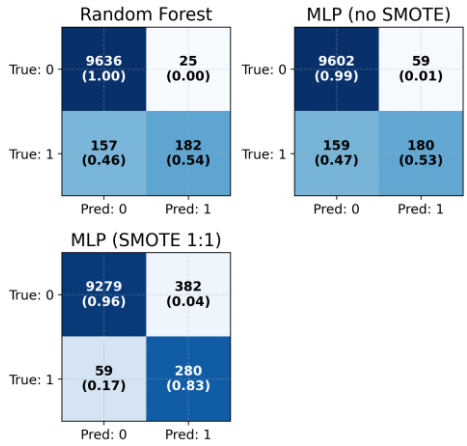

Figure 6. Aggregated confusion matrices for RF, MLP (no SMOTE), and MLP (SMOTE 1:1)

RF exhibits a conservative error profile, with a limited number of false positives but a non-negligible share of false negatives, indicating that some early or borderline degradation cases may remain undetected at this threshold. In contrast, the unbalanced MLP achieves higher sensitivity to failure events but generates a substantially larger number of false positives, which would translate into unnecessary inspections and increased maintenance workload. Introducing SMOTE alters this trade-off by improving coverage of minority-class failures, albeit at the expense of a higher false-positive rate, highlighting the classical recall–precision tension under severe class imbalance.

To further illustrate deployment-oriented threshold selection under a recall-driven policy, the MLP with SMOTE was additionally evaluated at the operating point corresponding to Recall ≥ 0.90. This regime was achieved at a probability threshold of 0.044. At this threshold, recall reached 0.903, while precision decreased to 0.204 and the false-positive rate increased to 0.124 (306 true positives, 33 false negatives, and 1196 false positives). These results reinforce that aggressive sensitivity targets substantially amplify false alarms, highlighting the operational trade-offs inherent in recall-oriented deployment strategies.

Beyond threshold-based error analysis, probability calibration is essential to determine whether predicted probabilities can be meaningfully interpreted as failure risk estimates. From a decision-theoretic perspective, well-calibrated probabilities are a prerequisite for Bayesian and cost-aware maintenance planning, where intervention decisions are guided by expected failure costs rather than binary predictions alone. Calibration quality was therefore assessed using the Brier score and ECE.

ECE was computed using 10 equal-width probability bins over the interval [0,1], following standard binning-based calibration assessment. For each bin, the absolute difference between the mean predicted probability and the empirical failure frequency was calculated and weighted by the proportion of samples in that bin. The final ECE value corresponds to the weighted average of these bin-level calibration gaps, consistent with standard formulations in the calibration literature.

To provide a robust assessment of probability reliability, ECE was evaluated at two complementary levels. First, fold-wise ECE values were calculated within each outer fold of the nested cross-validation procedure and are reported as mean ± standard deviation in Table 5, capturing calibration stability across different data splits. Second, ECE was also computed from pooled out-of-fold predictions to support the reliability diagrams shown in Figure 7, yielding lower absolute values due to increased sample support per probability bin. These two perspectives are complementary and jointly characterize both the stability and the global accuracy of probability estimates.

Although Table 5 and Figure 7 provide consistent evidence regarding calibration behavior, their interpretations reflect two complementary perspectives that may appear superficially contradictory. Table 5 reports fold-wise averaged calibration metrics (mean ± standard deviation across outer folds), capturing the expected calibration performance and its variability under different data partitions. In contrast, Figure 7 is based on pooled out-of-fold (OOF) predictions, providing a global visualization of calibration across the entire dataset.

This distinction explains why the unbalanced MLP achieves the lowest Brier score and ECE in Table 5, indicating superior average calibration performance, while the Random Forest appears visually closer to the diagonal in the reliability diagram. The latter reflects a more stable calibration behavior across probability ranges when predictions are aggregated, whereas the MLP exhibits localized deviations despite lower average error. From an operational perspective, such stability may be preferable when predicted probabilities are directly used to support threshold selection and cost-sensitive maintenance decisions.

Table 5 summarizes these fold-wise calibration results, including AP, Brier score, and ECE, expressed as mean ± standard deviation across the outer folds of the nested cross-validation procedure. These values capture not only average calibration quality but also the stability of these estimates across different data splits.

