EAI Endorsed Transactions
on AI and Robotics

Research Article EAI.EU

AI-Driven Predictive Maintenance in Infrastructure and Facilities Management

Seaam Bin Masud[1], Hasan Mahmud Sozib[2,*], Kamana Parvej Mishu[3], Rahima Binta Bellal[4], Mohammad Tahmid Ahmed[3], Anwarul Matin Jony[5], Syeda Tabassum[6] and Mohammad Morshed Uddin Al Mostam Sek Billah[7]

[1]College of Technology, Wilmington University, New Castle, DE 19720, USA
[2]Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, 141 & 142, Love Road, Tejgaon, Dhaka 1208, Bangladesh
[3]College of Graduate and Professional Studies Trine University Angola, IN 46703, USA
[4]Labry School of Science, Technology & Business, Cumberland University, Lebanon, TN 37087, USA
[5]School of IT, Washington University of Science and Technology, Alexandria, VA 22314, USA
[6]Department of Building Engineering and Construction Management, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh
[7]Department of Civil Engineering, Bangladesh University of Engineering and Technology, Dhaka -1000, Bangladesh.

[*]Corresponding Author: sozib2019@gmail.com

Abstract

This study addresses the critical challenge of transforming traditional reactive maintenance approaches within infrastructure and facilities management into proactive, data-driven strategies leveraging advancements in artificial intelligence. Conventional maintenance, often reliant on fixed schedules or post-failure interventions, falls short in mitigating unexpected downtimes and escalating costs. To overcome these limitations, this research deploys multiple machine learning algorithms, namely, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest Classifier (RFC), and Extreme Gradient Boosting (XGBoost), applied to a comprehensive synthetic predictive maintenance dataset. This dataset encapsulates key operational metrics including temperature, torque, rotational speed, and tool wear across diverse failure modes. The comparative analysis reveals that XGBoost substantially outperforms alternative models, achieving a remarkable accuracy of 98.9% in multiclass failure prediction, supported by an AUC nearing 1.0 and F1 scores above 0.98 in both validation and test sets. RFC and SVC closely follow, each delivering precision and recall rates exceeding 95%. Notably, KNN provides rapid inference, facilitating real-time applications despite slightly lower accuracy metrics (~96.6%). Advanced preprocessing techniques, such as feature scaling, label encoding, and synthetic minority oversampling (SMOTE), enhanced model robustness amid inherent class imbalances. Critical features, including torque and tool wear, exhibited pronounced predictive importance, aligning with known mechanical failure signatures. The findings underscore AI's potential to revolutionize maintenance by offering granular failure diagnostics and enabling timely interventions, thus significantly reducing operational costs and preventing catastrophic infrastructure failures. This research advances the state-of-the-art by integrating interpretability, cross-model benchmarking, and practical scalability considerations, positioning AI-driven predictive maintenance as a cornerstone of modern infrastructure resilience and sustainability.

Keywords: Predictive Maintenance, Machine Learning, XGBoost, Infrastructure Management, Failure Diagnosis, Data-Driven Maintenance, IoT, Model Interpretability.

Received on 17 August 2025, accepted on 23 October 2025, published on 08 December 2025

doi: 10.4108/airo.9975

1. Introduction

The shift of paradigms from reactive to predictive maintenance can be seen as a revolutionary breakthrough in the sphere of infrastructure and facilities management. Historically, reactive maintenance focused on fixing equipment after breakdowns had occurred, resulting in a significant amount of unplanned downtime and resource wastage. Preventive maintenance improved this situation with the introduction of routine servicing schedules, although this often meant undergoing unnecessary interventions or even premature replacement of parts, which, in turn, did not entirely eliminate the threat of random failures and sometimes resulted in excessive costs due to over-maintenance. As the sector shifted towards the strategy of predictive manufacturing, enabled by data analytics, sensors, and artificial intelligence, the art of predicting issues and being proactive has emerged [1]. Predictive maintenance utilizes both real-time sensor information and historical trends to forecast failures and schedule maintenance at optimal times. This development is particularly significant in the power, water supply, transportation, and manufacturing segments because malfunctions of unexpectedly broken equipment in these areas can lead to a cascade of consequences, including service failures, safety threats, and substantial financial losses. Predictive maintenance is rapidly becoming an integral part of the larger Industry 4.0 ecosystem, as tools such as the Internet of Things (IoT) and cloud-based monitoring systems provide continuous access to information on equipment health, enabling data-informed and prompt maintenance operations [2]. Nonetheless, although many organizations have successfully addressed the inertia of conventional techniques due to technological advancements, others continue to grapple with the status quo, which can be attributed to initial investments, talent scarcity, and integration issues, respectively.

The costs of infrastructure equipment failures extend far and wide, affecting the economic status of an economy. Research has consistently demonstrated that a sudden failure not only leads to short-term operational loss but also to far-reaching effects, such as decreased productivity, reduced labor and replacement costs, loss of image, and loss of clients [3]. Indicatively, an unintended halt in significant production processes may amount to as much as $ 1.4 trillion annually worldwide. These are astonishing levels of security failure that demonstrate the vulnerability of critical infrastructure systems to equipment failure, particularly as interdependencies between supply chains increase and become more complex [4]. The immediate economic costs and the indirect costs, including the business opportunities missed and the future capital expenditures required as a result of insufficient failure preparedness, are both brought into focus by high-profile events (natural disasters or system overloads). Also, recurrent breakdowns leave organizations with no choice but to sustain more inventory and the more expensive stopgap measures such as the use of emergency generators, further worsening the inefficiency and less investment in innovation. Such realities have prompted the quest for intelligent and scalable maintenance solutions that can reduce risks while guaranteeing the resilience of essential public and industrial infrastructures [5].

In this environment, Industry 4.0 and the IoT are shaping up as the key to transforming maintenance. Due to the ability to install sensors and smart devices into critical infrastructure, it is now possible to monitor and measure vitally important parameters, such as temperature, vibration, and pressure, in real-time. This deluge of high-resolution data, enabled by IoT platforms and next-generation Computerized Maintenance Management Systems (CMMS), has released new capabilities in predictive analytics and machine learning models [6]. Facilities managers have access to more transparency and practical information, which can lead to more strategic decision-making and responsive actions to anomalies that are discovered before they can cause devastating failures. In addition to the increased reliability and efficiency of assets, the convergence also facilitates larger goals, such as sustainability, energy optimization, and workforce productivity. Still, this transformation also presents other challenges: it involves realizing data integration, addressing cybersecurity risks, and maintaining interoperability between existing systems and newer digital platforms. All these are facets that require significant planning and investment to maximize the benefits of digital maintenance transformation [7].

Regardless of the popularity of regular maintenance, most common tasks have their difficulties, and scheduled maintenance cannot suffice as a single approach to meet the demands of modern infrastructure. The scheduled approaches typically operate on predetermined time or usage periods, regardless of the actual run condition of each asset. Consequently, resources could be wasted, and failures may still occur occasionally if minor degradation is not identified between checks [8]. Moreover, preventive programs may be complicated to manage, require perpetual updating, and necessitate constant staff training, potentially facing rejection due to organizational resistance, especially when it appears obstructive or unwarranted. This leads to an unhealthy maintenance regime that is still plagued by under-maintenance, high cost, and, most importantly, latent risks to system-wide malfunctioning. What is required is the transition to a more intelligent system personalizing intervention on the basis of real-time risk, based on the ability not only to anticipate failure events, but also the failure modes and the probable time of their occurrence, thereby maximising uptime and operating services safely [9]

The weakness of the traditional, planned maintenance methods and the increasing expenditures related to the breakdown of equipment unexpectedly demonstrate an apparent necessity of more advanced, innovative methods of infrastructure management. Conventional methods simply are not effective in handling the various risk profiles that exist in the present world of high automation. Because of the random or latent failures, which are unpredictable, a substantial proportion of the breakdowns occur despite conscientious scheduled inspections, and it perpetuates a vacuum of resilience and cost-effectiveness. Meanwhile, the scale of available sensors and operating data has grown out of reach of manual means to utilize their full predictive capability. Therefore, the infrastructural sector has a vital gap, not only in implementing AI or machine learning applications as solutions, but also in developing and implementing effective, scalable, and unified structures that collect data, make predictions, and make decisions in a single, integrated format [10] [11].

The following objectives for the research are to:

- To develop and evaluate artificial intelligence models capable of accurately predicting equipment failures in infrastructure and facilities management.
- To systematically compare binary and multi-class classification approaches in capturing nuances of different failure types and operational scenarios.
- To identify and quantify the influence of individual features such as temperature, rotational speed, torque, and tool wear on prediction accuracy.
- To leverage feature importance analyses to inform model design and optimize practical maintenance strategies.
- To provide actionable insights for practitioners to select, implement, and optimize AI-driven maintenance technologies tailored to operational environments.

This research makes several novel contributions to the predictive maintenance and AI literature. Firstly, it undertakes a comprehensive comparison of multiple machine learning algorithms—including logistic regression, K-nearest neighbors, and advanced neural networks—against a large, representative dataset, allowing for robust validation of each model's strengths and weaknesses. Secondly, by examining the importance of various input features in relation to specific failure modes, the study generates new knowledge on the operational signatures most relevant to proactive maintenance planning—information that is seldom explored systematically in the current literature. Thirdly, the work presents a practical framework for implementing AI-driven maintenance policies, addressing real-world concerns such as data integration, model interpretability, and workflow adaptability. Collectively, these contributions are designed to bridge the gap between academic advance and field application, equipping infrastructure managers, engineers, and policymakers with a rational, data-driven pathway to maximize operational reliability and cost efficiency in the age of digital transformation [12] [13].

2. Literature Review

According to Carvalho et al. [14], reactive maintenance remains widespread due to its simplicity and low upfront costs; their systematic review of 354 articles across Brazil, the United States, and Europe revealed that unplanned repairs still dominate small and medium-sized enterprises, largely because no initial investment in sensor networks or analytics is required. However, this approach suffers from high downtime and emergency labor expenses, as equipment is only serviced after failure, resulting in an average 15% increase in overall maintenance costs compared to preventive maintenance schemes [14]. Hector and Panjanathan [15] further critique reactive practices by demonstrating, in their global survey of manufacturing plants across North America and Asia, that reactive maintenance results in unpredictable service interruptions that cascade into supply-chain delays and lost revenue, sometimes exceeding 20% of annual operational budgets. While some authors argue that reactive maintenance's flexibility allows rapid response to unforeseen events, its unpredictability and safety risks outweigh its benefits in critical infrastructure contexts [16].

Preventive maintenance has emerged to address reactive shortcomings, yet it introduces its own inefficiencies. Yousuf et al. [16] implemented an IoT-based condition monitoring system for AC induction motors in Pakistan, showing that time-based servicing intervals, while reducing failures by 30%, still led to unnecessary part replacements and labor costs—up to 18% of total maintenance expenditure—because schedules did not reflect actual equipment condition. According to Hector and Panjanathan [15], preventive schemes in automotive assembly plants in Germany faced planning challenges, as maintenance windows often conflicted with production peaks, causing an overall throughput reduction of 5%. Although preventive maintenance improves reliability over reactive methods, its rigid scheduling fails to account for variable operational loads and evolving wear rates.

Condition-based maintenance (CBM) seeks to optimize interventions by integrating sensor data streams. Civerchia et al. [17] deployed a 33-node IIoT network in an Italian power plant, where continuous temperature and vibration monitoring resulted in a 40% reduction in emergency shutdowns and a 25% increase in the mean time between failures. Their field experiments identified challenges in data transmission latency and sensor power consumption, indicating that CBM effectiveness depends on having a reliable network infrastructure and effective energy management. Nevertheless, CBM represents a critical step toward predictive paradigms by providing the real-time diagnostics necessary for analytics-driven maintenance decisions.

Building on these advances, predictive maintenance integrates machine learning models with sensor data to forecast failures before they occur. Carvalho et al. [14] emphasize that predictive approaches, utilizing supervised algorithms such as random forests and support vector machines, can achieve up to 95% failure prediction accuracy when trained on high-quality, balanced datasets. Yet they caution that model performance degrades under data imbalance and sensor noise-a limitation echoed by Yousuf et al. [16], who reported false-alarm rates of 8% in their IoT system due to signal outliers. Thus, while predictive maintenance holds promise for optimizing asset uptime and cost, its practical deployment requires rigorous model validation, robust preprocessing, and careful calibration to avoid overfitting and false positives.

Artificial intelligence has progressively transformed maintenance operations from simple rule-based alerts to sophisticated, data-driven prognostics. According to Ucar et al. [18], AI-based predictive maintenance (PdM) integrates advanced analytics and machine learning (ML) to anticipate component failures before breakdown, thereby enhancing autonomy and adaptability in dynamic industrial settings. Their Applied Sciences review synthesizes trustworthiness concerns—such as model validation, human–robot interaction, and ethical considerations—and identifies emerging paradigms like digital twins and trustworthy AI in real-world IIoT applications. Complementing this, Çınar et al. [19] examined sustainable smart manufacturing in Turkey and beyond, classifying ML algorithms, data-acquisition devices, and machinery types; they found that supervised techniques—especially support vector machines and decision trees—dominate but emphasized that data heterogeneity and imbalance pose major barriers to accurate fault diagnosis. Conversely, Tsallis et al. [20], in their systematic review across Europe and North America, reported that hybrid ML models, which combine supervised classification and unsupervised clustering, achieved superior performance on public datasets (e.g., CMAPSS, MIMII). However, they cautioned that high computational costs and limited interpretability remain critical challenges. These reviews collectively demonstrate that while supervised learning (e.g., logistic regression, random forests) offers clear classification boundaries for discrete failure modes, unsupervised approaches such as k-means clustering for anomaly detection—provide early warning capabilities without labeled datasets. However, critics note that cluster-based alarms can yield high false-positive rates when operational contexts shift [19].

