

Machine Learning-Based Failure Detection in IoT-Enabled Engineering Systems: A Time-Aware and Imbalance-Sensitive Evaluation Using the MetroPT-3 Dataset

Original Research Article | Volume 1 | Issue 5 | 2026 | Article Number: 042

Accepted: 24 September 2026 | Published: 10 October 2026 | ISSN: 2979-8582 (Online)



Crossref : <https://doi.org/10.67693/BJCR-9KU8493Z>



Mustapha Ibrahim

Department of Computer Science, School of Science and Technology, The Federal Polytechnic Damaturu, Yobe State, Nigeria

Correspondence: Mustapha Ibrahim, mustee2004@yahoo.com

Abstract

The integration of Internet of Things (IoT) technologies and machine learning (ML) has advanced intelligent monitoring and predictive maintenance of engineering systems. However, failure detection using industrial sensor data remains challenging because failure events are rare compared with normal operation, resulting in highly imbalanced datasets where accuracy alone may provide misleading conclusions. This study presents a time-aware and imbalance-sensitive evaluation of Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost) classifiers for detecting failure-related conditions using the MetroPT-3 dataset. The dataset contains 1,516,948 multivariate time-series observations collected from a metro Air Production Unit (APU), including pressure, temperature, electrical, and digital control measurements. Documented failure periods were transformed into a binary classification problem, and a chronological 70:30 train-test split was applied to preserve temporal ordering and reduce information leakage. The test dataset contained 455,085 observations, including 1,622 failure cases (0.356%). Models were evaluated using accuracy, precision, recall, F1-score, balanced accuracy, MCC, PR-AUC, ROC-AUC, confusion matrices, and feature-importance analysis. Although all models achieved accuracy above 99%, failure-detection performance varied considerably. Decision Tree achieved 99.54% accuracy but detected only 0.86% of failures. Random Forest achieved 99.64% accuracy and zero recall at the default threshold but showed strong ranking capability with a ROC-AUC of 0.9973. XGBoost achieved the best minority-class performance with 60.74% precision, 39.40% recall, 47.79% F1-score, 0.6965 balanced accuracy, and 0.4877 MCC. Feature analysis identified DV_pressure, Oil_temperature, and TP2 as dominant predictors, with the first two contributing approximately 81.81% of XGBoost feature importance. The findings demonstrate that reliable failure detection requires temporal validation, imbalance-sensitive metrics, probability-based evaluation, and appropriate threshold selection rather than accuracy alone.

Keywords: Machine Learning, Failure Detection, Predictive Maintenance, Internet of Things, XGBoost, Random Forest, MetroPT-3, Class Imbalance, Condition Monitoring

1. INTRODUCTION

The rapid advancement of sensing technologies, communication networks, edge computing, and artificial intelligence has transformed traditional engineering assets into intelligent connected systems. Industrial equipment increasingly uses sensors to collect operational data, enabling real-time monitoring and data-driven decision-making. The integration of IoT technologies and machine learning supports condition monitoring, fault detection, and predictive maintenance. Measurements such as pressure, temperature, electrical current, vibration, and flow provide valuable information about equipment behaviour, allowing models to identify abnormal patterns and potential failures.

Predictive maintenance is essential in intelligent engineering systems because unexpected failures can cause production losses, increased maintenance costs, and safety risks. Unlike corrective maintenance after failure or preventive maintenance based on fixed schedules, predictive maintenance uses operational data to estimate equipment condition and enable timely intervention. Compressor systems are suitable for data-driven monitoring because their operating conditions are reflected through measurable physical variables. Recent studies have reported increased use of time-series analysis, anomaly detection, and machine-learning methods to improve compressor reliability (Forbicini et al., 2024).

The railway sector is another important application area, where failures in critical subsystems can affect reliability, safety, and maintenance efficiency. The MetroPT dataset, collected from an Air Production Unit (APU) in a metro system, provides a realistic benchmark for industrial anomaly detection and failure classification (Veloso et al., 2022). Machine-learning approaches have been widely applied in predictive maintenance. Decision Tree provides interpretable classification rules, Random Forest improves robustness through ensemble learning, and XGBoost offers strong nonlinear modelling capability. Previous studies have demonstrated the effectiveness of XGBoost for industrial classification due to its ability to capture complex sensor relationships (Salim et al., 2022).

However, failure detection remains challenging because failures are rare compared with normal operating conditions. Highly imbalanced datasets may cause models to favour the majority class, allowing classifiers to achieve high accuracy while failing to detect actual failures. Therefore, metrics such as precision, recall, F1-score, MCC, and PR-AUC are recommended for reliable rare-event evaluation (Imani et al., 2026). In addition, industrial sensor data are time-dependent, and random train-test splitting may introduce similar observations into training and testing sets, causing information leakage and optimistic performance estimates. Time-aware evaluation through chronological splitting provides a more realistic assessment of future operating conditions (Al-Ali & Alharbi, 2026).

Therefore, this study evaluates Decision Tree, Random Forest, and XGBoost for MetroPT-3 failure detection using a time-aware and imbalance-sensitive framework. The study examines model capability beyond accuracy by applying multiple performance metrics to assess rare failure identification.

