

RESEARCH ARTICLE

Explainable AI for Predictive Maintenance in Industry 4.0

Dinesh Bhagwan Hanchate^{1*}¹ Computer Science and Engineering, Dnyanshree Institute Engineering and Technology, Satara, Maharashtra, 415013, India*Corresponding Author Email: dinesh.hanchate@dnyanshree.edu.in**ARTICLE INFO****Article history**

Received 2026-01-09

Revised 2026-02-18

Accepted 2026-02-20

Keywords

Explainable AI

Predictive maintenance

Industry 4.0

SHAP

Machine learning

ABSTRACT

Predictive maintenance has become a cornerstone of Industry 4.0, where machine learning models analyze sensor data to anticipate equipment failures and reduce unplanned downtime. The most accurate models, however, are often opaque, and maintenance engineers are reluctant to act on predictions they cannot interpret. This tension between accuracy and interpretability limits industrial adoption of data-driven maintenance. This paper presents an explainable predictive maintenance framework that couples high-performing ensemble classifiers with post-hoc explanation methods—SHAP for global and local feature attribution and LIME for instance-level explanations—within a layered Industry 4.0 pipeline. The framework is specified formally and as pseudocode, covering data acquisition, preprocessing with class-imbalance handling, model training, and an explanation layer that exposes the drivers of each failure prediction to the engineer. Using an illustrative synthetic dataset modeled after industrial milling-machine telemetry, five classifiers are compared, and the selected gradient-boosting model is interpreted via global and local explanations. The results indicate strong predictive performance together with human-readable attributions, supporting explainable AI as a practical route to trustworthy predictive maintenance.

1. Introduction

The fourth industrial revolution, commonly termed Industry 4.0, has transformed manufacturing through the integration of cyber-physical systems, the Internet of Things, and data-driven intelligence (Lee et al., 2015). Modern production lines are instrumented with large numbers of sensors that continuously stream measurements of temperature, vibration, torque, rotational speed, and other operating variables. This abundance of data has enabled predictive maintenance, a strategy that forecasts equipment failures before they occur, allowing maintenance to be scheduled proactively rather than reactively or at fixed intervals. The economic case for predictive maintenance is substantial. Unplanned downtime remains one of the largest sources of lost productivity in manufacturing, and maintenance accounts for a material share of

operating costs (Carvalho et al., 2019; Zonta et al., 2020). Machine learning models, particularly ensemble methods and deep neural networks, have achieved high accuracy in failure prediction and remaining useful life estimation (Lei et al., 2018; Serradilla et al., 2022), and are now central to the smart-maintenance vision of Industry 4.0.

A critical obstacle remains. The most accurate models tend to be the least transparent: ensemble and deep learning models behave as black boxes whose internal reasoning is not directly accessible to a human. In high-stakes industrial settings, an engineer asked to take a machine offline or replace a component needs to understand why the model predicts an imminent failure, not merely that it does. Without this understanding, trust erodes, adoption stalls, and an unexplained prediction offers no guidance on which physical factor to inspect. Explainable artificial intelligence has emerged to close this gap by producing human-understandable accounts of model decisions (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020), through global methods that rank the overall influence of each feature and local methods that justify an individual prediction (Lundberg & Lee, 2017; Ribeiro et al., 2016).

The objective of this paper is to design and demonstrate an explainable predictive maintenance framework suited to Industry 4.0. The main contributions are threefold. First, we formulate a layered pipeline that connects sensor data acquisition and preprocessing to machine learning prediction and, crucially, to an explanation layer that surfaces the drivers of each prediction to the maintenance engineer. Second, we integrate SHAP and LIME for global and local explanation and describe the procedure as reproducible pseudocode. Third, we present an illustrative study using a synthetic milling-machine dataset, comparing five classifiers and interpreting the best model via global feature importance and local explanations for individual failure predictions.

The remainder of the paper is organized as follows. Section 2 reviews related work on predictive maintenance, machine learning, and explainable AI in industrial contexts. Section 3 presents the system model, problem formulation, and the proposed explainable framework. Section 4 reports and discusses the numerical results. Section 5 concludes with limitations, managerial insights, and future research directions.