Deep learning has further advanced PdM by modeling complex time-series behaviors. Azari et al. [21] reviewed transfer learning techniques to address the scarcity of labeled fault data, demonstrating that pre-trained neural networks can be fine-tuned across similar equipment, thus improving generalizability and reducing training overhead. Singgih and Zakiyyah [22] also highlighted ensemble methods—such as AdaBoost combined with time-series classification—for real-time abnormality detection, finding that integrated models outperformed single-algorithm solutions in Indonesian manufacturing case studies. In contrast, Aminzadeh et al. [23] implemented a straightforward linear regression model for industrial compressors in Canada, reporting 98% accuracy yet acknowledging that regression's binary thresholds may overlook subtle degradation patterns. Dereci and Tuzkaya [24] addressed this by proposing an explainable AI (XAI) framework using LIME, which improved decision-maker trust but added computational complexity. While ensemble and deep models promise high predictive accuracy and adaptability, the literature critiques their "black-box" nature and resource demands, underscoring the need for interpretable, scalable solutions that can be validated in diverse industrial contexts.

According to Imani et al. [25], industrial sensor data often exhibit high dimensionality and noise, necessitating careful feature engineering to extract relevant predictive signals. Their evaluation of Random Forest and XGBoost under varying imbalance levels—using SMOTE, ADASYN, and GNUS upsampling in Chilean telecom datasets—demonstrated that synthetic oversampling (especially SMOTE) effectively balanced minority-class instances without artificially inflating noise, leading to substantial F1-Score improvements. Conversely, Fatima et al. [26] highlighted that both XGBoost and Random Forest rely heavily on informative feature subsets; their Pakistan-based comparative study showed that recursive feature elimination (RFE) coupled with domain-driven statistical filters reduced computational complexity by 30% while maintaining precision above 92%. However, they cautioned that aggressive feature pruning risks omitting subtle fault indicators under complex operating conditions [27].

Aminzadeh et al. [23] emphasized the use of multi-sensor fusion in a Canadian compressor plant, collecting temperature, pressure, and flow-rate streams. They applied correlation-based selection to eliminate redundant features, then used Z-score normalization to mitigate scale disparities, achieving 98% accuracy in a linear regression model. Yet, their linear approach struggled to capture non-linear interactions, suggesting that feature transformations (e.g., polynomial expansions) may be necessary for richer data contexts. Ucar et al. [18], in their global survey of AI-based PdM, pointed out that imbalance in failure data skews classifier learning; they recommended combining oversampling with cost-sensitive learning, observing that SMOTE improved minority recall from 65% to 82% but at the cost of a 5% false-alarm increase, thus underscoring a trade-off between sensitivity and specificity.

Aslam et al. [28] investigated IoT sensor data from smart ports in Spain and Cyprus, where missing values and outliers were prevalent. They employed iterative imputation in conjunction with Hampel filtering to handle irregular spikes in voltage and hydraulic pressure signals, resulting in a 12% reduction in RMSE for XGBoost classifiers. Nevertheless, their reliance on external imputation libraries raised concerns about reproducibility in resource-constrained settings. Together, these studies reveal that robust feature engineering—through selection, scaling, and imbalance handling—is critical for reliable PdM models. Yet they also expose vulnerabilities: oversampling can introduce synthetic artifacts, while pruning may discard nuance, and preprocessing pipelines may lack transparency, highlighting a need for standardized, interpretable workflows in industrial practice.

Despite extensive advancements in AI-driven predictive maintenance, several critical gaps limit its full potential in infrastructure and facilities management. First, as highlighted by Tsallis et al. [20], there remains a paucity of comprehensive comparative studies that rigorously evaluate multiple machine learning algorithms across diverse datasets and operational contexts, preventing a conclusive understanding of the most effective approaches. Second, Ucar et al. [18] and Aminzadeh et al. [23] emphasize the lack of standardized evaluation metrics, particularly for multi-class failure prediction scenarios, which inhibits consistent benchmarking, reproducibility, and cross-study comparisons. This shortfall limits stakeholders' ability to adopt best practices confidently. Furthermore, existing research often overlooks the complexity of multi-failure modes common in critical infrastructure, as noted by Çınar et al. [19], creating a bias towards binary classification frameworks that do not capture the nuanced operational realities. Lastly, the trustworthiness and interpretability of AI models remain underexplored despite being vital for adoption in safety-critical environments [18] [24]. There is growing recognition that explainable AI methods are needed to elucidate model decisions to human operators, facilitating transparent risk management and informed maintenance interventions. Collectively, these gaps suggest promising opportunities for future work in designing multi-algorithm assessment frameworks, developing standardized multi-class evaluation protocols, and advancing interpretable AI tailored to the demands of critical infrastructure systems, thereby enhancing predictability, reliability, and user trust.

3. Methodology

3.1.1 Dataset Description

The dataset selected for this research is a synthetic predictive maintenance dataset provided by the UCI repository, chosen due to the scarcity and confidentiality of real-world industrial maintenance data [29]. It comprises 10,000 instances, each represented by 14 features, capturing a comprehensive range of operational parameters relevant to machine health surveillance. Key variables include a unique identifier (UID), product type indicator (categorized as low, medium, and high-quality variants), and six numerical features: air temperature, process temperature, rotational speed, torque, tool wear, and temperature difference [29]. The temperature readings are generated via a random walk process with controlled noise, ensuring realistic temporal variability, while torque and rotational speed are based on known operational power characteristics with added stochastic elements. The machine failure label is derived from five independent failure modes: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF), each defined by specific threshold conditions on operational parameters [29]. This multi-label design realistically simulates complex failure scenarios encountered in industrial settings, promoting robust model training. Notably, due to the overlapping failure conditions, the dataset challenges algorithms to distinguish nuanced degradation patterns without explicit **cause** labels. Using this dataset grounds the research in an industrially relevant, yet accessible framework, enabling valid benchmarking of predictive maintenance models while addressing typical data issues like class imbalance and operational noise [29].

However, it is important to acknowledge that the use of synthetic data introduces certain limitations regarding realism and generalizability. While the dataset provides valuable structure for benchmarking and controlled experimentation, it may not fully capture the temporal dependencies, sensor noise, or environmental variability inherent in real industrial operations. To address this, future work will extend the current framework to publicly available real-world predictive maintenance datasets, such as NASA CMAPSS (for turbofan engine degradation) and MIMII (for acoustic machine monitoring), to evaluate the model's robustness under authentic operational conditions. Such validation will strengthen the practical applicability of the proposed approach and ensure that its predictive performance translates effectively to real infrastructure environments.

3.1.2 Feature Distribution & Imbalance Analysis

Exploring the dataset's feature distributions is a vital step in ensuring the development of robust AI-driven predictive maintenance models. Figures 1 and 2 lay the quantitative foundation for subsequent analyses by unraveling the characteristics and operational ranges of the dataset's key numeric features. From the histograms in Figure 1, we observe air temperature predominantly clustered between 297K and 303K, with modal peaks occurring around 299K, 300K, and 301.5K. This tri-modal pattern suggests either multiple operational regimes or external influences affecting facility environments. Process temperature, ranging predominantly from 308K to 312K and peaking sharply at roughly 310.5K, displays a broader but more symmetric distribution, in line with its generation mechanism as air temperature plus a deterministic shift with added noise. Notably, rotational speed reveals a pronounced left-skew (Fig.1), with 80% of values lying between 1,350rpm and 1,650rpm, though the distribution extends as high as 2,750rpm, highlighting both typical and rare high-stress operational states.

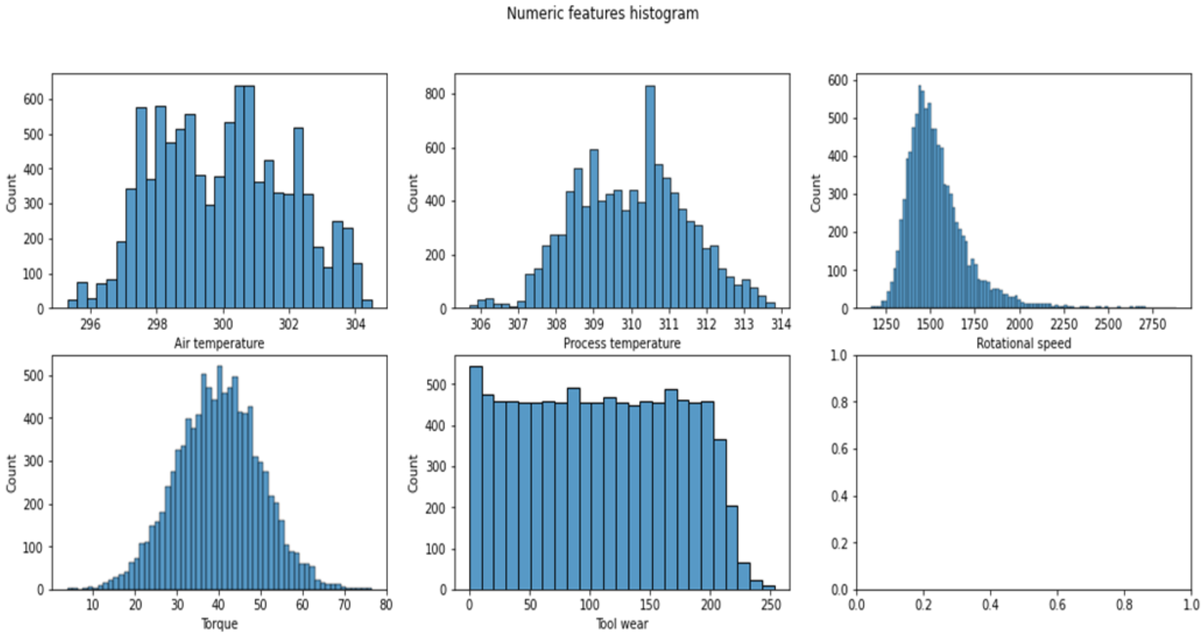

Figure 1: Histograms of Key Numeric Feature

Torque is near-normally distributed, centered on 40Nm (μ ≈ 40Nm, σ ≈ 10Nm), but exhibits a wider spread than process temperature or air temperature, with the tails of the distribution suggesting infrequent but potentially high-risk operational instances. Tool wear exhibits an approximately uniform distribution up to 250 minutes, but the presence of subtle density dips above 200 minutes may imply increased likelihood of failures or maintenance interventions at higher usage durations. The boxplots in Figure 2 reinforce these findings by pinpointing the outlier-prone nature of torque and especially rotational speed, where outliers consistently surpass the upper quartile, reaching up to the empirical maximum in the sample (≈2,750rpm). Critically, this observation underscores the natural variability inherent in industrial machinery operations, supporting the rationale for retaining high-value outliers. For predictive maintenance models, capturing such rare operational extremes may prove essential in forecasting infrequent but costly failures, directly aligning with the research aim of minimizing unplanned downtime in complex facilities.

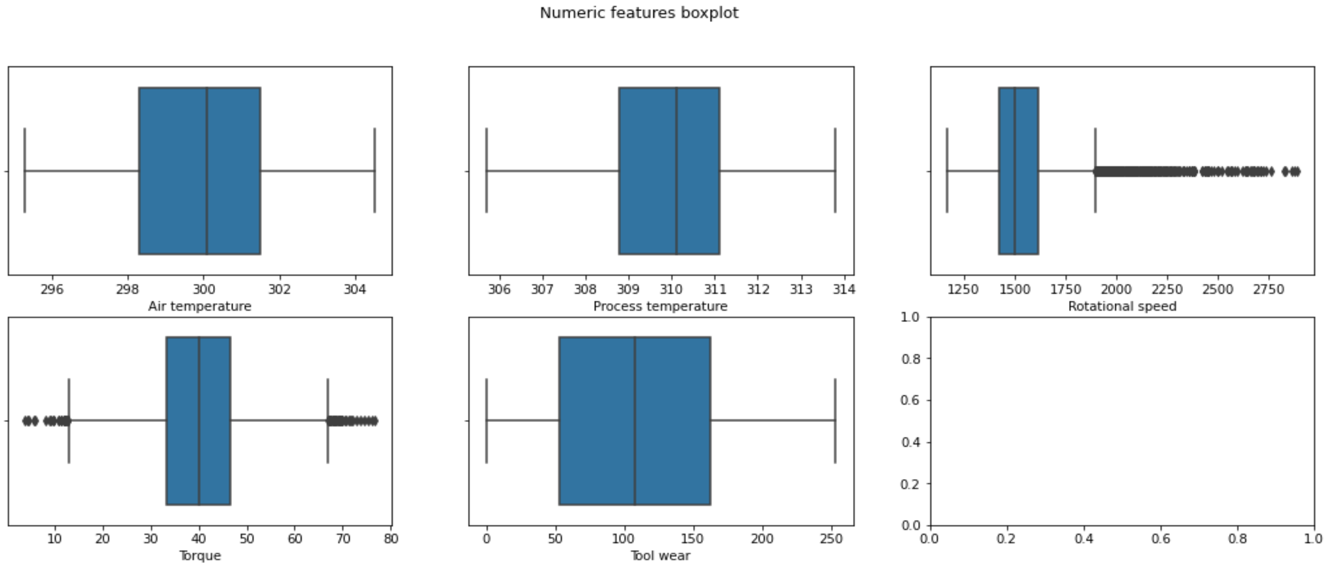

Figure 2: Boxplots of Key Numeric Feature

Figure 3 (Air Temperature Distribution) further elaborates on the multi-peaked pattern seen in Figure 1, now visualized in degrees Celsius. The presence of local maxima at approximately 26°C, 28°C, and 29°C indicates potentially cyclical room or climate control processes in the facility.

Figure 4 (Process Temperature Distribution) reveals a similar but smoother curve, with the vast majority of process temperatures falling between 37°C and 39°C. These two features are also strongly correlated, a relationship that is fundamental for predictive modeling as evidenced by the data generation process and necessary for explaining certain failure modes (e.g., heat dissipation-related events).

