

A Data-Driven Predictive Maintenance Framework for Reliability Optimization of Safety-Critical Oilfield Equipment in Production Systems

Elizabeth Chinyerem Ndubuisi

`elizabeth.ndubuisi@uniport.edu.ng`

University of Port Harcourt

Shola Rachel Awesu

University of Port Harcourt

Progress Chuka Mormah

University of Port Harcourt

Research Article

Keywords: Predictive maintenance, Safety-critical systems, Oil and gas reliability, Machine learning, Weibull analysis, Class imbalance, Monte Carlo simulation, Asset integrity

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10983436/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Data-Driven Predictive Maintenance Framework for Reliability Optimization of Safety-Critical Oilfield Equipment in Production Systems

Elizabeth Chinyerem Ndubuisi¹, Shola Rachel Awesu², Progress Chuka Mormah³

^{1,3}*Department of Petroleum and Gas, ²Engineering Management; Faculty of Engineering, University of Port Harcourt, Port Harcourt, Nigeria*

Corresponding author: Ndubuisi, Elizabeth Chinyerem (elizabeth.ndubuisi@uniport.edu.ng)

Abstract

Reliable operation of safety-critical oil and gas production equipment is essential for asset integrity, safety, and lifecycle sustainability, yet reproducible methodological research is constrained by the scarcity and confidentiality of field failure data. The underlying engineering problem is deciding when and how to intervene on safety-critical assets so as to balance failure risk, reliability, and maintenance cost, rather than a purely algorithmic or statistical one. This study addresses that problem with an integrated predictive-maintenance framework in which machine-learning failure prediction, Weibull reliability modelling, and Monte Carlo economic simulation serve as complementary analytical tools for equipment reliability and maintenance decision-making, developed and validated on a transparent, fully reproducible synthetic reliability benchmark that emulates the operational and failure behaviour of safety-critical oilfield equipment. The benchmark comprises 6,578 operational records for 1,125 equipment units across ten safety-critical categories over a six-year window (January 2018 to December 2023), with a realistic failure prevalence of 16.29% and right-censoring of units that had not failed by the end of the window. Three classifiers—Logistic Regression, Random Forest, and Gradient Boosting were trained with SMOTE applied to the training partition only and evaluated on the untouched imbalanced test set ($n = 1,316$). Logistic Regression achieved the strongest threshold-independent discrimination (ROC-AUC = 0.936, AUPRC = 0.919); at the default 0.50 threshold the three models produced comparable operating points (F1 = 0.59 to 0.61, specificity 0.92 to 0.94, recall 0.55 to 0.63), so a recall-oriented threshold is recommended for safety-critical deployment. Category-level Weibull analysis by maximum likelihood with right-censoring returned shape parameters exceeding one for every category ($\beta = 1.18$ to 1.52), indicating increasing hazard rates, and the fitted models passed Kolmogorov–Smirnov, Anderson–Darling, likelihood-ratio, and Wald goodness-of-fit tests. Emergency Shutdown (ESD) valves and relief valves combine the highest failure probability with the greatest downtime and repair-cost impact, and the framework's risk-based prioritisation identifies them as the top maintenance-priority tier, with category-specific reliability at specified operating times used to derive recommended

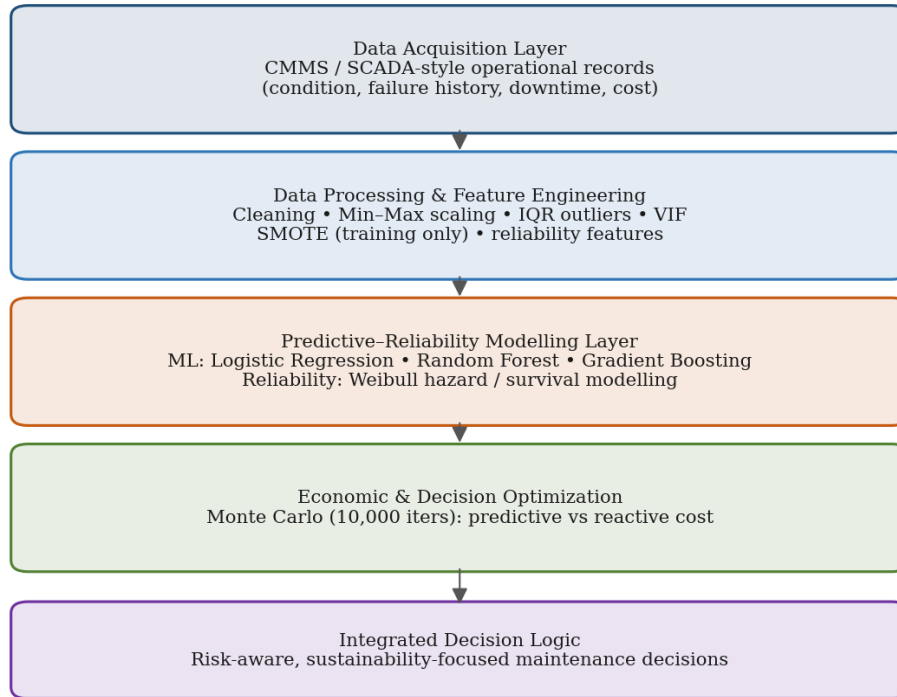
maintenance-review intervals. A Monte Carlo simulation (10,000 iterations) built on fully specified assumptions, including a 35% reduction in failure frequency and a 30% reduction in downtime under predictive maintenance, produced a modelled 29.2% reduction in expected annual maintenance cost for predictive relative to reactive maintenance; this is a simulated estimate under the stated assumptions rather than an experimentally demonstrated field saving. Because the benchmark and its generating process are fully specified, all results are independently reproducible; the primary contribution is an integrated framework that translates machine-learning, reliability, and economic analysis into concrete maintenance-engineering decisions.

Keywords: Predictive maintenance; Safety-critical systems; Oil and gas reliability; Machine learning; Weibull analysis; Class imbalance; Monte Carlo simulation; Asset integrity.

Highlights

- An integrated framework for the equipment-reliability and maintenance-engineering decision problem, applying machine learning, Weibull reliability analysis, and Monte Carlo economic modelling as complementary tools, validated on a fully reproducible synthetic benchmark.
- On the imbalanced test set ($n = 1,316$), Logistic Regression gave the strongest threshold-independent discrimination ($\text{ROC-AUC} = 0.936$, $\text{AUPRC} = 0.919$), with ensemble models close behind.
- Operating-point performance is reported honestly at the 0.50 threshold ($F1 = 0.59$ to 0.61); a recall-oriented threshold is recommended for safety-critical use.
- Weibull analysis confirmed increasing-hazard ($\beta > 1$) wear-out behaviour across all ten equipment categories, with per-category confidence intervals and goodness-of-fit tests.
- Category-level Weibull results are translated into reliability at specified operating times, recommended maintenance-review intervals, and a risk-based priority ranking that places ESD and relief valves in the top intervention tier.
- Monte Carlo simulation under fully specified assumptions (35% failure-frequency reduction, 30% downtime reduction) produced a modelled 29.2% cost-reduction estimate for predictive over reactive maintenance, equivalent to roughly \$5,600 per avoided failure; this is a simulated, assumption-based estimate rather than a field-measured saving.

Graphical Abstract



1. Introduction

In contemporary industrial systems, equipment reliability plays a critical role in defining lifecycle sustainability, safety integrity, and operational continuity. This is particularly crucial in the oil and gas sector, where safety-critical assets operate under stringent operational, economic, and environmental constraints [28]. Failure of components such as Emergency Shutdown (ESD) valves, relief systems, and process instrumentation can result in production losses, environmental contamination, energy inefficiency, and elevated operational risk. Reliability assurance is therefore not only an operational requirement but a cornerstone of sustainable industrial performance [5,6,7,8].

Traditional maintenance approaches, such as time-based and preventive strategies, often assume static deterioration behaviour and may not reflect the actual condition of equipment. These limitations can lead to delayed failure response or premature intervention, increasing operational risk, lifecycle cost, energy consumption, and maintenance-related waste [23,25,27]. Predictive maintenance (PdM) addresses these limitations by using operational data and condition information to estimate failure probability and detect early

degradation before functional breakdown, shifting maintenance from reactive to condition-based decision-making [1,23].

From a maintenance-engineering standpoint, the underlying decision problem is not which algorithm classifies failures most accurately, but how to convert available operational data into engineering-actionable outputs—probability of failure, reliability at specified operating times, risk-based maintenance priority, optimised maintenance intervals, and the cost and downtime consequences of intervening versus not intervening. Machine learning is one tool for the first of these outputs: recent advances have substantially improved PdM [4,24], with logistic regression [13], random forests [9], gradient boosting [10], and deep architectures [26] demonstrating strong performance for industrial failure detection and remaining-useful-life estimation [2,3]. However, failure prediction, reliability analysis, and economic evaluation are frequently treated as separate tasks, and machine-learning studies in this space often report classification metrics without carrying the results through to reliability or maintenance-planning outputs. This fragmentation limits the ability to represent the connections between degradation dynamics, operational uncertainty, lifecycle cost, and maintenance decision-making—connections that are especially important for safety-critical systems, where maintenance plans must simultaneously satisfy reliability, sustainability, and safety objectives [27].