1.1 Problem Statement

Although machine-learning approaches have shown strong potential for engineering condition monitoring, several challenges remain. First, many studies rely heavily on accuracy despite severe class imbalance, which may overestimate practical performance. Second, evaluation strategies that ignore temporal ordering may produce unrealistic results due to information leakage. Third, effective engineering deployment requires understanding which sensor variables contribute most strongly to model decisions. Therefore,

there is a need for systematic evaluation frameworks that simultaneously consider temporal validity, class imbalance, and engineering interpretation.

1.2 Aim of the Study

The aim of this study is to experimentally evaluate machine-learning models for failure detection in IoT-enabled engineering systems using a time-aware and imbalance-sensitive framework based on the MetroPT-3 dataset.

1.3 Research Objectives

The objectives of this study are to:

1. evaluate the effectiveness of Decision Tree, Random Forest, and XGBoost classifiers in identifying failure-related operating conditions in the MetroPT-3 dataset;
2. investigate whether high classification accuracy corresponds to effective failure detection under extreme class imbalance;
3. compare machine-learning models to determine which classifier provides the best balance between failure detection and false alarms; and
4. identify the sensor measurements that contribute most strongly to machine-learning-based failure classification.

1.4 Research Questions

This study addresses the following research questions:

RQ1: How effectively can Decision Tree, Random Forest, and XGBoost identify failure-related operating conditions in the MetroPT-3 dataset?

RQ2: Does high classification accuracy correspond to effective failure detection under extreme class imbalance?

RQ3: Which model provides the best balance between detecting failures and minimizing false alarms?

RQ4: Which sensor measurements contribute most strongly to machine-learning-based failure classification?

1.5 Research Contributions

The main contributions of this study are:

1. Time-aware evaluation: A chronological train-test strategy is applied to provide a realistic assessment of future-condition generalization.
2. Imbalance-sensitive analysis: The study demonstrates the limitations of accuracy-based evaluation under extreme class imbalance.
3. Comprehensive performance evaluation: Models are assessed using minority-sensitive metrics, probability-based evaluation, and confusion-matrix analysis.
4. Model interpretation: Feature-importance analysis identifies the sensor variables most influential in failure classification.

2. RELATED LITERATURE

2.1 IoT-Enabled Intelligent Engineering Systems

The development of Industrial Internet of Things (IIoT) technologies has transformed conventional engineering assets into connected and intelligent systems. Sensor networks now enable continuous monitoring of operational variables, including pressure, temperature, electrical current, vibration, flow

rate, and control signals. These measurements support condition monitoring by allowing machine-learning algorithms to identify abnormal operating patterns and provide early indications of equipment degradation. The integration of IoT sensing and artificial intelligence has therefore become a key approach for improving reliability, reducing downtime, and supporting predictive maintenance decisions.

Predictive maintenance is particularly important in industrial systems where unexpected failures may result in production losses, increased maintenance costs, and safety risks. By analysing historical and real-time sensor data, machine-learning models can identify complex relationships between operational conditions and failure events. Compressor-based systems represent an important application area because their operating conditions are strongly reflected through measurable physical parameters. Recent studies have reported increasing adoption of time-series analysis, machine learning, and anomaly detection techniques for improving compressor reliability and reducing maintenance requirements (Forbicini et al., 2024).

2.2 Machine Learning Approaches for Engineering Failure Detection

Machine-learning methods have become widely used in engineering failure detection because of their ability to model nonlinear relationships among multiple sensor variables. Decision Tree models provide an interpretable classification approach by generating decision rules through recursive feature splitting. However, individual trees may suffer from overfitting and sensitivity to noisy data. Random Forest improves upon individual decision trees through ensemble learning, where multiple trees are trained using different samples and feature subsets. This improves generalization, captures complex relationships, and provides feature-importance estimates. However, under severe class imbalance, Random Forest may still favour the majority class unless appropriate imbalance-handling strategies are applied.

Extreme Gradient Boosting (XGBoost) is a gradient-boosting algorithm that sequentially builds decision trees to correct previous prediction errors. Its advantages include nonlinear modelling capability, regularization, feature-interaction learning, and support for class-weight adjustment. Previous research has demonstrated the effectiveness of XGBoost for predictive-maintenance applications and industrial classification problems (Salim et al., 2022).

2.3 MetroPT Dataset and Previous Research

The MetroPT dataset is a benchmark dataset developed for predictive-maintenance and anomaly-detection research using real operational measurements from an Air Production Unit (APU) of a metro railway system. The dataset contains multivariate sensor measurements, including, pressure measurements; temperature measurements; electrical current measurements; digital control signals; and documented failure periods.

The availability of real industrial data makes MetroPT valuable for evaluating machine-learning approaches under realistic operating conditions. Veloso et al. (2022) introduced the dataset and demonstrated its suitability for predictive-maintenance research. Subsequent studies have investigated different approaches using MetroPT. Meira et al. (2023) developed a data-driven predictive-maintenance framework for railway systems, while deep-learning approaches such as LSTM autoencoders have been applied to detect abnormal operating patterns through reconstruction errors (Akshitha Mary A. C. et al., 2025). Hybrid machine-learning and deep-learning approaches have also been explored, with Peng et al. (2026) proposing IntelliMetro-Hybrid for metro predictive-maintenance applications.