Variant	AP	Brier score	ECE
RF	0.707 ± 0.061	0.034 ± 0.004	0.058 ± 0.006
MLP	0.747 ± 0.060	0.016 ± 0.002	0.010 ± 0.005
MLP (SMOTE 1:1)	0.730 ± 0.066	0.033 ± 0.007	0.042 ± 0.011

Table 5. AP, Brier score, and ECE for the most competitive models

As shown in Table 5, the unbalanced MLP achieves the lowest Brier score and ECE on average across folds, indicating strong calibration performance under the fold-wise evaluation protocol. Random Forest exhibits moderately higher Brier score and ECE values, but remains reasonably calibrated across most of the probability range, consistent with its conservative error profile. In contrast, the MLP with SMOTE displays a clear calibration degradation: although its AP remains competitive, both its Brier score and ECE increase substantially, reflecting systematic overconfidence, particularly in the upper probability range, as evidenced by deviations above the diagonal in the reliability diagram.

Figure 7 presents the reliability diagrams based on pooled OOF predictions. RF produces probability estimates closely aligned with the diagonal reference, indicating reliable confidence estimates across the probability range. The MLP without SMOTE shows systematic miscalibration, while the SMOTE-based variant alters the probability distribution but does not resolve calibration error: although it may reduce some mid-range deviations, it remains miscalibrated, especially at higher confidence levels, consistent with the higher Brier score and ECE reported in Table 5.

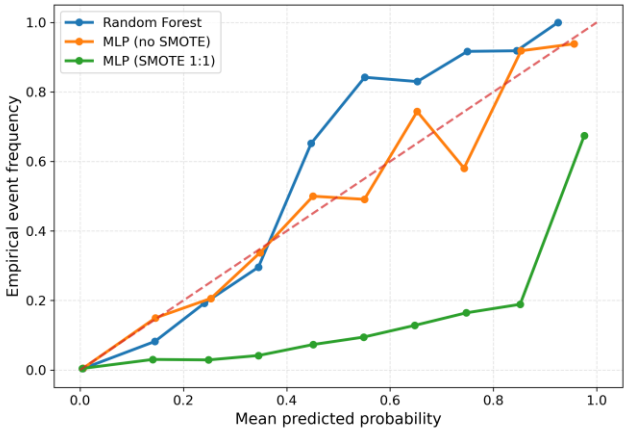

Figure 2. Calibration curves (outer-fold predictions)

To assess whether the observed differences reflect statistically significant performance gaps, Wilcoxon signed-rank tests were conducted on a fold-by-fold basis, comparing the MLP with SMOTE against RF. The tests do not support the hypothesis that MLP with SMOTE outperforms RF in ROC-AUC ($p \approx 0.90$) or AP ($p = 1.00$). Conversely, the Brier score is significantly higher for MLP with SMOTE ($p \approx 0.001$, alternative hypothesis "MLP+SMOTE > RF"), providing statistical evidence of inferior calibration.

Overall, these results underscore that effective PdM requires not only accurate failure ranking but also trustworthy probability estimates. From an operational and Bayesian decision-making perspective, miscalibrated probabilities can lead to suboptimal threshold selection, excessive false alarms, or delayed interventions. The combined analysis of confusion matrices, reliability diagrams, and calibration metrics therefore reinforces RF as the most reliable option when predicted probabilities are directly used to trigger maintenance actions within TPM-oriented frameworks.

4.4. Impact of Class Imbalance Strategy on MLP

The impact of class imbalance handling on neural network behavior is analyzed by comparing the MLP trained on the original imbalanced data with its SMOTE (1:1) counterpart. The results, summarized in Table 6, are computed from OOF predictions and isolate the effect of data-level resampling on discrimination, thresholded performance, and probability calibration.

Overall, SMOTE induces a pronounced shift in the operating profile of the MLP toward higher sensitivity to rare failure events. Macro-recall increases substantially, from 0.762 to 0.893, while AP remains broadly comparable (AP = 0.719 vs. 0.723). This indicates that resampling improves the network's ability to rank and detect minority-class observations without materially altering its global ranking performance.

