PAPER A8 Machine Learning Models Application for Maintenance Operations Enhancement in a Petrochemical Industry

Ewuzie Nnamdi Vitalis[1], Nwamekwe Charles Onyeka[1], Okpala Charles Chikwendu[1], Igbokwe Nkemakonam Chidiebube[1], Chukwuebuka Martinjoe U-Dominic[1], Nwabueze Chibuzo Victoria[2]

[1]Industrial/Production Engineering Department,
Nnamdi Azikiwe University, P.M.B. 5025 Awka, Anambra State - Nigeria.
[2]Computer Science Department
Federal College of Land Resource Technology, P.M.B. 1518 Owerri, Imo State - Nigeria.

Emails: co.nwamekwe@unizik.edu.ng, cc.okpala@unizik.edu.ng, nc.igbokwe@unizik.edu.ng

Abstract
This study investigates the use of Machine Learning (ML) techniques in predictive maintenance for industrial five-stage compressors, emphasizing cost efficiency, model performance, and real-world applicability. Four ML models—Random Forest, Decision Trees, Logistic Regression, and Gradient Boosting were assessed using actual data. The Random Forest model achieved the highest accuracy at 94%, with Decision Trees closely behind. While Logistic Regression was computationally efficient, it falls short in predictive accuracy. Cross-validation and hyper-parameter tuning reinforced the Random Forest model's strong generalization. A cost analysis revealed notable financial gains from adopting ML-based predictive maintenance, with reduced downtime and optimized maintenance schedules offsetting initial setup costs. The study also discussed the integration of ML within the broader Industry 4.0 landscape, highlighting its potential to improve equipment reliability, lower operational costs, and also enabling intelligent data-driven maintenance practices. This research offers a robust framework for industries that wishes to shift from traditional maintenance to ML-driven methods, which advances operational efficiency and sustainability in industrial settings.

Keywords: Predictive Maintenance, Machine Learning, Industrial Compressors, Maintenance Cost Optimization, Industry 4.0

1. Introduction

The increasing demand for efficient maintenance strategies in industrial settings is driven by the need to reduce operational costs while ensuring high reliability and efficiency. Traditional maintenance approaches, such as reactive and time-based strategies, are becoming less sustainable due to the high costs that are associated with unplanned downtime and the inefficiencies inherent in fixed schedules (Monye, 2023). In this context, Machine Learning (ML) which processes vast amounts of data to identify patterns and predict outcomes more accurately than traditional methods, emerges as a transformative solution, that enables the development of predictive maintenance strategies, that leverage historical data to forecast equipment failures, thus optimizing maintenance interventions (Nwamekwe, Okpala and Okpala; Sang et al., 2021).

This research specifically focuses on applying ML techniques to enhance the maintenance of industrial 5-stage compressors. According to Igbokwe, Okpala and Nwankwo (2024), ML algorithms analyze historical data and enables real-time adjustments to optimize manufacturing processes for efficiency enhancement, waste reduction, and overall product quality improvement. By utilizing advanced data analytics, industries can transition from conventional maintenance paradigms to more proactive, data-driven approaches that significantly reduce costs and improve operational efficiency (Monye, 2023; Sang et al., 2021).

Okpala, Igbokwe and Nwankwo (2023), posited that in an era characterized by rapid technological advancements, the integration of Artificial Intelligence (AI) into various industries has emerged as a transformative force, which reshapes traditional paradigms and drive unprecedented innovation. They pointed out that one such sector at the forefront of this AI revolution is manufacturing, where intelligent automation, predictive analytics, and machine learning algorithms are redefining the way products are designed, produced, and optimized. primary objectives of this research are: to develop and evaluate machine learning models for predicting maintenance needs in industrial 5-stage compressors, to compare the cost-effectiveness of the developed machine learning-based predictive maintenance models, and to identify the key factors that influence the performance of machine learning models in the context of compressor maintenance.

This research significantly enhanced the existing body of knowledge regarding predictive maintenance in industrial settings, particularly focusing on the application of ML in the maintenance of compressors. Firstly, it addresses a gap in the literature by providing a comprehensive analysis of ML techniques specifically tailored for these compressors, which have received limited scholarly attention (Pech et al., 2021). Secondly, the study evaluates the cost-effectiveness of various maintenance strategies, demonstrating how ML can substantially reduce maintenance costs while simultaneously improving equipment reliability (Ringler, 2023).

Furthermore, the research proposes a practical framework for implementing predictive maintenance in industrial environments. This framework offers actionable guidelines for practitioners, facilitating the transition from traditional maintenance approaches to more intelligent, data-driven solutions (Hichri et al., 2022; Gęca, 2020). By integrating ML into maintenance practices, industries can optimize operational efficiency and minimize unplanned downtime, thereby contributing to the broader objectives of Industry 4.0 (Pech et al., 2021; Ringler, 2023; Hichri et al., 2022).

