

Explainable Uncertainty-Aware XGBoost for Industry 4.0 IoT: Risk-Aware Shop Floor Optimization

Shaikh Shehab Ahamed, Rashid Abrar Tanzil, Md. Mahmudul Hasan Ratul, Prosenjit Sarker, Sawon Mukherjee and Fawjia Islam

Department of Industrial Engineering and Management
Khulna University of Engineering & Technology
Bangladesh

shehabahamed2003@gmail.com, rashidabrar@pm.com,
mahmudulhasanratul2002@gmail.com, sarker2111035@stud.kuet.ac.bd,
sawonmukherjee1@gmail.com, islam2111018@stud.kuet.ac.bd

Abstract

Industry 4.0 plants collect rich IoT data, but many predictive models remain black boxes and give only point estimates. This limits trust and can lead to risky operating decisions. We propose a unified framework that makes high-accuracy gradient boosting both explainable and uncertainty aware, and then converts its outputs into actionable shop-floor settings. We train paired XGBoost models for regression KPIs and a maintenance-risk classifier using group-aware cross validation to avoid unit leakage. We explain each prediction with SHAP values to reveal the key controllable levers. We add uncertainty quantification for both task types. We use conformal prediction intervals for continuous KPIs and Wilson score bounds for calibrated class probabilities. We then run a multi-objective decision optimizer that minimizes energy and completion time while constraining worst-case risk using upper confidence bounds, not mean probabilities. On a distributed manufacturing dataset (3,741 samples, 10 units), the framework achieves stable generalization with about 90% interval coverage across major KPIs. The optimizer reduces completion time and energy with only a negligible downtime trade-off, and the explanations show that control and policy variables dominate over passive sensor flags.

Keywords

Explainable AI Manufacturing, XGBoost Uncertainty Quantification, SHAP Predictive Maintenance, Industry 4.0 Optimization, Risk-Aware Decision Framework

1. Introduction

The rise of Industry 4.0 has revolutionized factories with smart sensors and data analytics (Tao et al., 2018). Predictive models for quality and maintenance are now common. However, a critical gap remains. Many high-accuracy models are "black-boxes," offering no explanations for their predictions. This lack of transparency hinders trust and adoption by plant managers who need to understand the 'why' behind an alert (Rudin, 2019).

Furthermore, even interpretable models often lack a measure of their own reliability. They provide a single prediction without confidence intervals. This is risky for operational decisions. Engineers need to know not just what might fail, but how uncertain that prediction is. Few frameworks combine explainability with robust uncertainty quantification (UQ) for manufacturing data (Levi et al., 2022). This integration is essential for robust, safety-aware industrial AI (Ben-Tal et al., 2009).

This study tackles this dual gap. We propose an explainable AI (XAI) framework built on gradient-boosted models. Our method provides clear predictions and explains them using SHAP values. Crucially, it also quantifies prediction uncertainty for both regression and classification tasks. Finally, it integrates these outputs into a decision optimization module. This module finds operating settings that balance quality, energy use and risk.

We bridge data science with practical shop-floor decision-making. The primary objectives are:

1. To develop explainable gradient-boosted models for quality and condition prediction.
2. To integrate uncertainty quantification, providing prediction intervals and confidence bounds.

3. To perform decision optimization that uses explanations and uncertainty to recommend safe, efficient operating regions.

The paper is structured as follows: Section II details the literature review. Section III methodology. Section IV provides experimental setup. Section V results & discussion. Section VI concludes with contributions and practical implications.

2. Literature Review

Data-driven predictive models have become central to smart manufacturing, yet most high-accuracy models operate as black boxes that provide only point estimates (Tao et al., 2018). Rudin argues that this opacity undermines trust in high-stakes industrial settings and advocates for models whose reasoning can be directly interrogated (Rudin, 2019). SHAP (SHapley Additive exPlanations) addresses this by providing consistent, locally faithful feature attributions for any tree-based model, and has become the standard post-hoc explanation method for gradient-boosted ensembles (Lundberg & Lee, 2017).