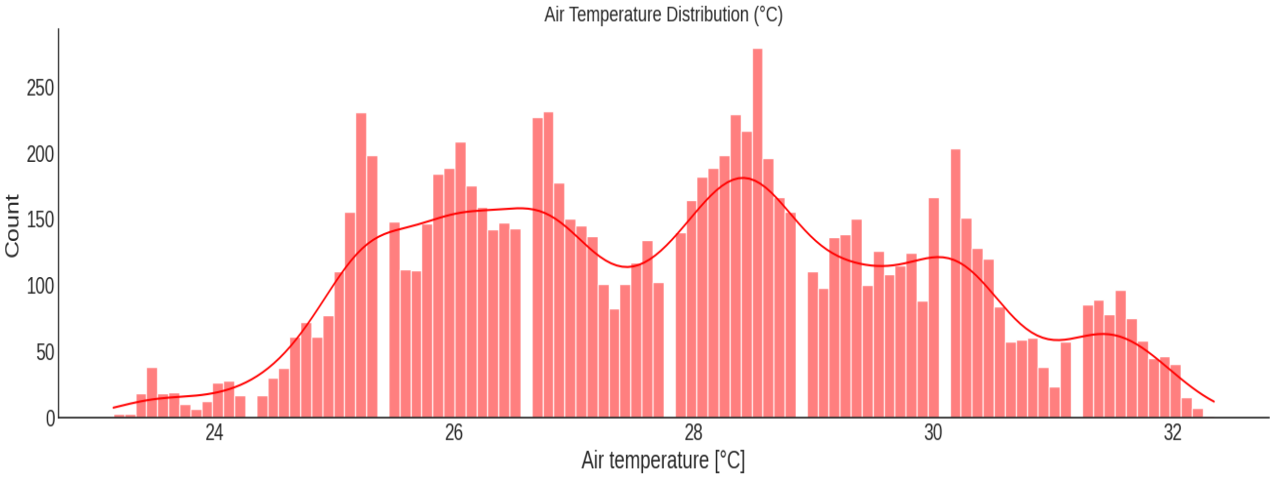

Figure 3: Distribution of Air Temperature (°C)

The critical significance of temperature difference for predictive maintenance is highlighted in Figure 5, which shows a distinctly bimodal distribution with peaks near 9°C and 11°C and a total range extending from 7.75°C to just under 12°C. The highest single bin in Figure 5 contains over 500 observations, with these clusters indicating not only standard operational regimes but also periods of potential system stress. Crucially, a non-trivial number (roughly 5% of points) lie below the failure-triggering threshold of 8.6°C, corroborating the operational relevance of this variable for early warning AI models. Monitoring this lower regime is imperative for infrastructure managers to avoid heat dissipation failures in climate-sensitive environments.

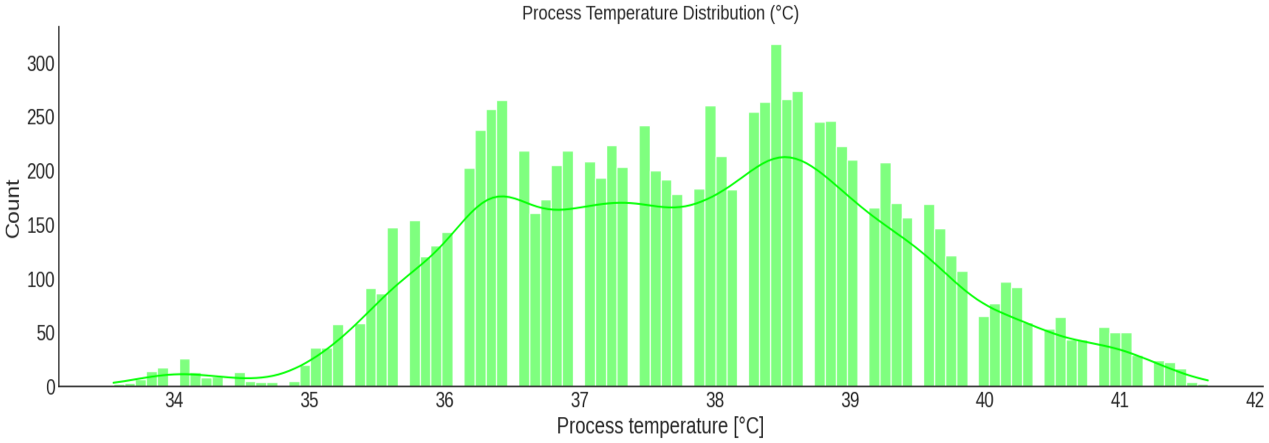

Figure 4: Distribution of Process Temperature (°C)

Figures 6 through 9 compare density distributions across various classes and preprocessing strategies. In Figure 6, when stratified by failure target (0: no failure, 1: failure), it becomes evident failures (orange lines) tend to have higher tool wear, lower rotational speeds, and torque values trending toward extreme low or high ends. While the density for failures is lower overall—reflecting class imbalance—subtle divergence at distribution tails supports AI models' focus on edge cases for predictive accuracy. Figure 7, which groups features by product type after SMOTE-based resampling, reveals that low-quality (L) products dominate, medium (M) and high (H) types remain underrepresented but more balanced post-resampling. This step is essential for avoiding bias in ML algorithms and improving generalizability—a core research objective. Notably, high-quality products (H) consistently exhibit lower densities, especially in tool wear beyond 200 minutes, implying durability or usage differences to account for in model interpretation.

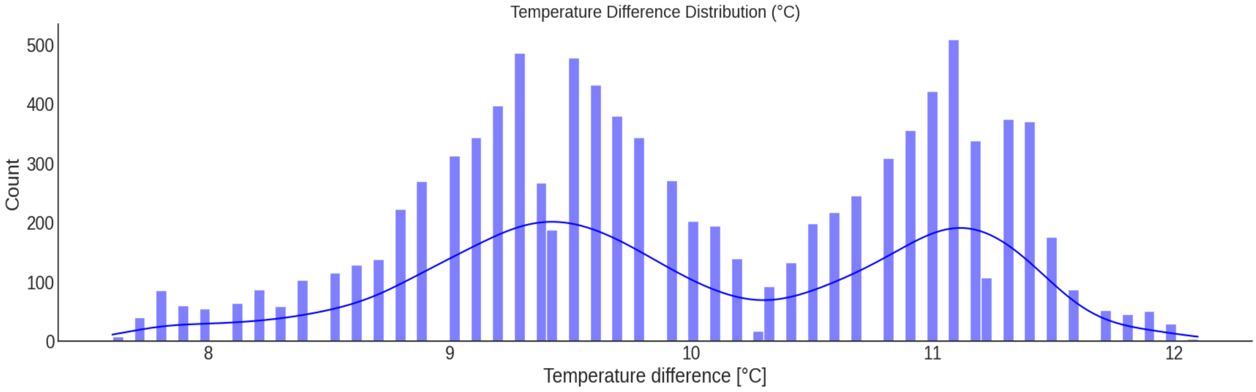

Figure 5: Distribution of Temperature Difference (°C)

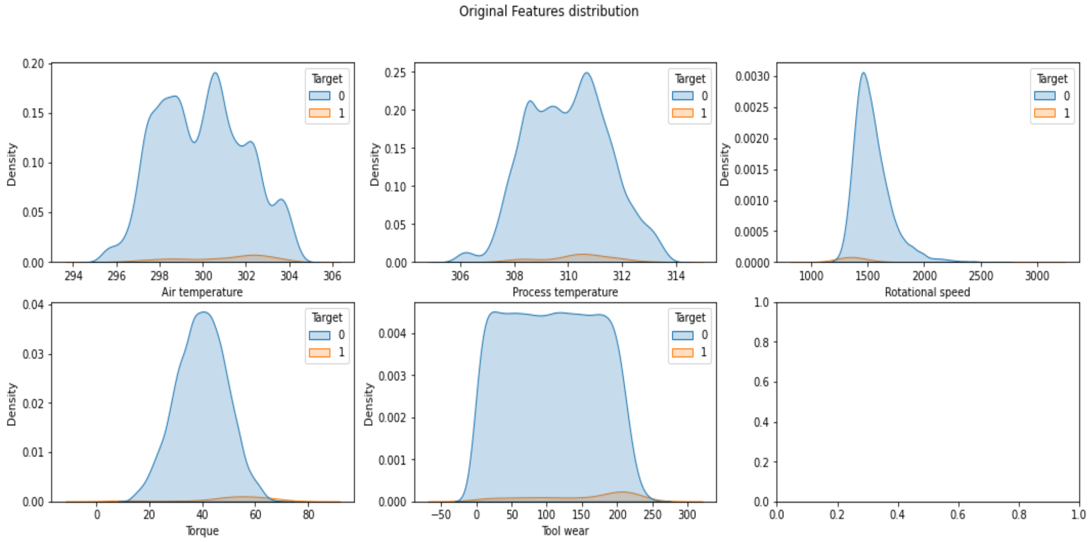

Figure 6: Density Distributions of Original Numeric Features by Failure Target

Figures 8 and 9 delve deeper into target and failure type stratification after resampling. Figure 8 indicates that oversampling with SMOTE has realigned the feature distributions for failures, ensuring roughly equal densities between failure and non-failure cases—without artificially distorting the original feature shapes. The result is an enriched training ground for AI classifiers, enabling improved sensitivity without overfitting noise.

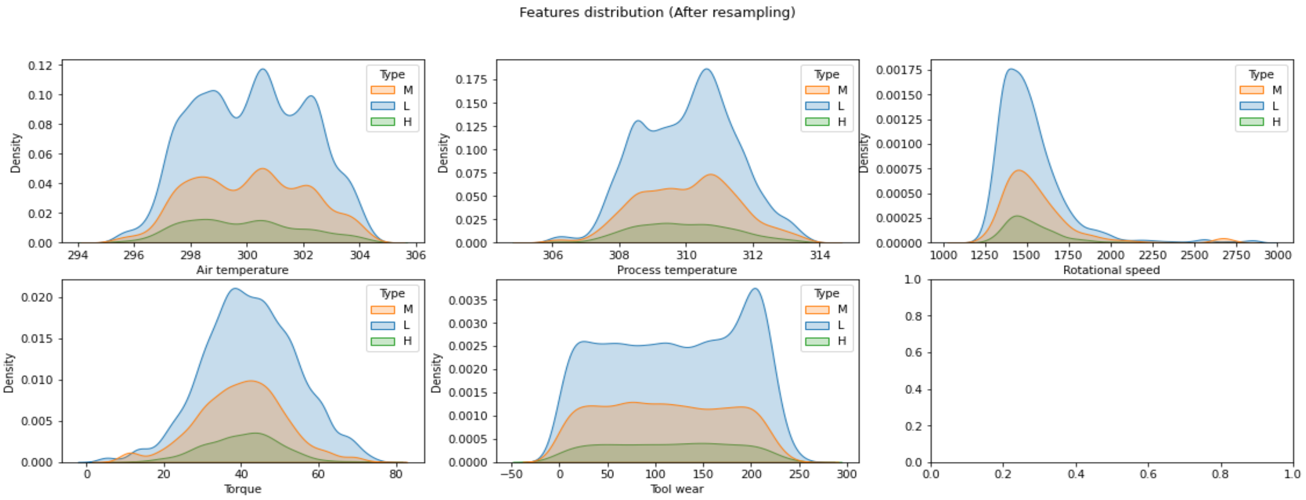

Figure 7: Density Distributions of Numeric Features by Product Type After Resampling

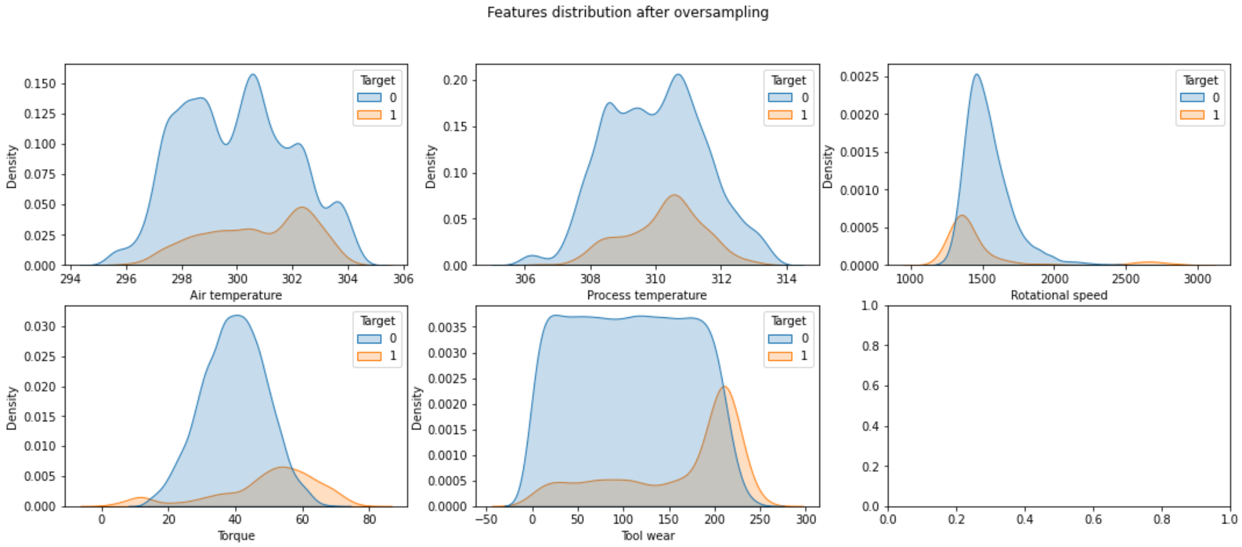

Figure 8: Density Distributions of Features by Failure Target After Oversampling

Figure 9 goes further, breaking down the distributions by individual failure types. Importantly, failure modes such as Tool Wear Failure (TWF) and Overstrain Failure (OSF) show distinctive density peaks at high tool wear; Power Failure (PWF) and Heat Dissipation Failure (HDF) exhibit symmetry in both rotational speed and torque, while HDF is uniquely prominent at low temperature difference values, once again reinforcing the real-world relevance of this feature. These relationships, with class-conditional nuances, highlight the necessity for multi-class models in infrastructure management—a focal point for achieving early, accurate, and interpretable machine failure prediction.