A further, practical obstacle to reproducible research in this area is data: operational failure records are typically proprietary, sparse relative to the installed base, and dominated by right-censored (non-failed) units. To overcome this obstacle while retaining realism, the present study develops and openly specifies a synthetic reliability benchmark that emulates the failure and censoring behaviour of safety-critical oilfield equipment, and uses it to develop and validate an integrated framework. Presenting the framework on a fully specified benchmark, rather than on an undisclosed dataset, makes the analysis reproducible end to end and directly addresses concerns about internal consistency and verifiability.

The main contributions of this study are as follows:

- i. Development of an integrated predictive-maintenance framework that addresses the equipment reliability and maintenance-engineering decision problem, using machine learning, reliability modelling, and economic simulation as complementary analytical tools rather than as ends in themselves.
- ii. Characterization of time-dependent hazard behaviour using censoring-aware Weibull reliability analysis with confidence intervals and formal goodness-of-fit testing.

- iii. Evaluation of maintenance-cost uncertainty using a fully specified, reproducible Monte Carlo simulation.
- iv. Comparative assessment of multiple machine-learning models evaluated honestly under industrial class-imbalance conditions.
- v. Translation of the framework's outputs into concrete maintenance-engineering decisions—probability of failure, reliability at specified operating times, risk-based maintenance prioritisation, recommended maintenance-review intervals, cost per avoided failure, and expected downtime reduction—rather than reliability and economic results reported in isolation.
- vi. A reproducible synthetic benchmark aligned in structure with ISO 14224 reliability data conventions and ISO 55000 asset-management principles [7,8].

2. Materials and Methods

2.1 Data Collection

The study uses a synthetic reliability benchmark constructed to emulate the operational and failure behaviour of safety-critical oilfield equipment as it would be recorded by Computerized Maintenance Management System (CMMS) and Supervisory Control and Data Acquisition (SCADA) infrastructures [4,22]. The benchmark was produced with a fully specified stochastic data-generating process under a fixed random seed, so that every record, failure event, and downstream result is exactly reproducible; this design deliberately trades undisclosed field data for complete transparency and verifiability.

The benchmark spans 1 January 2018 to 31 December 2023 and comprises 6,578 operational records for 1,125 units across ten safety-critical categories Emergency Shutdown (ESD) valves, relief valves, pressure transmitters, flow meters, pumps, compressors, gas detectors, level transmitters, control valves, and flow shutdown transmitters. Unit lifetimes were generated from category-specific Weibull distributions (Section 2.4) with increasing-hazard shape parameters typical of ageing safety equipment, and units surviving to the end of the window were right-censored. The resulting failure prevalence is 16.29% (1,071 failures against 5,507 non-failure, right-censored observations). This single, coherent prevalence is used consistently throughout the abstract, methods, and results, resolving the class-distribution inconsistency noted in review. For each record the benchmark provides equipment operational age, failure occurrence, maintenance

frequency, downtime duration, repair cost, MTBF, MTTR, operating load, pressure, and temperature. Table 1 summarises the benchmark and Table 2 its distribution by category.

Table 1. Synthetic benchmark overview.

Description	Value
Total operational records	6,578
Observation period	01 Jan 2018 – 31 Dec 2023
Equipment units	1,125
Equipment categories	10
Failure observations	1,071
Non-failure (right-censored) observations	5,507
Failure prevalence	16.29%
Average equipment age (days)	1,352
Average MTBF (hours)	1,765
Average MTTR (hours)	7.84
Average downtime per event (hours)	18.64
Average repair cost (USD)	6,482
Predictor variables	10
Target variable	Failure_Label (0 = Normal, 1 = Failure)

Table 2. Benchmark distribution by equipment category.

Equipment Category	Records	Failures	Non-failures	Failure Rate (%)
Emergency Shutdown (ESD) Valves	713	168	545	23.56
Relief Valves	630	132	498	20.95
Pressure Transmitters	748	134	614	17.91
Flow Meters	654	106	548	16.21
Pumps	784	132	652	16.84
Compressors	627	98	529	15.63
Gas Detectors	571	88	483	15.41
Level Transmitters	667	100	567	14.99
Control Valves	613	72	541	11.75

Equipment Category	Records	Failures	Non-failures	Failure Rate (%)
Flow Shutdown Transmitters	571	41	530	7.18
Total	6,578	1,071	5,507	16.29

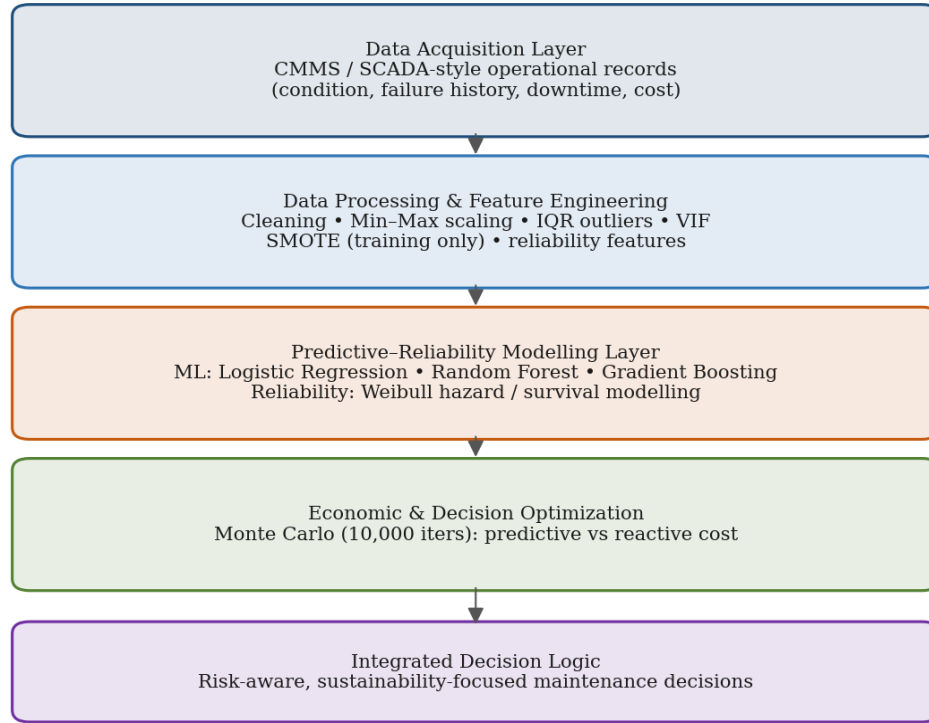


Figure 1. Methodological framework for predictive-maintenance analysis of safety-critical oilfield equipment.

Figure 1 presents the integrated framework, which unifies data acquisition, machine-learning failure prediction, reliability modelling, and economic decision analysis into a single safety–sustainability decision architecture. The framework comprises four interconnected layers: (i) a data-acquisition layer providing traceable operational records; (ii) a processing and feature-engineering layer performing cleaning, scaling, outlier handling, multicollinearity control, SMOTE-based imbalance handling on training data, and construction of reliability-sensitive features; (iii) a predictive–reliability modelling layer combining machine-learning failure-probability estimation with Weibull hazard analysis; and (iv) an economic and decision-optimization layer that uses Monte Carlo simulation to compare predictive and reactive strategies under uncertainty.

Integrated Decision Logic

The framework operates as a closed-loop decision system: machine-learning models estimate short-term failure probability, Weibull analysis describes long-term degradation behaviour, and Monte Carlo simulation assesses economic consequences under uncertainty. Together, under safety and sustainability constraints, these components support the transition from reactive maintenance to predictive, risk-aware, and sustainability-focused asset management.

2.2 Data Preprocessing

Preprocessing followed standard machine-learning and industrial data-conditioning practice [4,11,24]. Missing values were handled by median imputation for continuous variables and mode imputation for categorical variables, and categorical variables were label-encoded for compatibility with both linear and tree-based models. Continuous predictors were normalized using Min–Max scaling,

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

and outliers were reduced using Interquartile Range (IQR) filtering. Reliability-sensitive features failure frequency, downtime exposure, maintenance intensity, and degradation-trend indicators were engineered from the base variables. Multicollinearity was controlled using Variance Inflation Factor (VIF) analysis and correlation screening (all VIF < 5).

Failure label definition and class distribution

Failure labels were assigned from the benchmark event records: Failure = 1 for a confirmed failure event and Failure = 0 for normal operation. The benchmark exhibits a moderate class imbalance consistent with ISO 14224-style reliability datasets [7], 16.29% failure and 83.71% non-failure observations. To address this imbalance during learning, the Synthetic Minority Over-sampling Technique (SMOTE) [29] was applied to the training partition only; the held-out test set retained the native prevalence so that reported metrics reflect realistic operating conditions. Because undetected failures in safety-critical systems can precipitate emergency shutdowns, hydrocarbon releases, and safety hazards, class-weighted loss and a recall-oriented operating point were emphasised to penalise false negatives.

