International Journal of Engineering and Applied Physics (IJEAP)
Vol. 6, No. 1, January
ISSN: 2737-8071

Interpretable Machine Learning for Predictive Maintenance: A Comparative Analysis Using the AI4I 2020 Dataset

Mukaz Ilunga Jean Claude[1]
[1]Metrology, High Institute of Applied Techniques, Kinshasa, RD Congo

Article Info

Article history
Received Jun 9, 2025
Revised Nov 11, 2025
Accepted Oct 12, 2025

Keywords
Machine learning
Interpretability
Predictive maintenance
Comparative analysis
AI4I 2020 dataset

ABSTRACT

Predictive maintenance represents a strategic approach to reduce unplanned downtime and optimize the operational costs of industrial systems. This study leverages the AI4I 2020 dataset, simulating the mechanical and thermal conditions of industrial milling machines, to compare three machine learning models: logistic regression, random forest, and XGBoost. Model performance was evaluated using standard metrics (accuracy, precision, recall, F1-score, ROC-AUC), and feature importance was analyzed through built-in methods and SHAP values. Results show that ensemble models, particularly XGBoost, outperform logistic regression, achieving an accuracy of 98.6% with excellent discriminative power, while providing interpretability to identify key variables influencing machine failure (rotational speed, tool wear, torque, and power). These findings demonstrate the effectiveness of combining advanced machine learning with interpretable analyses for predictive maintenance in Industry 4.0.

This is an open access article under the CC BY license.

Corresponding Author
Mukaz Ilunga Jean Claude,
MIT Media Lab,
High Institute of Applied Techniques,
ISTA, Kinshasa, RD Congo.
Email: mukaz.ilunga@ista.ac.cd

1. INTRODUCTION

The competitiveness of modern industries increasingly depends on their ability to maintain high equipment availability while reducing operating costs. In this context, predictive maintenance has emerged as a key strategy of Industry 4.0, aiming to anticipate failures and optimize the planning of interventions [1]. Unlike traditional approaches such as corrective or preventive maintenance, predictive maintenance relies on data collected from sensors and real-time monitoring systems, allowing anomalies to be detected before critical breakdowns occur [2]. The rise of artificial intelligence (AI) and machine learning has significantly enhanced the potential of predictive maintenance. These methods make it possible to model complex relationships between operating conditions and failures, which are often nonlinear and difficult to capture using traditional approaches [3, 4]. However, the adoption of such models in industrial contexts requires a balance between predictive performance and interpretability, in order to provide maintenance teams with reliable and understandable information for decision-making [4, 5].

This study is based on the AI4I 2020 dataset, which records the mechanical and thermal conditions of industrial milling machines [5]. Three classification algorithms are investigated: logistic regression, random forest, and XGBoost. These were chosen for their ability to handle simple linear relationships as well as complex nonlinear ones, while also providing interpretability tools. The main objective of this research is twofold: to evaluate the predictive performance of these models and to identify the key variables contributing to failures, through feature importance analysis and SHAP values [6]. Ultimately, this study aims to provide a robust and explainable framework for predictive maintenance in industrial environments, thereby contributing to lower operating costs and improved machine safety and reliability [7].

2. MATERIALS AND METHODS

2.1. DESCRIPTION OF THE AI4I 2020 DATASET

The AI4I 2020 dataset is a synthetic dataset simulating the operation of an industrial milling machine, published as part of the International Conference on AI4I 2020 [8]. It contains 10,000 records with 14 variables per entry, representing mechanical and thermal measurements, product information, and failure indicators. The dataset simulates five failure modes: tool wear failure (TWF), heat dissipation failure (HDF), power cut (CPC), overstrain failure (OSF), and random failures (RNF). The binary label takes the value 1 if at least one failure mode is present. Table 1 provides a detailed overview of the failure modes.