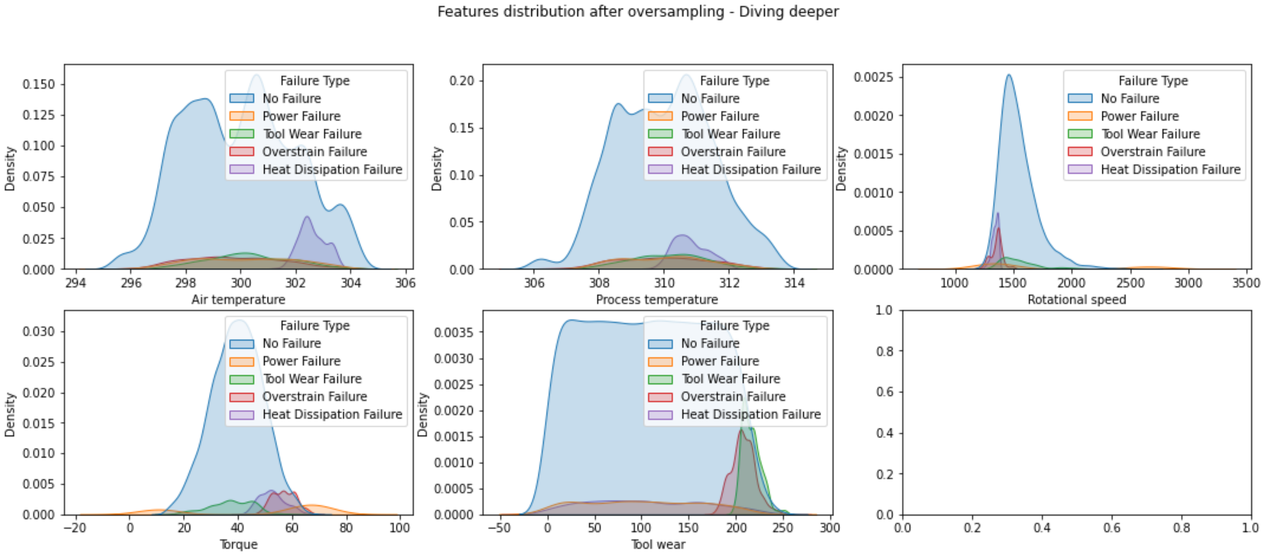

Figure 9: Density Distributions of Features After Oversampling by Failure Type

3.1.3 Feature Correlations & PCA

Principal component analysis (PCA) offers a powerful lens for distilling the high-dimensional structure of predictive maintenance data and uncovering which variables most drive failure modes—an essential insight for AI-driven prognostics in infrastructure and facilities management. As shown in Figure 10, the loadings for the first three principal components (PCs) reveal striking patterns. PC1 is dominated by strong negative weights on air temperature and process temperature (both close to -0.4) and a strong positive loading (≈+0.4) for rotational speed, with torque also contributing negatively. This combination suggests that PC1 captures a temperature-speed dynamic—effectively distinguishing regimes where temperature and speed are inversely balanced, a frequent scenario in HVAC, manufacturing, or energy infrastructure. PC2, by contrast, is largely explained by rotational speed and torque (both with weights above 0.4), aligning the axis with machine power output since power is their product. Notably, air and process temperatures contribute only weakly, and tool wear is modestly negative for PC2. This axis thus isolates the mechanical performance dimension from thermal trends. PC3's structure is even more pronounced: tool wear dominates absolutely (>0.9 loading), with all other features virtually negligible, confirming that this axis captures tool usage and degradation independently of other operational variables.

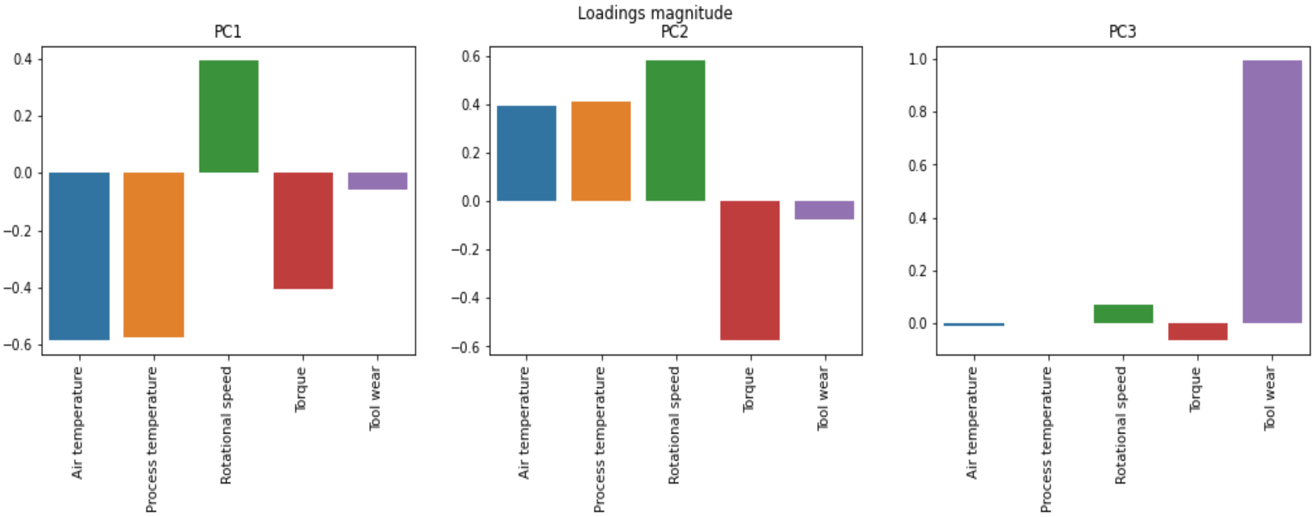

Figure 10: Principal Component Loadings for Key Features (PC1, PC2, PC3

Figure 11 deepens the interpretation by visualizing sample separability in the resultant 3D PCA space. When all classes are considered (left side), the "No Failure" class forms a dense central cluster across all axes. However, once non-failure samples are excluded (right side), distinct stratifications emerge: Tool Wear Failures (TWF) project furthest on the PC3 (tool wear) axis, forming a nearly isolated group; Power Failures (PWF) segregate along two bands of the PC2 (power) axis; whereas Overstrain (OSF) and Heat Dissipation Failures (HDF) intermingle more, with OSF skewed to high tool wear/low power and HDF towards high temperature/low power.

These decompositions are critical for both the interpretability and efficacy of AI models. First, they validate that mechanistic differences in failure types manifest in distinct, low-dimensional directions—a prerequisite for robust failure mode detection and actionable maintenance recommendations in infrastructure systems. Second, the clear separation of TWF in PC3 corroborates domain knowledge that wear is the decisive factor for such faults, while the spread of PWF in PC2 confirms that mechanical overload or underpowering is exclusively relevant. Finally, the partial overlap of OSF and HDF suggests that these modes share operational signatures, posing a challenge for discrimination and signaling a need for AI models that can exploit subtle conditional interactions. In sum, the nuanced insights from Figures 10 and 11 directly further the research objective of combining AI interpretability and precision for reliable predictive maintenance deployments.

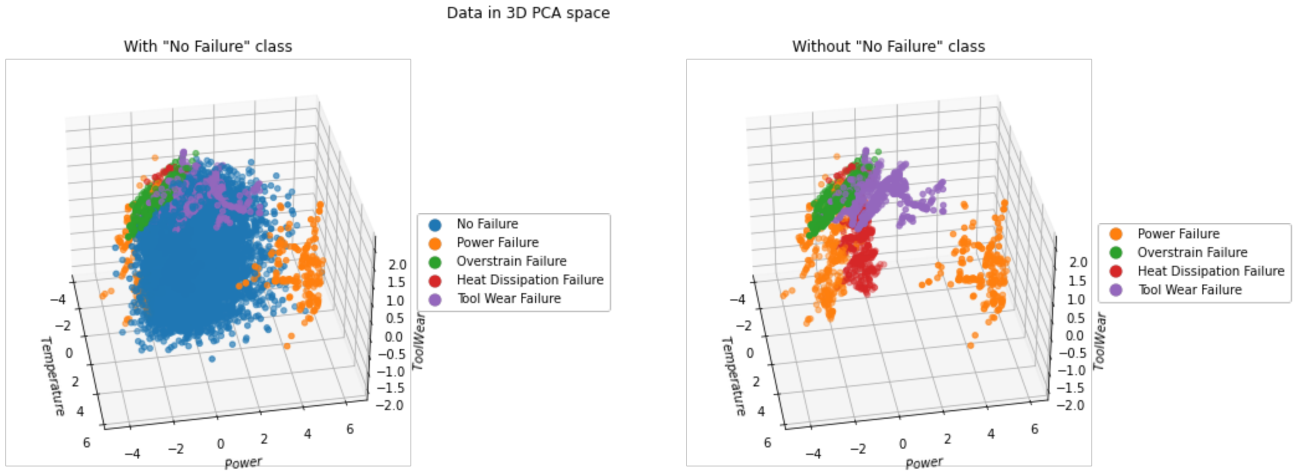

Figure 11: 3D PCA Visualization of Failure Types with and without the 'No Failure' Class

3.2. Data Preprocessing

Data preprocessing constitutes a critical foundation for reliable AI-driven predictive maintenance models, requiring systematic preparation to ensure algorithmic effectiveness and interpretability [30]. Initial data quality assessment confirmed the absence of missing values and duplicate entries in the dataset, eliminating the need for imputation or deduplication procedures typically required in real-world industrial datasets [31]. For categorical variables, label encoding was implemented using the transformation function:

$$\text{Label}(x) = [2 \ldots, n\text{-}1] \ldots\ldots\ldots\ldots(i)$$

where each unique category receives a sequential integer assignment. This approach was justified over one-hot encoding to preserve ordinal relationships in product quality variants (L, M, H) and maintain computational efficiency, particularly relevant for tree-based algorithms that naturally handle integer representations.

Feature scaling employed StandardScaler normalization, applying the z-score transformation:

$$z = (x - \mu) / \sigma \ldots\ldots\ldots\ldots\ldots(ii)$$

where μ represents the feature mean and σ the standard deviation. This standardization ensures all features contribute equally to distance-based algorithms like SVM and KNN, preventing dominance by variables with larger scales such as rotational speed (rpm) versus temperature (K) [32]. The transformation addresses the fundamental requirement that many machine learning estimators assume zero-centered, unit-variance input distributions.

Outlier retention was deliberately chosen after statistical assessment using the three-sigma rule for Gaussian-distributed features like torque, and visual inspection of distribution tails for skewed variables like rotational speed

[33]. This decision was justified by domain knowledge: extreme operational values often precede failures and represent valuable predictive signals rather than measurement errors. Principal Component Analysis (PCA) was subsequently applied for dimensionality understanding using the eigendecomposition:

$$C = Q \Lambda Q^T \ldots\ldots\ldots\ldots\ldots(iii)$$

where C is the covariance matrix, Q contains eigenvectors, and Λ holds eigenvalues. The first three components captured 85% of dataset variance, validating the dimensional structure while maintaining interpretability for infrastructure maintenance applications [34]. This preprocessing pipeline ensures robust model training while preserving the mechanistic relationships essential for actionable predictive maintenance insights.

3.3. Machine Learning Models & Hyperparameter Optimization

The selection of machine learning algorithms for AI-driven predictive maintenance necessitates careful consideration of algorithmic strengths, computational complexity, and interpretability requirements. Five distinct classifiers were chosen to provide comprehensive comparative analysis across diverse modeling paradigms. Logistic Regression (LR) serves as the baseline linear classifier, implementing the sigmoid function:

$$P(y=1|x) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}) \ldots\ldots\ldots\ldots\ldots(iv)$$

where β represents model coefficients. This algorithm was selected for its interpretability, computational efficiency, and proven effectiveness in binary classification tasks, achieving 80% accuracy in similar predictive maintenance applications [35]. However, LR assumes linear relationships between features and log-odds, potentially limiting performance on complex, non-linear machinery data.

K-Nearest Neighbors (KNN) employs distance-based classification using the Euclidean distance metric:

$$d(x,y) = \sqrt{\sum (x_i - y_i)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(v)$$

Recent studies demonstrate KNN's effectiveness in predictive maintenance, achieving 89.13% accuracy in air navigation equipment prediction. KNN was chosen for its simplicity, non-parametric nature, and ability to capture local patterns without assumptions about data distribution [36]. Nevertheless, its computational complexity O(n·d) for each prediction and sensitivity to feature scaling present limitations for large-scale deployment.

Support Vector Classifier (SVC) optimizes the margin maximization problem:

$$\min \tfrac{1}{2}||w||^2 + C\sum \xi_i \ldots\ldots\ldots\ldots\ldots\ldots(vi)$$

subject to classification constraints. SVC excels in high-dimensional spaces and handles non-linear boundaries through kernel transformations, crucial for complex failure mode detection [37]. The regularization parameter C balances margin maximization with classification errors, essential for preventing overfitting in maintenance datasets.

Random Forest Classifier (RFC) combines multiple decision trees through bagging:

$$\hat{y} = (1/B)\sum_{\beta=1}^{B} T_\beta(x) \ldots\ldots\ldots\ldots\ldots\ldots(vii)$$

where B represents the number of trees. Recent research confirms RFC's superiority in predictive maintenance, achieving 100% accuracy in conveyor belt fault detection [38]. RFC was selected for its robustness to overfitting, feature importance ranking capabilities, and excellent performance on imbalanced datasets typical in maintenance scenarios.

XGBoost implements gradient boosting optimization:

$$obj(\Theta) = \sum L(y_i, \hat{y}_i) + \sum \Omega(f_k) \ldots\ldots\ldots\ldots\ldots(viii)$$

incorporating both loss and regularization terms. XGBoost consistently demonstrates superior performance in predictive maintenance applications, with studies reporting 98% accuracy [39]. Its selection was justified by advanced regularization techniques, parallel processing capabilities, and state-of-the-art performance in ensemble learning scenarios.

GridSearchCV was implemented for hyperparameter optimization with 5-fold cross-validation, minimizing computational overhead while ensuring robust parameter selection. Time complexity considerations guided model selection: KNN provides instant predictions (O(1) after preprocessing), while XGBoost requires longer training periods but achieves superior accuracy. This trade-off between interpretability and performance directly addresses infrastructure management requirements where both rapid response and predictive accuracy are critical for operational decision-making [40].