Explainability alone is insufficient for risk-aware operation: point estimates carry no reliability signal. Levi et al. showed that poorly calibrated uncertainty outputs can be actively misleading in safety-critical regression tasks, and advocated for coverage-based evaluation of prediction intervals (Levi et al., 2022). For tool wear prediction specifically, Lee et al. demonstrated that prediction intervals improve operational reliability, though their study used quantile regression forests rather than higher-accuracy gradient boosters (Lee et al., 2025). For binary classification, Nemani et al. provided a comprehensive treatment of UQ for both regression and classification in engineering health prognostics, showing that raw model confidence scores systematically diverge from true posterior probabilities and must be calibrated before use in risk decisions (Nemani et al., 2023).

Despite these advances, no existing framework jointly integrates SHAP-based explainability with formally calibrated uncertainty bounds for both regression and classification tasks within a manufacturing setting (Levi et al., 2022). Furthermore, production decision optimizers typically consume deterministic predictions, ignoring the safety buffers that uncertainty intervals provide. Robust optimization theory establishes the principle of constraining worst-case rather than expected outcomes (Ben-Tal et al., 2009), yet this has not been applied to shop-floor optimization driven by gradient-boosted models. Our framework closes these gaps by coupling SHAP explanations and statistically grounded confidence bounds with a multi-objective optimizer that enforces hard constraints on worst-case maintenance risk.

3. Methodology

We propose a machine learning model that predicts both continuous production metrics and binary outcomes from manufacturing data. The model uses XGBoost, an explainable and uncertainty-aware machine learning algorithm, for predicting continuous values such as energy consumption and machine wear, and for binary classification tasks such as predicting whether maintenance is required.

3.1 Model Architecture

We use XGBoost (eXtreme Gradient Boosting) as the principal machine learning algorithm. It is highly efficient with tabular data and captures complex patterns. It also offers interpretability and uncertainty quantification. The expression for XGBoost is given by Equation (1).

$$\mathcal{L}(\theta) = \sum_{i=1}^n \mathcal{L}(y_i, \hat{y}_i) + \Omega(f) \quad \dots (1)$$

Here, $\mathcal{L}$ is the loss function, quantifying the difference between the true labels y_i and the predicted outputs $\hat{y}_i$, and $\Omega(f)$ represents the regularization term which is to be defined.

For regression tasks, we use the squared error loss function given by Equation (2).

$$\mathcal{L}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad \dots (2)$$

Here, y_i is the true value and $\hat{y}_i$ is the predicted value for the i -th instance.

For classification tasks, we use the logistic loss function as defined in Equation (3).

$$\mathcal{L}(y_i, \hat{y}_i) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad \dots (3)$$

Here, y_i represents the true label for the i -th sample, where the label is either 0 or 1 and $\hat{y}_i$ is the predicted probability that the i -th sample belongs to class 1.

The regularization term $\Omega(f)$ is defined to prevent overfitting and is given by Equation (4).

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad \dots (4)$$

Here, γ controls the complexity of the tree by penalizing the number of leaves T , w_j is the weight of the j -th leaf, and λ is the regularization coefficient. The model is trained by minimizing this objective function across multiple boosting rounds, where each round adds a new decision tree that corrects the errors of the previous one.

3.2 Data Preprocessing

The data is split into training and test sets using Group K-Fold Cross-Validation, ensuring that data from the same unit does not appear in both the training and validation sets. This prevents overfitting and ensures better generalization. The preprocessing steps are illustrated in Figure 1. These preprocessing steps are applied only to the training data, ensuring no leakage from the test data.