2. Literature Review

The conceptual foundation of this work lies in the architecture of Industry 4.0, in which cyber-physical systems synchronize the physical factory floor with a computational layer, enabling self-monitoring and autonomous decision-making (Lee et al., 2015). Within this paradigm, predictive maintenance has been studied extensively. Systematic reviews of machine learning for maintenance map the techniques applied to failure prediction and condition monitoring and consistently report the dominance of data-driven methods (Carvalho et al., 2019). Broader reviews of predictive maintenance in Industry 4.0 emphasize the integration of the Internet of Things, cyber-physical systems, and the multidisciplinary nature of the field (Zonta et al., 2020). Surveys of deep learning describe how architectures map onto the stages of a maintenance program,

from anomaly detection to remaining useful life estimation (Serradilla et al., 2022). Dedicated reviews of machinery health prognostics structure the problem into data acquisition, health-indicator construction, health-stage division, and life prediction (Lei et al., 2018).

The machine learning methods underpinning these systems are well established. Ensemble approaches based on bagging and boosting, including random forests and gradient-boosted trees, are consistently reported as strong performers on industrial tabular data because they combine accuracy with robustness (Breiman, 2001; Chen & Guestrin, 2016). Their strength, however, comes at the cost of transparency, which motivates a parallel literature on explainability. General surveys of explainable AI define the field, organize its methods into taxonomies, and frame the challenges of building responsible and human-centric systems (Barredo Arrieta et al., 2020; Adadi & Berrada, 2018). A prominent perspective argues that for high-stakes decisions, one should prefer inherently interpretable models over post-hoc explanations of black boxes, a tension that recurs throughout the industrial literature (Rudin, 2019).

Several explanation techniques have become standard tools. Additive feature-attribution based on Shapley values provides a unified and theoretically grounded account of how each feature contributes to a prediction (Lundberg & Lee, 2017), local surrogate modeling explains individual predictions of any classifier by approximating it locally with an interpretable model (Ribeiro et al., 2016), gradient-based localization produces visual explanations for convolutional networks (Selvaraju et al., 2017), and counterfactual methods show how an input would need to change to alter the outcome (Mothilal et al., 2020). These methods form the toolkit that industrial applications adapt to their needs.

The convergence of these threads is an active research area. Surveys of artificial intelligence and explainable AI in Industry 4.0 catalog the methods adopted and identify explainability as a key enabler for adoption in high-stakes industry applications (Ahmed et al., 2022). Applied studies demonstrate the value of explanation in maintenance directly: explainable models and explanatory interfaces have been built and evaluated on realistic predictive-maintenance data (Matzka, 2020); unsupervised and semi-supervised deep learning has been combined with explanation to detect, diagnose, and attribute faults in press machines and rotating machinery (Serradilla et al., 2021; Brito et al., 2022); Shapley-based attribution has been used to identify the most informative condition-monitoring features for bearing fault diagnosis (Brusa et al., 2023); and comparative studies have applied multiple explanation methods, including SHAP, LIME, and partial dependence, to interpret failure-prediction models for rotating machines (Gawde et al., 2024). Across this body of work, two observations recur: accurate failure prediction in Industry 4.0 is overwhelmingly data-driven, and explanation is increasingly treated as a requirement rather than an option. What remains comparatively underdeveloped is a single, reproducible treatment that integrates a layered Industry 4.0 pipeline, a benchmarked model choice and both global and local explanations into a single engineer-facing framework. The framework developed here is positioned to address that gap.

3. Methodology

3.1 System Model

The proposed framework is organized as a layered pipeline, shown in Figure 1. At the base, industrial assets are instrumented with IoT sensors that stream operating measurements. A data-acquisition and preprocessing layer cleans the data, engineers features, and addresses class imbalance, since failure events are rare relative to normal operation. A machine learning layer trains failure-prediction models on the prepared data. The distinctive element is the explanation layer, which sits between the model and the human: it produces global explanations that rank the overall influence of each feature and local explanations that justify individual predictions. These explanations feed into a decision-support layer that provides the maintenance engineer with a basis for trust, diagnostic insight, and scheduling guidance, and the engineer's feedback, in turn, informs later iterations of the pipeline.