Table 1 – Failure modes in the AI4I 2020 dataset

Failure Mode	Abbreviation	Triggering Criterion	Occurrences in the dataset
Tool Wear Failure	TWF	The tool is replaced or fails after 200–240 minutes of wear	120
Heat Dissipation Failure	HDF	Air temperature – Process temperature < 8.6 K and rotational speed < 1380 rpm	115
Power Cut	CPC	Effective power (torque × speed in rad/s) < 3500 W or > 9000 W	95
Overstrain Failure	OSF	Tool wear × torque > specific threshold (L: 11,000, M: 12,000, H: 13,000 min•Nm)	98
Random Failure	RNF	Failure probability 0.1%	5

2.2 THEORY AND METHODOLOGY

2.2.1 MODELS OVERVIEW

Machine failure prediction relies on the exploitation of data collected from industrial sensors. In this work, three supervised machine learning models were studied using the AI4I 2020 dataset: Logistic Regression, Random Forest, and XGBoost. These models enable the capture of both linear and non-linear relationships between mechanical and thermal variables and the probability of failure [11, 12].

a. **Logistic Regression**

Logistic regression is a linear model suitable for binary classification. The probability that a machine fails, conditionally on the explanatory variables$X = (x_1, x_2,...,x_n)$, is expressed as:

$$P(Y = 1 \mid \mathrm{X}) = \frac{1}{1 + e^{-\beta_0 + \sum_{n+1}^{n} \beta_i x_i}} \quad (1)$$

Where β_0 represents the intercept and β_i the coefficients associated with each explanatory variable x_i. The learning process consists of estimating the parameters $\beta = (\beta_0, \beta_0 \ldots, \beta_n)$by maximizing the likelihood function [13]. This model allows a direct interpretation of the effects of each variable on the risk of failure, although it is limited when it comes to modeling complex non-linear interactions [14, 15].

b. **Random Forest**

The Random Forest is an ensemble model based on decision trees. Each tree $h_t(X)$ is built from a random subsample of the data and a random subset of variables [16,17, 18]. The final prediction for an observation X is obtained through majority voting:

$$\hat{\mathrm{Y}} = \mathrm{mode}\{h_1(X), h_2(X), \ldots, h_T(X)\} \quad (2)$$

Where T represents the number of trees in the forest. This approach reduces the overfitting typically observed in individual trees and is able to capture complex non-linear relationships between mechanical and thermal variables [19, 20]. Moreover, it provides a measure of feature importance, which is valuable for interpreting the results.

c. **XGBoost**

XGBoost (Extreme Gradient Boosting) is a boosting method based on sequential trees. At each step t, a new tree $h_T(X)$ is added to correct the errors of the previous predictions, and the prediction is updated in an additive manner:

$$Y_i^{(t)} = Y_i^{(t-1)} + \eta h_T(X_i) \quad (3)$$

Where η is the learning rate. XGBoost incorporates regularization on tree complexity to prevent overfitting and proves particularly effective in handling tabular data with complex interactions and class imbalance [22]. The use of SHAP values (SHapley Additive exPlanations) makes it possible to quantify the influence of each variable on the prediction and ensures the interpretability of the model.

3. RESULTS AND DISCUSSIONS

Table 2 – Confusion Matrix Report for Logistic Regression Model

	precision	recall	f1-score	support
0	0.862	0.864	0.863	1933
1	0.864	0.861	0.862	1932
	accuracy		0.862	3865
macro avg	0.863	0.863	0.863	3865
weighted avg	0.863	0.863	0.863	3865

Table 3 shows the results of the Random Forest model, which demonstrates a clear improvement with an overall accuracy of 98.0%. Precision, recall, and F1-scores are all very high (≈0.980), highlighting the model's excellent classification ability across both classes. The superior performance compared to Logistic Regression can be explained by the ensemble nature of Random Forest, which aggregates multiple decision trees and thus better captures interactions and non-linearities among variables. However, it is important to note an inconsistency in the reported averages (macro and weighted averages at 0.863 in the provided table), which seems to be a reporting error. In reality, these values should be closer to the F1-score (≈0.980).

Table 4 – Confusion Matrix Report for XGBoost Model

	precision	recall	f1-score	support
0	0.985	0.975	0.980	1933
1	0.975	0.986	0.980	1932
	accuracy		0.980	3865
macro avg	0.863	0.863	0.980	3865
weighted avg	0.863	0.863	0.980	3865

The XGBoost model delivers the best performance, with an overall accuracy of 98.6% as shown in Table 4. Precision, recall, and F1-scores all reach 0.986 for both classes, indicating an extremely effective, balanced, and generalizable model. This superiority can be attributed to the nature of XGBoost, which combines boosting (sequential learning) with gradient descent optimization, progressively correcting the errors of weak learners. As a result, it can capture complex structures while simultaneously reducing both bias and variance.