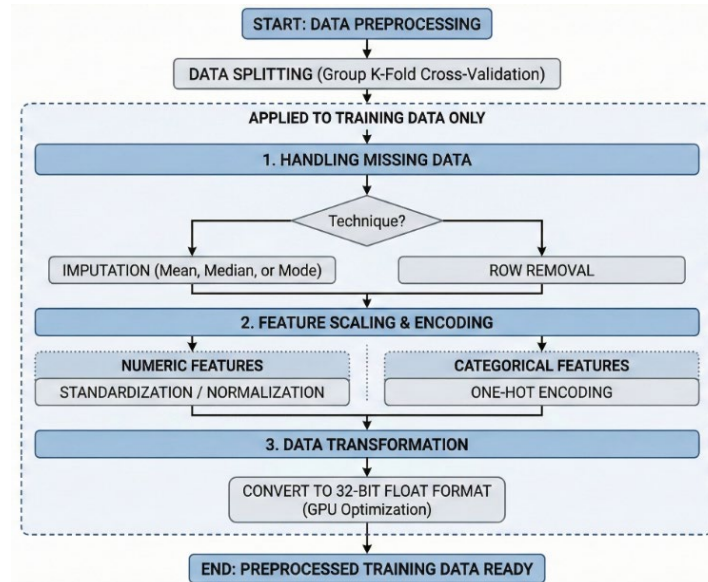


Figure 1. Schematic representation of the data preprocessing step.

3.3 Model Training

The model is trained using XGBoost, leveraging cross-validation and early stopping to ensure robustness and avoid overfitting. A step-by-step breakdown of the training procedure as an algorithm is as follows:

Input: Dataset $\mathcal{D}$, Hyperparameter Space θ , Early Stopping Rounds $N = 100$, Number of Folds $K = 5$.
Output: Optimized ModelBundle $\mathcal{B}$ (containing Regression and Classification models), Performance Summaries $\mathcal{S}$.

1. Initialize $\mathcal{B} \leftarrow \emptyset$, $\mathcal{S} \leftarrow \emptyset$.
2. Data Partitioning (Group K-Fold): Divide the dataset $\mathcal{D}$ into K folds based on unique unit IDs to ensure that data from the same unit never overlaps between the training and validation sets.
3. Hyperparameter Optimization: Execute Random or Grid Search over θ (including Learning Rate, Max Depth, and Tree Count) to identify the optimal configuration.
4. For each fold $k \in \{1, \dots, K\}$:
 - a. Split: Assign D_{train} and D_{val} according to the k -th group split.
 - b. Configuration: Initialize the training configuration parameters, setting the device (GPU/CPU), number of parallel jobs, and tree method to hist.
 - c. Gradient Boosting: Train the XGBoost model on D_{train} . Monitor validation error on D_{val} at each iteration.
5. Early Stopping Rule: If the validation error does not improve for N consecutive rounds, terminate training and revert to the best iteration.
6. Return $\mathcal{B}$, $\mathcal{S}$.

3.4 Explainable AI

To ensure that the predictions of the model are interpretable, SHAP (SHapley Additive exPlanations) is used to explain the decisions of the model. SHAP values provide an explanation for each individual prediction. For a given prediction $\hat{y}_i$, SHAP decomposes the output as Equation (5).

$$\hat{y}_i = \phi_0 + \sum_{j=1}^M \phi_j \quad \dots (5)$$

Here, ϕ_0 is the base value (expected output without any features), and ϕ_j is the SHAP value for feature j , which indicates its contribution to the prediction. Each SHAP value ϕ_j may be defined by Equation (6).

$$\phi_j = \mathbb{E}_{S \sim \mathcal{F}(x_j)} [f(S \cup x_j) - f(S)] \quad \dots (6)$$

Here, $\mathbb{E}$ denotes the expected value, S is a subset of features excluding x_j , and $f(S)$ is the prediction when feature x_j is excluded. Once the model is trained and optimized, it is deployed for making predictions on new data. The ModelBundle containing the trained model is saved to disk and can be reloaded for inference.