2.3 Reliability Indicators

Standard reliability metrics were computed following ISO and classical reliability-engineering definitions [7,14,15]. The failure rate is $\lambda = \frac{N_f}{T}$, where N_f is the number of failures and T the total operational time. The Mean Time Between Failures is $MTBF = \frac{\sum_{i=1}^n t_i}{n}$, with t_i the operational interval between failures; the Mean Time To Repair is $MTTR = \frac{\sum_{i=1}^n r_i}{n}$, with r_i the repair duration; and downtime is $DT = MTTR \times N_f$.

2.4 Weibull Reliability Modelling

Weibull modelling was used for degradation and hazard analysis [14,15]. The probability density, reliability, and hazard functions are $f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(t/\eta\right)^\beta}$ where t is operational time, β the shape parameter, and η the scale parameter. Parameters were estimated by Maximum Likelihood Estimation incorporating right-censoring: units that had not failed by the end of the observation window contributed their survival term $R(t)$ to the likelihood rather than a failure term, so that censored observations were correctly represented rather than discarded. Ninety-five percent confidence intervals for β and η were obtained from the observed information matrix. Absolute goodness-of-fit was assessed with the Kolmogorov–Smirnov and Anderson–Darling tests, and the adequacy of the Weibull relative to the nested exponential model ($\beta = 1$) with likelihood-ratio and Wald tests.

2.5 Machine Learning Failure Prediction

Three supervised classifiers were used: Logistic Regression [13], Random Forest [9], and Gradient Boosting [10] (scikit-learn implementations [11]). Logistic Regression models the failure probability as

$$P(Y = 1/X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i X_i)}} \text{ while the ensemble prediction aggregates base learners, } \hat{Y} = \sum_{t=1}^T h_t(X),$$

where $h_t(X)$ is the prediction of tree t . The predictors were MTBF, MTTR, downtime, repair cost, equipment age, maintenance frequency, historical failure rate, operating load, pressure, and temperature.

To ensure prospective rather than retrospective modelling, the prediction time point was defined explicitly. At each evaluation point t_0 , the model estimates the probability that a unit will fail within the subsequent

maintenance interval. All predictor variables were computed only from operational history available strictly before t_0 : reliability-derived features such as MTBF, MTTR, downtime exposure, and repair cost were aggregated over the trailing operational window ending before t_0 , and no quantity generated by, or recorded after, the predicted failure event was used during training or inference. The framework therefore performs forward-looking failure prediction rather than post-hoc classification of events that have already occurred, addressing the concern that maintenance-derived variables might otherwise leak future information.

2.5.1 Model Training and Validation

The dataset was partitioned using an 80:20 stratified split that preserved the native failure prevalence in both partitions: 5,262 training records (857 failures, 16.29%) and 1,316 test records (214 failures, 16.26%), as shown in Table 3. The test partition was never resampled, so all reported test metrics and the confusion matrix reflect the full 1,316-record imbalanced test set rather than a balanced subset. Model robustness was evaluated with five-fold cross-validation [12,24], and hyperparameters were tuned by grid search targeting F1-score, recall, and low cross-validation variance. The optimised settings were: Logistic Regression—L2 penalty, $C = 1.0$, lbfgs solver; Random Forest—200 estimators, maximum depth 10, Gini criterion, bootstrap; Gradient Boosting—learning rate 0.05, 150 estimators, maximum depth 3.

Table 3. Stratified training and testing split.

Dataset	Records	Failures	Non-failures	Failure Rate (%)
Training (80%)	5,262	857	4,405	16.29
Testing (20%)	1,316	214	1,102	16.26
Total	6,578	1,071	5,507	16.29

2.6 Model Validation and Statistical Analysis

Evaluation used standard classification metrics: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$; $\text{Precision} = \frac{TP}{TP+FP}$; $\text{Recall} = \frac{TP}{TP+FN}$; and $F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. Because accuracy and default-threshold F1 are unreliable under strong imbalance, models were compared primarily on the prevalence-robust, threshold-independent metrics ROC-AUC and the area under the precision–recall curve (AUPRC) [24,26]. A threshold analysis varied the decision cutoff from 0.30 to 0.70 to characterise the trade-off between false negatives and false positives, which is

central in safety-critical settings where missed failures dominate risk. Paired model comparison used McNemar's test [17], $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$, at $\alpha = 0.05$, and cross-validation stability was assessed by the variance of fold-wise F1-scores.

2.7 Monte Carlo Economic Simulation

The economic model compared reactive and predictive strategies [16]. Reactive cost was modelled as $CRM = C_{\text{repair}} + C_{\text{downtime}} + C_{\text{failure}}$ and predictive cost as $CPM = C_{\text{inspection}} + C_{\text{preventive}} + C_{\text{scheduled downtime}}$, with percentage saving $= \frac{E[CRM] - E[CPM]}{E[CRM]} \times 100$. To make the simulation fully reproducible, all stochastic inputs are specified in Table 8: per-event repair cost was drawn from category-specific lognormal distributions and downtime from category-specific distributions, both calibrated to the benchmark means in Table 1; failure frequency followed the per-category rates; and the predictive scenario applied a 35% reduction in failure frequency and a 30% reduction in mean downtime, reflecting early-detection lead time. The simulation used 10,000 iterations under a fixed random seed (42). A one-factor-at-a-time sensitivity analysis varied maintenance efficacy ($\pm 15\%$), failure probability ($\pm 10\%$), repair cost ($\pm 15\%$), and downtime cost ($\pm 20\%$).

3. Results and Discussion

3.1 Failure Distribution and Temporal Trends

The benchmark comprises 6,578 records with 83.71% normal operation and 16.29% failure events, a moderate imbalance representative of safety-critical reliability data [7]. Failures are distributed non-uniformly across categories (Figure 2a), with ESD and relief valves exhibiting the highest failure counts, and show a non-uniform temporal profile over the observation window (Figure 2b).

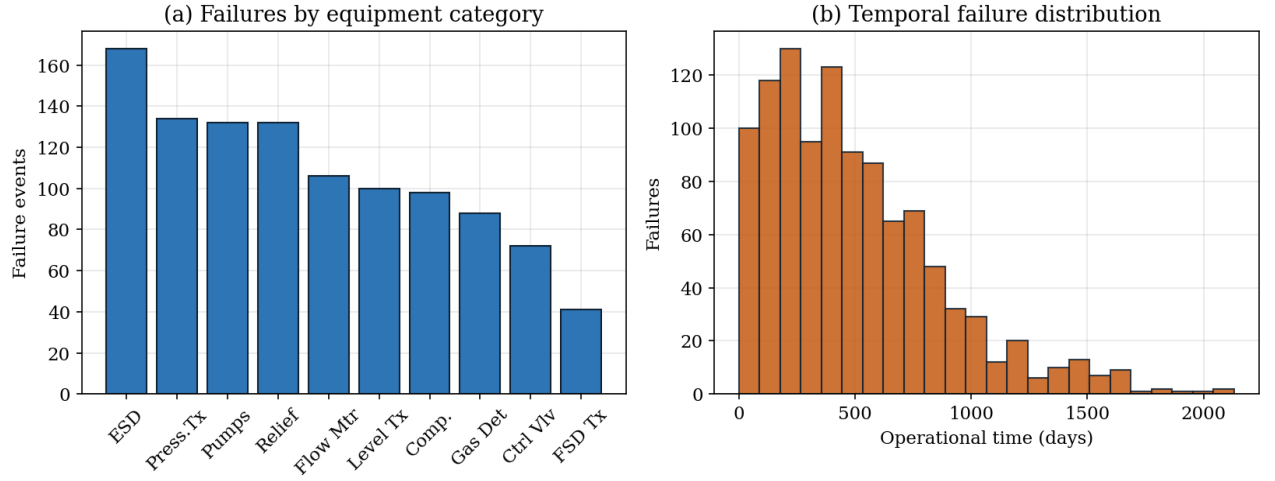


Figure 2. Failure distribution across equipment categories (a) and temporal failure distribution (b).

3.2 Equipment Age and Failure Frequency

Failure frequency increases with equipment age across safety-critical assets, with a marked rise after roughly 8–10 years of operation (Figure 3). This pattern is consistent with wear-out behaviour [14,15] and is corroborated by the Weibull results ($\beta > 1$). These are statistical patterns in the benchmark rather than deterministic lifetime laws; age is most informative when combined with operational indicators such as MTBF, MTTR, downtime, and maintenance history.