However, these recall gains are accompanied by clear trade-offs. Macro-precision decreases from 0.868 to 0.708 and macro-F1 drops from 0.806 to 0.768, indicating that the additional detected failures are achieved at the cost of a larger false-positive burden. From a threshold-based operational standpoint, this implies that SMOTE shifts the classifier toward a more aggressive decision regime, favoring coverage of rare failures over false-alarm control.

Metric	MLP (no SMOTE)	MLP (SMOTE 1:1)	Δ (SMOTE − no SMOTE)
AP	0.719	0.723	0.004
Recall (macro)	0.762	0.893	0.131
Precision (macro)	0.868	0.708	−0.160
F1 (macro)	0.806	0.768	−0.038
Brier score	0.016	0.033	0.017
ECE	0.003	0.040	0.037

Table 6. Impact of SMOTE (1:1) on MLP performance based on pooled outer-loop OOF predictions

From a probabilistic decision-making perspective, the calibration metrics in Table 6 reveal an important side effect of resampling. While SMOTE improves recall, it substantially degrades probability reliability: the Brier score more than doubles (from 0.016 to 0.033), and ECE increases by an order of magnitude (from 0.003 to 0.040). This pattern indicates systematic misalignment between predicted probabilities and empirical event frequencies, consistent with the overconfidence observed in the reliability diagrams discussed in Section 4.3.

Such behavior has direct implications for Bayesian decision-making in PdM. When predicted probabilities are used to approximate posterior failure risk, miscalibration can distort expected cost calculations, leading either to overly frequent preventive actions or to poorly chosen decision thresholds. In this sense, SMOTE should be interpreted primarily as a mechanism to prioritize detection coverage (recall) rather than to improve the trustworthiness of predicted probabilities.

While Table 6 provides the numerical comparison, Figure 8 offers a compact visual summary of these trade-offs across discrimination (AP), thresholded performance (macro-precision, recall, and F1 at a 0.50 cutoff), and calibration quality (Brier score and ECE). Together, they reinforce that the principal benefit of SMOTE lies in increased sensitivity to rare failures, whereas both precision and calibration tend to deteriorate.

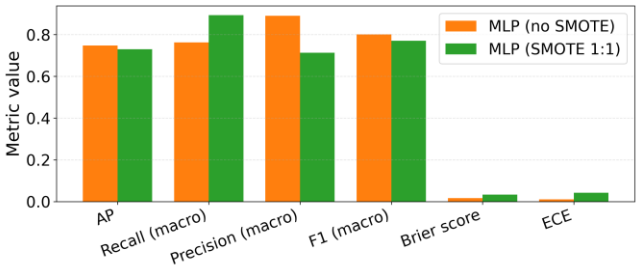

Figure 3. Impact of SMOTE (1:1) on MLP performance (outer-loop OOF predictions)

Importantly, these results frame the interpretation of SMOTE in operational terms. The resampled MLP may be advantageous in safety-critical contexts where missed failures are unacceptable and additional inspections are tolerable. Conversely, in resource-constrained maintenance environments where false alarms carry significant cost, the degradation in calibration and precision limits the practical attractiveness of aggressive resampling strategies.

4.5. False Positives and Operational Relevance

While discrimination- and recall-oriented metrics describe a model's ability to identify rare failure events, their practical relevance ultimately depends on how prediction errors translate into operational costs. In PdM, false positives are not benign: they correspond to unnecessary inspections, premature interventions, production interruptions, and increased workload for maintenance teams. Consequently, model assessment must be framed within a decision-theoretic perspective, where predicted probabilities are mapped to actions based on expected cost rather than solely on threshold-independent performance metrics.

From a Bayesian decision-making standpoint, maintenance actions are triggered by comparing the estimated posterior failure probability to a cost-dependent decision threshold that reflects the relative costs of false negatives (missed failures) and false positives (unnecessary interventions). Under severe class imbalance, even modest miscalibration can substantially distort expected-cost calculations, causing small threshold adjustments to produce disproportionate increases in false alarms. As a result, probability calibration and threshold stability become central determinants of operational viability.