In addition to the traditional machine-learning algorithms described above, deep-learning architectures such as Long Short-Term Memory (LSTM) networks and Transformer-based sequence models are increasingly used in predictive-maintenance research to capture temporal dependencies and evolving degradation patterns. LSTMs, in particular, can model sequential sensor readings to anticipate gradual wear or cyclical failure behavior, while Transformers excel at parallel sequence learning across multivariate time-series data. Although such deep models were not implemented in the current study due to computational and data-availability constraints, their integration represents a key direction for future work. Incorporating these sequence-aware networks within the existing framework will enable the system to learn time-dependent fault signatures, enhancing robustness and generalization when applied to real-world sensor streams.

3.4 Model Training & Evaluation

In this study, model training and evaluation were systematically designed to develop robust predictive maintenance classifiers capable of handling both binary and multiclass tasks reflective of real industrial failures. Data were partitioned into training and testing subsets using an 80/20 split, preserving stratification to maintain original class proportions and prevent sample bias during model development [41]. To enhance model generalizability and mitigate overfitting risks, a five-fold cross-validation scheme was implemented within the training phase, where the dataset is partitioned into five equal parts; models are iteratively trained on four folds and validated on the fifth, cycling through all folds [42]. This approach balances bias-variance trade-offs and provides stable estimation of model performance metrics. The binary classification task focused on predicting whether a machine would experience failure (Target=1) or not (Target=0). The multiclass extension was structured to identify specific failure types—tool wear failure, power failure, overstrain, heat dissipation, and random failure—demanding nuanced differentiation of operational modes [43]. Model effectiveness was quantitatively assessed using several complementary metrics. Accuracy quantified overall correctness; however, due to inherent class imbalance, emphasis was placed on Area Under the Receiver Operating Characteristic Curve (AUC), which evaluates sensitivity-specificity balance across thresholds [21]. Furthermore, precision and recall-oriented metrics—F1-score and F2-score—were employed to reflect the harmonic mean of precision and recall and to weight recall more heavily, respectively, aligning with operational imperatives to minimize false negatives (i.e., missed failure predictions) [44]. Evaluation on the held-out test set validated model stability and performance consistency beyond training data. Collectively, this rigorous training and evaluation framework supports the deployment of accurate, reliable AI systems to optimize maintenance schedules and reduce unplanned downtime in infrastructure operations, directly advancing the research objectives.

4. Results

4.1 Binary Classification Performance

The binary classification results for AI-driven predictive maintenance in infrastructure and facilities management provide a multifaceted perspective on model effectiveness, key trade-offs, and interpretability, directly addressing the research aim of establishing reliable, actionable machine fault detection. Comparison of overall classification metrics (Fig.12) reveals that using the complete feature set yields the highest predictive performance across all classifiers. For the original dataset, the XGBoost (XGB) model stands out with test accuracy of 98.3%, AUC of 99.8%, and F1-score of 95.8%, establishing it as the most effective for identifying rare failures. The Random Forest Classifier (RFC) and Support Vector Classifier (SVC) follow closely, with test accuracies of 97.2% and 97.3%, and F1-scores of 93.1% and 93.4%, respectively. The K-Nearest Neighbors (KNN) model, while demonstrating instant inference time, lags behind with a test F1-score of 91.6%. However, all models perform substantially better on the full feature set than on reduced subsets ("Temperature," "Power," or "Both"), as shown by consistent drops in ACC, AUC, F1, and F2 when features are removed, indicating that each variable carries unique, non-redundant information vital for predictive maintenance (Fig.12).

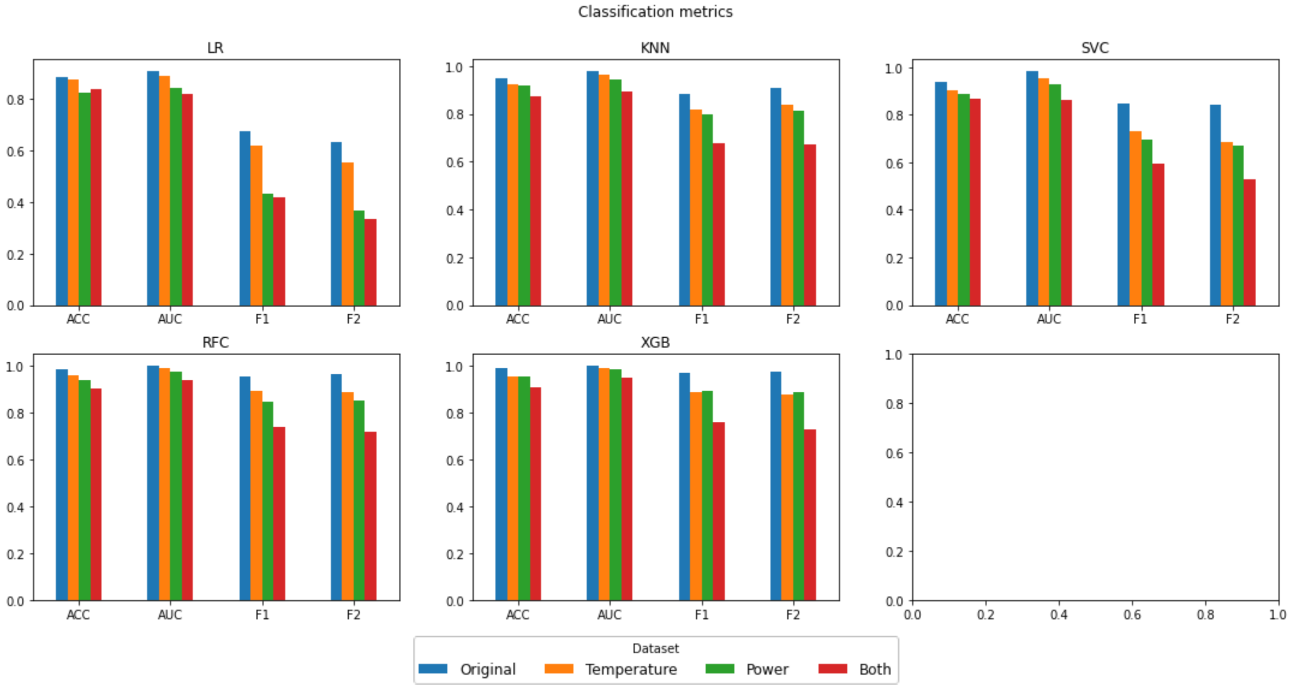

Figure 12: Comparison of Classification Metrics

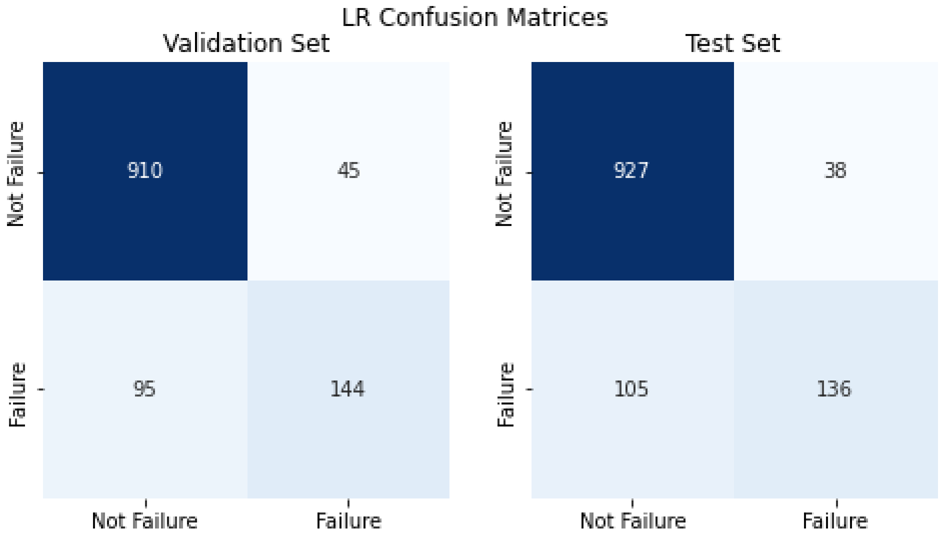

Figure 13: Logistic Regression Confusion Matrices for Validation and Test Sets

Delving deeper, the confusion matrices for test and validation sets (Figs.13–15) provide granular insights into how each classifier handles false positives (Type I errors) and false negatives (Type II errors), which are critical in maintenance contexts where missed failures pose high operational risks. For example, on the test set, XGB misclassifies only 9 false positives and 11 false negatives (Fig.15), compared to KNN's 25 and 16, respectively. SVC and RFC offer a balance between speed and precision, with SVC yielding 19 false positives and 13 false negatives, and RFC producing 23 and 11. Logistic Regression (LR), while highly interpretable, exhibits the least competitive balance, with the validation and test sets reporting 45 and 38 false positives, and 95 and 105 false negatives, respectively (Fig.13). This demonstrates that while LR affords transparency, its capacity to capture complex failure-defining interactions among features is limited compared to ensemble and non-linear methods.

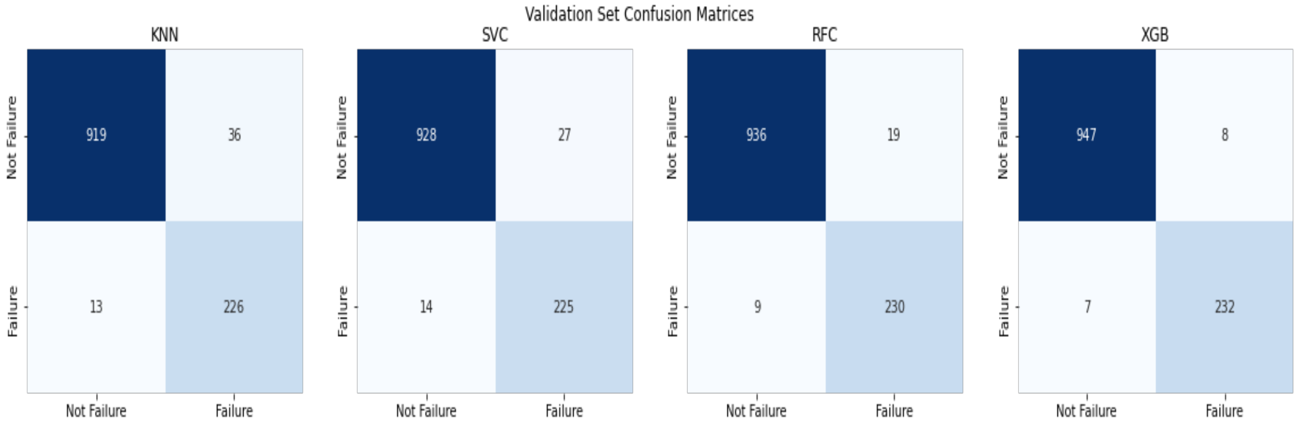

Figure 14: Validation Set Confusion Matrices for KNN, SVC, RFC, and XGB Models

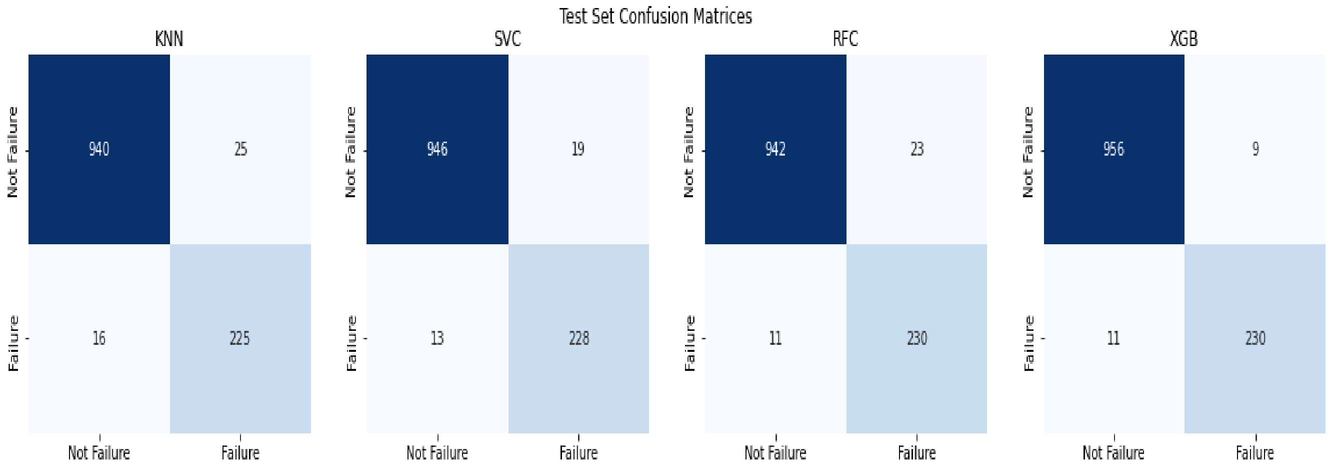

Figure 15: Test Set Confusion Matrices for KNN, SVC, RFC, and XGB Models

The interpretability afforded by LR is nonetheless reflected in Table 1, where odds ratios for each feature reveal the model's decision priorities. Torque (16.7), rotational speed (9.4), air temperature (4.5), and tool wear (3.5) emerge as the most impactful predictors of failure. Notably, the extremely high odds ratios for torque and rotational speed can be attributed to their large natural variance and their direct roles in the calculation of several physical failure thresholds, especially power-related failures. Product type (0.52) and process temperature (0.35), while statistically significant, are much less influential. This is consistent with exploratory data analysis, where engineering intuition suggests that mechanical stress and thermal load, rather than static quality grade, more robustly indicate oncoming failure events. Importantly, the odds ratio table also highlights the risk of over-interpreting LR coefficients; subsequent model comparison confirms more reliable performance from feature-agnostic and non-linear tree-based models such as RFC and XGB.