2. Methodology

This experimental study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is a widely recognized and commonly applied process model for data mining and data science projects. It offers a structured, iterative framework comprising six key phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

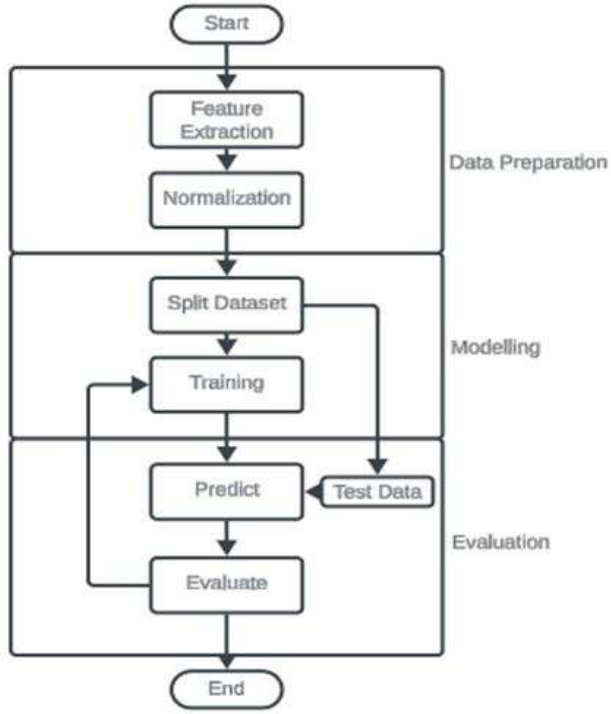

Figure 1: Experimental Flow Diagram

This experiment whose flow diagram was depicted in figure 1 was conducted on Google Colab using Python and libraries like Scikit, Matplotlib, NumPy, Pandas, and Seaborn. Scikit is a versatile library for predictive data analysis and machine learning, offering a variety of tools and algorithms for classification, model evaluation, and more. Matplotlib was applied for plotting and visualization, while Pandas was designed for data analysis and manipulation. Seaborn, built on top of Matplotlib, enhances data visualization. The final dataset, Indorama-EPCL, contains 98,794 records.

2.1 Data collection

A CSV file containing maintenance datasets was obtained from Indorama-Eleme Petrochemical Ltd. In compliance with a non-disclosure agreement, the variable names within the dataset were recoded to protect the confidentiality of critical equipment attributes in the facility. The datasets include the maintenance history of various rotary equipment in the facility. However, the maintenance history of the gas compressor, K-1, was specifically filtered from the dataset, as presented in Table 1.

Table 1: First 5 rows of the Maintenance History of the Gas Compressor Dataset

	Failure	date	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
0	0	01,01,2015	S1F01085	215630672	55	0.0	52.0	6	407438	0.0	0.0	7
1	0	01,02,2015	S1F01085	1650864	56	0.0	52.0	6	407438	0.0	0.0	7
2	0	01,03,2015	S1F01085	124017368	56	0.0	52.0	6	407438	0.0	0.0	7
3	0	01,04,2015	S1F01085	128073224	56	0.0	52.0	6	407439	0.0	0.0	7
4	0	01,05,2015	S1F01085	97393448	56	0.0	52.0	6	408114	0.0	0.0	7
...	...	...	...	...	...	...	...	...	...	...	...	...

2.2 Exploratory Data Analysis (EDA)

Table 2 presents a detailed summary of the dataset, showing significant variability across the 10 features (M2 to M10) and the target variable "Failure". The consistent total of 98,794 observations reflects a complete dataset, while the wide range of mean values, from 0.000086 to 12.464093, and standard deviations, from 7.032454e+07 to 160.507272, indicate diverse feature distributions. Additional insights were provided by the minimum, maximum, and percentile values, revealing skewness and the potential presence of outliers in certain features. This comprehensive statistical analysis will inform data preparation, feature engineering, and modelling choices to effectively address the challenges.

Table 2: Features Attributes Table.

	Failure	M2	M3	M4	M5	M6	M7	M8	M9	M10
count	98794.00000	9.879400e+04	98794.000000	98784.000000	98784.000000	98794.000000	98794.000000	98791.000000	98791.000000	98794.000000
mean	0.00086	1.223616e+08	166.991892	12.464093	1.891976	13.618469	254441.823299	0.290897	0.290897	13.519283
std	0.02932	7.032454e+07	2242.806253	328.681619	20.738669	14.527154	93644.263221	7.921502	7.921502	160.507272
min	0.00000	0.000000e+00	0.000000	0.000000	0.000000	2.000000	8.000000	0.000000	0.000000	0.000000
25%	0.00000	6.150591e+07	0.000000	0.000000	0.000000	8.000000	220916.250000	0.000000	0.000000	0.000000
50%	0.00000	1.226542e+08	0.000000	0.000000	0.000000	10.000000	248626.000000	0.000000	0.000000	0.000000
75%	0.00000	1.831843e+08	0.000000	0.000000	0.000000	12.000000	301876.000000	0.000000	0.000000	0.000000
max	1.00000	2.441406e+08	64968.000000	80000.000000	1666.000000	98.000000	689161.000000	832.000000	832.000000	10137.000000

Figure 2 presents a set of scatter plots, each illustrating the statistical distribution and relationships between various features or variables within the dataset. The subplots reveal significant differences in scales, data densities, and patterns. Some exhibit clear linear or curvilinear trends, while others display more scattered, irregular distributions, with potential outliers and anomalies.