However, direct comparison between existing studies remains difficult due to differences in, failure definitions; preprocessing procedures; feature-engineering approaches; data-splitting strategies; balancing methods; and evaluation metrics. Therefore, reported accuracy values alone are insufficient for determining practical failure-detection capability.

2.4 Class Imbalance in Engineering Failure Detection

Class imbalance remains a major challenge in machine-learning-based failure detection. Industrial equipment normally operates under healthy conditions, while failures occur rarely. Consequently, failure datasets often contain a large majority of normal observations and a limited number of failure cases. Under such conditions, accuracy can provide misleading results. For example, when failures represent only 0.356% of observations, a classifier predicting all samples as normal can achieve approximately 99.64% accuracy while failing to detect any failures. Therefore, evaluation of rare-event classification requires metrics that consider minority-class performance, including precision, recall, F1-score, balanced accuracy, Matthews Correlation Coefficient (MCC), and Precision-Recall Area Under Curve (PR-AUC). Imani et al. (2026) demonstrated that ROC-AUC alone may be misleading in highly imbalanced problems and recommended complementary evaluation measures for rare-event detection.

2.5 Time-Aware Evaluation of Sensor Data

Industrial sensor data are inherently temporal because measurements collected within short periods are often correlated. Random train-test splitting may introduce similar observations into both training and testing datasets, resulting in information leakage and overly optimistic performance estimates. Time-aware evaluation methods, such as chronological splitting, provide a more realistic assessment by training models on historical observations and testing them on future conditions. Al-Ali and Alharbi (2026) emphasized the importance of temporal validation and reliable evaluation procedures in predictive-maintenance research. Therefore, this study applies chronological train-test separation to improve the reliability of model evaluation.

2.6 Explainability and Feature Importance

In engineering applications, predictive performance alone is insufficient because maintenance decisions require understanding the variables influencing model predictions. Feature-importance analysis provides insight into which sensor measurements contribute most strongly to classification outcomes. However, feature importance indicates model dependence rather than direct causality. A highly influential variable means that the model relies strongly on that feature, but it does not prove that the variable directly causes failure. Recent MetroPT studies have demonstrated the value of combining anomaly detection with explainable artificial intelligence methods, such as SHAP, to improve engineering interpretation and support decision-making (Davari et al., 2024).

2.7 Research Gap

Although previous studies demonstrate the potential of machine learning for MetroPT-based predictive maintenance, several limitations remain. First, many studies focus primarily on accuracy without adequately considering extreme class imbalance. Second, random data-splitting approaches may not represent future operational conditions. Third, differences in preprocessing, feature engineering, and evaluation strategies make comparison between studies difficult. Therefore, further research is required to integrate, temporal validation; imbalance-sensitive evaluation; probability-based assessment; confusion-matrix analysis; and engineering interpretation. This study addresses these gaps by providing a time-aware comparison of Decision Tree, Random Forest, and XGBoost models for failure detection under severe class imbalance.

2.8 Previous Studies

Several studies have investigated machine-learning and deep-learning approaches using the MetroPT dataset for predictive maintenance and anomaly detection. However, direct comparison remains challenging because existing studies differ in target definition, preprocessing procedures, validation

strategies, and evaluation metrics. To position the contribution of this study within existing MetroPT-based predictive-maintenance research, a comparative analysis of previous approaches is presented in Table 1. The comparison highlights differences in objectives, modelling strategies, evaluation methodologies, and reported outcomes. Since previous studies employ different problem formulations, data-processing procedures, validation strategies, and performance metrics, direct numerical comparison should be interpreted with caution. The main distinction of the present study is its focus on realistic failure detection under severe class imbalance through chronological evaluation and imbalance-sensitive performance assessment.

Table 1. Comparison of the Proposed Study with Previous MetroPT-Based Research

Study	Approach	Objective	Evaluation Approach	Reported Performance	Limitation Compared with Current Study
Veloso et al. (2022)	MetroPT dataset introduction	Dataset development for predictive maintenance research	Dataset analysis and anomaly investigation	Demonstrated suitability of MetroPT for industrial monitoring	Did not provide a comprehensive imbalance-sensitive ML benchmark
Meira et al. (2023)	Data-driven predictive-maintenance framework	Identification of abnormal APU behaviour	Machine-learning-based predictive maintenance evaluation	Demonstrated potential of ML for railway maintenance	Different experimental design prevents direct metric comparison
Salim et al. (2022)	XGBoost-based predictive maintenance	Failure classification	Conventional classification evaluation	Reported strong XGBoost performance	Did not specifically address severe class imbalance and temporal leakage
Akshitha Mary A. C. et al. (2025)	LSTM autoencoder	Anomaly detection using reconstruction error	Deep-learning anomaly detection	Demonstrated ability to identify deviations from normal behaviour	Focused on anomaly detection rather than binary failure classification
Peng et al. (2026)	IntelliMetro-Hybrid (ML + DL)	Metro predictive maintenance	Hybrid model evaluation	Reported high predictive performance	Different architecture and evaluation methodology
This study	Decision Tree, Random Forest, XGBoost	Failure-related condition classification	Chronological split + imbalance-sensitive evaluation	XGBoost: 60.74% precision, 39.40% recall, 47.79% F1-score, 0.4877 MCC, 0.9927 ROC-AUC	Focuses specifically on realistic failure detection under extreme imbalance