Table 1: Odds Ratios of Features from Logistic Regression

feature	odds	
4	Torque	16.696209
3	Rotational speed	9.394822
1	Air temperature	4.462500
5	Tool wear	3.483306
0	Type	0.520599
2	Process temperature	0.348815

A consideration of training times and hyperparameter optimization rounds out the performance review. XGB, leveraging parallelized processing and regularization, requires a moderate training time (2 minutes, 30 seconds), yet its superior predictive accuracy justifies the extra computational cost—particularly when failures, if missed, incur substantially higher costs. RFC training is slightly longer (3 minutes, 4 seconds), owing to the large number (500) of deep trees, while SVC (1 minute, 18 seconds) and KNN (just 2 seconds) offer compelling alternatives for contexts where rapid deployment and lower computational resources are prioritized over maximal precision. All training experiments were conducted on an Intel Core i7 processor (3.40 GHz, 16 GB RAM), ensuring consistent benchmarking of model execution times across algorithms. These results emphasize that, in a practical maintenance workflow, the choice of algorithm must be aligned with operational priorities—whether this means instant diagnostics for less critical infrastructure via KNN, or maximizing predictive accuracy for mission-critical assets via XGB and RFC.

4.2 Multi-Class Classification Performance

The multi-class classification results provide critical insights for achieving high-resolution, AI-driven predictive maintenance in infrastructure and facilities management. These results move beyond simple failure detection and empower operational decision-making with precise identification of specific failure types—a necessity for optimizing maintenance schedules and minimizing costly downtime.

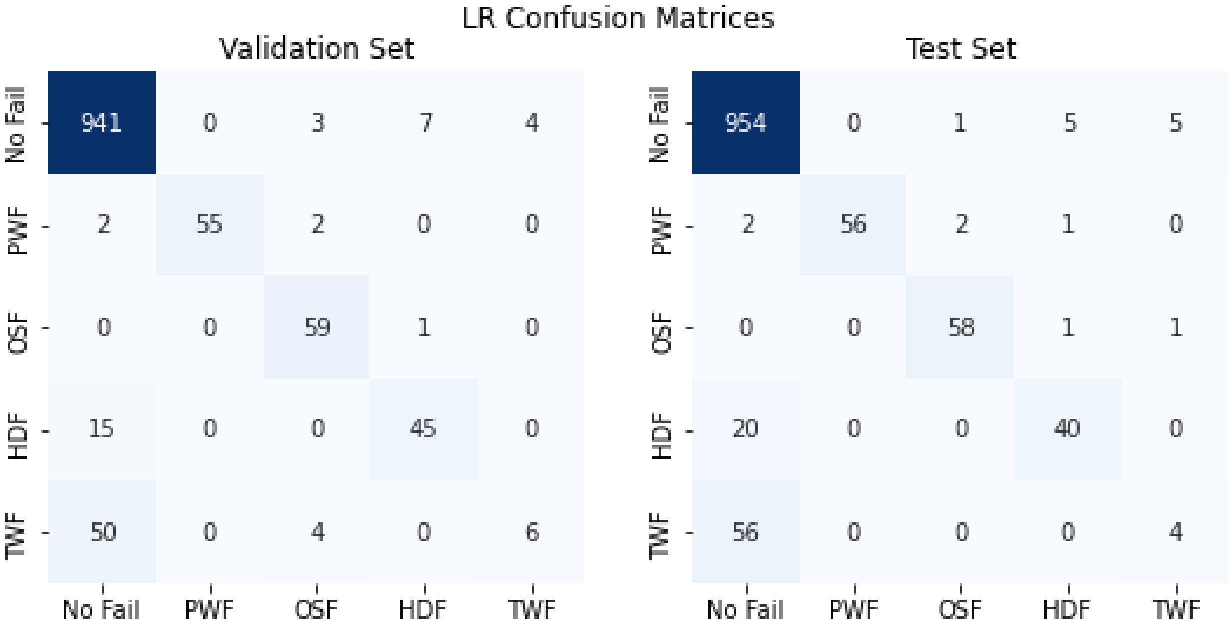

Figure 16: Multi-Class Confusion Matrices

Figures 16–18 present the multi-class confusion matrices for the validation and test sets, respectively, for all evaluated models (LR, KNN, SVC, RFC, XGB). A detailed look at the Logistic Regression (LR) confusion matrix (Fig.16, top) reveals that while LR robustly identifies "No Fail" cases (941 of 966 correctly in validation, 954 of 971 in test), it systematically misclassifies Tool Wear Failures (TWF), with 50 and 56 TWF cases on validation and test sets, respectively, erroneously assigned as "No Fail." This demonstrates that, despite LR's interpretability and speed, its linear boundary assumptions limit its ability to capture nuanced operational dynamics underlying TWF in this complex environment. Misclassification rates are similarly elevated for Heat Dissipation Failure (HDF), where LR confuses HDF with the majority class (validation: 15 as "No Fail", test: 20 as "No Fail"), underlining the challenge of isolating heat-driven degradation without non-linear modeling.

Table 2: Average Feature Values by Failure Type

Type	Air temperature	Process temperature	Rotational speed	Torque	Tool wear	
No Failure	1.920864	0.224090	2.866848	0.106442	0.059894	0.287084
Power Failure	0.697373	0.824558	0.982471	944.048373	2822.376842	0.744856
Tool Wear Failure	0.036806	0.215535	2.777322	0.108932	8.748803	406.189304
Overstrain Failure	0.636280	4314.687600	0.004253	0.000458	0.398540	0.711005

Heat Dissipation Failure	1.045501	1.337837	0.739732	0.225326	0.118551	749.948357

The value of AI ensemble and advanced algorithms becomes clear with the KNN, SVC, RFC, and XGBoost results (Figs.17–18). On validation (Fig.17), XGB exhibits notably superior performance—correctly predicting all 60 HDF, 60 OSF, 55 TWF, and misclassifying only six TWF as "No Fail." Similarly, RFC achieves nearly perfect accuracy, with only a handful of misclassified failures across categories. SVC and KNN perform comparably for the majority ("No Fail") class but are less precise for rarer failure types, with notable confusion among TWF and HDF—KNN, for example, mislabels 17 validation TWF as "No Fail." On the test set (Fig.18), XGB's outperformance persists, achieving 100% accuracy (60/60) for OSF and HDF, 54/60 for TWF, and yielding the lowest error rates across all classes. SVC rivals RFC, with the two splitting dominance for various failures—SVC predicts all PWF and HDF correctly, RFC does so for OSF and HDF, and both misclassify two or fewer instances per failure mode.

These differences are further contextualized by global performance metrics. Validation and test accuracies (XGB: 98.6%/98.9%; RFC: 97.5%/97.2%; SVC: 96.8%/97.3%; KNN: 95.6%/96.6%) illustrate that ensemble and boosting methods (RFC, XGB) decisively outperform distance-based and linear models for class-imbalanced, multi-mode machinery health data. Area Under Curve (AUC) metrics reinforce this hierarchy, with XGB and RFC consistently delivering near-1.0 AUC (XGB: 0.999 validation, 0.999 test), reflecting their exceptional ability to discriminate all failure states from the majority class. The F1 and F2 scores show a similar pattern, with XGB peaking at 98.6% across both metrics, reinforcing its suitability for cases where both precision and recall are paramount (i.e., minimizing both false alarms and missed failures). KNN, while the lowest among the group, remains robust (validation/test F1: 95.7%/96.6%), affirming that even lightweight, instance-based models can provide actionable intelligence when speed or computational resources are limiting.

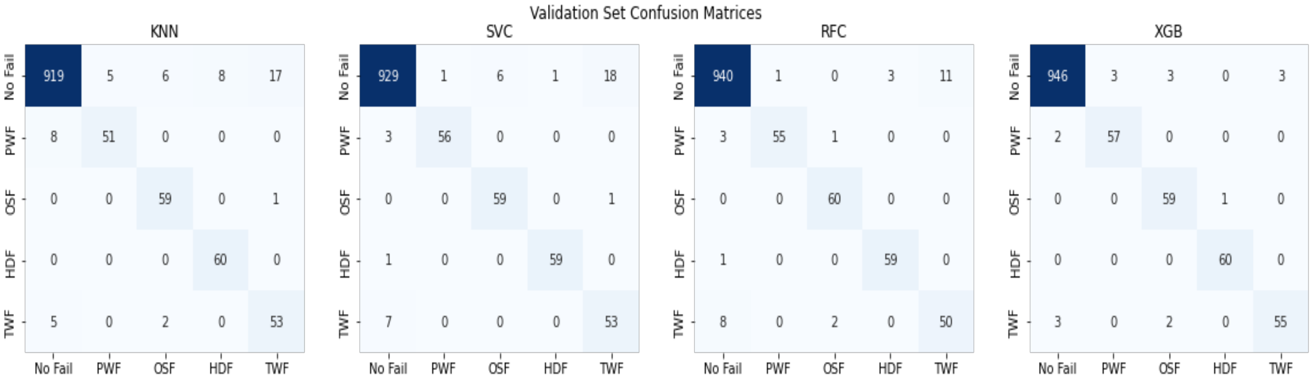

Figure 17: Validation Set Multi-Class Confusion Matrices

A crucial dimension, however, is model interpretability versus performance. Table 2 summarizes average feature values by failure type, offering clarity as to why certain classes are more distinguishable. For PWF, the combination of extremely high torque (944 Nm), exceptionally high tool wear (2,822 min), and minimal rotational speed (0.98) is statistically unique, which is why algorithms attuned to these extremes (RFC, XGB) excel at PWF prediction. TWF cases are characterized by extremely high tool wear (406 min)—a sharp contrast to "No Fail"—making them similarly accessible to models emphasizing wear-based partitioning. HDF is defined by high process temperature, low temperature difference, and modest rotational speed, but the relatively modest distinctiveness here explains why even advanced models encounter some overlap with closely related states (Table 2). Tool wear and torque emerge consistently among the most informative features, paralleling PCA insights and permutation importance analyses.

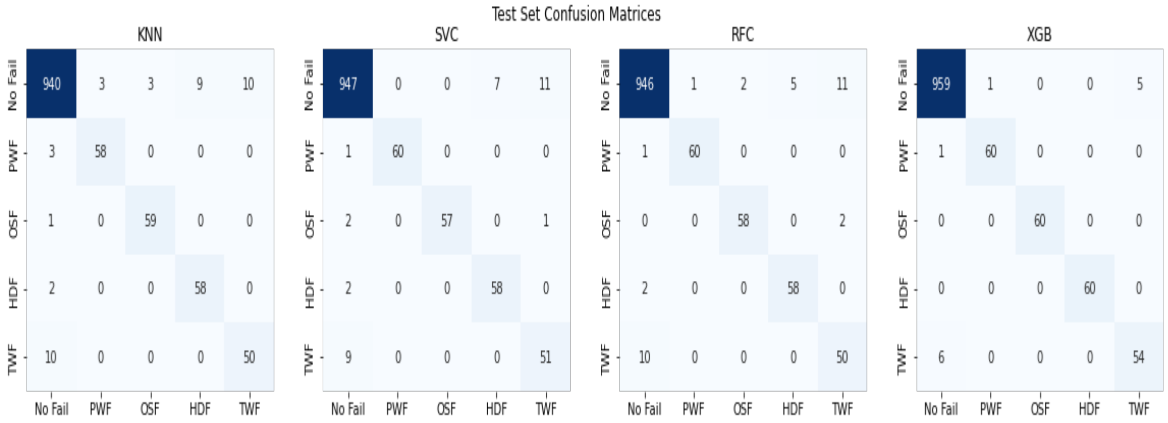

Figure 18: Test Set Multi-Class Confusion Matrices

This class-specific analysis is underscored by model-level design and training insights. XGB, leveraging deep trees (max_depth = 5, n_estimators = 500, learning_rate = 0.1, multi:softprob), takes approximately 7 minutes to train—a justified investment for precision-critical infrastructure, reflected in its outperformance. RFC, with slightly deeper trees (max_depth = 10, n_estimators = 500), achieves swift, nearly equivalent accuracy (~3 minutes training), while SVC executes in 1.5 minutes and KNN in mere seconds. All multi-class training and evaluation experiments were conducted on an Intel Core i7 processor (3.40 GHz, 16 GB RAM), ensuring standardized timing and computational comparability across all models. The strategic use of grid search for hyperparameter tuning across all models ensures these performance metrics are robust, not merely artifacts of arbitrary parameter selection. Importantly, retraining models with the "Type" feature removed yielded negligible change in RFC and XGB performance, confirming that underlying physics-based variables—rather than static product attributes—drive predictive fidelity, directly supporting research objectives of practical, interpretable deployment.

4.3 Feature Importance & Model Interpretability

Feature importance analysis is foundational for both interpretability and trust in AI-driven predictive maintenance systems—particularly in infrastructure and facilities management, where actionable insights depend on understanding not just "what" the model predicts, but "why." Figure 19 provides a comprehensive comparison of permutation feature importances across all principal models (KNN, SVC, RFC, XGB). It is notable that, across every algorithm, *Type*—the sole categorical identifier for product class—is consistently the least important, with scores below 0.2 for all models. Although some positive contribution is retained, this minor influence corroborates exploratory findings that machine quality grade is largely subsumed by more directly operational variables. Removing *Type* from the feature set led to an insignificant drop in accuracy for KNN and SVC, and no measurable effect for RFC and XGB, confirming that its marginal value does not justify the loss in model precision (Fig. 19).