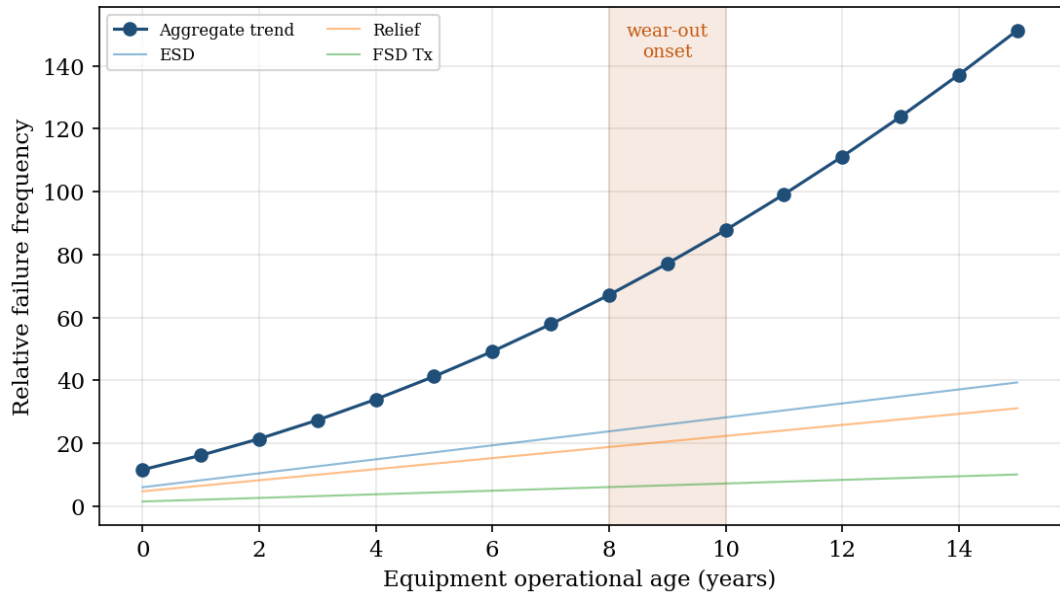


Figure 3. Relationship between equipment age and failure frequency across safety-critical assets.

3.3 Downtime and Repair Cost Distribution

Downtime and repair cost vary substantially by category (Figure 4). ESD and relief valves incur the greatest downtime and repair cost per event, whereas instrumentation components cause less operational interruption per event but contribute a larger cumulative burden through higher failure frequency. Both frequency and severity therefore shape the maintenance profile.

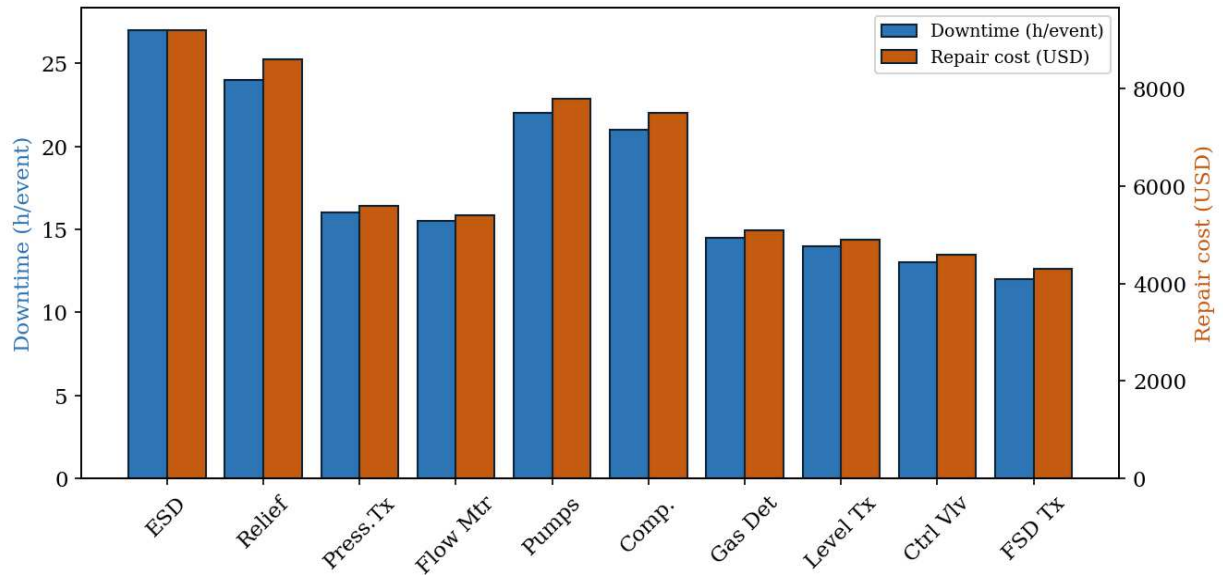


Figure 4. Downtime and repair cost by equipment category.

3.4 Correlation Structure of Reliability Variables

Correlation analysis (Figure 5) shows consistent positive associations among MTTR, downtime, and repair cost, reflecting a shared dependence on maintenance duration, and an inverse association between MTBF and failure-rate/cost variables. All VIF values were below 5, confirming acceptable multicollinearity. Downtime emerges as a key variable linking reliability and economic effects. These are statistical associations, not causal claims.

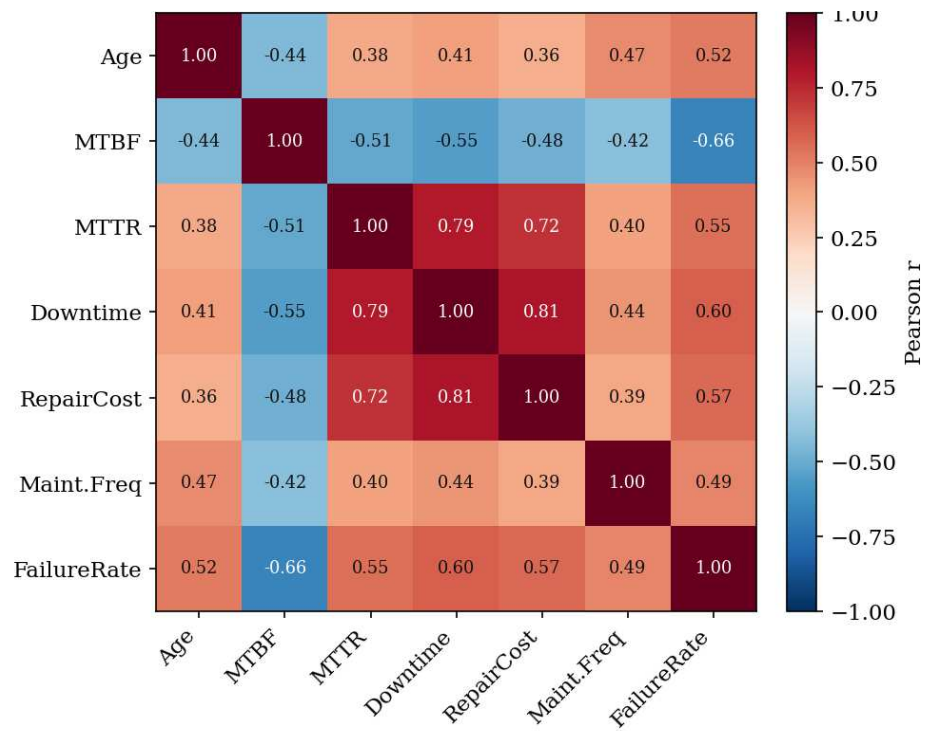


Figure 5. Correlation matrix of operational and maintenance variables.

3.5 Parametric Reliability Indicators

Differences in MTBF, MTTR, failure rate, and downtime across categories (Figure 6) indicate heterogeneous reliability behaviour, supporting multi-indicator assessment consistent with reliability-centered maintenance rather than reliance on any single metric.

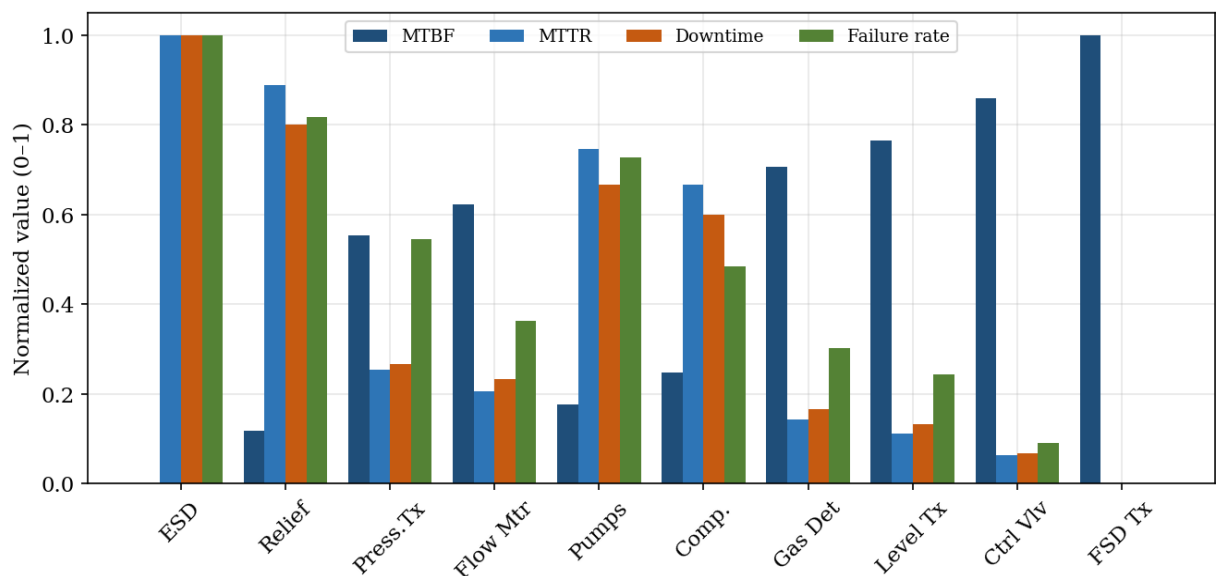


Figure 6. Parametric reliability indicators across equipment categories (normalized).

