Paper ID: 513

An Analysis of Machine Maintenance Prediction Using Machine Learning

Protik Barua[1], Md. Miraiz Hossain[2], Kazi Naimur Rahman[3], S. M. Fahim Faisal[4]
[1-4]Department of Mechatronics and Industrial Engineering, Chittagong University of Engineering & Technology, Bangladesh
E-mail: u1709018@student.cuet.ac.bd, miraizhossain@cuet.ac.bd, naimur@cuet.ac.bd, fahimfaisal@cuet.ac.bd

Abstract

Predictive maintenance is essential for optimizing machine performance, reducing downtime, and minimizing repair costs. This study presents an analysis of machine maintenance prediction using machine learning, with the objective of forecasting machine failures and scheduling maintenance effectively. Utilizing the Kaggle dataset "Machine Predictive Maintenance Classification," the process involved data collection, feature extraction, and preprocessing, which included the removal of missing data, application of oversampling for dataset balancing, and the use of KDE plots and correlation heatmaps to explore feature relationships. Dimensionality reduction was performed through Principal Component Analysis (PCA), and the dataset was divided into 90% training and 10% testing sets. Five machine learning algorithms XGBoost, Support Vector Machine (SVM), Random Forest, Linear Regression, and Gaussian Naive Bayes were implemented to construct predictive models. XGBoost outperformed other models, achieving 98.7% accuracy on the validation set and 98.3% on the test set. These results demonstrate the effectiveness of XGBoost in predicting machine failures and estimating maintenance timing, providing a robust machine learning framework to improve predictive maintenance strategies.

Keywords: Machine Learning, XGBoost, Random Forest, Linear Regression, Gaussian Naive Bayes, Predictive Maintenance, SVM

1. Introduction

The maintenance of industrial machinery is essential for ensuring operational efficiency, reliability, and safety across various sectors, including manufacturing, transportation, and healthcare. Maintenance strategies have been either reactive, where repairs are made after a failure occurs or preventive, where maintenance is scheduled at regular intervals. However, both approaches often result in unnecessary downtime, excessive costs, and resource inefficiencies. Predictive maintenance (PdM) offers a more effective solution by leveraging data-driven insights to anticipate equipment failures before they occur. This minimizes unplanned downtime, optimizes maintenance schedules, and reduces overall operational costs.

Machine learning plays a pivotal role in enabling predictive maintenance. By analyzing large volumes of historical and real-time data, ML algorithms can identify patterns, predict equipment failures, and offer actionable insights with high accuracy [1]. The ability of Machine Learning to automate and enhance predictive maintenance processes has revolutionized maintenance strategies. Despite, significant advancements in ML-based PdM, there is a notable gap in the literature specifically focusing on its application in automotive machinery systems. While both predictive maintenance and ML have been covered extensively in separate review papers, there is no comprehensive survey that explores their intersection for automotive systems. This gap is critical, given the growing complexity and reliance on automotive machinery, where optimizing performance and reducing downtime is paramount. Addressing this gap serves as the primary motivation for this thesis. This research contributes to two keyways. First, it will review and categorize existing studies on ML-based PdM for automotive systems, analyzing the use cases and machine learning techniques employed. Second, it will identify open challenges and future research directions. Helping to guide further advancements in the field. Additionally, the thesis will explore key components of predictive maintenance models, such as data collection, preprocessing, feature selection, and algorithm choice, while evaluating model performance, accuracy, and limitations. Finally, practical applications of the findings will be discussed, particularly how these predictive models can be implemented to enhance maintenance practices in the automotive industry.