3. MATERIALS AND METHODS

3.1 Research Design

This study adopts a quantitative experimental research design to evaluate machine-learning models for failure detection in IoT-enabled engineering systems. The research workflow consisted of data acquisition, preprocessing, failure-label construction, feature preparation, chronological data splitting, model training, performance evaluation, and feature-importance analysis. The objective was not only to identify the model with the highest classification accuracy but also to determine whether the models could effectively detect

rare failure-related operating conditions under severe class imbalance.

The overall workflow is summarised as:

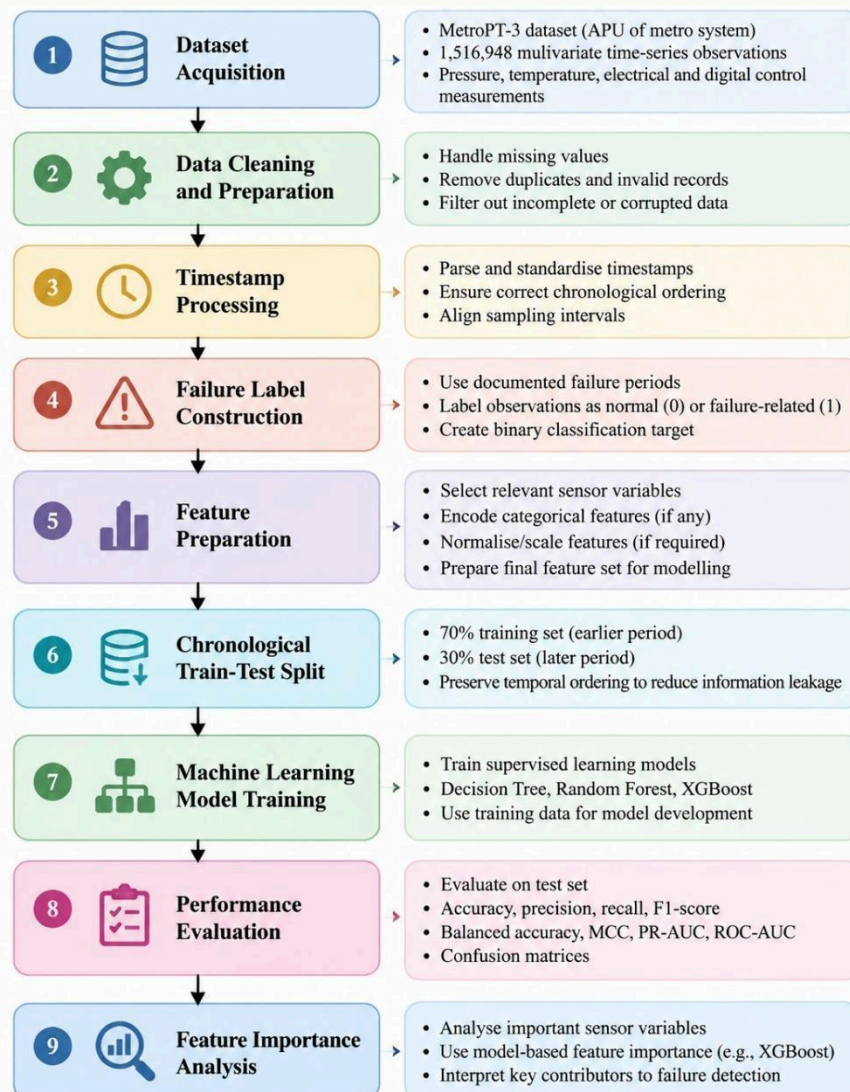


Figure 1: Research framework

7. EXPERIMENTAL DESIGN

7.1 Chronological Train-Test Partition

Because MetroPT-3 is a time-series dataset, chronological splitting was applied instead of random sampling. The dataset was divided as follows training 70% and testing 30%. The earliest observations were used for training, while later observations were reserved for testing. This approach reflects a practical predictive-maintenance scenario where models are trained using historical data and evaluated on future operating conditions.

7.2 Test Dataset Composition

The final test dataset contained, normal condition 453,463, failure-related conditions, 1622 making a total of 455,085. The failure prevalence was:

$$\frac{1622}{455085} \times 100 = 0.356\%$$

Therefore, the dataset represents an extremely imbalanced classification problem.