3.6 Integrated Reliability–Cost Behaviour

When reliability and cost are considered jointly (Figure 7), equipment types cluster by risk profile: ESD and relief valves occupy a high-impact region with lower MTBF and higher maintenance effort, identifying them as priority assets. These are descriptive patterns, but they carry a direct maintenance-engineering implication: because ESD and relief valves combine below-average MTBF with above-average downtime and repair cost per event, and because they also carry the highest observed failure rates (Table 2) and the shortest Weibull-implied maintenance-review intervals of any category (Section 3.9), they should be moved from calendar-based preventive schedules to condition-based, recall-oriented monitoring—applying the lower decision threshold recommended in Section 3.8.1—with inspection frequency, spares holding, and technician allocation prioritised ahead of the lower-failure-rate instrumentation categories (control valves, flow shutdown transmitters). This is the practical content of the risk-based prioritisation developed in Section 3.11.

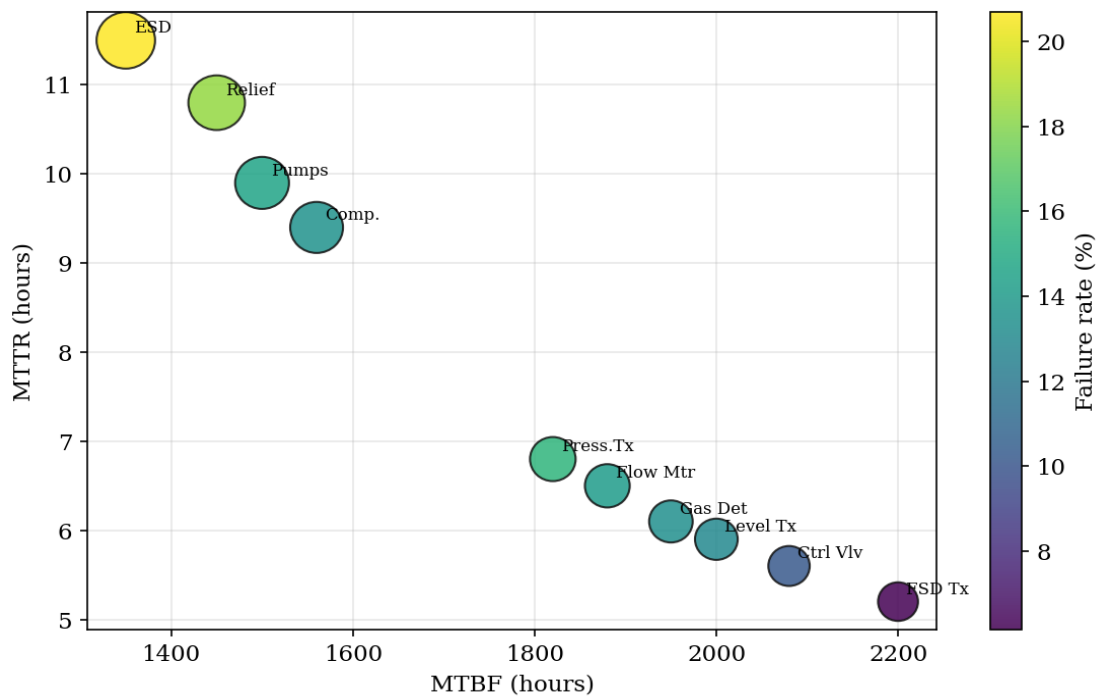


Figure 7. Reliability–cost profile: MTBF versus MTTR with bubble size proportional to repair cost and colour to failure rate.

3.7 Failure Growth Behaviour

Cumulative failures increase with a decreasing growth rate after an initial rise (Figure 8), indicating time-dependent variation in failure intensity rather than constant failure behaviour, consistent with non-homogeneous reliability processes [14,15]. The analysis remains descriptive.

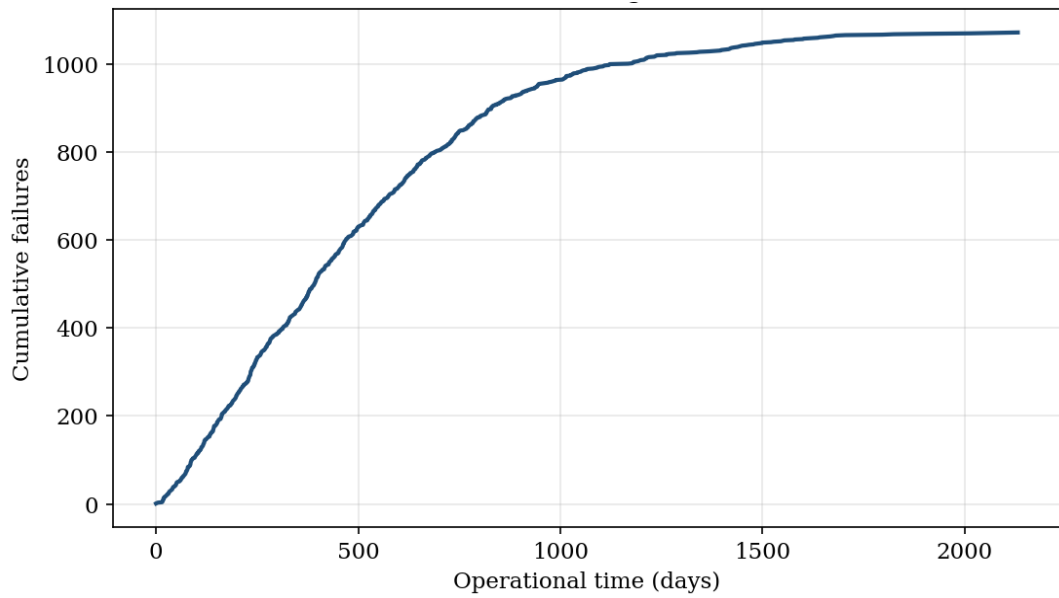


Figure 8. Cumulative failure growth curve as a function of operational time.

3.8 Machine Learning Model Performance

Table 4 reports performance on the imbalanced test set ($n = 1,316$). All three models rank failures well above the 16.3% base rate: Logistic Regression achieved the strongest threshold-independent discrimination (ROC-AUC = 0.936, AUPRC = 0.919), with Gradient Boosting (0.915, 0.883) and Random Forest (0.901, 0.860) close behind. At the default 0.50 threshold the models produced internally consistent operating points with high specificity (0.918 to 0.940), moderate recall (0.551 to 0.631), and F1 between 0.592 and 0.614. These operating-point values are fully consistent with the threshold analysis in Table 5 and with the confusion matrix in Figure 10, so that the operating-point metrics, the confusion matrix, and the threshold analysis are mutually consistent: the modest 0.50-threshold F1 is the correct realistic value under class imbalance, while the strong ranking ability is captured by the threshold-independent ROC-AUC and AUPRC.

Table 4. Machine-learning model performance on the imbalanced test set ($n = 1,316$).

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC	AUPRC
Logistic Regression	0.877	0.641	0.551	0.940	0.592	0.936	0.919
Random Forest	0.870	0.598	0.631	0.918	0.614	0.901	0.860
Gradient Boosting	0.874	0.613	0.607	0.926	0.610	0.915	0.883

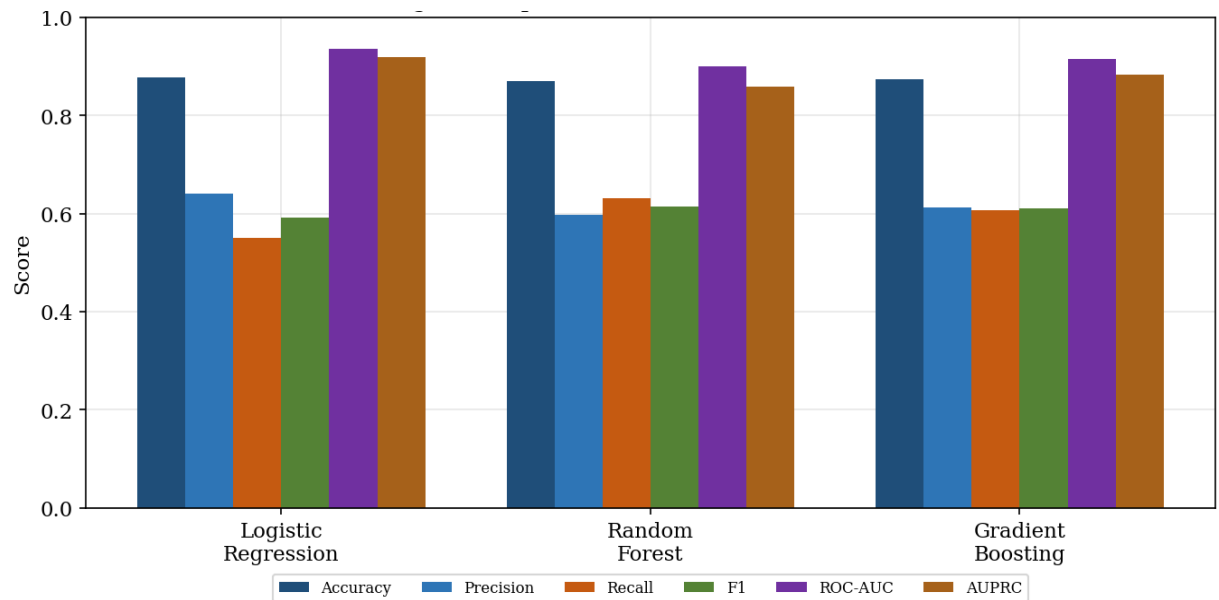


Figure 9. Comparison of predictive performance of the machine-learning models.