3.5 Uncertainty Quantification

Uncertainty quantification (UQ) allows the system provide a range of probable outcomes. For regression tasks, such as predicting energy consumption or machine wear, the model calculates a Prediction Interval (PI). Given a miscoverage level α (where $\alpha = 0.1$ for a 90% confidence level), the interval is defined in Equation (7).

$$PI = [\hat{y}_a - \epsilon, \hat{y}_a + \epsilon] \quad \dots (7)$$

Here, $\hat{y}_a$ is the predicted mean, and ϵ represents the margin of error derived from the distribution of residuals during the calibration phase. For classification tasks we use the Wilson Score Interval to provide statistically grounded probability bounds. Unlike standard intervals, the Wilson score remains stable even with small sample sizes or probabilities near 0 or 1. The lower and upper bounds of the probability p are given by Equation (8).

$$w = \frac{1}{1 + \frac{z^2}{n}} \left(\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z^2}{4n^2}} \right) \quad \dots (8)$$

Here, $\hat{p}$ is the predicted probability from the XGBoost model, n is the number of instances in the calibration bin, and z is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution. These intervals allow the system to identify not just the most likely outcome, but also the worst-case scenario for high-stakes variables like machine failure.

3.6 Decision Optimization

The optimization process evaluates a candidate space of N potential configurations for a given instance. The objective is to find a set of lever values that minimizes a multi-objective cost function, J . This function allows for a customized balance between energy efficiency, production throughput, and mechanical reliability based on user-defined weights (w). It is defined by Equation (9).

$$J = w_{energy} \cdot \hat{y}_{energy} - w_{completion} \cdot \hat{y}_{completion} + w_{failure} \cdot \hat{y}_{failure_risk} \quad \dots (9)$$

Instead of relying on mean probabilities, which can be misleading if the model is uncertain, the system imposes limits on the upper bounds of risk variables. For instance, to minimize the chance of unexpected downtime, a constraint is applied to the upper bound of the maintenance probability (τ) as follows:

$$\hat{p}(\text{maintenance})_{upper} \leq \tau_{maintenance}$$

By optimizing against these worst-case statistical bounds, the system provides a safety buffer. This ensures that the recommended operational settings remain reliable even when the production environment is unstable or data is noisy.

4. Experimental Setup

The core experimental setup is summarized in Table 1.

Table 1. Summary of the experimental setup for the proposed methodology

Component	Experimental Setup
Dataset	distributed_manufacturing_dataset.csv with 3,741 samples, 23 columns, and 10 unique units (Unit ID).
Data Split Protocol	Held-out test split for final reporting, plus GroupKFold cross-validation on the training set (up to 5 folds), grouping by Unit ID to reduce unit-level leakage.
Early Stopping	Within each CV fold, an inner early-stopping split is created (group-aware when possible) with test_size = 0.2 and early_stopping_rounds = 100.
Model Family	XGBoost models for regression and classification; tree_method = hist, max_bin = 256, n_jobs = 4, device = auto, random seed = 42.
Uncertainty and Calibration	Regression: conformal prediction intervals from out-of-fold residuals ($\alpha = 0.1 \rightarrow 90\%$ PI), with logit/expit transform to keep bounds in $[0, 1]$.

	Classification: cross-fitted isotonic calibration, plus bin-wise Wilson intervals ($\alpha = 0.1$, 10 bins); decision threshold selected by F_1 .
--	---

5. Results and Discussion

The results are produced through a single computational workflow: the trained uncertainty-aware XGBoost model generates (i) held-out test metrics and uncertainty coverage, (ii) sensitivity evidence via one-at-a-time “what-if” perturbations, (iii) SHAP global/local explanations to attribute changes to specific inputs, and (iv) a Pareto search that evaluates candidate lever settings using risk-aware bounds (upper90) for decision robustness. Key baseline-versus-optimized outcomes and their corresponding controllable lever settings are summarized in Table 2. This makes the study computational rather than illustrative: each numeric value reported below is a direct output of the trained model and the optimization routine applied to the same set of controllable levers.