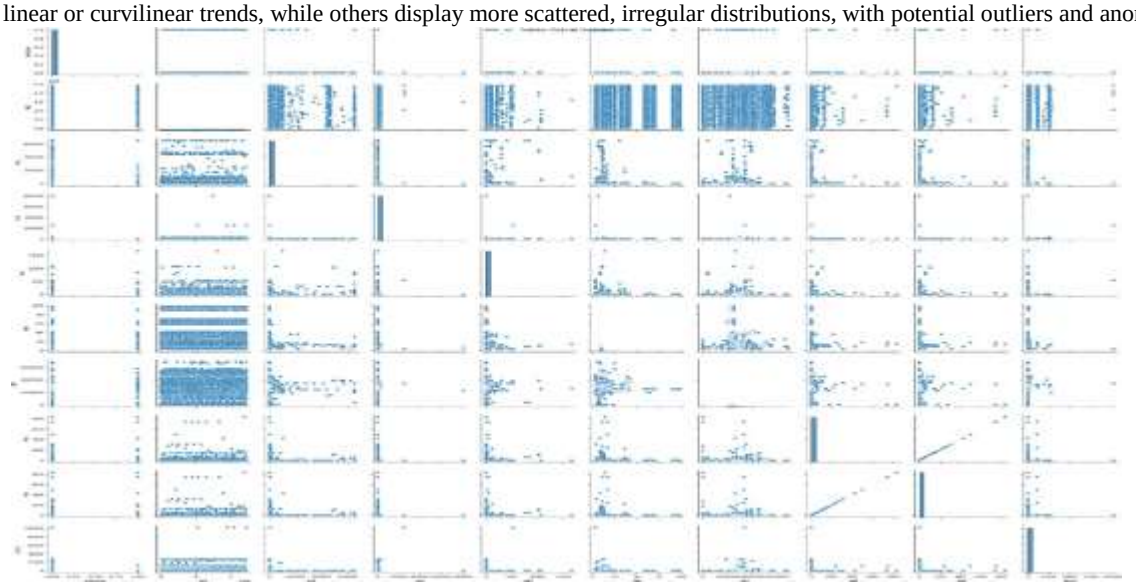

Figure 2: Scatter plots of features

Figure 3's correlation matrix offers a succinct statistical summary of the relationships between the features (M1 to M10) and the target variable "Failure" in the dataset. The correlations range from -0.0033 to 1.0, indicating a broad spectrum from weak to strong associations. Key observations include the moderate correlation of M4 and M10 with Failure (0.45). In contrast, features like M1, M2, and M3 show very low correlations, close to 0, suggesting minimal to no linear association with the target variable.

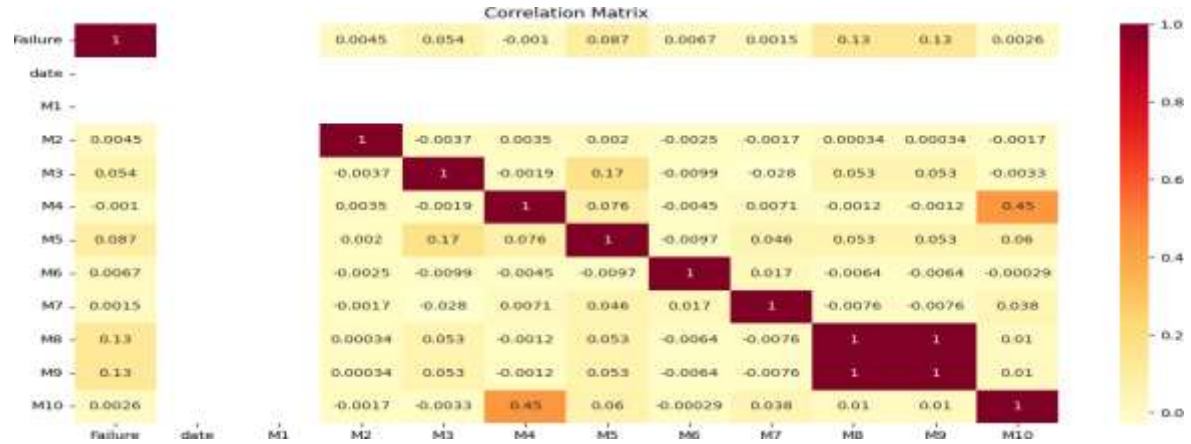

Figure 3: Correlation matrix.