The results presented in Sections 4.3 and 4.4 highlight this mechanism. Models exhibiting poorer calibration, particularly the SMOTE-enhanced MLP, tend to assign overly confident probabilities to uncertain operating regimes. When such scores are translated into decision thresholds, the resulting operating points are highly sensitive to threshold selection, amplifying false-positive rates as sensitivity is increased. Although resampling improves recall, the accompanying degradation in probability reliability limits its effectiveness when decisions are guided by expected cost rather than by recall alone.

In contrast, the RF model demonstrates a more conservative and stable decision profile. Its comparatively smoother calibration behavior enables predicted probabilities to be translated into operational thresholds with greater robustness across a range of cost assumptions. From an expected-loss perspective, this stability mitigates the rapid escalation of false alarms that would otherwise erode maintenance efficiency and OEE.

These findings reinforce that improvements in recall achieved through aggressive imbalance-handling strategies must be interpreted considering downstream operational implications. In safety-critical environments where the cost of missed failures dominates, recall-oriented models may still be justified despite higher false-positive rates. However, in typical industrial contexts with constrained maintenance resources, models that balance discrimination with reliable probability estimates offer a more sustainable and cost-effective path to deployment. Under such conditions, RF emerges as the most operationally robust option among the evaluated approaches.

4.6. Feature Importance Analysis

Feature importance analysis was conducted exclusively for the RF model due to its inherent interpretability as a tree-based ensemble algorithm. Unlike neural networks, whose interpretability typically relies on post-hoc methods such as SHAP or LIME, RF provides a direct and well-established mechanism for assessing the relative contribution of input variables through its internal splitting structure. Although post-hoc explainability techniques could also be applied to the MLP, they introduce additional computational overhead and methodological variability, which falls outside the scope of establishing transparent and reproducible baselines.

Figure 9 reports the global feature importance scores computed using the mean decrease in impurity (Gini

importance) criterion. It is well recognized that impurity-based importance measures may exhibit bias toward variables with higher variance or a larger number of potential split points. Accordingly, the reported importances should be interpreted as relative indicators of contribution within the trained model, rather than as causal measures of variable influence.

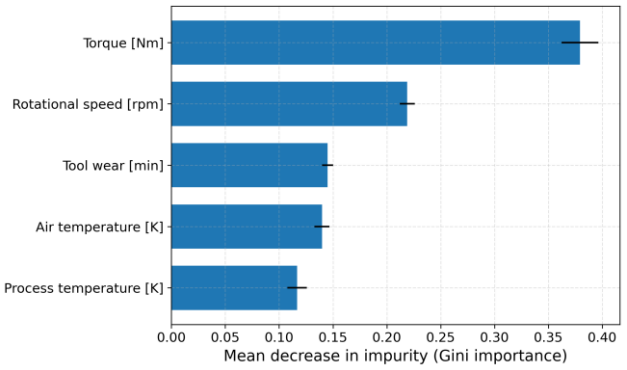

Figure 9. Random Forest global feature importance

Despite these known limitations, the importance ranking proved qualitatively stable across the outer cross-validation folds, with Torque and Tool Wear consistently emerging as the most influential predictors. This stability suggests that the dominance of these variables is not driven by a single data partition or random fluctuation, but reflects persistent patterns learned by the ensemble across independent training subsets.

From a domain perspective, the identified importance hierarchy aligns closely with physical intuition and prior findings in the predictive maintenance literature (Bezerra et al., 2024). Torque directly reflects mechanical load and frictional resistance experienced by the motor, serving as an early indicator of abnormal operating conditions. Tool Wear captures cumulative degradation effects over operational cycles, making it a natural proxy for progressive failure risk. Process temperature also exhibits substantial relevance, consistent with the well-established association between thermal stress and mechanical failures in industrial motors. Rotational speed and air temperature display lower relative importance but still contribute indirectly to model performance, particularly through interactions with the dominant variables.

Overall, while impurity-based feature importance should not be interpreted causally, the combination of ranking stability across folds and alignment with known degradation mechanisms supports the validity of the RF as an interpretable and operationally meaningful baseline. These results reinforce that transparent models can provide actionable insights for maintenance decision support without requiring complex post-hoc explainability techniques.