7.3 Evaluation Metrics

Since accuracy alone is unreliable under extreme imbalance, multiple evaluation metrics were applied. The models were evaluated using, Accuracy; Precision; Recall; F1-score; Balanced Accuracy; Matthews Correlation Coefficient (MCC); Precision-Recall Area Under Curve (PR-AUC); Receiver Operating Characteristic Area Under Curve (ROC-AUC); and Confusion matrices. ROC-AUC and PR-AUC were calculated using predicted probabilities rather than only binary predictions, allowing assessment of model ranking capability independent of the selected classification threshold.

8. EXPERIMENTAL RESULTS

8.1 Overall Classification Performance

The classification results of the evaluated models are presented in Table 3. Although all models achieved accuracy values above 99%, substantial differences were observed in their ability to identify failure-related conditions.

Table 3. Classification Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Dummy Classifier	0.996436	0.0000	0.0000	0.0000
Decision Tree	0.995364	0.027132	0.008631	0.013096
Random Forest	0.996431	0.0000	0.0000	0.0000
XGBoost	0.996932	0.607414	0.393958	0.477936

The results demonstrate that accuracy alone does not provide a reliable indication of failure-detection capability. The Dummy Classifier achieved 99.64% accuracy by predicting all observations as normal, despite detecting no failures. Similarly, Random Forest achieved comparable accuracy but failed to identify failure observations at the default threshold. In contrast, XGBoost achieved the strongest minority-class performance, with substantially higher precision, recall, and F1-score. This indicates a better balance between detecting failures and controlling false alarms.

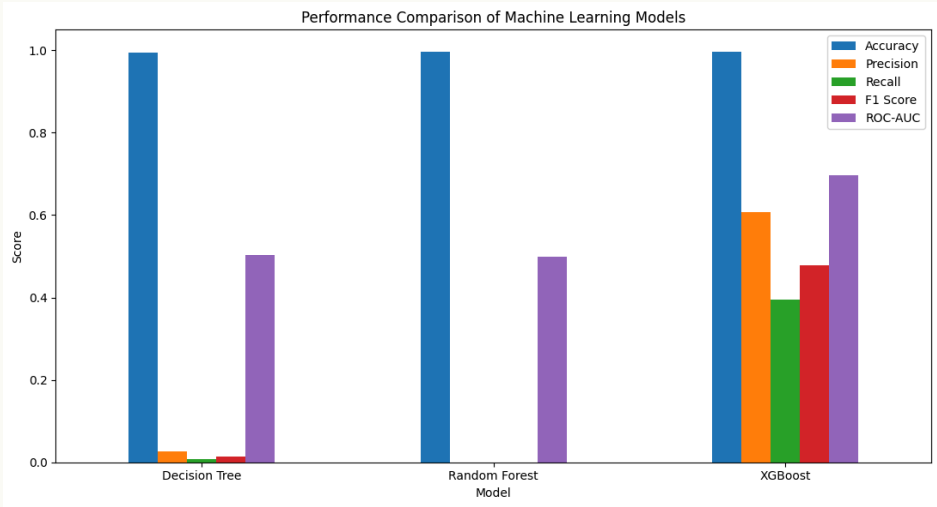


Figure 1. Comparison of minority-class performance among machine-learning models

8.2 Majority-Class Baseline Analysis

A Dummy Classifier was included as a baseline to demonstrate the impact of extreme class imbalance. The baseline achieved, Accuracy = 99.6436%; Recall = 0%; and F1-score = 0%. Although the accuracy appears high, the model provided no practical failure-detection capability because all observations were classified as normal.

8.3 Decision Tree Performance

The Decision Tree model achieved, Accuracy = 99.5364%; Precision = 2.71%; Recall = 0.86%; and F1-score = 1.31%.

The confusion matrix is shown in Table 4.

Table 4. Decision Tree Confusion Matrix

	Predicted Normal	Predicted Failure
Actual Normal	452,961	502
Actual Failure	1,608	14

The model correctly identified only: $\frac{14}{1622} \times 100 = 0.86\%$ of failure observations. Although the Decision Tree achieved high overall accuracy, its low recall indicates limited usefulness for practical failure detection.