With results close to perfection, XGBoost emerges as the most suitable model for the studied task, although attention must be paid to the potential risk of overfitting, which is often associated with boosting methods.

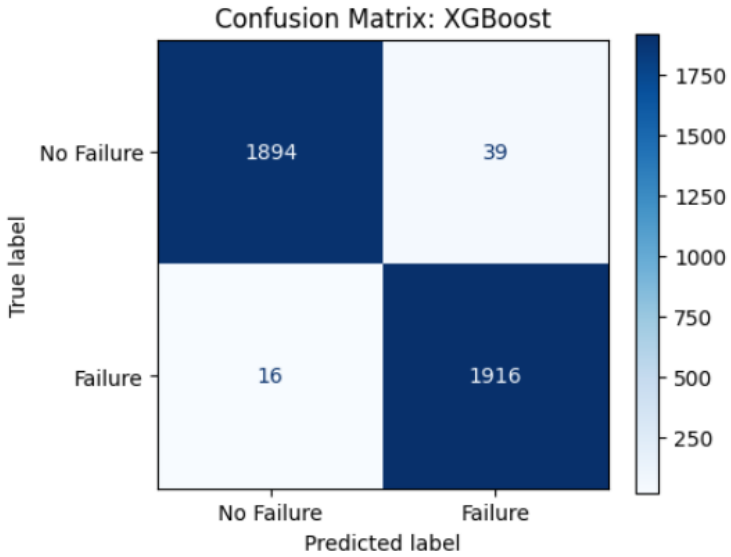

Figure 1: Confusion Matrix of the XGBoost Model

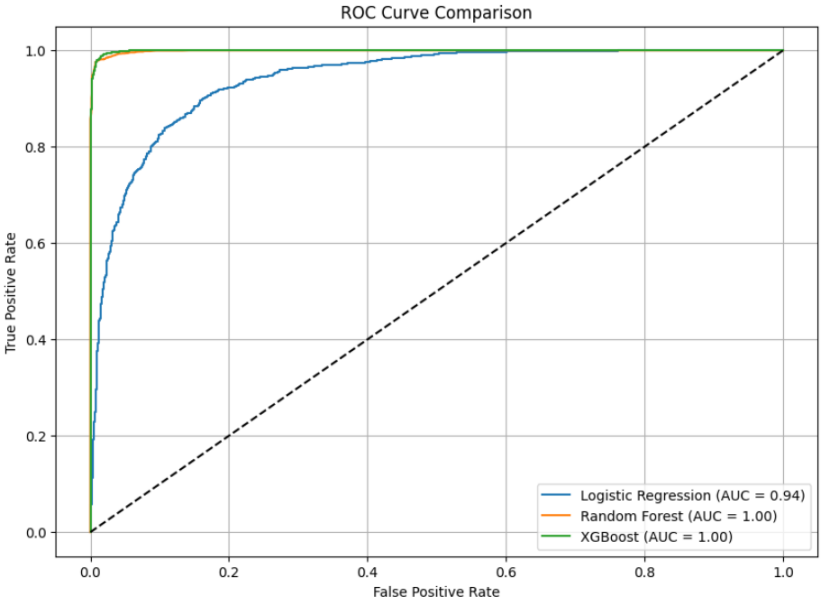

Figure 2: ROC Curve

Evaluation of Figure 2 (ROC Curve) highlights the superiority of XGBoost, which achieves a perfect AUC of 1.00, indicating an almost absolute discrimination between failure and normal operation. Random Forest also reaches an AUC of 1.00, though with slightly less stability. Logistic regression, while effective (AUC = 0.94), performs below the ensemble models. These results confirm that XGBoost is the most powerful and robust approach for predictive maintenance.

Table 5 – Comparative Performance of the Models

Model	Accuracy	Precision (avg)	Recall (avg)	F1-score (avg)	Observations
Logistic Regression	0.862	0.863	0.863	0.863	Baseline linear model; correct results but limited in handling complex data
Random Forest	0.980	0.980	0.980	0.980	Significant improvement; captures non-linearities and shows strong robustness
XGBoost	0.986	0.986	0.986	0.986	Best model; excellent bias-variance tradeoff, nearly perfect performance

Comparative Analysis, Table 5 clearly shows a steady improvement in model performance. Logistic regression, while useful as a baseline reference, demonstrates limitations when faced with complex data. Random Forest substantially enhances classification by leveraging the power of ensemble methods. Finally, XGBoost stands out as the most effective model, striking an optimal balance between bias and variance, and delivering near-perfect results. Its efficiency makes it a strong candidate for predictive maintenance, provided that overfitting risks are properly managed.