4.7. Comparative Analysis with Recent Literature

Recent advances in predictive maintenance research using the AI4I dataset and related benchmarks have extended PdM beyond classical machine learning baselines. Temporal deep learning models incorporating attention mechanisms have reported gains in AUROC and F1 by explicitly modeling multivariate temporal dependencies (Liu & Su, 2024). Ordinal reformulations of the PdM problem, such as OPMEB, leverage the ordered nature of degradation states and outperform nominal multiclass baselines in structured degradation scenarios (Yürek & Birant, 2024). In parallel, continual learning frameworks address non-stationarity by mitigating catastrophic forgetting and sustaining performance under concept drift, offering increased robustness in dynamic industrial environments (Benatia et al., 2025). Finally, privacy-preserving approaches based on multi-key homomorphic encryption enable collaborative model training while protecting sensitive operational data, an increasingly relevant requirement in industrial consortia (Kang et al., 2024).

While these approaches demonstrate important methodological advances, they also introduce additional complexity, data requirements, and implementation challenges. In contrast, the present study deliberately focuses on establishing a transparent and reproducible baseline under severe class imbalance, emphasizing rigorous validation, probability calibration, and operational interpretability rather than architectural novelty.

Table 7 summarizes the qualitative comparison between the baseline models evaluated in this work and representative advanced PdM approaches, highlighting key trade-offs across methodological dimensions.

Approach	Simplicity	Reproducibility	Temporal Modeling	Robustness to Drift	Privacy
Baseline (RF, MLP)	Simple	High	None	Sensitive to drift	None
Temporal Attention	More complex	High	Captures sequences	Still limited	None
Ordinal PdM (OPMEB)	More complex	High	Exploits ordinal degradation	Limited	None
Continual Learning	More complex	Partial	Can include sequences	Handles drift/forgetting	None
Privacy-preserving	More complex	Partial	Depends on base model	Collaborative updates	Secure data sharing

Table 7. Comparative analysis of baseline models and advanced PdM approaches across key attributes

Rather than competing with these advanced methods, the contribution of this work lies in providing a statistically rigorous and operationally grounded reference point. By explicitly addressing class imbalance, calibration quality, false-positive burden, and interpretability, the RF and MLP baselines evaluated here offer meaningful insight into the practical trade-offs that underpin PdM deployment. They clarify how gains in discrimination or recall may be offset by degraded calibration or increased operational cost, dimensions that are often underemphasized in purely performance-driven comparisons.

As such, the baseline established in this study serves as a useful and transparent benchmark against which more complex approaches, such as temporal deep learning (Liu & Su, 2024), ordinal degradation modeling (Yürek & Birant, 2024), continual adaptation under drift (Benatia et al., 2025), and privacy-preserving frameworks (Kang et al., 2024), can be meaningfully and fairly evaluated. This positioning reinforces the role of transparent, reproducible baselines as an essential foundation for advancing PdM research that is not only accurate, but also reliable and operationally actionable.

5. DISCUSSION

This section synthesizes the main findings of the study, connecting model performance with practical implications for PdM within TPM frameworks. Beyond numerical evaluation, the discussion emphasizes how discrimination, calibration, and error trade-offs jointly support data-driven maintenance decisions, while also acknowledging key limitations and outlining directions for future research.

The experimental findings highlight distinct operational profiles for the evaluated models under severe class imbalance. RF achieves the most robust balance between discrimination and calibration reliability, making it suitable for cost-sensitive maintenance planning where predicted probabilities directly inform intervention decisions. In contrast, the MLP combined with SMOTE prioritizes detection sensitivity at the expense of calibration quality, representing a viable alternative when missed failures carry disproportionately high cost.

From a TPM perspective, the consistent identification of torque and tool wear as dominant predictors (Section 4.6) reinforces their utility as actionable condition indicators. Rather than prescribing fixed decision rules, the calibrated probability estimates provided by interpretable models like RF can inform operator dashboards, condition-monitoring thresholds, and focused improvement activities, thereby contributing to enhanced equipment availability and OEE.

Some limitations of the present work should be acknowledged. The analysis relies on a synthetic and controlled benchmark dataset (AI4I) and addresses a binary failure classification task, which may not fully reflect the complexity of real industrial environments. Although AI4I captures key statistical properties of industrial processes and provides a controlled setting for methodological evaluation, real-world deployments typically involve additional challenges such as sensor noise, signal drift, delayed or uncertain failure labeling, and evolving operating regimes. These factors may affect both model discrimination and probability calibration, potentially requiring periodic recalibration and continuous performance monitoring in practice.