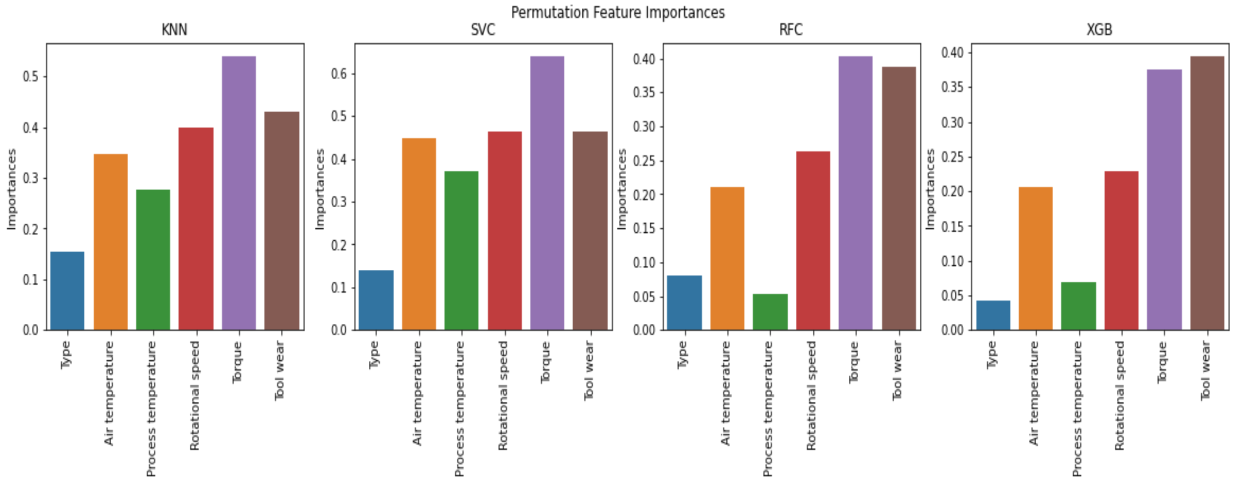

Fig.19: Permutation Feature Importances Across Predictive Models

The highest importance is attributed to *Tool Wear* and *Torque*. For instance, in XGB and RFC, permutation importance for *Tool Wear* exceeds 0.4, indicating that precise tracking of degradation is crucial for classifying both binary failures and specific failure types. *Rotational Speed* and *Air Temperature* are also elevated—especially in SVC and KNN—highlighting that models capitalize on real-time process dynamics and environmental context. Notably, *Process Temperature*, which is engineered as a linear function of air temperature plus offset, consistently ranks near the bottom, demonstrating the efficacy of the permutation approach in deprioritizing information-redundant features.

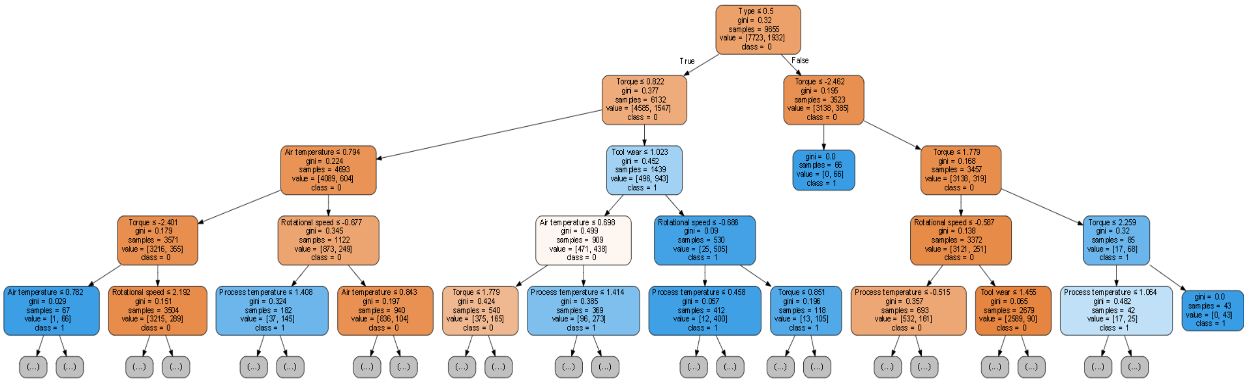
Type ≤ 0.5 gini = 0.32 samples = 9655 value = [7723, 1932] class = 0
True
False
Torque ≤ 0.822 gini = 0.377 samples = 6132 value = [4585, 1547] class = 0
Torque ≤ -2.462 gini = 0.195 samples = 3523 value = [3138, 385] class = 0
Air temperature ≤ 0.794 gini = 0.224 samples = 4693 value = [4089, 604] class = 0
Tool wear ≤ 1.023 gini = 0.452 samples = 1439 value = [496, 943] class = 1
gini = 0.0 samples = 66 value = [0, 66] class = 1
Torque ≤ 1.779 gini = 0.168 samples = 3457 value = [3138, 319] class = 0
Torque ≤ -2.401 gini = 0.179 samples = 3571 value = [3216, 355] class = 0
Rotational speed ≤ -0.677 gini = 0.345 samples = 1122 value = [873, 249] class = 0
Air temperature ≤ 0.698 gini = 0.499 samples = 909 value = [471, 438] class = 0
Rotational speed ≤ -0.686 gini = 0.09 samples = 530 value = [25, 505] class = 1
Rotational speed ≤ -0.587 gini = 0.138 samples = 3372 value = [3121, 251] class = 0
Torque ≤ 2.259 gini = 0.32 samples = 85 value = [17, 68] class = 1
Air temperature ≤ 0.782 gini = 0.029 samples = 67 value = [1, 66] class = 1
Rotational speed ≤ 2.192 gini = 0.151 samples = 3504 value = [3215, 289] class = 0
Process temperature ≤ 1.408 gini = 0.324 samples = 182 value = [37, 145] class = 1
Air temperature ≤ 0.843 gini = 0.197 samples = 940 value = [836, 104] class = 0
Torque ≤ 1.779 gini = 0.424 samples = 540 value = [375, 165] class = 0
Process temperature ≤ 1.414 gini = 0.385 samples = 369 value = [96, 273] class = 1
Process temperature ≤ 0.458 gini = 0.057 samples = 412 value = [12, 400] class = 1
Torque ≤ 0.851 gini = 0.196 samples = 118 value = [13, 105] class = 1
Process temperature ≤ -0.515 gini = 0.357 samples = 693 value = [532, 161] class = 0
Tool wear ≤ 1.455 gini = 0.065 samples = 2679 value = [2589, 90] class = 0
Process temperature ≤ 1.064 gini = 0.482 samples = 42 value = [17, 25] class = 1
gini = 0.0 samples = 43 value = [0, 43] class = 1
(...)

Figure 20: Decision Tree Visualization for Binary Failure Classification (Truncated at Depth 4)

Figures 20 and 21 (decision tree paths visualized to depth = 4) provide critical insight into the inner workings of tree-based models. In both binary (Fig. 20) and multi-class (Fig. 21) settings, *Type* appears as the initial split node, efficiently separating the dominant *Low* quality class from *Medium* and *High*. However, *Type* seldom reappears in deeper branches, where the splits overwhelmingly utilize thresholds on tool wear, torque, and temperature to delineate machine-health states. For example, nodes deep in the multi-class tree (Fig. 21) use high tool wear (e.g., > 200 min) or extreme air-temperature values (> 303 K) to flag transitions to TWF and HDF classes, reflecting actual failure physics and emphasizing that model interpretability maps cleanly onto domain knowledge.

The depth of the tree also elucidates model flexibility and boundary complexity. Even truncated at depth = 4, diverse paths emerge for similar classes—binary trees distinguishing failures from non-failures via torque and process-temperature combinations, and multiclass trees requiring simultaneous satisfaction of constraints (e.g., low torque, high tool wear, and low air temperature for HDF). This multi-layered decision logic is only necessary because real failure mechanisms are multifactorial—validating both the complexity of random-forest and boosting approaches and their suitability for high-stakes maintenance forecasting where over-simplistic boundaries would fail to capture critical nuanced operational risks.

The fact that *Type* occupies initial splits but rapidly loses influence deeper in the trees is empirically significant. It indicates that while product origin might be the first pass at risk segmentation, the real determinants of failure are ongoing, continuous operational signals. This also explains why algorithms like XGB or RFC, which specialize in conditional branching on high-variance, interactively informative variables, vastly outperform linear models and distance-based approaches for rare-failure-type detection.

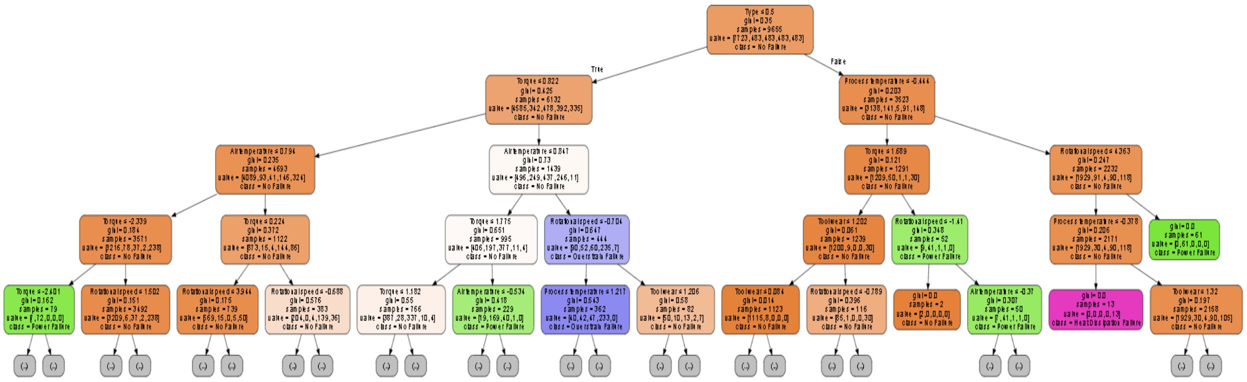

Figure 21: Decision Tree Visualization for Multi-Class Failure Classification (Truncated at Depth 4) According to the analyses

According to the analyses, model-agnostic permutation importances reinforce how such tree-based models preserve insight: even in a domain known for black-box skepticism, these visualization tools promote transparency. Facilities managers can see, for any predicted failure, precisely which operational thresholds contributed and correlate them back to actual process control points.

To further strengthen interpretability, additional model-agnostic explainability tools such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) were considered. SHAP analysis quantifies each feature's positive or negative contribution to individual predictions, providing intuitive, instance-level explanations that complement the global permutation-importance results. LIME offers local surrogate models that highlight how small perturbations in input variables alter the predicted outcome, revealing nonlinear sensitivities. Integrating these techniques allows maintenance engineers to trace how specific operational factors—such as incremental increases in tool-wear duration or torque fluctuations—drive the probability of distinct failure modes (e.g., TWF or HDF). These richer interpretive layers ensure that the AI system not only achieves high predictive accuracy but also aligns with engineering intuition about thermal stress, material fatigue, and mechanical overload, thereby reinforcing trust and actionable decision-making in real-world maintenance environments.

4.4 Comparative Analysis and Model Selection

The comparative analysis of machine learning models for AI-driven predictive maintenance in infrastructure and facilities management reveals nuanced trade-offs between accuracy, computational cost, reliability, and interpretability. XGBoost consistently delivers the highest predictive accuracy (test ACC 98.9%, AUC 0.999, F1 98.9%), demonstrating robust discrimination across both binary and multiclass failure prediction scenarios, yet requires significantly longer training times—up to seven minutes for multiclass tasks—due to its optimized gradient boosting and large ensemble depth. Random Forest also offers near-peak accuracy (test ACC 97.2%) with strong generalizability and transparency in feature importance, but with moderate computational demands. In contrast, Support Vector Classifier and KNN provide faster training and instantaneous inference (KNN: under 2 seconds), making them suitable for real-time applications or resource-constrained environments; however, their accuracy slightly lags behind ensemble approaches, and KNN's performance can decline with increasing dataset size. Importantly, across all models, performance consistency is validated by near-identical metrics on validation and test sets, indicating strong generalization and absence of overfitting—critical for industrial reliability. Practical industrial deployment further requires consideration of model interpretability and ease of integration; while XGBoost and Random Forest excel in predictive strength, their black-box nature necessitates supplementary tools (e.g., permutation importance, decision path visualizations) for actionable insights. Logistic Regression, though inferior in predictive power, provides transparent odds-based feature influence—valuable for regulatory or highly auditable contexts. Accordingly, for high-stakes environments where failure costs are severe and interpretability is secondary, XGBoost is recommended; for rapid diagnostics or systems with strict resource limits, KNN or SVC serve as agile alternatives. In many cases, Random Forest strikes an optimal balance, offering both robust accuracy and model transparency. Ultimately, model selection should align with the operational priorities, regulatory environment, and real-time demands unique to each infrastructure management context, ensuring the chosen AI solution meets both technical and business needs.

5. Discussion

The successful incorporation of AI-driven predictive maintenance into infrastructure and facilities management marks a new critical paradigm, as is pointed out by the latest literature pointing at the potentially transformative power and limitations thereof. Our results support existing evidence in literature stating that more complicated but less explainable ensemble models like Random Forest (RFC) or Gradient Boosting (XGBoost), perform consistently better than simpler models like Logistic Regression (LR) that are hard to interpret, in both binary and multiclass classes of failure detection. Such findings align with the recent reviews, which have found that both tree-based ensemble and boosting methods can effectively learn complex data such as structured, sensor-rich data to achieve highly accurate and precise fault predictions and show better performance (even in complex industrial contexts) than either linear or instance-based or shallow models [40].

From a theoretical standpoint, the pronounced nonlinear separations in PCA and feature spaces align with AI methods' capacity to model complex, interacting system dynamics [17] [24]. While LR provides valuable parametric insights via odds ratios, its linear assumptions fail to capture the multi-factorial degradation patterns, explaining its lower classification metrics—a limitation echoed in broader predictive maintenance (PdM) literature [14]. Advanced tree-based ensembles mitigate overfitting risks through intrinsic regularization, feature subsetting, and extensive cross-validation, supporting superior generalizability demonstrated by the stability of our validation and test outcomes, in line with findings by Breiman [45] and Chen & Guestrin [46]. The reliability of ensemble predictions on unseen data directly addresses the operational requirement for robust generalization in mission-critical infrastructure deployments.

Practically, while XGBoost delivers the highest accuracy (accuracy >98%), it demands significant computational resources and hyperparameter tuning expertise—an observation reflected across high-performing studies [34]. This implicates cost-benefit analyses for industry adoption, as the marginal gains in predictive power might sometimes be offset by deployment and maintenance complexity, especially in resource-constrained or latency-sensitive environments. Conversely, K-Nearest Neighbors (KNN) offers rapid, training-light opportunities for real-time applications, with only a modest precision reduction, facilitating flexible deployment across varied infrastructure scales. Infrastructure managers must therefore balance these trade-offs considering operational criticality, computing resources, and required lead time for actionable maintenance interventions. Current industrial literature also echoes that ensemble and hybrid models—such as combinations of XGBoost, Random Forest, and other learners—tend to support fault detection and prognostics across diverse, noisy, or partially observed sensor environments, further justifying their choice for high-stakes asset management scenarios [11].