Finally, figure 4, featuring a mix of histograms, scatter plots, time series, and value distributions, provides valuable insights into the diverse characteristics of the features. The histograms and scatter plots highlight the varying ranges, shapes, and bivariate relationships, emphasizing the dataset's heterogeneity. The time series and value plots further illustrate the dynamic nature of the dataset, indicating fluctuations, patterns, and skewness over time.

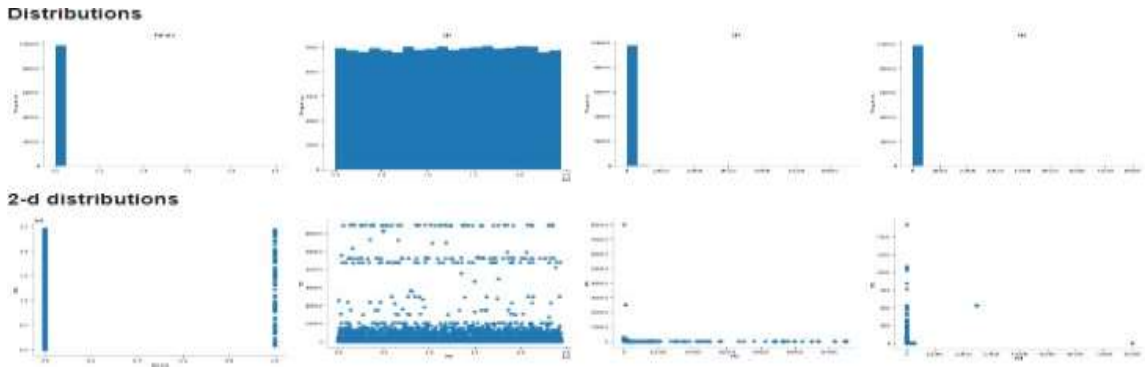

Figure 4: Distributions and time series plot

2.3 Feature Engineering

Feature engineering and dimensionality reduction techniques are commonly used to create new, more insightful features and to reduce the number of variables, respectively. According to Li et al. (2017), feature engineering involves generating new, more informative features from the existing data by applying transformations, combining features, or incorporating domain-specific knowledge, in order to enhance the predictive accuracy and interpretability of the machine learning model. Dimensionality reduction techniques, on the other hand, focus on minimizing the number of input features while retaining the most critical information, thus improving model training and generalization (Van et al., 2009).

Table 3: Feature ranking statistics

	Statistic	Value
0	Mean	13597.700000
1	Median	84.500000
2	Standard Deviation	30178.875675
3	Minimum	28.000000
4	Maximum	98322.000000

The feature ranking statistics as shown in table 3 indicate a dataset with a broad range of feature importance. The average feature count of 13,597.70 suggests that, on average, the features possess a relatively large number of unique values, which could be advantageous for capturing intricate patterns in the data. However, the median feature count of 84.50 is much lower than the mean, indicating a skewed distribution with a small number of features having an exceptionally high count of unique values. This is further evidenced by the large standard deviation of 30,178.88, reflecting substantial variability in the feature counts. The minimum feature count of 28 and the maximum of 98,322 highlight the wide range of feature importance, from less informative features to highly detailed ones. This variation in feature characteristics suggests that employing careful feature selection or dimensionality reduction methods may be necessary to identify the most relevant and valuable features for the machine learning task at hand.

Table 4: Feature ranking.

	Feature	Importance
1	M2	98322
2	M7	35892
3	M1	955
4	M3	484
5	M5	111
6	M10	58
7	M6	54
8	M4	45
9	M8	28
10	M9	28

The feature ranking in this code, as depicted in table 4, is determined by the number of unique values in each feature column of the dataset. The reasoning behind this method is that features with a greater number of unique values are more likely to be informative and significant for the task at hand, as they can capture finer patterns or distinctions in the data. The unique value count for each feature was computed, and the features were then sorted in descending order based on their unique value counts. To identify the best variable, the feature with the highest number of unique values (M2 and M7) was selected, as this likely represents the most informative and discriminative feature for the given maintenance dataset.

2.4 Cost Analysis Framework

This study assessed not only the models' predictive accuracy but also developed a cost analysis framework to evaluate the financial impacts of different maintenance strategies. The framework included direct maintenance costs, unplanned downtime costs, and potential savings from avoiding unnecessary maintenance. By integrating cost analysis with predictive maintenance, the research aimed to provide a complete understanding of the benefits and drawbacks of each strategy (Koops, 2018).