Table 2. Baseline and optimized KPIs with corresponding control-lever settings

Category	Variable	Baseline	Optimized	Δ (Optimized – Baseline)	% Δ
Output KPI	Completion Time	133.576	130.268	–3.308	–2.48%
Output KPI	Energy Used	28.378	28.316	–0.062	–0.22%
Output KPI	Downtime	11.649	11.719	+0.070	+0.60%
Control lever	Resource Allocated	–1.28	8.93	+10.21	—
Control lever	Priority Level	3	1	–2	—
Control lever	Fuzzy Tuned K_d	0.126	0.420	+0.294	+233%
Control lever	Fuzzy Tuned K_i	0.400	0.307	–0.093	–23.3%
Control lever	PID Output	0.532	0.746	+0.214	+40.2%

The held-out evaluation demonstrates that the framework provides usable predictive accuracy with quantified reliability. As illustrated in Figure 2, the test metrics reported for continuous targets show both error magnitude and approximately 90% prediction-interval coverage, indicating that the model provides not only point estimates but decision-safe bounds: Completion Time (RMSE 32.74, MAE 28.02, PI coverage 0.902), Energy Used (RMSE 4.91, MAE 3.97, PI coverage 0.890), Downtime (RMSE 5.86, MAE 5.03, PI coverage 0.894), Machine Wear (RMSE 5.99, MAE 4.81, PI coverage 0.893), Predictive Failure Score (RMSE 0.285, MAE 0.258, PI coverage 0.903), and Error (RMSE 0.320, MAE 0.255, PI coverage 0.948). Operationally, this matters because the optimizer and managers can act on upper90 risk bounds instead of optimistic point predictions.

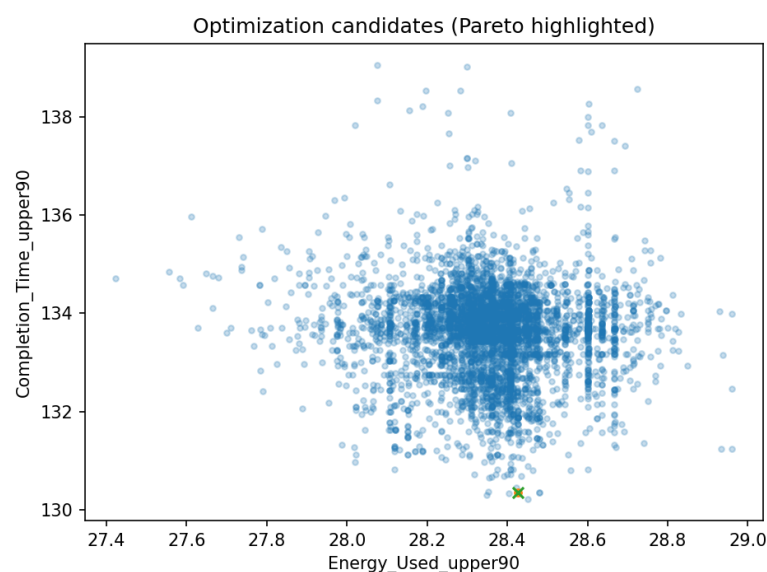


Figure 2. Risk-aware pareto optimization for predictive framework.

The optimization outcome is then reported as a baseline vs optimized comparison that is directly traceable to controllable inputs. In Figure 2, the recommended policy changes Resource Allocation from $-1.28 \rightarrow 8.93$ and Priority Level from $3 \rightarrow 1$, producing a computed improvement in key outcomes: Completion Time $133.576 \rightarrow 130.268$ (-3.308) and Energy Used $28.378 \rightarrow 28.316$ (-0.062), with a small trade-off Downtime $11.649 \rightarrow 11.719$ ($+0.070$). The same solution indicates controller-level adjustments (Fuzzy Tuned Kd $0.126 \rightarrow 0.420$; Ki $0.400 \rightarrow 0.307$; PID Output $0.532 \rightarrow 0.746$) that align with the sensitivity evidence in Figure 3, where one lever is varied at a time to verify which inputs materially move each KPI under fixed context (i.e., showing marginal cause–effect rather than correlation).