2. Literature Review

In manufacturing sectors Predictive maintenance has become a central method for boosting operational efficiency and decreasing downtime. The introduction of machine learning has further transformed by enabling the study of machine data to estimate maintenance needs effectively. Mutsuddi et al. [2] offered a detailed evaluation of the state-of-the-art methodologies utilized for predictive maintenance in manufacturing businesses. Their work emphasized several methodologies, including anomaly identification, fault diagnostics, and time-series analysis. They also presented examples and case studies that illustrated both the advantages and limits of these strategies. The authors underlined that typical maintenance techniques, frequently focused on periodic inspections, may be expensive and inefficient. Susto et al. [3] introduced a multiple classifier for PdM, addressing the rising requirement to decrease downtime and related costs. They highlighted the problem of developing "health factors," which are quantitative indicators representing the state of systems connected to maintenance concerns. Their suggested technique includes numerous categorization modules that give distinct performance trade-offs, enabling dynamic decision-making for maintenance management. This technique illustratesed how high-dimensional and filtered data may be efficiently employed in a cost-based maintenance decision system. Zhang et al. [4] studied the notion of cleaner production as a strategy for manufacturing firms to achieve sustainable production. They suggested a big data-based analytics architecture for product lifecycle management, which merges big data analytics with service-driven patterns. This architecture enhances the accessibility and availability of data relevant to product lifecycles, hence increasing decision-making in production and maintenance operations. Qinming et al. [5] focused on improving maintenance procedures with a stochastic dynamic programming model that accounts for numerous failure modes and spare parts inventories. By applying a hidden semi-Markov model with an aging component, they identified multi-failure situations and include spare parts inventory costs into the maintenance model. Their paper offered a double-layer dynamic programming technique to attain low overall costs, illustrating its application through a case study. Ming-Yi et al. [6] presented a modularized architecture for predictive maintenance scheduling. This framework blends real-time prognostics information with conventional preventive maintenance models, exhibiting its use for multi-component systems. Their numerical example assesses the dynamic updating features of preventive maintenance schedules, showing the frameworks potential to support predictive maintenance plans across multiple sectors. Ameeth et al. [7] explored the effect of the Industrial Internet of Things (IIoT) in boosting predictive maintenance procedures. They applied Autoregressive Integrated Moving Average forecasting using time-series data acquired from sensors in a slitting machine. Their findings underlined the usefulness of machine learning in quality management and cost reduction within industrial operations. Kausha et al. [8] focused on automating machine monitoring systems using Machine Learning. They addressed the absence of accurate automation systems by integrating the Binary Decision Tree algorithm with CART technique to anticipate production machine accuracy. Their study emphasized the integration of energy meter characteristics for increased performance reporting. Archit et al. [9] explored the ramifications of PdM within the framework of Industry 4.0. They emphasized the constraints of traditional maintenance techniques and the promise of PdM to decrease downtime costs while maximizing equipment use. Their study stresses the relevance of data-driven strategies in forecasting equipment breakdowns and supporting sustainable practices in production. Chaitali et al. [10] suggested a system that uses IIoT technology for predictive maintenance. By monitoring machine health using various data collecting methods, they try to discover probable breakdowns before they occur. Their technique employs sensor data visualization tools and real-time computer forecasts to boost maintenance procedures in industrial environments. Chienyem et al. [11] explored into the practical application of predictive maintenance in Awka Metropolis, Nigeria. By utilizing decision trees and recurrent neural networks, they proved the usefulness of their predictive maintenance model in forecasting equipment breakdowns. Their findings suggested great potential for cost reductions and operational efficiency through proactive maintenance measures. Marina et al. [12] created a Random Forest-based system for electric motors, emphasizing its efficiency in minimizing costly failures through precise predictions based on sensor data. Charlene [13] did a desktop study that found gaps in previous research, stressing the necessity for hybrid techniques and IoT integration to improve maintenance schedules and boost data quality. In summary, the literature demonstrates a growing corpus of research focusing on the application of machine learning in predictive maintenance across numerous sectors. These studies together revealed the transformational influence of data-driven techniques on maintenance tactics, underlining the necessity for further research and integration of advanced analytics in industrial operations.

3. Methodology

In this work, machine learning technique is employed for machine failure prediction. "Fig. 1. Workflow of Machine Predictive Maintenance Using ML" depicts the whole workflow of machine predictive maintenance. Several procedures are taken into consideration for the goal of machine predictive maintenance such as dataset collecting, pre-processing, PCA implementation, dataset splitting, model selection, performance assessment. After the examination of performance matrices, the best model may be picked to propose for future development.