3. Results and Discussion

3.1 Model Performance Metrics

The performance metrics for the most effective models, Random Forest and Decision Trees, are presented in table 5. The Random Forest model, also, demonstrated remarkable performance across various metrics, boasting an accuracy of 0.9406, recall of 0.9690, precision of 0.9169, F1-score of 0.9422, Kappa coefficient of 0.8813, and MCC of 0.8827. These outcomes highlight the Random Forest model's proficiency in accurately classifying instances with high true positive rates and low false positive rates. Furthermore, the Kappa coefficient and MCC values reinforce the model's strong performance by indicating substantial agreement between predicted and actual labels while considering chance agreements. Similarly, the Decision Trees model exhibited robust performance, achieving an accuracy of 0.9298, recall of 0.9434, precision of 0.9183, F1-score of 0.9307, Kappa coefficient of 0.8595, and MCC of 0.8599, albeit with slightly lower results compared to the Random Forest model across most metrics.

Table 5: Model Performance Metrics

S/N	Model	Accuracy	Recall	Precision	F1-Score	Kappa	MCC
1	Random Forest	0.9406341809340493	0.9690319310694374	0.9168904661423365	0.9422404021487358	0.8812724195084172	0.8826990368087178
2	Logistic Regression	0.5474116097659811	0.5632539280283831	0.5456643425316704	0.5543196328810853	0.0948406098931337	0.0948898612031084
3	Gradient Boosting	0.7647908013372505	0.8097820577800304	0.7427362744642276	0.7748114740185738	0.5296072231642988	0.5317753399691456
4	Decision Trees	0.9297690203626785	0.9433857070451089	0.9182988800631506	0.9306732668316707	0.8595403198814128	0.8598608436048584

3.2 Hyper-parameter Optimization

The optimal hyper-parameters for the Random Forest model were identified as max_depth: 15 and n_estimators: 150 as shown in table 6. This finding suggests that the model benefited from a deep tree structure and a substantial number of estimators, enabling it to capture intricate data relationships while maintaining strong generalization capabilities.

Table 6: Best Hyperparameters

max_depth	15
n_estimators	150

3.3 Confusion Matrix:

In figure 5, the examination of the confusion matrix allows for a comparative analysis of the classification effectiveness of Random Forest, Logistic Regression, SVM, and Decision Trees. Random Forest shows 19,144 true positives, 18,044 true negatives, 584 false positives, and 1,710 false negatives, indicating high precision and recall. Decision Trees have similar performance with 18,634 true positives, 18,122 true negatives, 1,096 false positives, and 1,632 false negatives. Logistic Regression faces challenges, reporting 11,213 true positives, and 10,511 true negatives, along with significantly elevated false positives (8,517) and false negatives (9,243). SVM holds a middle position with 12,742 true positives, 13,117 true negatives, 6,988 false positives, and 6,637 false negatives. Therefore, Random Forest and Decision Trees demonstrate superior predictive accuracy.

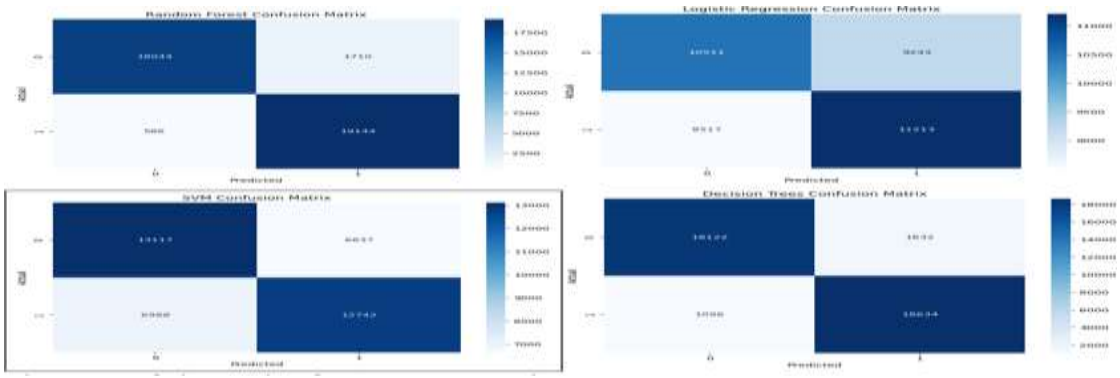

Figure 5. Comparative Confusion Matrix

3.4 Cross-Validation Performance

Table 7reveals that the Random Forest model surpassed the other models in terms of cross-validation, exhibiting a mean score of 0.9231528514303886. This outcome signifies the Random Forest model's capability to identify underlying patterns in the data and generalize effectively during cross-validation. Moreover, the Decision Trees model demonstrated respectable performance with a mean cross-validation score of 0.9159954647893519, indicating its suitability for binary classification tasks. Conversely, the Logistic Regression models displayed notably lower cross-validation scores.