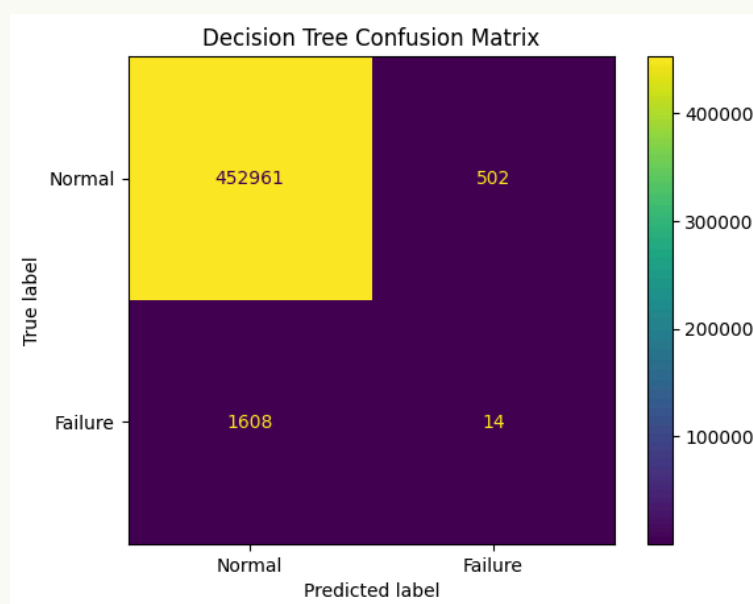


Figure 2: Decision Tree Confusion Matrix

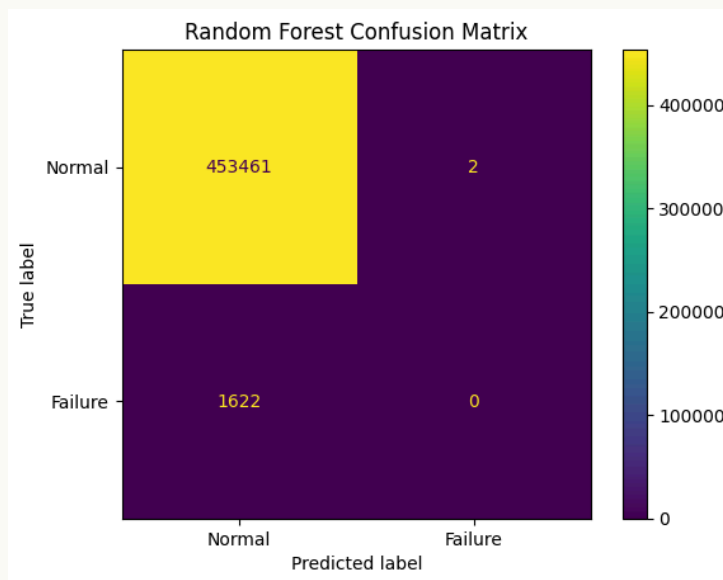
8.4 Random Forest Performance

Random Forest achieved: Accuracy = 99.6431%; Precision = 0%; Recall = 0%; and F1-score = 0%. The confusion matrix is shown in Table 5.

Table 5. Random Forest Confusion Matrix

	Predicted Normal	Predicted Failure
Actual Normal	453,461	2
Actual Failure	1,622	0

At the default classification threshold of 0.5, the model classified almost all observations as normal. However, probability-based evaluation revealed that the model retained strong ranking capability.



8.5 Probability-Based Evaluation

To provide a more complete assessment, additional metrics based on predicted probabilities were calculated.

Table 6. Advanced Performance Metrics

Model	Balanced Accuracy	MCC	PR-AUC	ROC-AUC
Decision Tree	0.503762	0.013324	0.003768	0.503762
Random Forest	0.499998	-0.000125	0.365583	0.997289
XGBoost	0.696524	0.487741	0.481962	0.992712

The advanced metrics provide additional insight into model behaviour beyond binary classification results. Decision Tree achieved performance close to random classification, with balanced accuracy of 0.5038 and ROC-AUC of 0.5038.

Random Forest produced an important observation. Although the default classification results showed zero recall, the model achieved:

- ROC-AUC = 0.9973;
- PR-AUC = 0.3656.

This indicates that Random Forest learned useful probability separation between normal and failure observations. The limitation was mainly associated with the classification threshold rather than the model's ability to rank failure likelihood. These results highlight the importance of threshold optimisation and probability calibration in highly imbalanced maintenance applications.

8.6 XGBoost Performance

XGBoost achieved the strongest overall performance among the evaluated models. The model obtained: Balanced Accuracy = 0.6965; MCC = 0.4877; PR-AUC = 0.4820; and ROC-AUC = 0.9927. The confusion matrix is presented in Table 7.

Table 7. XGBoost Confusion Matrix

	Predicted Normal	Predicted Failure
Actual Normal	453,050	413
Actual Failure	983	639

The model detected: $\frac{639}{1622} \times 100 = 39.40\%$ of failure observations. The precision value of 60.74% indicates that approximately six out of ten predicted failures were correct. Compared with the other evaluated models, XGBoost provided the best balance between failure detection capability and false-alarm control.

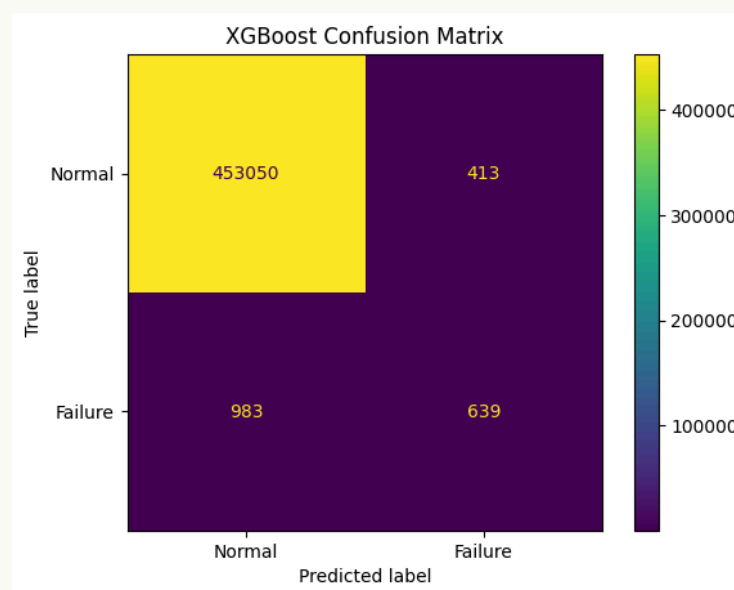


Figure 4: XGBoost Confusion Matrix

10. DISCUSSION

The results show that evaluating machine-learning models for industrial failure detection requires consideration of class imbalance, temporal characteristics, and operational objectives. Although all models achieved accuracy above 99%, their ability to detect failure conditions differed significantly. Therefore, the discussion focuses on minority-class detection, false-alarm control, and engineering interpretation.