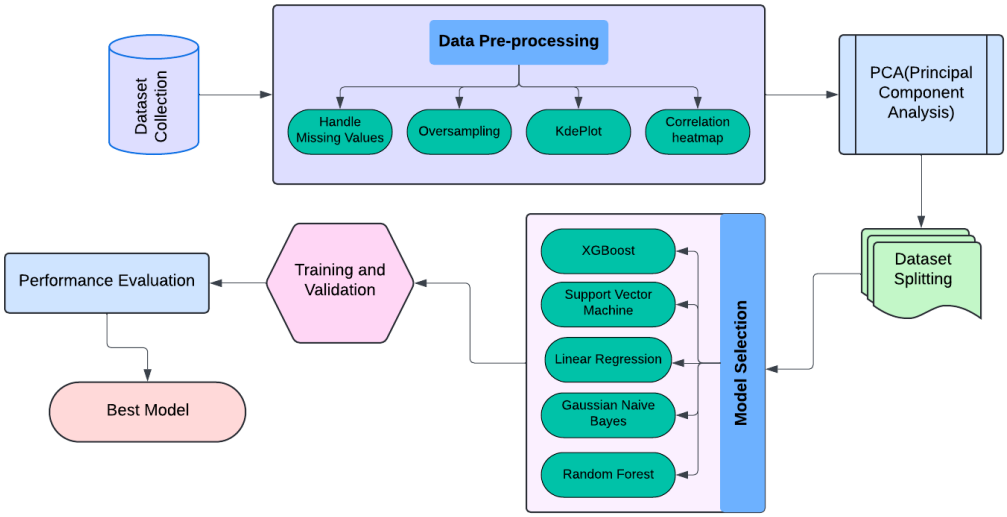


Fig. 1. Workflow of Machine Predictive Maintenance Using ML

Dataset Collection and Data Features

The dataset utilized in this study is the "Machine Predictive Maintenance Classification" from Kaggle. The dataset consists of 10,000 data points with 14 attributes. Key features include UID (a unique identifier), productID (indicating product quality—Low, Medium, or High), air temperature (generated with a random walk process around 300 K), process temperature (air temperature plus 10 K), rotational speed (calculated from power with noise), torque (normally distributed around 40 Nm), and tool wear (adjusted by product quality). The dataset comprises a 'machine failure' label, indicating if a failure occurred, and multiple failure types. There are two targets: 'Failure or Not' and 'Failure Type', with care made to avoid utilizing one target as a feature to prevent data leaking.

Data Preprocessing

In the dataset used for predictive maintenance, a significant portion of the data includes missing or incomplete information regarding the failure types. Specifically, some records contain the target value of 0, indicating a "random failure," while others contain the target value of 1, representing "no failure." These data points do not provide a clear or specific cause of failure, making them unsuitable for accurate machine learning predictions. The presence of such incomplete data can lead to lower model accuracy as the model cannot effectively learn from these ambiguous instances. To address this issue, approximately 27% of the dataset must be removed during data preprocessing with missing or unclear target values. This step ensures that the model is trained on high-quality data, which contains specific and relevant failure types, thereby improving the model's performance and accuracy in predicting machine failures. By eliminating this noisy data, the predictive model can focus on learning from meaningful patterns and relationships in the dataset, leading to more reliable and precise predictions.

Oversampling

In the dataset utilized for this study, the distribution of machine failure causes is uneven, which might severely influence the performance of machine learning models by biasing them toward the majority class. To overcome this, oversampling is performed, a strategy that entails reproducing instances of minority classes and perhaps lowering instances of majority classes to attain balance. The original dataset shows 29% power failure, 34% heat dissipation failure, 14% tool wear failure, and 24% overstrain failure. Through oversampling, each failure type is modified to reflect 25% of the sample, guaranteeing a balanced distribution. This balance increases the model's capacity to anticipate all failure kinds properly, resulting to more dependable and robust predictive maintenance results.

Kdeplot

In this study, two visualization approaches KDE plot, and correlation heatmap are applied to better comprehend the dataset. A KDE (Kernel Density Estimation) plot is used to produce smooth curves for each data point. It employs kernel functions to estimate the probability density, and by summing the kernel functions for each data point, a smooth curve is formed after preprocessing the data. This graphic helps to determine the distribution and patterns within the data. “Fig. 2. Kdeplot of the dataset after data pre-processing” depicts the kdeplot which shows the smooth curve of every data point.