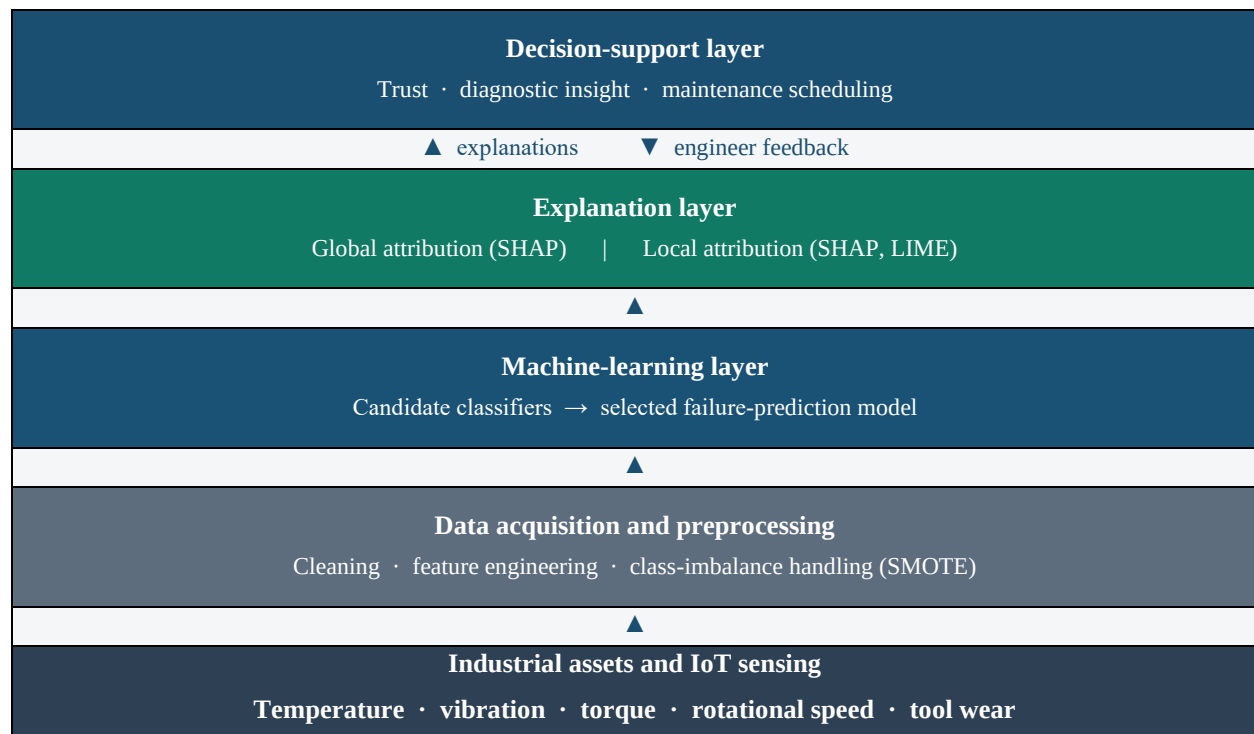


Figure 1. Layered explainable predictive maintenance pipeline for Industry 4.0.

3.2 Problem Formulation

Failure prediction is formulated as a binary classification problem. Each observation is a vector of sensor-derived features describing a machine at an operating cycle, and the target indicates whether a failure occurs. A model maps the feature vector to an estimated failure probability, and a threshold converts this probability into a maintenance decision. The model is trained by minimizing a regularized loss over the training set, as expressed in Eq. (1), where the loss penalizes disagreement between predicted and true labels and the regularization term controls model complexity.

$$\text{Objective} = \sum_i L(y_i, f(x_i)) + \Omega(f) \quad (1)$$

To explain a prediction, the framework uses additive feature attribution. The prediction is decomposed into a baseline value plus a sum of contributions, one per feature, where each contribution is the feature's Shapley value computed over feature subsets, as shown in Eqs. (2) and (3). The Shapley value fairly distributes the difference between the prediction and the baseline across the features, assigning each feature an importance for that specific prediction.

$$f(x) = \phi_0 + \sum_j \phi_j \quad (2)$$

$$\phi_j = \sum_{S \ni j} \frac{|S|! (n - |S|)!}{n!} [f(S \cup \{j\}) - f(S)] \quad (3)$$

Global importance is obtained by averaging the magnitude of each feature's local contributions across the dataset, yielding a ranking of overall feature influence that summarizes model behavior.