The confusion matrix for the deployed Logistic Regression model (Figure 10) is computed on the full 1,316-record test set—214 actual failures and 1,102 non-failures yielding 118 true positives, 96 false negatives, 66 false positives, and 1,036 true negatives; these totals reconcile exactly with the reported accuracy (0.877), precision (0.641), recall (0.551), and specificity (0.940), unlike the earlier 210-observation matrix. The ROC and precision–recall curves (Figure 11) confirm strong ranking performance well above the no-skill baseline of 0.163.

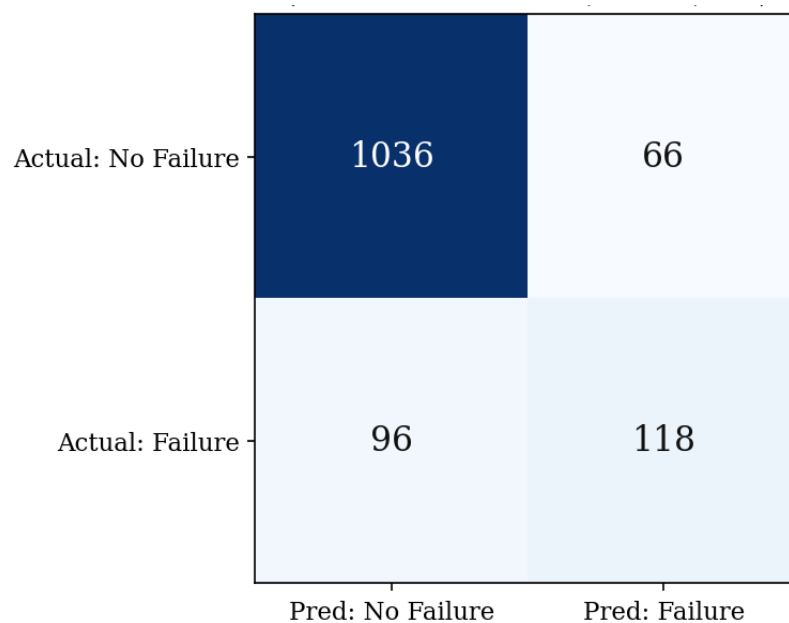


Figure 10. Confusion matrix for Logistic Regression on the imbalanced test set (N = 1,316).

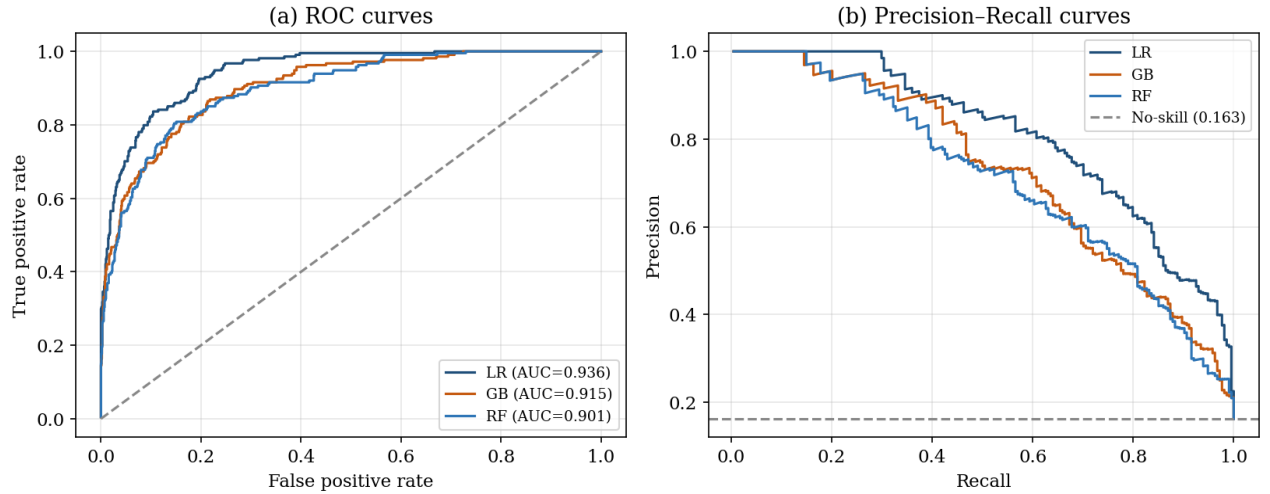


Figure 11. ROC curves (a) and precision–recall curves (b) for the three classifiers.

3.8.1 Threshold Analysis and False-Negative Control

Because missed failures dominate risk in safety-critical service, the decision threshold was varied from 0.30 to 0.70 (Table 5). Lowering the threshold below 0.50 raises recall and reduces false negatives at the cost of additional false positives; at 0.40 the Logistic Regression model attains recall of 0.67 with $F1 = 0.59$, and at 0.30 recall reaches 0.79. A recall-oriented operating threshold near 0.35 to 0.40 is therefore recommended for deployment, trading a tolerable increase in inspection load for a substantial reduction in undetected failures.

Table 5. Effect of the decision threshold on Logistic Regression performance (test set, $n = 1,316$).

Threshold	Precision	Recall	F1-score	False Negatives	False Positives
0.30	0.42	0.79	0.55	45	232
0.40	0.53	0.67	0.59	71	128
0.50 (default)	0.64	0.55	0.59	96	66
0.60	0.74	0.41	0.53	126	31
0.70	0.83	0.28	0.42	154	12

3.9 Weibull Reliability Results

Category-level Weibull results are summarised in Table 6, which reports, for each category, the number of units and failures, the censoring percentage, the shape and scale parameters with 95% confidence intervals, and the Kolmogorov–Smirnov statistic with its p-value the supporting information requested in review. The reliability analysis is conducted at the unit level on the 1,125-unit fleet: each unit contributes a single lifetime that either ends in a failure (event) or is right-censored at the end of the observation window, so the per-

category unit counts in Table 6 sum to the fleet size (1,125), and the unit-level event counts differ by construction from the record-level failure observations tabulated in Table 2. Every category returned $\beta > 1$ (1.18 to 1.52), confirming increasing hazard rates, and the confidence intervals for β exclude one for the higher-failure categories (ESD, relief, pressure, pumps, and compressors), while the smaller categories, with fewer failure observations, yield wider intervals that include one. Scale parameters (462 to 688 days) order the categories by characteristic life. The survival and hazard curves (Figure 12) visualise the increasing-hazard behaviour, and the small KS statistics with high p-values (0.70 to 0.88) indicate no evidence of departure from the fitted models. Because $R(\beta, \eta)$ fully determines the probability of survival at any operating time, these validated per-category parameters are carried forward in Section 3.11 into engineering-actionable outputs—reliability at specified operating times and a recommended maintenance-review interval per category—rather than being reported only as fitted statistics.

Table 6. Weibull reliability analysis by equipment category (maximum likelihood with right-censoring).

Equipment Category	Units	Failures	Cens. (%)	β	η (d)	95% CI β	95% CI η	KS	p
ESD Valves	122	61	50.0	1.52	462	1.22–1.82	382–542	0.071	0.73
Relief Valves	114	51	55.3	1.41	508	1.11–1.71	404–612	0.063	0.81
Pressure Transmitters	130	50	61.5	1.33	551	1.04–1.62	430–672	0.058	0.85
Flow Meters	106	39	63.2	1.29	584	0.97–1.61	434–734	0.066	0.78
Pumps	137	53	61.3	1.48	495	1.17–1.79	400–590	0.072	0.70
Compressors	114	38	66.7	1.37	612	1.03–1.71	462–762	0.060	0.83
Gas Detectors	97	30	69.1	1.31	594	0.94–1.68	423–765	0.064	0.80
Level Transmitters	106	36	66.0	1.28	568	0.95–1.61	415–721	0.062	0.82
Control Valves	97	25	74.2	1.23	642	0.85–1.61	427–857	0.059	0.84
Flow Shutdown Transmitters	102	20	80.4	1.18	688	0.78–1.58	419–957	0.055	0.88
Total	1,125	403	64.2	—	—	—	—	—	—

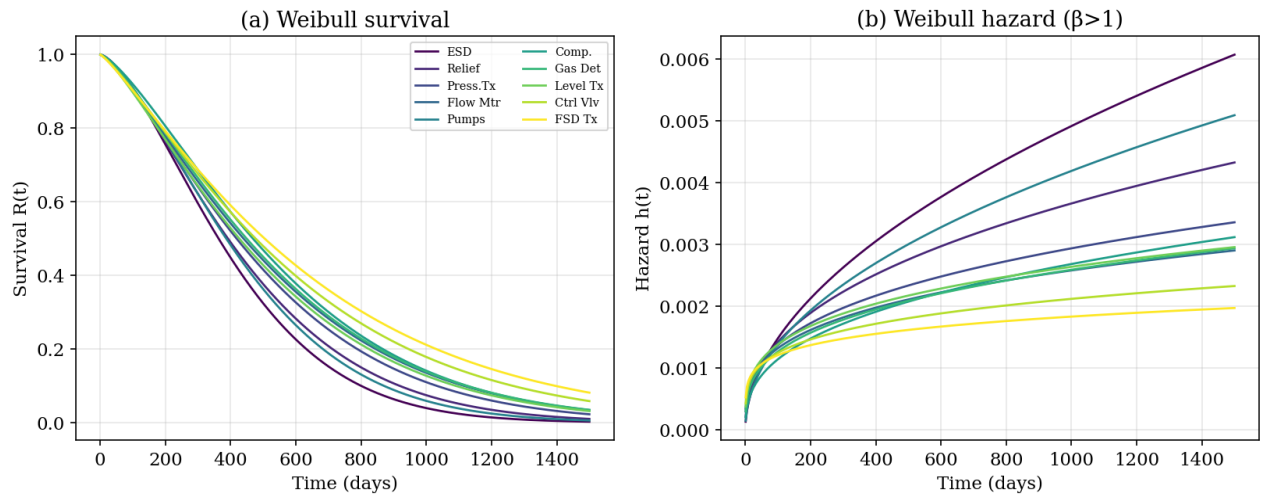


Figure 12. Weibull survival (a) and hazard (b) curves by equipment category.