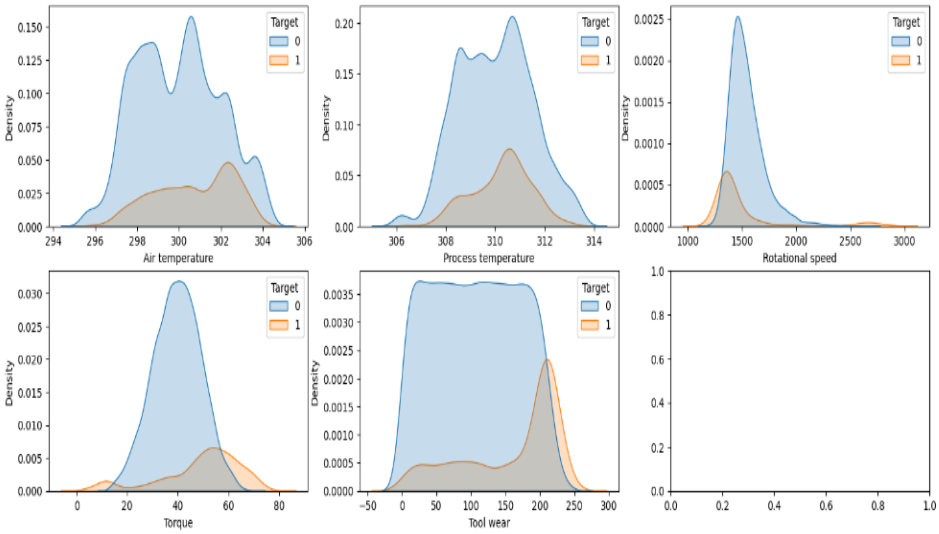


Fig. 2. Kdeplot of the dataset after data pre-processing

Implementation of PCA

Principal Component Analysis (PCA) is implemented in this research to reduce the dimensionality of the high-dimensional dataset. PCA condenses complex data into a smaller set of key components that capture the maximum variance within the dataset, thereby simplifying the data while retaining the most important information.

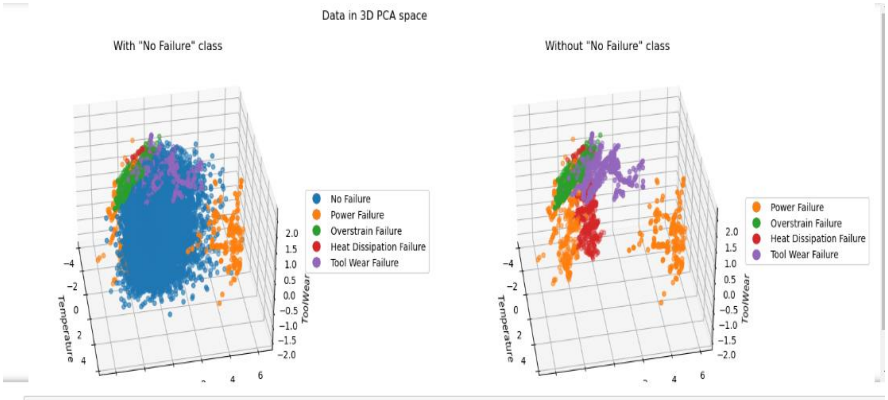


Fig. 3. Data segmentation after implementation of PCA

PCA helps to identify patterns and deviations from normal behavior, making it easier for machine learning models to process the data efficiently by focusing on the principal components that explain the highest variance. PCA aids in eliminating redundant or less informative features, improving the overall performance of predictive models by reducing noise and computational complexity while maintaining the core structure of the data. “Fig. 3. Data segmentation after implementation of PCA” represents the data segmentation after PCA implementation.

Dataset Splitting

In this study, two visualization approaches KDE plot, and correlation heatmap are applied to better comprehend the dataset. A KDE (Kernel Density Estimation) plot is used to produce smooth curves for each data point. It employs kernel functions to estimate the probability density, and by summing the kernel functions for each data point, a smooth curve is formed after preprocessing the data. This graphic helps to determine the distribution and patterns within the data. “Fig. 2. Kdeplot of the dataset after data pre-processing” depicts the kdeplot which shows the smooth curve of every data point.