3.3 Proposed Explainable Framework

The framework proceeds in four stages. First, sensor data are acquired and preprocessed: features are cleaned and scaled, and the rare-failure class is balanced via oversampling (Chawla et al., 2002) to prevent the model from being biased toward the majority class. Second, several classifiers are trained and compared, and the best-performing model, a gradient-boosted tree ensemble in this study (Chen & Guestrin, 2016), is selected. Third, the selected model is explained globally using SHAP (Lundberg & Lee, 2017) to reveal which sensor variables most strongly influence failure predictions. Fourth, for any individual prediction of interest, local explanations are generated using SHAP and LIME (Ribeiro et al., 2016), so that the engineer can see which feature values drove that specific prediction toward or away from failure. The procedure is summarized in the following pseudocode.

Algorithm 1: Explainable Predictive Maintenance Framework

```

Input: sensor dataset D, candidate models M, threshold tau
Preprocess D: clean, engineer features, scale
Balance training set with oversampling (e.g., SMOTE)
for each model m in M do
    train m on balanced training set
    evaluate m (accuracy, precision, recall, F1, AUC)
end for
select best model f* by validation F1 / AUC
// Global explanation
compute SHAP values over the test set
rank features by mean |SHAP value|    (global importance)
// Local explanation (per prediction)
for a prediction of interest x do
    p = f*(x)
    if p >= tau then flag for maintenance
    compute SHAP/LIME attributions phi_j for x

```

```

    present ranked feature contributions to the engineer
end for
Output: predictions, global ranking, per-prediction explanations

```

Model performance is assessed with standard classification metrics. Because failures are rare, accuracy alone is misleading, so precision, recall, and the F1-score, defined below as the harmonic mean of precision and recall, are emphasized alongside the area under the ROC curve.

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (4)$$

3.4 Experimental Setup

Because access to a production line with labeled failure data is impractical, the framework is evaluated on an illustrative synthetic dataset modeled on industrial milling-machine telemetry (Matzka, 2020), with features for air temperature, process temperature, rotational speed, torque, tool wear, and a product-quality variant, and a rare binary failure target. The results reported in the next section are therefore synthetic and intended to demonstrate the framework rather than to report measurements from a live deployment. The dataset characteristics and the experimental configuration are summarized in Tables 1 and 2.

Table 1. Synthetic dataset characteristics (illustrative, modeled on milling-machine telemetry).

Attribute	Description
Records (illustrative)	10,000 operating cycles
Air temperature [K]	Ambient temperature (random-walk, $\sigma = 2$ K)
Process temperature [K]	Air temperature + 10 K offset
Rotational speed [rpm]	Derived from power (2860 W) with noise
Torque [Nm]	Normally distributed ($\mu = 40$, $\sigma = 10$)
Tool wear [min]	Increases with quality variant H/M/L
Machine failure	Binary target ($\approx 3.5\%$ failure rate)
Failure modes	TWF, HDF, PWF, OSF, RNF

Table 2. Experimental configuration.

Setting	Value
Train/test split	80% / 20% (stratified)
Cross-validation	5-fold
Class imbalance handling	SMOTE oversampling on the training set
Hyperparameter tuning	Grid search
XAI methods	SHAP (global + local), LIME
Evaluation metrics	Accuracy, precision, recall, F1, AUC

4. Numerical Results and Discussion

This section reports the illustrative results. Classifier performance is compared first, the selected model is then interpreted through global and local explanations, and the implications are discussed. All figures and tables are generated from the synthetic model described in Section 3.4.

4.1 Model Performance

Figure 2 compares accuracy, precision, recall, and F1-score across five classifiers, and Table 3 adds the area under the ROC curve. The gradient-boosted ensemble (XGBoost) achieves the strongest results across all metrics, followed by the random forest. The linear and single-tree baselines trail, particularly on recall, the metric of greatest operational concern, because a missed failure is far more costly than a false alarm.