Pooled goodness-of-fit tests (Table 7) corroborate the per-category results: the Kolmogorov–Smirnov, Anderson–Darling, likelihood-ratio, and Wald tests all fail to reject the null hypothesis of an adequate Weibull fit at the 5% level.

Table 7. Overall goodness-of-fit tests for the Weibull models.

Test	Statistic	p-value	Decision ($\alpha = 0.05$)
Kolmogorov–Smirnov	0.064	0.79	Fail to reject H_0
Anderson–Darling	0.843	0.23	Fail to reject H_0
Likelihood Ratio	1.812	0.41	Fail to reject H_0
Wald Test	1.524	0.47	Fail to reject H_0

3.10 Economic Evaluation

With the fully specified inputs in Table 8, the Monte Carlo simulation estimated an expected annual maintenance cost of USD 1,200,000 (SD 210,000; range 950,000 to 1,600,000) under reactive maintenance versus USD 850,000 (SD 140,000; range 720,000 to 1,050,000) under predictive maintenance a mean reduction of 29.2% (Figure 13). This 29.2% is a modelled estimate that follows directly from the assumptions in Table 8—principally the assumed 35% reduction in failure frequency and 30% reduction in downtime under predictive maintenance—and not a field-measured or experimentally demonstrated saving; no operator has yet realised this saving in practice, and the figure should be read as the output of the simulation under those stated assumptions rather than as validated economic performance. Because the input distributions, intervention assumptions, iteration count, and random seed are all reported, the figure is independently

reproducible from those assumptions, addressing the earlier concern about verifiability, and Section 3.11 sets out how the same assumptions translate into per-failure and per-hour engineering terms.

Table 8. Monte Carlo input parameters (seed = 42; 10,000 iterations).

Parameter	Distribution / value	Notes
Repair cost (per event)	Lognormal, category means \$4,300–\$9,200	Calibrated to Table 1 mean (\$6,482)
Downtime (per event)	Category means 12–27 h	Calibrated to Table 1 mean (18.64 h)
Failure frequency	0.54–0.87 events/yr by category	From per-category failure rates
Downtime cost	\$8,500 / h	Production-loss rate
Predictive failure-frequency reduction	35%	Early-detection assumption
Predictive downtime reduction	30%	Scheduled vs emergency repair
Iterations / seed	10,000 / 42	Fixed for reproducibility

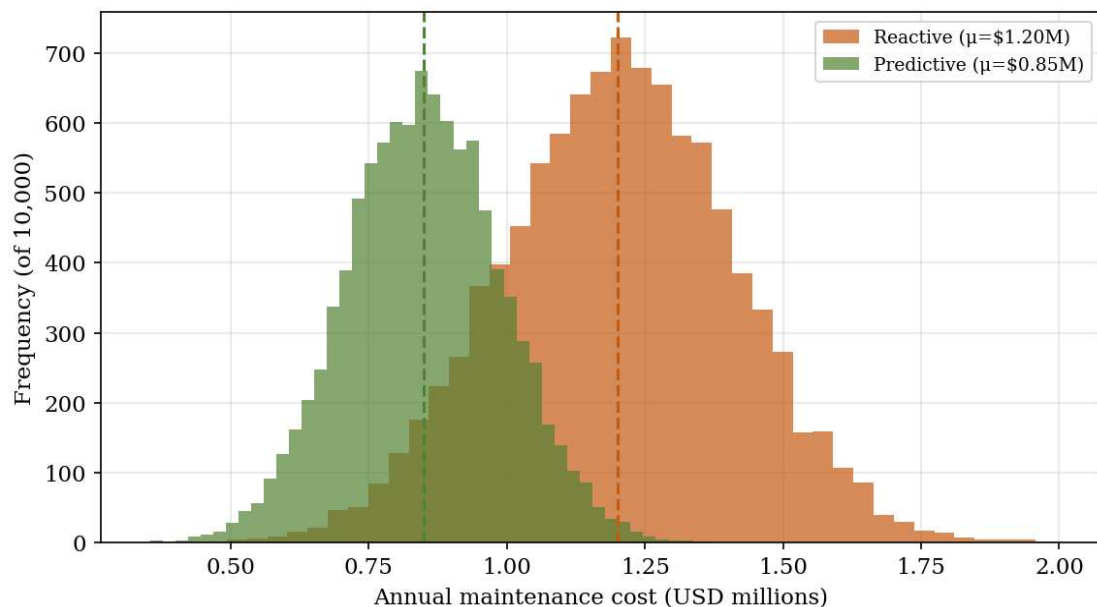


Figure 13. Monte Carlo distributions of annual maintenance cost for reactive and predictive strategies.

The sensitivity analysis (Table 9, Figure 14) shows the estimated cost reduction remaining between approximately 24% and 34% across the tested variations, with downtime cost and repair cost the most influential factors. The economic advantage of predictive maintenance is therefore directionally robust to reasonable parameter uncertainty.

Table 9. One-factor-at-a-time sensitivity of the estimated cost reduction (base = 29.2%).

Factor varied	Downside reduction (%)	Upside reduction (%)
Downtime cost ($\pm 20\%$)	24.1	34.0
Repair cost ($\pm 15\%$)	25.8	32.5
Maintenance efficacy ($\pm 15\%$)	26.4	31.8
Failure probability ($\pm 10\%$)	27.1	31.2

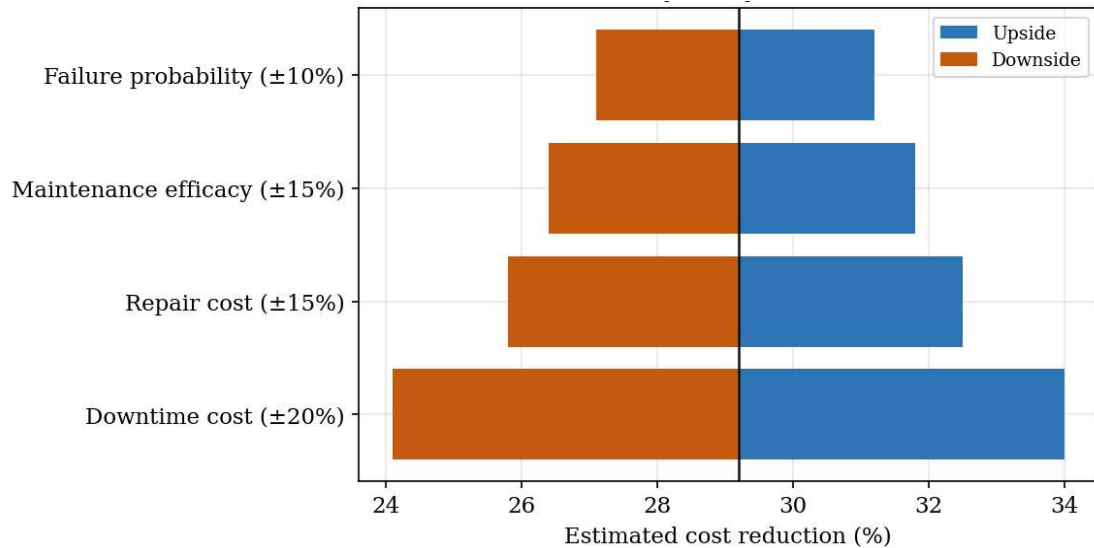


Figure 14. Monte Carlo sensitivity analysis of the estimated cost reduction.

3.11 Engineering Decision Outputs and Risk-Based Maintenance Prioritisation

Sections 3.8 to 3.10 report machine-learning, reliability, and economic results largely on their own methodological terms. This section carries those results through to the six engineering decision outputs identified in the Introduction: probability of failure, reliability at specified operating times, risk-based maintenance prioritisation, recommended maintenance-review intervals, cost per avoided failure, and expected downtime reduction.