In addition, class imbalance was handled exclusively through SMOTE; alternative strategies such as cost-sensitive learning, adaptive resampling, or ensemble-based balancing could further modulate the trade-off between recall and false alarms. Third, the problem formulation was restricted to binary classification, whereas multiclass degradation modeling or Remaining Useful Life (RUL) prediction may yield richer operational insights. Finally, interpretability analysis focused on impurity-based feature importance from the RF model; extending interpretability to neural networks through post-hoc methods such as SHAP or Integrated Gradients could enhance transparency, albeit at the cost of additional methodological complexity.

These limitations also delineate clear paths for future research. Future work may explore post-hoc probability calibration techniques, such as Platt scaling and isotonic regression, to further improve the reliability of predicted failure probabilities without altering the underlying classifiers. Promising directions include the integration of temporal deep learning architectures with attention mechanisms to model sequential dependencies in sensor data (Liu & Su, 2024), ordinal formulations of degradation processes such as OPMEB to exploit ordered failure states (Yürek & Birant, 2024), continual learning frameworks to adapt to evolving operating conditions and mitigate catastrophic forgetting (Benatia et al., 2025), and privacy-preserving strategies such as multi-key homomorphic encryption to enable collaborative model development without exposing sensitive operational data (Kang et al., 2024).

6. CONCLUSION

This study establishes a reproducible and calibration-aware baseline for predictive maintenance using the AI4I dataset, with explicit attention to severe class imbalance, probability reliability, and operational decision relevance. The results demonstrate that RF achieves the most robust balance among the evaluated approaches between performance and calibrated probability estimates, exhibiting stable behavior across decision thresholds. In contrast, the MLP combined

with SMOTE represents a viable alternative in recall-oriented scenarios, albeit with clear trade-offs in calibration quality and false-positive burden. These findings confirm that model suitability in PdM is inherently context-dependent and should be guided by how predictive performance aligns with operational cost structures and maintenance priorities.

This work highlights that probability calibration and false-positive control are critical for translating predictive outputs into maintenance decisions. By explicitly evaluating calibration error and threshold sensitivity, the study demonstrates that well-calibrated probability estimates are a prerequisite for risk-based planning within TPM frameworks.

Framed within a TPM perspective, the proposed baseline illustrates how interpretable ML models can support operator-driven monitoring, autonomous maintenance activities, and focused improvement initiatives aimed at enhancing OEE. Rather than prescribing fixed decision rules, the framework provides calibrated risk estimates and interpretable signals that enable data-driven maintenance planning under realistic operational constraints.

In this context, the effective integration of predictive models into maintenance routines also depends on broader organizational capabilities related to strategy, knowledge management, and dynamic capabilities, as evidenced in recent empirical studies on innovation management in digital ecosystems (Pinto et al., 2025).

Despite its limitations, this study offers a rigorous and operationally grounded reference for predictive maintenance research and practice. By integrating discrimination performance, probability calibration, and decision relevance within a transparent and reproducible framework, the proposed baseline enables fair comparison with more complex approaches. Future developments should build upon this foundation to address temporal dynamics, richer degradation modeling, and deployment constraints, ensuring that advances in PdM remain not only accurate, but also reliable, interpretable, and practically actionable.

NOMENCLATURE

ACC	Accuracy
AI4I (2020)	AI4I 2020 PdM Dataset (UCI ML Rep.)
AI4I-PMDI	AI4I PdM Dataset with Irregularities (real-world extension)
ANN	Artificial Neural Network
AP	Average Precision
AUROC	Area Under the Receiver Operating Characteristic Curve
Bi-LSTM	Bidirectional Long Short-Term Memory
BO	Bayesian Optimization
ECE	Expected Calibration Error
FP / FN	False Positive / False Negative
F1	F1-score (harmonic mean of precision and recall)
k-NN	k-Nearest Neighbors
ML	Machine Learning
MLP	Multilayer Perceptron
OEE	Overall Equipment Effectiveness
OPMEB	Ordinal Predictive Maintenance with Ensemble Binary Decomposition
OOF	*outer-loop out-of-fold*
OVA / OVO	One-Versus-All / One-Versus-One
PdM	Predictive Maintenance
RF	Random Forest
ROC	Receiver Operating Characteristic
RUL	Remaining Useful Life
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TPM	Total Productive Maintenance
TPR / TNR	True Positive Rate (Sensitivity/Recall) / True Negative Rate (Specificity)
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting algorithm

REFERENCES

Ahuja, I. P. S., & Khamba, J. S. (2008). Total productive maintenance: literature review and directions. *International journal of quality & reliability management*, 25(7), 709-756. doi: 10.1108/02656710810890890.