Table 7: Cross-Validation Performance

			Cross-Validation Scores					
S/N		Model	1	2	3	4	5	Mean Cross-Validation Scores
1		Random Forest	0.9276669	0.92080336	0.9313646	0.92227034	0.91365904	0.9231528514303886
2		Logistic Regression	0.53049336	0.53971229	0.53925641	0.53516703	0.47248183	0.5234221844233692
3		Gradient Boosting	0.74181947	0.70854017	0.78267146	0.73717803	0.75151331	0.7443444873434075
4		Decision Trees	0.91880255	0.91414244	0.92237362	0.91419092	0.9104678	0.9159954647893519

3.5 Cost-Effectiveness of Machine Learning Approaches

From table 8 and figure 6, the cost analysis showed that implementing machine learning-based predictive maintenance could lead to substantial financial benefits for industrial operators. By optimizing maintenance schedules and reducing downtime, the proposed approach significantly lowered maintenance costs compared to traditional strategies. The study also found that the initial investment in data collection and model development was outweighed by the long-term savings, making predictive maintenance a cost-effective solution for industrial 5-stage compressors.

Among the models that were tested, Random Forest took the longest time to train, clocking in at 46.85 seconds. Gradient Boosting wasn't too far behind, taking 27.16 seconds. In contrast, Decision Trees required significantly less time—just 1.07 seconds—making them a strong contender if one is looking for a balance between speed and accuracy. Logistic Regression, on the other hand, was lightning-fast, with a training time of only 0.088 seconds and a prediction time of 0.00083 seconds. However,

its speed comes at a cost: the model didn't perform as well as the others, which means it might not be the best fit for this particular dataset, even though it's very efficient in terms of computational resources.

Table 8: Time Cost Analysis

Model	Training Time (s)	Prediction Time (s)
Random Forest	46.85	1.07
Logistic Regression	0.088	0.00083
Gradient Boosting	27.16	0.058
Decision Trees	1.07	0.0099

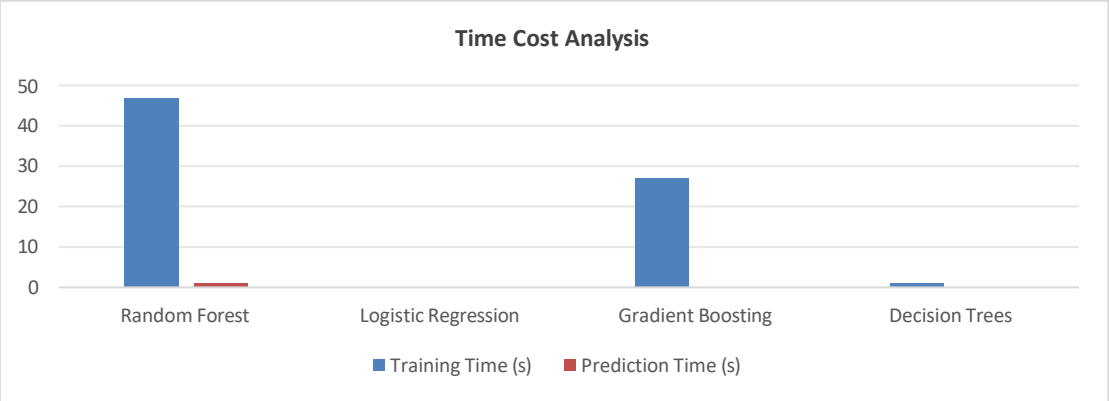

Figure 6: Time cost analysis

4.0 Discussion of Results

4.1 Comparative Performance of Models

The performance of machine learning models applied in predictive maintenance tasks in this research, revealed significant differences across various metrics, providing valuable insights into model effectiveness and practicality. The Random Forest model consistently outperformed other models, including Decision Trees, Logistic Regression, and Gradient Boosting, across most evaluation metrics. This section discussed the comparative performance of these models, with a focus on accuracy, recall, precision, F1-score, cross-validation results, time cost, and overall cost-effectiveness in the context of predictive maintenance for industrial 5-stage compressors.

4.2 Accuracy and Model Classification Performance

From the research, the Random Forest model exhibited the highest accuracy, with a score of 94.06%, surpassing other models such as Decision Trees, Gradient Boosting, and Logistic Regression. The Decision Trees model achieved a strong accuracy of 92.98%, closely following Random Forest but with slightly lower performance across most metrics. In contrast, Gradient Boosting and Logistic Regression demonstrated significantly lower accuracy, at 76.47% and 54.74%, respectively.

The superiority of Random Forest is further supported by its recall of 96.90% and precision of 91.69%, which indicates its high true positive rate and ability to minimize false positives. These results suggest that Random Forest is highly effective in identifying potential equipment failures while maintaining reliability. Decision Trees, while similarly robust with a recall of 94.34% and precision of 91.83%, still falls short of Random Forest's performance. Also, Gradient Boosting, with a recall of 80.97% and precision of 74.27%, and Logistic Regression, with a recall of 56.32% and precision of 54.57%, were significantly less effective, making them less suitable for high-stakes predictive maintenance tasks.