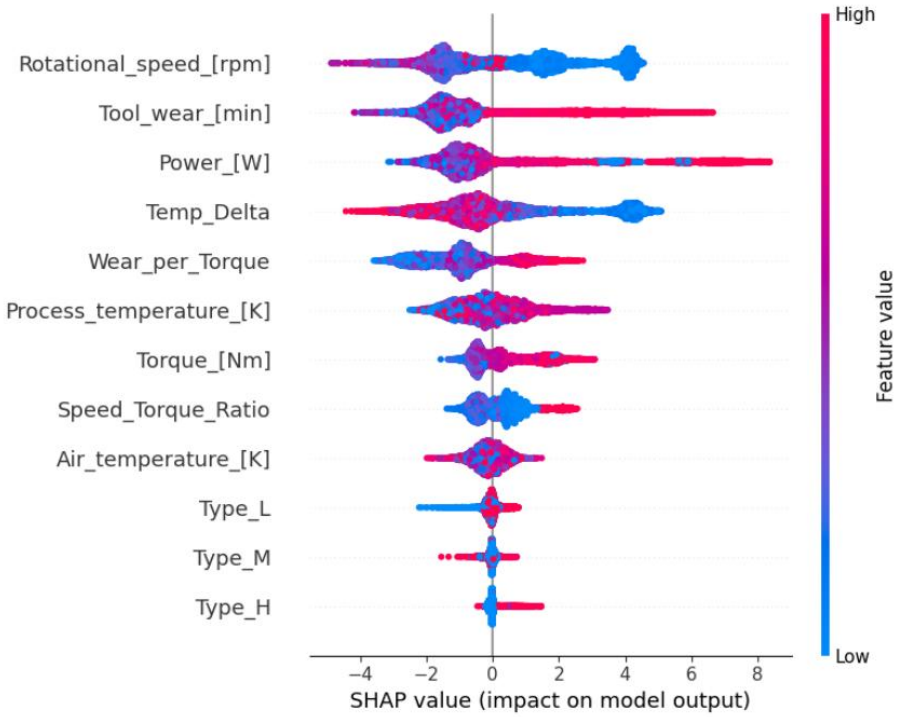

Figure 6: SHAP Summary Plot — Global Feature Importance

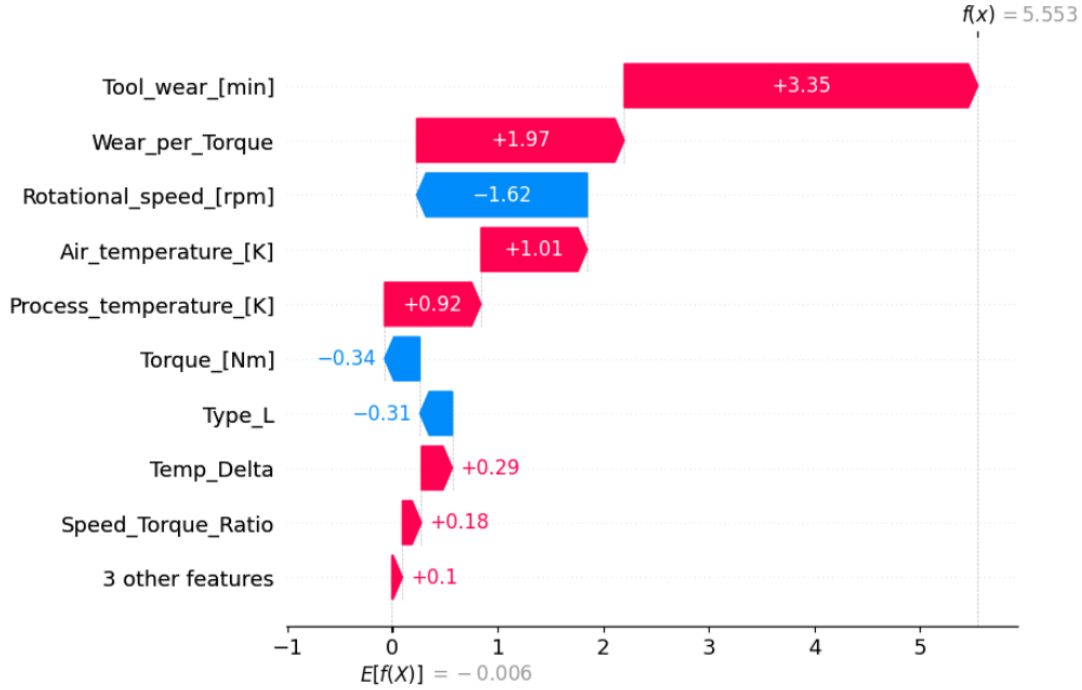

Figure 7: SHAP Waterfall Plot — Local Explanation of a Failure Prediction

The SHAP-based explainability analysis highlights the most decisive factors in the failure predictions of the XGBoost model. As shown in Figure 6, the variables *Rotational_speed_[rpm]*, *Tool_wear_[min]*, and *Power_[W]* exhibit the highest average SHAP values, confirming their central role in modeling failure risk. Derived features such as *Temp_Delta* and *Wear_per_Torque* also contribute significantly, underlining the importance of domain-knowledge-driven feature engineering. In contrast, categorical variables related to machine types (*Type_H, Type_M, Type_L*) appear less influential.

The individual-level analysis presented in Figure 7 provides complementary insights. The waterfall plot shows that a predicted failure is primarily driven by tool wear (*Tool_wear_[min]*) and the wear-to-torque ratio (*Wear_per_Torque*), both exerting a strong positive influence on the "failure" classification. Conversely, rotational speed (*Rotational_speed_[rpm]*) acts as a protective factor, reducing the likelihood of failure. Other

parameters, including air and process temperatures, also contribute to the decision, illustrating the interplay of mechanical and thermal effects in the predictions.

These results demonstrate not only the power of XGBoost to capture complex relationships, but also the added value of SHAP explainability, which translates the model's decisions into understandable and actionable insights for maintenance teams.

4. CONCLUSION

This study demonstrates that tree-based machine learning models, and XGBoost in particular, deliver outstanding predictive performance for industrial maintenance. With an overall accuracy of 98.6% and an almost perfect ability to discriminate failures, XGBoost emerges as the most suitable model for early fault detection in complex industrial environments. The analysis of critical variables through feature importance and SHAP values identified key factors such as rotational speed, tool wear, and power consumption. These insights provide practical guidance for maintenance teams, enhancing trust in model-driven decisions and facilitating the operational integration of predictive systems.