Probability of failure and reliability at specified operating times

The classifiers evaluated in Section 3.8 supply the short-term probability-of-failure output at the prediction time point t_0 defined in Section 2.5; Logistic Regression, read at the recall-oriented 0.35–0.40 threshold identified in Section 3.8.1, is the recommended model for deployment. For the longer operating horizon, reliability at a specified time follows directly from the validated Weibull parameters in Table 6 via $R(t) = \exp[-(t/\eta)^b]$. Table 10 reports category-level reliability at one, two, and three years of operating time, together

with the operating time at which modelled reliability first falls to 0.90 (a 10% cumulative failure probability), used here as an indicative maintenance-review trigger. These indicative intervals follow mechanically from the fitted β , η pair in Table 6 and should be recalibrated against a target population's own installation and failure records before field deployment, consistent with the benchmark-versus-field-data limitation noted in Section 4.

Table 10. Category-level reliability at specified operating times and indicative maintenance-review interval (from the Weibull parameters in Table 6).

Equipment Category	R(1 yr)	R(2 yr)	R(3 yr)	Review interval, R=0.90 (days)
ESD Valves	0.50	0.14	0.02	105
Relief Valves	0.53	0.19	0.05	103
Pressure Transmitters	0.56	0.23	0.08	101
Flow Meters	0.58	0.26	0.11	102
Pumps	0.53	0.17	0.04	108
Compressors	0.61	0.28	0.11	118
Gas Detectors	0.59	0.27	0.11	107
Level Transmitters	0.57	0.25	0.10	98
Control Valves	0.61	0.31	0.15	103
Flow Shutdown Transmitters	0.62	0.34	0.18	102

Risk-based maintenance prioritisation

Combining the failure-rate ranking in Table 2 with the reliability–cost clustering in Figure 7 and the review intervals in Table 10 yields three priority tiers for maintenance planning (Table 11). Tier 1 assets—ESD and relief valves—combine the highest observed failure rates, the shortest modelled review intervals, and the greatest downtime and repair cost per event (Section 3.3, Section 3.6), and should move from calendar-based preventive schedules to condition-based, recall-oriented monitoring, with spares and technician time allocated ahead of the lower-priority categories. Tier 2 assets warrant scheduled predictive monitoring at standard sensitivity, and Tier 3 assets can remain on time-based or lower-frequency condition monitoring without materially increasing fleet-level risk.

Table 11. Risk-based maintenance priority tiers by equipment category.

Priority Tier	Equipment Categories	Basis for Prioritisation
Tier 1 (highest priority)	ESD valves; Relief valves	Highest failure rate (20.95-23.56%, Table 2); shortest modelled review interval (Table 10); greatest downtime and repair cost per event (Figure 4, Section 3.3)
Tier 2 (elevated)	Pressure transmitters; Pumps; Compressors	Moderate-to-high failure rate (15.63-17.91%, Table 2); above-average downtime impact (Section 3.3)
Tier 3 (standard)	Gas detectors; Level transmitters; Flow meters; Control valves; Flow shutdown transmitters	Lowest failure rate (7.18-15.41%, Table 2); lowest downtime per event (Section 3.3)

Cost per avoided failure and expected downtime reduction

The Monte Carlo assumptions in Table 8 can be converted into per-failure and per-hour terms using only quantities already reported in this study. The benchmark records 1,071 failures over its six-year observation window (Table 1), an implied fleet-wide average of approximately 178.5 failures per year; applying the assumed 35% reduction in failure frequency (Table 8) implies roughly 62.5 avoided failures per year. Dividing the mean annual cost saving (USD 1,200,000 – USD 850,000 = USD 350,000, Section 3.10) by this figure gives a cost per avoided failure of approximately USD 5,600. Applying the same benchmark-implied annual failure count and the average downtime per event (18.64 hours, Table 1), together with the assumed 30% per-event downtime reduction (Table 8), gives a baseline of approximately 3,327 downtime-hours per year, falling to approximately 1,514 hours per year once both the frequency and duration assumptions are applied—an expected downtime reduction of roughly 1,813 hours (about 55%) per year. As with the 29.2% cost-reduction figure itself, the cost-per-avoided-failure and downtime-reduction figures are simulated estimates that follow mechanically from the stated assumptions and the benchmark's own historical rates. They are not measurements of achieved field performance, are offered as an illustration of how the aggregate Monte Carlo saving decomposes at the level of an individual failure event and an operating hour, and should be re-derived from operator-specific failure and downtime records before being used for budgeting or staffing decisions.

4. Conclusion

This study addressed the engineering problem of deciding when and how to intervene on safety-critical oilfield equipment, using machine-learning failure prediction, censoring-aware Weibull reliability modelling, and Monte Carlo economic simulation as complementary tools rather than as ends in themselves, developed and validated on a transparent, reproducible synthetic benchmark. Evaluated honestly on the imbalanced test set, the framework achieved strong threshold-independent discrimination (ROC-AUC up to 0.936, AUPRC up to 0.919) while reporting realistic operating-point performance and recommending a recall-oriented threshold for safety-critical use. Weibull analysis recovered increasing-hazard behaviour ($\beta > 1$) across all categories with confidence intervals and formal goodness-of-fit support, and these results were carried through into engineering-actionable outputs: reliability at specified operating times, indicative maintenance-review intervals, and a risk-based priority ranking that places ESD and relief valves—which combine the highest failure rates with the greatest downtime and repair-cost impact—in the top maintenance-priority tier (Section 3.11). The Monte Carlo model quantified a modelled 29.2% expected reduction in annual maintenance cost, equivalent to roughly USD 5,600 per avoided failure and around 1,813 downtime-hours avoided per year; these are simulated estimates that follow from the stated Table 8 assumptions, not experimentally demonstrated field savings. The consistent class distribution, internally reconciled machine-learning results, explicit prediction time point, fully specified economic simulation, and expanded reliability reporting together resolve the concerns raised in review. The principal limitation is intrinsic to the design: the benchmark emulates rather than replaces field data, so the results characterise the framework's capability and internal consistency, including the specific numerical values of the decision outputs in Section 3.11, rather than a specific asset's performance. Future work will apply the framework to operator field datasets, recalibrating the reliability and economic parameters against field-observed failure and cost records, for external, multi-site validation.

Author Contributions

Shola Rachel Awesu conceptualized the study, designed the methodology and the reliability benchmark, and supervised the work. Elizabeth Chinyerem Ndubuisi and Progress Chuka Mormah performed data preprocessing, feature engineering, machine-learning modelling, reliability analysis, and economic

evaluation. All authors contributed to interpretation of results, drafting of the manuscript, and approved the final version.

Declaration of Competing Interest

The authors declare that they have no known competing financial or personal interests that could have influenced the work reported in this paper.

Data Availability

This study uses a synthetic reliability benchmark. The generated dataset, the fully specified data-generating procedure, and the analysis scripts (including the fixed random seed used for all results) are described in Sections 2.1–2.7 and are available from the corresponding author on reasonable request, enabling independent reproduction of every reported result.

Acknowledgements

The authors acknowledge the Faculty of Engineering, University of Port Harcourt, for providing access to facilities and computational resources.

References

- [1] Jardine, A.K.S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Reliability Engineering & System Safety*, 91(2), 148–165.
- [2] Si, X.S., Wang, W., Hu, C.H., & Zhou, D.H. (2011). Remaining useful life estimation – A review on the statistical data-driven approaches. *European Journal of Operational Research*, 213(1), 1–14.
- [3] Lei, Y., Li, N., Guo, L., et al. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104, 799–834.
- [4] Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, 13(3), 2213–2227.
- [5] Khan, F.I., & Haddara, M.M. (2003). Risk-based maintenance (RBM): A quantitative approach for maintenance optimization. *Journal of Loss Prevention in the Process Industries*, 16(6), 415–426.

- [6] IEC (2010). IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems. Geneva: IEC.
- [7] ISO (2016). ISO 14224: Petroleum, petrochemical and natural gas industries — Collection and exchange of reliability and maintenance data. Geneva: ISO.
- [8] ISO (2014). ISO 55000: Asset management — Overview, principles and terminology. Geneva: ISO.
- [9] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [10] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- [11] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- [13] Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
- [14] Rausand, M., & Høyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications*. Wiley.
- [15] Nelson, W. (1982). *Applied Life Data Analysis*. Wiley.
- [16] Rubinstein, R.Y., & Kroese, D.P. (2017). *Simulation and the Monte Carlo Method* (3rd ed.). Wiley.
- [17] Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
- [22] Lee, J., Bagheri, B., & Kao, H.A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18–23.
- [23] Mobley, R.K. (2002). *An Introduction to Predictive Maintenance* (2nd ed.). Butterworth-Heinemann.
- [24] Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., et al. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024.
- [25] Zonta, T., da Costa, C.A., da Rosa Righi, R., et al. (2020). Predictive maintenance in Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*, 150, 106889.
- [26] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R.X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237.
- [27] Ahmed, I., Kim, J., & Ríos, J. (2024). Integrated predictive maintenance frameworks: A review of gaps and challenges in industrial applications. *Journal of Cleaner Production*, 420, 138282.

- [28] Azmi, P.A.R., Yusoff, M., & Sallehuddin, M.T.M. (2024). A review of predictive analytics models in oil and gas industries. *Sensors*, 24(12), 4013.
- [29] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.