Alshkeili, H. M. H. A., Almheiri, S. J., & Khan, M. A. (2025). Privacy-Preserving Interpretability: An Explainable Federated Learning Model for Predictive Maintenance in Sustainable Manufacturing and Industry 4.0. *AI*, 6(6), 117. doi: 10.3390/ai6060117.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115. doi: 10.1016/j.inffus.2019.12.012.

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

Bezerra, F. E., Oliveira Neto, G. C. D., Cervi, G. M., Francesconi Mazetto, R., Faria, A. M. D., Vido, M., ... & Amorim, M. (2024). Impacts of feature selection on predicting machine failures by machine learning algorithms. *Applied Sciences*, 14(8), 3337. doi: 10.3390/app14083337.

Benatia, M. A., Hafsi, M., & Ayed, S. B. (2025). A continual learning approach for failure prediction under non-stationary conditions: Application to condition

monitoring data streams. *Computers & Industrial Engineering*, 204, 111049. doi: 10.1016/j.cie.2025.111049.

Çiftpınar, A. B., Kanar, P., & Cıcek, Z. I. E. (2025). Failure Prediction Using Ensemble Learning: A Comparative Study with Synthetic and Real-World Datasets. *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi*, 25(4), 785-797. doi: 10.35414/akufemubid.1571811.

Esteban, A., Cano, A., Ventura, S., & Zafra, A. (2025). Simultaneous fault prediction in evolving industrial environments with ensembles of Hoeffding adaptive trees. Applied Intelligence, 55(13), 930. https://doi.org/10.1007/s10489-025-06786-7.

Gomaa, A. H. (2025). Advancing Total Productive Maintenance in Smart Manufacturing: From Methodology to Implementation. *International Journal of Smart Manufacturing*. doi: 10.70322/ism.2025.10019.

Kamel, H. (2022, July). Artificial intelligence for predictive maintenance. In Journal of Physics: Conference Series (Vol. 2299, No. 1, p. 012001). *IOP Publishing*. doi: 10.1088/1742-6596/2299/1/012001.

Kang, D. H. E., Kim, D., Song, Y., Lee, D., Kwak, H., & Anthony, B. W. (2024). Harnessing the potential of shared data in a secure, inclusive, and resilient manner via multi-key homomorphic encryption. *Scientific Reports*, 14(1), 13626. doi: 10.1038/s41598-024-63393-1.

Liu, C. L., & Su, H. C. (2024). Temporal learning in predictive health management using channel-spatial attention-based deep neural networks. *Advanced Engineering Informatics*, 62, 102604. doi: 10.1016/j.aei.2024.102604.

Lyubchyk, L., Grinberg, G., & Yamkovyi, K. (2023, October). Machine Learning-Based Predictive Maintenance using Data Aggregation via Regularized Clustering. *In 2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 1-6). IEEE. doi: 10.1109/DESSERT61349.2023.10416542.

Muchiri, P., & Pintelon, L. (2008). Performance measurement using overall equipment effectiveness (OEE): literature review and practical application discussion. *International journal of production research*, 46(13), 3517-3535. doi: 10.1080/00207540601142645.

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015, February). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1). doi: 10.1609/aaai.v29i1.9602.

Nunes, P., Santos, J., & Rocha, E. (2023). Challenges in predictive maintenance–A review. *CIRP Journal of Manufacturing Science and Technology*, 40, 53-67. doi: 10.1016/j.cirpj.2022.11.004.

Pinto, S. D. L., Muniz Jr, J., Freitas, C. R. D., & Dale Luche, J. R. (2025). A Framework for the Innovation Management Capacity: Empirical Evidence from the Porto Digital Cluster in Brazil. Administrative Sciences, 15(5), 191. doi:10.3390/admsci15050191.