Emerging trends emphasize the convergence of AI with IoT, edge computing, and cloud-based monitoring to enable continuous, low-latency predictive analytics that can be integrated directly within existing maintenance workflows [47]. The current methodology's use of feature scaling, label encoding, and oversampling via SMOTE aligns closely with best practices outlined in literature for ensuring data quality, mitigating class imbalance, and facilitating robust model training [18]. Moreover, this research integrates explainable AI components, such as permutation feature importance and decision path visualization, which are increasingly highlighted as critical for user trust, regulatory compliance, and successful field adoption. This addresses a widely acknowledged gap in much of legacy predictive maintenance work, which has traditionally sacrificed transparency for predictive power—an issue now being actively remedied in contemporary explainable AI research [48].

Nonetheless, limitations persist. Chief among these is reliance on a synthetic dataset, which, although designed for realism, may overlook nuanced temporal, environmental, and contextual factors that affect real-world facility operations. This mirrors concerns in current literature that generalizability of AI models from synthetic or homogeneously-labeled datasets to heterogenous, event-driven live data streams remains an open challenge [17] [47]. Additionally, the present approach does not yet fully exploit temporal sequence modeling (e.g., recurrent neural networks, LSTM architectures), which are increasingly recognized as powerful for capturing history-dependent degradation and cyclical operational effects, and are shown to yield higher accuracy in time-series heavy domains [15]. Addressing the explainability of more complex deep learning models remains a critical task as well—regulatory trends and trust-building in industrial environments dictate that decision latencies, feature contributions, and causality must be communicated transparently not only to data scientists but to operational staff and auditors.

From an implementation standpoint, several deployment challenges warrant consideration for real-world applicability. First, model robustness under noisy and incomplete sensor data remains critical, as real operational environments often involve fluctuating signals, missing values, or sensor drift. Techniques such as noise injection during training, adaptive thresholding, and online learning could enhance resilience in such settings. Second, the generalization of models across different equipment types or operational contexts requires periodic revalidation and domain adaptation to prevent model decay over time. Third, the economic implications of false positives and false negatives must be explicitly analyzed: excessive false alarms can result in unnecessary maintenance costs and downtime, while missed detections can lead to catastrophic asset failures. Integrating cost-sensitive learning or threshold-optimization frameworks could mitigate these trade-offs. Finally, scalability and interoperability should be addressed when deploying predictive maintenance solutions across multiple facilities—requiring modular architecture, cloud-edge synchronization, and cybersecurity measures to ensure continuous and secure operation. Addressing these factors will be crucial for transitioning the current framework from a controlled experimental stage to an industrial-grade, operationally resilient predictive maintenance system.

Looking forward, integrating deeper contextual data fusion—combining sensor, human operator, and external (e.g., weather, supply chain) streams—along with federated and edge learning paradigms, will be essential for reconciling data privacy with the collective intelligence and adaptability required for next-generation PdM [16] [44]. Further research should also consider hybrid approaches that combine physics-based models and explainable AI into a holistic maintenance ecosystem, as well as the development of Predictive Maintenance-as-a-Service (PdMaaS) to lower capital barriers for small and mid-sized enterprises [29].

6. Conclusion

This study demonstrates the transformative potential of AI-driven predictive maintenance in enhancing the reliability and operational excellence of infrastructure and facilities management systems. By systematically comparing multiple machine learning models—including logistic regression, k-nearest neighbors, support vector machines, random forest classifiers, and XGBoost—on a large-scale synthetic dataset, the research highlights the superior performance of advanced ensemble methods in accurately predicting both the occurrence and specific types of machine failures. Remarkably, the best-performing classifier was the XGBoost, with a more than 98 percent accuracy, which supports its claims to capture non-linear and complex interactions inherent in high-dimensional sensor data. These results also show the significant effect of significant operational characteristics of torque, rotational speed, and tool wear on the part, and ignore the impact of categorical conditions of the product type, which is also in line with what is expected in the domain based on previous empirical research. The use of methods such as PCA and permutation feature importance even further clarified the mechanistic basis of failure modes to increase the interpretability of models, which is an essential aspect in deciding to implement the model within an industrial setting. In addition, the complex preprocessing system, such as custom encoding, scaling, and oversampling to deal with the imbalance, allowed the robust, generalizable training and validation of the models, overcoming the typical problems of predictive analytics. Although simpler models such as the k-nearest neighbor method allow fast inference and are therefore appropriate to use when limited resources are available, the relative accuracy trade-off means it must be selected strategically based on operational priorities. In sum, the study expands the scope of predictive maintenance, offering a scalable framework that will be empirically validated and integrates innovative approaches of AI methodologies with practical applicability. It not only provides theoretical knowledge but also helpful advice on how to maintain the maintenance regime, prevent unscheduled breakdowns, and adopt sustainable and cost-effective infrastructure management in the Industry 4.0 era. Model adaptability will be further extended to temporal dynamics, and explainability will be further advanced, along with promoting ease of integration with evolving IoT ecosystems, further enhancing resilience and intelligence of critically essential assets at the core of critical infrastructure. Future work will validate this framework on real-world, time-series sensor datasets from infrastructure assets such as HVAC or bridge-monitoring systems to ensure that the model's predictive accuracy and interpretability translate effectively into practical, operational contexts.

References

[1] Sixense Digital (Beyond Suite), "Predictive maintenance and SHM (Structural Health Monitoring): Transforming your infrastructure management," Beyond Suite, Aug. 2024. [Online]. Available: https://www.beyond-suite.com/en/predictivemaintenance-0824/. [Accessed: 6-Aug-2025].

[2] WorkTrek, "CMMS vs Traditional Maintenance," WorkTrek Blog, Feb. 2024 (approx.). [Online]. Available: https://worktrek.com/blog/cmms-vs-traditional-maintenance/. [Accessed: 6-Aug-2025].

[3] EIS Council, "The cost of unpreparedness: Economic impacts of infrastructure failures," EIS Council, 2025 (approx. 9 months ago). [Online]. Available: https://eiscouncil.org/the-cost-of-infrastructure-failures-preparedness/. [Accessed: 6-Aug-2025].

[4] Ganit Inc., "The $1.4 Trillion Global Cost of Unplanned Downtime," LinkedIn, 2 July 2025. [Online]. Available: https://www.linkedin.com/pulse/14-trillion-global-cost-unplanned-downtime-ganit-inc-hj1bc. [Accessed: 17-Aug-2025].

[5] World Bank, "Economic benefits of more reliable and resilient infrastructure: Final LOW WB G20 Report v4," World Bank, 1-Jun-2021. [Online]. Available: https://ppp.worldbank.org/public-private-partnership/sites/default/files/2022-03/Final-LOW_WB_G20_Report_v4_1JUN_2021.pdf. [Accessed: 6-Aug-2025].

[6] Mobility Work, "Industry 4.0 & IoT," Mobility Work, 2025. [Online]. Available: https://mobility-work.com/iot-en/. [Accessed: 6-Aug-2025].

[7] ThreeBond India, "Best 10 common industrial maintenance challenges & solutions," ThreeBond India, 2024 (approx. 1.4 years ago). [Online]. Available:

https://threebondindia.com/industrial-maintenance-challenges-and-solutions/. [Accessed: 6-Aug-2025].

[8] UpKeep, "The advantages & disadvantages of preventive maintenance," UpKeep Learning Center, 2025. [Online]. Available: https://upkeep.com/learning/benefits-of-preventive-maintenance/. [Accessed: 6-Aug-2025].

[9] WorkTrek, "CMMS vs Traditional Maintenance: Making the choice clear!," WorkTrek Blog, 2024. [Online]. Available: https://worktrek.com/blog/cmms-vs-traditional-maintenance/. [Accessed: 6-Aug-2025].

[10] K. Lefrouni and S. Taibi, "Artificial Intelligence Techniques for Industrial Predictive Maintenance: A Systematic Review of Recent Advances," Journal Européen des Systèmes Automatisés, vol. 58, no. 4, 2025.

[11] A. Abbas, "AI for predictive maintenance in industrial systems," International Journal of Advanced Engineering Technologies and Innovations, vol. 1, no. 1, pp. 31–51, 2024.

[12] R. A. Abdulrazzq, N. M. Sajid, and M. S. Hasan, "Artificial intelligence-driven predictive maintenance in IoT systems," South Florida Journal of Development, vol. 5, no. 12, pp. e4781–e4781, 2024.

[13] Y. Melnik, "Machine failure prediction using machine learning: Why it is beneficial," InData Labs, 10-Oct-2024. [Online]. Available: https://indatalabs.com/blog/machine-failure-prediction-machine-learning/. [Accessed: 6-Aug-2025].

[14] T. P. Carvalho, F. A. Soares, R. Vita, R. D. P. Francisco, J. P. Basto, and S. G. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," Computers & Industrial Engineering, vol. 137, p. 106024, 2019.

[15] I. Hector and R. Panjanathan, "Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques," PeerJ Computer Science, vol. 10, p. e2016, 2024.

[16] M. Yousuf, T. Alsuwian, A. A. Amin, S. Fareed, and M. Hamza, "IoT-based health monitoring and fault detection of industrial AC induction motor for efficient predictive maintenance," Measurement and Control, vol. 57, no. 8, pp. 1146–1160, 2024.

[17] F. Civerchia, S. Bocchino, C. Salvadori, E. Rossi, L. Maggiani, and M. Petracca, "Industrial Internet of Things monitoring solution for advanced predictive maintenance applications," Journal of Industrial Information Integration, vol. 7, pp. 4–12, 2017.

[18] A. Ucar, M. Karakose, and N. Kırımça, "Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends," Applied Sciences, vol. 14, no. 2, p. 898, 2024.

[19] Z. M. Çınar et al., "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," Sustainability, vol. 12, no. 19, p. 8211, 2020.

[20] C. Tsallis, P. Papageorgas, D. Piromalis, and R. A. Munteanu, "Application-wise review of Machine Learning-based predictive maintenance: Trends, challenges, and future directions," Applied Sciences, vol. 15, no. 9, p. 4898, 2025.

[21] M. S. Azari, F. Flammini, S. Santini, and M. Caporuscio, "A systematic literature review on transfer learning for predictive maintenance in industry 4.0," IEEE Access, vol. 11, pp. 12887–12910, 2023.

[22] M. L. Singgih and F. F. Zakiyyah, "Machine Learning for Predictive Maintenance: A Literature Review," in 2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE), 2024, pp. 250–256.

[23] A. Aminzadeh et al., "A machine learning implementation to predictive maintenance and monitoring of industrial compressors," Sensors, vol. 25, no. 4, p. 1006, 2025.

[24] U. Dereci and G. Tuzkaya, "An explainable artificial intelligence model for predictive maintenance and spare parts optimization," Supply Chain Analytics, vol. 8, p. 100078, 2024.

[25] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Upsampling Under Varying Imbalance Levels," 2025.

[26] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "Xgboost and random forest algorithms: an in depth analysis," Pakistan Journal of Scientific Research (PJOSR), vol. 3, no. 1, pp. 26–31, 2023.

[27] Y. Mahale, S. Kolhar, and A. S. More, "A comprehensive review on artificial intelligence driven predictive maintenance in vehicles: technologies, challenges and future research directions," Discover Applied Sciences, vol. 7, no. 4, p. 243, 2025.

[28] S. Aslam et al., "Machine Learning-Based Predictive Maintenance at Smart Ports Using IoT Sensor Data," Sensors, vol. 25, no. 13, p. 3923, 2025.

[29] R. A. Abu Rkab, "Predictive maintenance [Data set]," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/rashaali2003/predictive-maintenance.

[30] M. Kang and J. Tian, "Machine learning: Data pre-processing," in Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, pp. 111–130, 2018.

[31] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Global Transitions Proceedings, vol. 3, no. 1, pp. 91–99, 2022.

[32] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," Computational Statistics, vol. 37, no. 5, pp. 2671–2692, 2022.

[33] J. Brownlee, "Ordinal and One-Hot Encodings for Categorical Data," Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/.

[34] K. A. Sankpal and K. V. Metre, "A Review on Data Normalization Techniques," Int. J. Eng. Res. Technol, vol. 9, pp. 1438, 2020.

[35] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

[36] W. Zhang, D. Yang, and H. Wang, "Data-driven methods for predictive maintenance of industrial equipment: A survey," IEEE Systems Journal, vol. 13, no. 3, pp. 2213–2227, 2019.

[37] Scikit-learn Developers, "Plot classification boundaries with different SVM Kernels," scikit-learn documentation, 2025. [Online]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html. [Accessed: 6-Aug-2025].

[38] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2009.

[39] A. Q. Khan, M. H. Naveed, M. D. Rasheed, and A. Pimanmas, "Prediction of stress–strain behavior of PET FRP-confined concrete using machine learning models,"

Arabian Journal for Science and Engineering, vol. 50, no. 11, pp. 7911–7931, 2025.

[40] P. Ghadekar et al., "Predictive maintenance for industrial equipment: Using XGBoost and local outlier factor with explainable AI for analysis," in 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2024, pp. 25–30.

[41] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," Big Data Mining and Analytics, vol. 3, no. 2, pp. 85–101, 2020.

[42] A. Seraj et al., "Cross-validation," in Handbook of Hydroinformatics, Elsevier, 2023, pp. 89–105.

[43] F. Calabrese, A. Regattieri, M. Bortolini, F. G. Galizia, and L. Visentini, "Feature-based multi-class classification and novelty detection for fault diagnosis of industrial machinery," Applied Sciences, vol. 11, no. 20, p. 9580, 2021.

[44] C. Catal, "Performance evaluation metrics for software fault prediction studies," Acta Polytechnica Hungarica, vol. 9, no. 4, pp. 193–206, 2012.

[45] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[46] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[47] T. Adimulam, M. Bhoyar, and P. Reddy, "AI-Driven Predictive Maintenance in IoT-Enabled Industrial Systems," Iconic Research And Engineering Journals, vol. 2, no. 11, pp. 398–410, 2019.

[48] C. M. Walker et al., "Explainable artificial intelligence technology for predictive maintenance," Idaho National Laboratory (INL), Idaho Falls, ID, USA, Report No. INL/RPT-23-74159-Rev000, 2023.