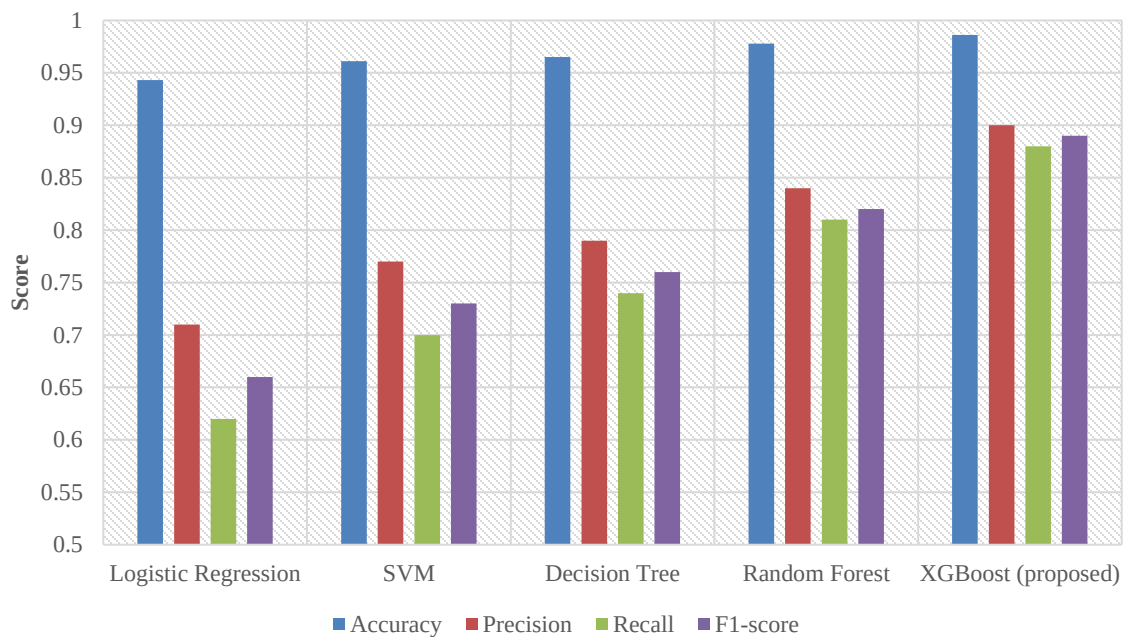


Figure 2. Classifier performance comparison (illustrative).

Table 3. Performance comparison across models (illustrative).

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.943	0.71	0.62	0.66	0.91
SVM	0.961	0.77	0.70	0.73	0.94
Decision Tree	0.965	0.79	0.74	0.76	0.93
Random Forest	0.978	0.84	0.81	0.82	0.97
XGBoost (proposed)	0.986	0.90	0.88	0.89	0.99

4.2 Global Explanation

Mean absolute SHAP values rank torque and tool wear as the dominant drivers of the predicted failure probability, with rotational speed next and the temperature channels and quality variant contributing comparatively little (Figure 3). That ordering matches the mechanics of load and progressive wear on a

milling tool, which gives the attribution a degree of face validity and, at fleet scale, indicates which sensor streams deserve the closest operational attention.