Prashanth, B. S., Manoj Kumar, M. V., Almuraqab, N., & Puneetha, B. H. (2025). Leveraging Safe and Secure AI for Predictive Maintenance of Mechanical Devices Using Incremental Learning and Drift Detection. *Computers, Materials & Continua*, 83(3), 4979–4998. doi: 10.32604/cmc.2025.060881.

Presciuttini, A., Cantini, A., & Portioli-Staudacher, A. (2024, November). From Explanations to Actions: Leveraging SHAP, LIME, and Counterfactual Analysis for Operational Excellence in Maintenance Decisions. *In 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-6). IEEE. doi: 10.1109/ICECCME62383.2024.10796106.

Ronzoni, N., De Marco, A., & Ronchieri, E. (2022, July). Predictive Maintenance Experiences on Imbalanced Data with Bayesian Optimization Approach. *In International Conference on Computational Science and Its Applications* (pp. 120-137). Cham: Springer International Publishing. doi: 10.1007/978-3-031-10536-4_9.

Shivaramu, P. (2025). Optimizing Manufacturing Processes with Predictive Maintenance Using Machine Learning and Lean Six Sigma. *SSRN Electronic Journal*. doi: 10.2139/ssrn.5161097.

Tortorella, G., Saurin, T. A., Fogliatto, F. S., Tlapa, D., Moyano-Fuentes, J., Gaiardelli, P., ... & Forstner, F. F. (2022). The impact of Industry 4.0 on the relationship between TPM and maintenance performance. *Journal of Manufacturing Technology Management*, 33(3), 489-520. doi: 10.1108/JMTM-10-2021-0399.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432. doi: 10.1371/journal.pone.0118432.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Computational Intelligence Magazine, 13(4), 59-76. doi: 10.1109/MCI.2018.2866730.

UCI Machine Learning Repository. (2020). AI4I 2020 predictive maintenance dataset. *Irvine, CA: University of California, School of Information and Computer Science*. doi: 10.24432/C5HS5C.

Yürek, O. E., & Birant, D. (2024). A new approach: ordinal predictive maintenance with ensemble binary decomposition (OPMEB). *Turkish Journal of Electrical Engineering and Computer Sciences*, 32(4), 534-554. doi: 10.55730/1300-0632.4086.

Zonta, T., Da Costa, C. A., da Rosa Righi, R., de Lima, M. J., Da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & industrial engineering*, 150, 106889. doi: 10.1016/j.cie.2020.106889

BIOGRAPHIES

José Roberto Dale Luche is a Professor in the Department of Production Engineering at the Faculty of Engineering and Sciences of Guaratinguetá, São Paulo State University (FEG/UNESP), Brazil. He received the Ph.D. and M.Sc. degrees in Production Engineering with an emphasis on Operations Research and holds a specialization in Database Systems. He earned undergraduate degrees in Production Engineering and Systems Analysis and Development, as well as a Licentiate degree in Mathematics (R2). His research focuses on Mixed Reality, Educational Technologies, Information Systems, and Operations Research.

Blaha Gregory Correia dos Santos Goussain is a Professor in the Department of Production Engineering at the Faculty of Engineering and Sciences of Guaratinguetá, São Paulo State University (FEG/UNESP), Brazil. He received the Ph.D. degree in Engineering, including a doctoral research internship (PDSE/CAPES) at the University of Tennessee, Knoxville, USA, and the M.Sc. degree in Engineering. He also holds specializations in Didactic-Pedagogical Processes for Distance Learning and in Production Management. He earned a B.Sc. degree in Production Engineering and a Licentiate degree in Mathematics. His research focuses on Production Engineering, Machine Learning, Neuroscience, and Applied Statistics.

Claudia Regina de Freitas is a Professor in the Department of Production Engineering at the Faculty of Engineering and Sciences of Guaratinguetá, São Paulo State University (FEG/UNESP), Brazil. She received the Ph.D. degree in Public Health from the University of Campinas (UNICAMP) and the M.Sc. degree in Sciences. She earned undergraduate degrees in Psychology and in Pedagogy. Her research focuses on Public Health, Psychology, and Education.