10.1 Accuracy Is Insufficient for Rare Failure Detection

The results confirm that accuracy alone is unsuitable for highly imbalanced failure-detection problems. Due to the dominance of normal operating conditions, models can achieve high accuracy while providing limited failure-detection capability. The Dummy Classifier achieved 99.64% accuracy by predicting all observations as normal but detected no failures. Similarly, Random Forest achieved comparable accuracy but failed to identify failures at the default threshold. These findings demonstrate that accuracy can overestimate model performance. Therefore, predictive-maintenance applications should prioritise imbalance-sensitive metrics, including precision, recall, F1-score, MCC, and PR-AUC.

10.2 Importance of Minority-Class Evaluation

Minority-class metrics provided a clearer assessment of practical model performance. Decision Tree achieved only 0.86% recall, indicating poor failure detection despite high overall accuracy. Random Forest also achieved zero recall at the default threshold; however, its probability-based results showed strong separation capability between normal and failure conditions. This highlights the importance of evaluating both classification results and probability estimates. XGBoost achieved the strongest minority-class performance, with 60.74% precision, 39.40% recall, and 47.79% F1-score. Therefore, XGBoost provided the best balance between failure detection and false-alarm reduction among the evaluated models.

10.3 Probability Ranking Versus Classification Decisions

An important finding was the difference between probability ranking and final classification decisions. Random Forest achieved a ROC-AUC of 0.9973 despite producing zero detected failures at the default threshold. This indicates that the model learned meaningful differences between normal and failure observations, but the selected threshold was unsuitable for the highly imbalanced dataset. Therefore, future predictive-maintenance systems should consider threshold optimisation, probability calibration, and cost-sensitive decision strategies rather than relying on default thresholds.

10.4 XGBoost Performance and Engineering Significance

XGBoost achieved the strongest overall performance among the evaluated models, providing the best balance between failure detection and false alarms.

The model achieved:

- Balanced Accuracy = 0.6965;
- MCC = 0.4877;
- PR-AUC = 0.4820; and
- ROC-AUC = 0.9927.

The confusion matrix showed 639 correctly detected failures, 413 false alarms, and 983 missed failures. Although the model detected 39.40% of failure events, it significantly outperformed the other approaches under the same evaluation conditions. These results indicate that XGBoost can effectively extract patterns from industrial sensor data under severe class imbalance. However, further improvements are required before safety-critical deployment, including temporal feature engineering, sequential models, anomaly detection, and ensemble strategies.

10.5 Importance of Pressure and Temperature Measurements

Feature-importance analysis identified DV_pressure, Oil_temperature, and TP2 as the most influential variables across the ensemble models. For XGBoost, DV_pressure contributed 67.58% and Oil_temperature contributed 14.24%, together accounting for approximately 81.81% of total feature importance. This suggests that pressure and temperature measurements provide important indicators for

distinguishing failure-related conditions in the Air Production Unit. However, feature importance represents model reliance rather than causation. Further analysis using explainable AI methods, such as SHAP, is recommended to better understand individual predictions.

10.6 Comparison with Previous Research

Previous MetroPT studies have demonstrated the potential of machine-learning and deep-learning methods for predictive maintenance. LSTM-based anomaly detection approaches have identified deviations from normal operation, while hybrid ML-DL frameworks have reported strong predictive performance (Akshitha Mary A. C. et al., 2025; Peng et al., 2026). However, direct comparison remains difficult because studies differ in: failure definitions; labelling methods; preprocessing; balancing strategies; validation approaches; and evaluation metrics.

Unlike studies focusing mainly on predictive accuracy or anomaly detection, this research evaluates failure classification under severe imbalance using chronological validation, probability-based assessment, and minority-class metrics. Therefore, the main contribution of this study is demonstrating that reliable industrial failure detection requires appropriate evaluation strategies rather than accuracy alone. The findings highlight the importance of combining temporal validation, imbalance-sensitive metrics, and model interpretation for predictive-maintenance applications.

11. LIMITATIONS OF THE STUDY

Although this study provides a comprehensive evaluation of machine-learning-based failure detection under time-aware and highly imbalanced conditions, several limitations should be acknowledged. First, the study focuses on failure detection rather than early failure prediction. The models identify observations occurring within documented failure periods but do not predict failures before they occur. In practical predictive-maintenance applications, early warning capability is essential for preventing unexpected equipment downtime. Future studies should therefore extend this framework by introducing prediction horizons, such as detecting failures several minutes or hours before occurrence. Second, only three tree-based machine-learning models were evaluated: Decision Tree, Random Forest, and XGBoost. Although these models provide strong baseline performance for industrial classification problems, other approaches may capture additional temporal and degradation patterns. The current study also does not investigate deep-learning-based sequential models or hybrid anomaly-detection approaches.