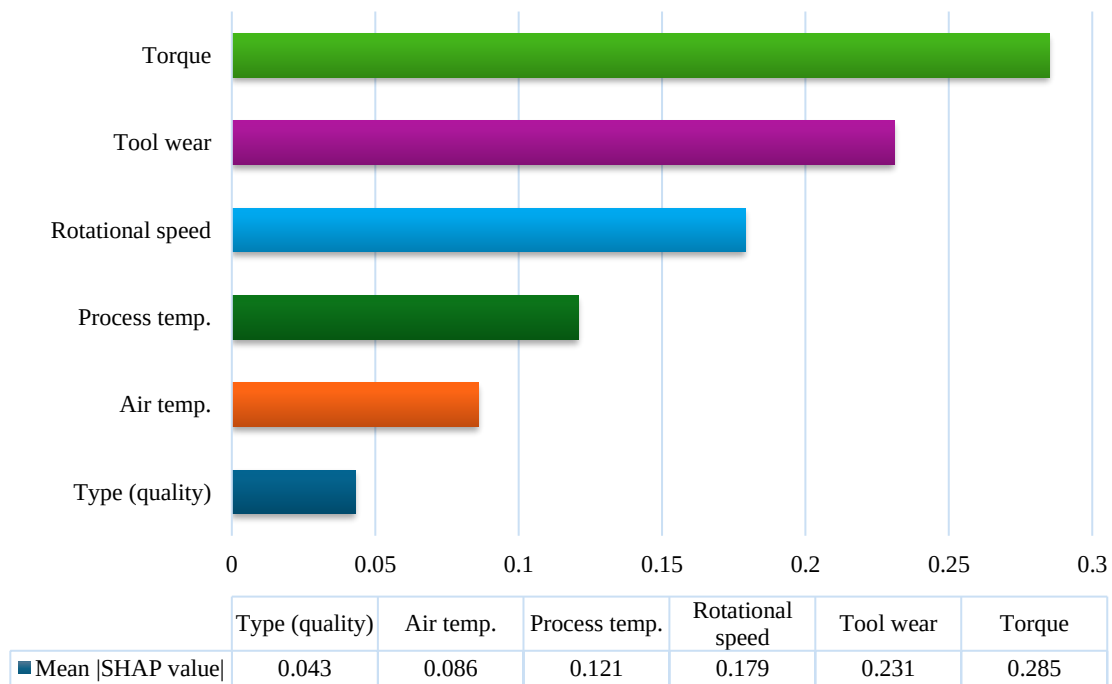


Figure 3. Global feature importance from mean absolute SHAP values (illustrative).

4.3 Local Explanation

For a machine assigned a high failure probability, the local SHAP attribution assigns most of the risk to high torque and elevated tool wear, both of which push the prediction toward failure, while rotational speed exerts a modest opposing effect (Figure 4). The same two variables that dominate the global ranking therefore reappear as the decisive drivers of this particular alert, which is useful: it shows that the case is not an outlier driven by an idiosyncratic sensor spike. Temperature and the quality variant contribute little on this instance, so they do not justify an immediate thermal inspection. The result converts an opaque score into an operational cue. The engineer can see that wear and load are driving the warning and can target the intervention, inspect the tool, reduce the cut, or schedule a change-out, rather than treating every alert as a generic stop.