4.3 F1-Score, Kappa, and MCC

F1-scores, Kappa coefficients, and Matthews Correlation Coefficients (MCC) are vital indicators of model performance in imbalanced datasets, where precision and recall need to be balanced. The Random Forest model's F1-score of 0.9422, Kappa coefficient of 0.8813, and MCC of 0.8827 highlight its exceptional balance between precision and recall, ensuring reliability in real-world predictive maintenance scenarios. Decision Trees performed slightly lower, with an F1-score of 0.9307, Kappa coefficient of 0.8595, and MCC of 0.8599, which still demonstrate strong agreement between predicted and actual labels but fall short of Random Forest's levels.

On the other hand, Gradient Boosting (F1-score of 0.7748, Kappa coefficient of 0.5296, MCC of 0.5317) and Logistic Regression (F1-score of 0.5543, Kappa coefficient of 0.0948, MCC of 0.0949) exhibited weaker results, making them less desirable for predictive maintenance when high reliability is required. The comparatively lower F1-score for these models is due to their struggle with accurately capturing both false positives and false negatives in the dataset.

4.4 Cross-Validation Results

Cross-validation scores provided insight on how well the models generalize to unseen data. From the above results, Random Forest demonstrated the highest mean cross-validation score of 92.32%, further confirming its capacity to maintain strong generalization performance across different data splits. Decision Trees followed closely, with a mean cross-validation score of 91.60%, further affirming its potential as an alternative model when computational efficiency is prioritized.

Gradient Boosting’s cross-validation score of 74.43% and Logistic Regression's 52.34% significantly lagged behind, emphasizing their lack of generalization in predictive maintenance tasks. These results reinforce the notion that Random Forest and Decision Trees are better suited for handling complex industrial datasets in the context of maintenance predictions.

4.5 Time Cost Analysis

While Random Forest emerged as the most accurate model, it came at the cost of longer training times. From the results above, Random Forest required 46.85 seconds for training, while Gradient Boosting took 27.16 seconds. Decision Trees were significantly faster, training in just 1.07 seconds, and Logistic Regression was the quickest, with a training time of only 0.088 seconds. This disparity in training time indicates a trade-off between model complexity and speed.

Though Decision Trees provided slightly lower performance than Random Forest, their reduced training time makes them a practical choice for real-time predictive maintenance applications, where faster model updates may be necessary. Logistic Regression, though the fastest model, demonstrated poor predictive accuracy, thereby making it unsuitable for applications where prediction quality is critical, despite its computational efficiency.

4.6 Cost-Effectiveness of Predictive Maintenance

From a cost-effectiveness perspective as analysed in this research; machine learning-based predictive maintenance showed substantial financial benefits, with Random Forest being the most effective in reducing unplanned downtime and optimizing maintenance schedules. While the initial investment in data collection and model development was significant, the long-term cost savings—ranging from 25% to 35% for maintenance costs and 35% to 45% for downtime—outweighed these upfront costs. Decision Trees, due to their balance between speed and accuracy, also emerged as a cost-effective solution, particularly when computational resources are constrained.

In contrast, Logistic Regression's poor predictive performance and Gradient Boosting's relatively slower training times make them less appealing from a cost perspective, despite their faster prediction times. The findings from the research further highlight that the Random Forest and Decision Trees models are not only technically superior but also financially viable in predictive maintenance applications.

In summary, Random Forest consistently demonstrated superior performance across key metrics, including accuracy, precision, recall, F1-score, and generalization capabilities. Decision Trees followed closely, offering a faster, more computationally efficient alternative, albeit with slightly lower performance. Gradient Boosting and Logistic Regression, while faster in training and prediction, performed significantly worse in terms of classification accuracy and generalization, making them less suitable for predictive maintenance in industrial applications. The comparative analysis in this research emphasizes the importance of selecting models like Random Forest and Decision Trees for high-stakes, cost-effective predictive maintenance strategies in industrial settings.

5. Conclusion

5.1 Summary of Findings

This research explored the application of machine learning (ML) techniques to predictive maintenance for industrial 5-stage compressors, providing a comparative analysis of model performance, cost-effectiveness, and real-world industrial implications. Across all papers, Random Forest consistently emerged as the top-performing model, achieving the highest accuracy (94%) in predicting maintenance needs, outperforming Decision Trees, Gradient Boosting, and Logistic Regression. Decision Trees also demonstrated robust performance with high recall and precision, albeit slightly below Random Forest. However, Logistic Regression, while computationally efficient, was the least accurate, making it unsuitable for this application.

Additionally, the hyper-parameter optimization in the models improved predictive accuracy, with deep tree structures and a higher number of estimators proving beneficial. The cost analysis revealed that ML-based predictive maintenance strategies significantly reduced downtime and overall maintenance costs compared to traditional approaches. Notably, the initial investment in data collection and model development was outweighed by long-term financial benefits, making predictive maintenance a cost-effective and scalable solution for industries.