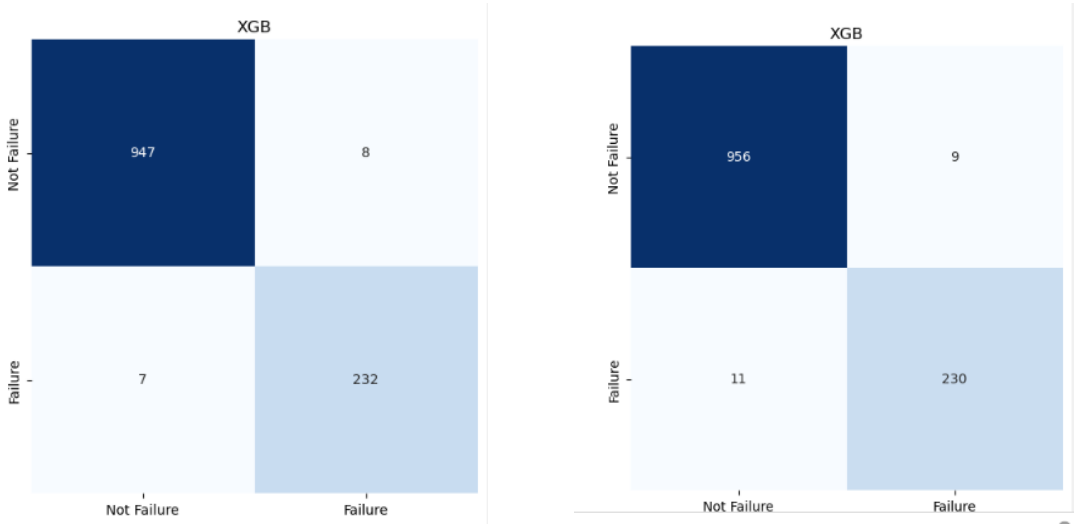
Model Implementation

Five machine learning techniques are applied for predictive maintenance model implementation: XGBoost, Support Vector Machine (SVM), Random Forest, Linear Regression, and Gaussian Naive Bayes. Each method is applied to the preprocessed and balanced dataset to forecast machine breakdowns. XGBoost is utilized for its high performance and efficiency in handling large datasets, SVM is known for its effectiveness in high-dimensional spaces, Random Forest provides robustness through ensemble learning, Linear Regression offers a simple and

interpretable model, and Gaussian Naive Bayes is used for its fast and probabilistic approach. These different algorithms allow for a full assessment of performance to select the best suited model for anticipating machine maintenance needs.

4. Experimental Results

After the application of five machine learning algorithms, confusion matrices are generated. A confusion matrix is a performance measuring tool used in machine learning, notably in classification tasks. It is a table that compares the actual target values (ground truth) with the values predicted by the model. From the comparison of five algorithms, it can be established that XGBoost delivers the superior performance. The confusion matrices of XGBoost method for validation and test score are presented in "Fig. 4. Confusion Matrices of XGBoost Model for (a) Validation; (b) Test".



(a) (b)

Fig. 4. Confusion Matrices of XGBoost Model for (a) Validation; (b) Test

Similarly, confusion matrices for SVM, Naive bayes, Random Forest and Linear regression have been obtained. But greater performance has been achieved via XGBoost model for both validation set and test set. Based on the confusion matrices, certain factors have been examined for choosing the optimal model such as accuracy, AUCvalue, F1-score and F2-measure. "Table 1. Performance Evaluation of different ML models for validation set" illustrates the comparison of five machine learning algorithms using the parameters of accuracy, AUC value, f1-score and f2-measure for validation score. Similarly, "Table 2. Performance Evaluation of different ML models for Test set" provides the comparison of five machine learning algorithms using the parameters of accuracy, AUC value, f1-score and f2-measure for test score.