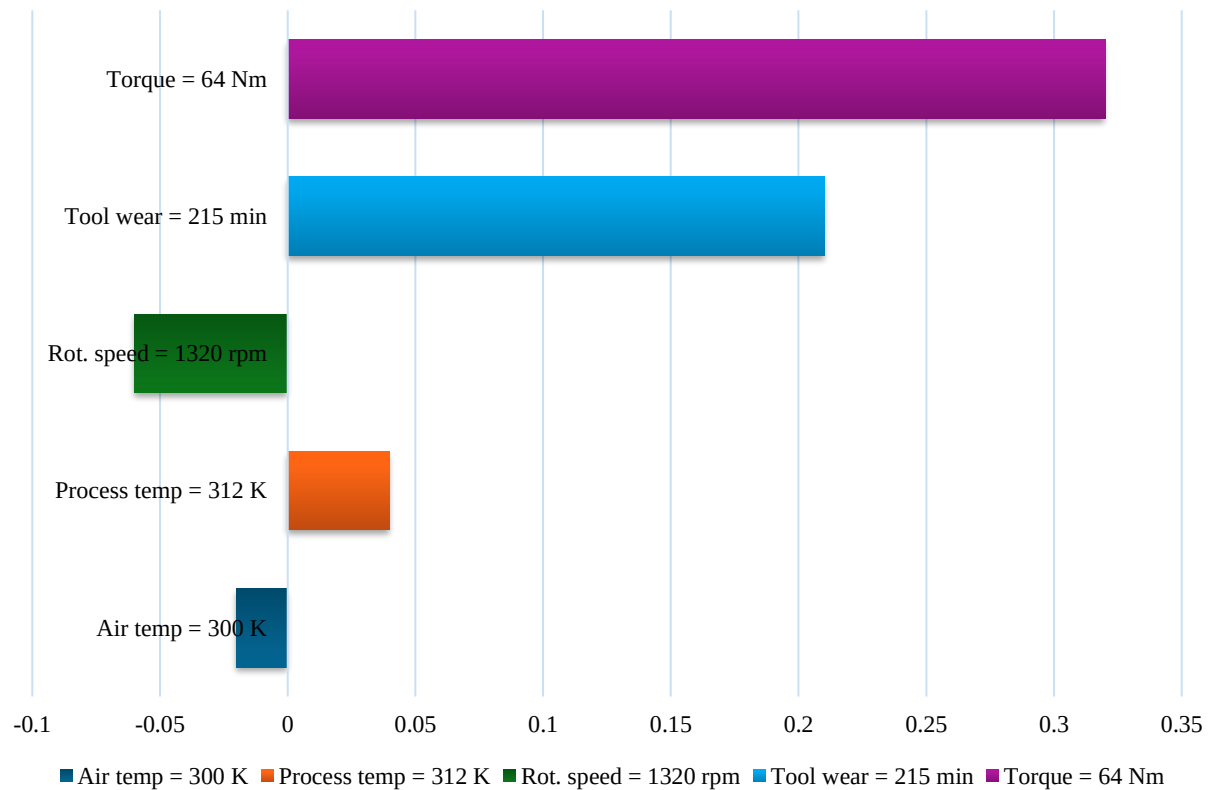


Figure 4. Local SHAP explanation for an individual failure prediction (illustrative).

The illustrative results support the central premise of the framework: a high-performing black-box model and a human-understandable explanation are not mutually exclusive when post-hoc attribution is layered onto an accurate predictor. Three implications follow for Industry 4.0 deployment. First, the combination of strong recall and clear explanations addresses the two barriers that most often stall adoption, namely missed failures and lack of trust. Second, global and local explanations serve complementary roles: the global ranking guides where to invest sensing and monitoring effort, while local explanations make individual alerts usable at the point of decision. Third, because the explanation layer is model-agnostic, the framework can accommodate later improvements in the underlying predictor without changing the engineer-facing interface. The principal limitation of this analysis is that the data are synthetic; the reported magnitudes should not be read as forecasts for any specific asset. Post hoc explanation methods also have well-known caveats, including sensitivity to feature correlations and the gap between attribution and causation, so explanations should inform engineering judgment rather than replace it. Validation on real industrial data is required before performance or explanation-quality claims can be made.

5. Conclusion

This paper presented an explainable predictive maintenance framework for Industry 4.0 that couples accurate machine learning classifiers with post hoc explanations via SHAP and LIME within a layered,

engineer-facing pipeline. An illustrative study on a synthetic milling-machine dataset showed that a gradient-boosted ensemble achieved the strongest predictive performance among five classifiers, and that global and local explanations rendered its predictions interpretable, attributing failure risk primarily to torque and tool wear in a manner consistent with engineering intuition.

The study has limitations. The evaluation relies on synthetic data, so the reported magnitudes are illustrative rather than predictive; failure prediction was treated as a static binary classification rather than remaining-useful-life estimation; and post hoc explanation methods remain sensitive to feature correlations and do not establish causation. From a managerial standpoint, the results suggest that manufacturers can adopt accurate, data-driven maintenance without sacrificing the transparency engineers need to act, that explanations can support governance and training around AI-assisted decisions, and that the choice between global and local explanations should follow from whether the decision is fleet-level or asset-level.

Future research should validate the framework on real industrial datasets and production telemetry, extend it from binary failure classification to remaining useful life prediction with explanation, and compare post hoc explanations against inherently interpretable models in high-stakes maintenance settings. Incorporating counterfactual explanations that tell an engineer what change would avert a predicted failure, integrating the framework with real-time IoT streams and digital twins, and conducting human-factors studies of how engineers actually use explanations are further promising directions.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.
- Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8), 5031–5042.
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brito, L. C., Susto, G. A., Brito, J. N., & Duarte, M. A. V. (2022). An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *Mechanical Systems and Signal Processing*, 163, 108105.
- Brusa, E., Cibrario, L., Delprete, C., & Di Maggio, L. G. (2023). Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring. *Applied Sciences*, 13(4), 2038.
- Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024.
- 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.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- Gawde, S., Patil, S., Kumar, S., Kamat, P., Kotecha, K., & Alfarhood, S. (2024). Explainable predictive maintenance of rotating machines using LIME, SHAP, PDP, ICE. *IEEE Access*, 12, 29345–29361.

- 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.
- Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104, 799–834.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.
- Matzka, S. (2020). Explainable artificial intelligence for predictive maintenance applications. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)* (pp. 69–74). IEEE.
- Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). ACM.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE.
- Serradilla, O., Zugasti, E., Ramirez de Okariz, J., Rodriguez, J., & Zurutuza, U. (2021). Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data. *Applied Sciences*, 11(16), 7376.
- Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. *Applied Intelligence*, 52(10), 10934–10964.
- Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*, 150, 106889.