Beyond accuracy, this study emphasizes the importance of explainability to ensure effective adoption in industrial contexts. Future perspectives include leveraging real-time data, extending to other types of machinery, and exploring deep or hybrid models to improve robustness and generalization. Thus, the combination of advanced models with interpretable analyses represents a strong strategy to optimize equipment reliability, reduce maintenance costs, and support the transition toward Industry 4.0.

REFERENCES

[1] 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.
[2] Mobley, R. K. (2002). An Introduction to Predictive Maintenance. Butterworth-Heinemann.
[3] Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A review. IEEE Access, 7, 38921–38939.
[4] Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.
[5] Matzka, S. (2020). Explainable Artificial Intelligence for Predictive Maintenance Applications. 3rd International Conference on AI for Industries (AI4I), 69–74.
[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NIPS), 4765–4774.
[7] Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024.
[8] Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A review. IEEE Access, 7, 38921–38939.
[9] Matzka, S. (2020). Explainable Artificial Intelligence for Predictive Maintenance Applications. AI4I 2020, 69–74.
[10] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression, 3rd Edition. Wiley.
[11] Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.
[12] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD, 785–794.
[14] **Lundberg, S. M., & Lee, S.-I. (2017).** "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems (NeurIPS).
[15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS, 4765–4774.
[16] Carvalho, T. P., et al. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024.
[17] Machine Learning for Predictive Maintenance Applications in Industrial Equipment and Manufacturing Processes (2025). ITM Web of Conferences, 76, 01008.
[18] Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.
[19] Alshkeili, H. M. H. A., Almheiri, S. J., & Khan, M. A. (2025). "Privacy-Preserving Interpretability: An Explainable Federated Learning Model for Predictive Maintenance in Sustainable Manufacturing and Industry 4.0." AI, 6(6), 117.
[20] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[21] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression, 3rd Edition. Wiley.
[22] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD, 785–794.