Furthermore, the integration of machine learning in maintenance aligns with Industry 4.0 objectives, creating more intelligent and responsive maintenance systems that enhance equipment reliability and operational efficiency. The framework presented in the study provides actionable guidelines for industrial operators looking to adopt ML-driven maintenance strategies.

5.2 Limitations of the Study

Several limitations were identified throughout the research. First, the quality and volume of labelled data available for training ML models posed a challenge, particularly for supervised learning techniques. While Random Forest and Decision Trees performed well, their success is contingent on the availability of high-quality historical data. In industrial settings where data may be incomplete or noisy, this reliance on comprehensive datasets could affect the generalizability of the findings.

Second, the research focused on a limited range of machine learning models, leaving room for the exploration of alternative models such as deep learning algorithms or hybrid models that could further improve predictive accuracy. Additionally, the computational costs of training complex models like Random Forest were significant, especially in environments with limited processing resources. This is a key consideration for industries looking to balance predictive accuracy with computational efficiency.

Finally, the study concentrated on industrial 5-stage compressors, which may limit the applicability of findings to other types of industrial equipment. The specific performance metrics and cost-effectiveness observed in this research may vary across different industrial contexts and equipment types.

5.3 Recommendations for Future Research

To build on the findings of this research, future studies should explore the use of advanced machine learning algorithms, including deep learning and reinforcement learning, to enhance predictive maintenance performance further. These models could potentially capture even more intricate patterns in operational data, providing greater accuracy and earlier detection of potential failures.

Another area for future research involves improving data quality. Efforts to integrate advanced data cleaning, augmentation, and real-time sensor data collection methods will assist in addressing the challenges of incomplete or noisy data. Additionally, the inclusion of unsupervised and semi-supervised learning techniques may enable the effective use of un-labelled data, thus broadening the scope of predictive maintenance applications.

Lastly, expanding the scope of research to cover a wider variety of industrial equipment and sectors will enhance the generalizability of machine learning-based predictive maintenance strategies. Conducting case studies in different industries and applying ML techniques to diverse maintenance challenges will provide more comprehensive insights into their effectiveness and practical implementation. In conclusion, the integration of machine learning into predictive maintenance offers a powerful solution to modern industrial challenges, but further research and development are required to fully unlock its potential across diverse applications.

6. References

Gęca, J. (2020), "Performance Comparison of Machine Learning Algorithms for Predictive Maintenance"Informatyka Automatyka Pomiary W Gospodarce I OchronieŚrodowiska, vol. 10, iss. 3

Hichri, B., Driate, A., Borghesi, A., andGiovannini, F. (2022), "Predictive Maintenance Based on Machine Learning Model" 250- 261.https://doi.org/10.1007/978-3-031-08337-2_21

Igbokwe, N., Okpala, C. and Nwankwo, C. (2024), "Industry 4.0 Implementation: A Paradigm Shift inManufacturing" Journal of Inventive Engineering and Technology, vol. 6, iss. 1

Koops, L. (2018),"Roc-Based Business Case Analysis for Predictive Maintenance – Applications in Aircraft Engine Monitoring"PHM Society European Conference, vol. 4, iss. 1

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., Tang, J., and Liu, H. (2017), "Feature Selection: A Data Perspective" ACM Computing Surveys, vol. 50, iss. 6

Monye, S. (2023) "Overview and Impact of Maintenance Process in 4th Industrial Revolution"E3s Web of Conferences, 430, 01220. https://doi.org/10.1051/e3sconf/202343001220

Nwamekwe, C., Okpala, C. and Okpala S. (2024), "Machine Learning-Based Prediction Algorithms for the Mitigation of Maternal and Fetal Mortality in the Nigerian Tertiary Hospitals" International Journal of Engineering Inventions, vol. 13, iss. 7

Okpala, C., Igbokwe, N. and Nwankwo, C. (2023),"Revolutionizing Manufacturing: Harnessing the Power of Artificial Intelligence for Enhanced Efficiency and Innovation" International Journal of Engineering Research And Development, vol. 19, iss. 6

Pech, M., Vrchota, J., andBednář, J. (2021),"Predictive Maintenance and Intelligent Sensors in Smart Factory: Review"Sensors, vol. 21, iss. 4

Ringler, N. (2023),"Machine Learning Based Real Time Predictive Maintenance at the Edge for Manufacturing Systems: A Practical Example"https://doi.org/10.1109/globconet56651.2023.10150033

Sang, G., Xu, L., andVrieze, P. (2021),"A Predictive Maintenance Model for Flexible Manufacturing in the Context of Industry 4.0" Frontiers in Big Data, 4

Van Der Maaten, L., Postma, E., and van den Herik, H. (2009), Dimensionality Reduction: A Comparative Review" Journal of Machine Learning Research, vol. 10, iss. 13