Third, the current methodology uses individual sensor observations without extensive temporal feature engineering. However, equipment degradation is often represented through gradual changes over time rather than isolated measurements. The absence of temporal features may limit the ability of the models to identify early degradation patterns. Fourth, although chronological train-test separation was applied to reduce information leakage, the evaluation relied on a single temporal split. Additional validation strategies are required to examine model stability across different operational periods. Furthermore, the models were evaluated offline, and real-time deployment factors such as computational cost, prediction latency, and integration with IoT monitoring platforms were not investigated.

12. FUTURE RESEARCH DIRECTIONS

Based on the limitations identified, several research directions are recommended. Future studies should extend the current failure-detection framework toward early-warning predictive maintenance by introducing prediction windows before failure occurrence. This would enable models to identify degradation trends and provide actionable maintenance alerts before system failure. Additional machine-learning approaches should also be investigated, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), transformer-based time-series models, and hybrid anomaly-detection

frameworks. These approaches may improve performance by capturing sequential dependencies and long-term operational patterns within sensor data.

Future research should incorporate temporal feature engineering methods, including rolling averages, moving standard deviations, signal gradients, lag variables, and sequence-based representations. Such features may provide additional information about gradual equipment deterioration. Further evaluation should also consider more rigorous validation approaches, such as rolling-window validation and forward-chaining evaluation, to assess model robustness across different operational conditions. In addition, real-time deployment studies should examine edge computing requirements, processing efficiency, and integration with industrial IoT architectures.

The Random Forest results also highlight the importance of threshold optimisation and probability calibration. Future systems should investigate cost-sensitive decision strategies that consider the operational consequences of missed failures and false alarms. Finally, explainable artificial intelligence methods, including SHAP, LIME, and permutation importance, should be integrated to improve engineering interpretation and increase confidence in machine-learning-based maintenance decisions.

13. CONCLUSION

This study presented a time-aware and imbalance-sensitive evaluation of machine-learning models for detecting failure-related operating conditions in an IoT-enabled engineering system using the MetroPT-3 dataset. Three tree-based classifiers, Decision Tree, Random Forest, and XGBoost, were evaluated using chronological data separation and multiple imbalance-sensitive performance metrics. The results demonstrated that accuracy alone is insufficient for assessing rare-event engineering failure detection. Although all models achieved accuracy above 99%, their ability to identify failure conditions differed significantly when evaluated using recall, F1-score, MCC, PR-AUC, and ROC-AUC. The Dummy Classifier achieved 99.64% accuracy while detecting no failures, demonstrating the limitations of majority-class prediction. Decision Tree achieved high accuracy but showed poor failure-detection capability. Random Forest demonstrated strong probability-ranking ability with a ROC-AUC of 0.9973, despite producing zero recall at the default classification threshold. This finding highlights the importance of separating probability estimation capability from final classification decisions. XGBoost achieved the strongest overall performance, providing the best balance between failure detection and false-alarm reduction with a precision of 60.74%, recall of 39.40%, F1-score of 47.79%, balanced accuracy of 69.65%, and MCC of 0.4877. Feature-importance analysis identified DV_pressure, Oil_temperature, and TP2 as the most influential sensor variables, indicating that pressure and temperature measurements contain valuable information for distinguishing failure-related operating conditions. This study demonstrates that reliable machine-learning-based failure detection requires a combination of temporal validation, imbalance-sensitive evaluation, probability-based analysis, threshold optimisation, and model interpretation. The proposed framework provides a practical baseline for future predictive-maintenance research and supports the development of more reliable IoT-enabled intelligent engineering systems.

References

- Akshitha Mary, A. C., et al. (2025). Predictive modeling for metro performance using MetroPT-3 dataset.
- Al-Ali, A., & Alharbi, A. (2026). Temporally rigorous and traceable predictive maintenance evaluation.
- Benmansour, A., et al. (2026). Predictive maintenance in industrial systems using machine learning.
- Davari, M., Veloso, B., et al. (2024). Detecting and explaining anomalies in the air production unit of a

metro system.

Forbicini, A., Vago, R., et al. (2024). Time-series analysis in compressor-based machines: A survey.

Imani, A., Joudaki, M., et al. (2026). Why ROC-AUC is misleading for highly imbalanced data.

Meira, A., Veloso, B., et al. (2023). Data-driven predictive maintenance framework for railway systems.

Peng, W., et al. (2026). IntelliMetro-Hybrid: A machine learning and deep learning framework for metro predictive maintenance.

Salim, R., Hebri, A., et al. (2022). Classification predictive maintenance using XGBoost.

Veloso, B., Ribeiro, R., et al. (2022). The MetroPT dataset for predictive maintenance.



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>)