Table 1. Performance Evaluation of different ML models for validation set

ML Models	Accuracy	AUC	F1-score	F2-measure
SVM	0.966	0.987	0.916	0.983
Random Forest	0.977	0.977	0.954	0.953
XGBoost	0.987	0.999	0.969	0.970
Linear Regression	0.883	0.905	0.673	0.629
Naive Bayes	0.860	0.880	0.592	0.537

Table 2. Performance Evaluation of different ML models for Test set

ML Models	Accuracy	AUC	F1-score	F2-measure
SVM	0.973	0.992	0.934	0.941
Random Forest	0.972	0.997	0.931	0.945
XGBoost	0.983	0.998	0.958	0.956
Linear Regression	0.881	0.917	0.655	0.598
Naive Bayes	0.865	0.902	0.605	0.550

6. Conclusion

After applying five Machine Learning models and comparing with them, it can be stated that XGBoost model delivers the best outcome for machine's preventative maintenance. Compared to XGBoost, random forest and support vector machine method lag small bit. Linear regression and naive bayes have delivered the worst outcome for this dataset.

References

[1] N. M. Abdulkareem and A. M. Abdulazeez, "Machine Learning Classification Based on Random Forest Algorithm: A Review," 2021, doi: 10.5281/zenodo.4471118.

[2] S. Mutsuddi, "Machine Learning for Predictive Maintenance in Manufacturing Industries," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 4, pp. 1175–1181, Apr. 2023, doi: 10.22214/ijraset.2023.50098.

[3] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," *IEEE Trans. Ind. Inform.*, vol. 11, no. 3, pp. 812–820, Jun. 2015, doi: 10.1109/TII.2014.2349359.

[4] Y. Zhang, S. Ren, Y. Liu, and S. Si, "A Big Data Analytics Architecture for Cleaner Manufacturing and Maintenance Processes of Complex Products," *J. Clean. Prod.*, vol. 142, pp. 626–641, 2017.

[5] Q. Liu, M. Dong, and Y. Peng, "A Dynamic Predictive Maintenance Model Considering Spare Parts Inventory Based on the Hidden Semi-Markov Model," *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 227, no. 9, pp. 2090–2103, 2013.

[6] M.-Y. You and G. Meng, "A Modularized Framework for Predictive Maintenance Scheduling," *Proc. Inst. Mech. Eng. O J. Risk Reliab.*, vol. 226, no. 4, pp. 380-391, July 2012, doi: 10.1177/1748006X11431209.

[7] A. Kanawaday and A. Sane, "Machine learning for predictive maintenance of industrial machines using IoT sensor data," *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2017, pp. 87-90, doi: 10.1109/ICSESS.2017.8342870.

[8] K. I. Masani, P. Oza, and S. Agrawal, "Predictive Maintenance and Monitoring of Industrial Machine using Machine Learning," *Scalable Computing: Practice and Experience*, vol. 20, no. 4, pp. 665–676, Dec. 2019, doi: 10.12694/scpe.v20i4.1585.

[9] A. P. Kane, A. S. Kore, A. N. Khandale, S. S. Nigade, and P. P. Joshi, "Predictive Maintenance using Machine Learning," *arXiv*, May 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.09402.

[10] C. Patil, S. K. Jadhav, A. L. Bardiya, and M. P. Raverkar, "Machine Learning-Based Predictive Maintenance of Industrial Machines," *Int. J. Comput. Trends Technol.*, vol. 71, no. 3, pp. 50-56, Feb. 2023, doi: 10.14445/22312803/IJCTT-V71I3P108.

[11] C. N. Okeke, O. F. Oluwatobi, U. U. Rita, G. U. Nwankwo, O. M. Chiadikobi, and others, "Predictive Maintenance of Industrial Equipment Using Machine Learning in Industrial Environment of Awka Metropolis, Nigeria," *Int. J. Sci. Technol. Res. Arch.*, vol. 5, no. 2, pp. 80-90, Oct. 2023, doi: 10.53771/ijstra.2023.5.2.0080.

[12] M. Paolanti, L. Romeo, A. Felicetti, J. Loncarski, and others, "Machine Learning Approach for Predictive Maintenance in Industry 4.0," in *Proc. 2018 14th IEEE/ASME Int. Conf. Mechatronic Embedded Syst. Appl. (MESA)*, Oulu, Finland, 2018, pp. 1-6, doi: 10.1109/MESA.2018.8449150.

[13] C. Magena, "Machine Learning Models for Predictive Maintenance in Industrial Engineering," *Int. J. Comput. Eng.*, vol. 6, no. 3, pp. 1-14, July 2024, doi: 10.47941/ijce.2137.