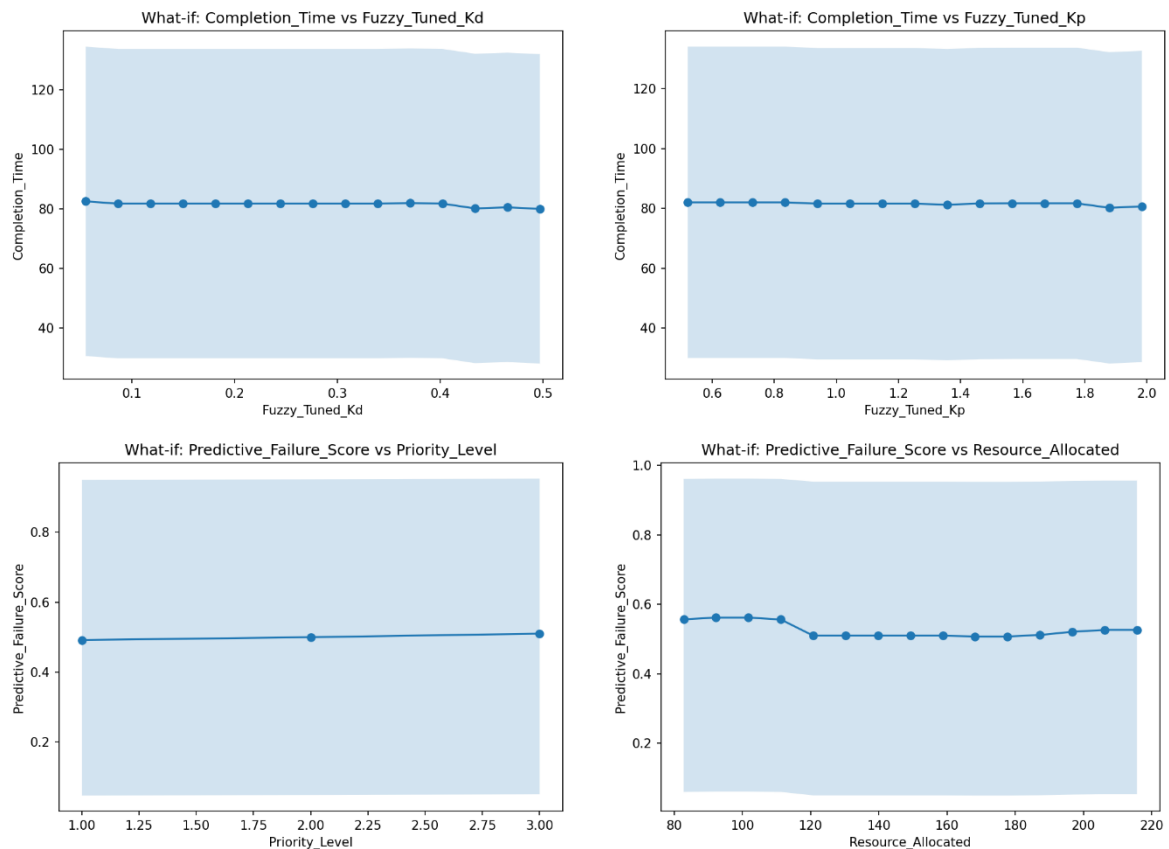


Figure 3. What-If sensitivity analysis of optimization candidates.

Finally, the figures are interpreted (not merely referenced) to connect “why” to “what changed.” In Figure 4, the global SHAP summaries show that KPI behavior is dominated by operational and control levers (e.g., resource allocation, PID/fuzzy gains, temperature, process conditions) rather than passive sensor-status flags, which validates why the Pareto search changes these levers. The local SHAP example provides a concrete, computed attribution: Maintenance Required shifts from 0.262 to -0.262 , driven primarily by Temperature (-0.55) with compensating increases from Fuzzy Tuned Kp ($+0.22$), Ki ($+0.16$), and Kd ($+0.13$)—demonstrating that the model’s decisions are explainable at the instance level and consistent with the global patterns. Together, Figure 3 (marginal sensitivity) and Figure 4 (feature attribution) close the loop between optimization recommendations and interpretable evidence, showing that the baseline-to-optimized improvements are computational outputs supported by explainable drivers, not narrative assertions.

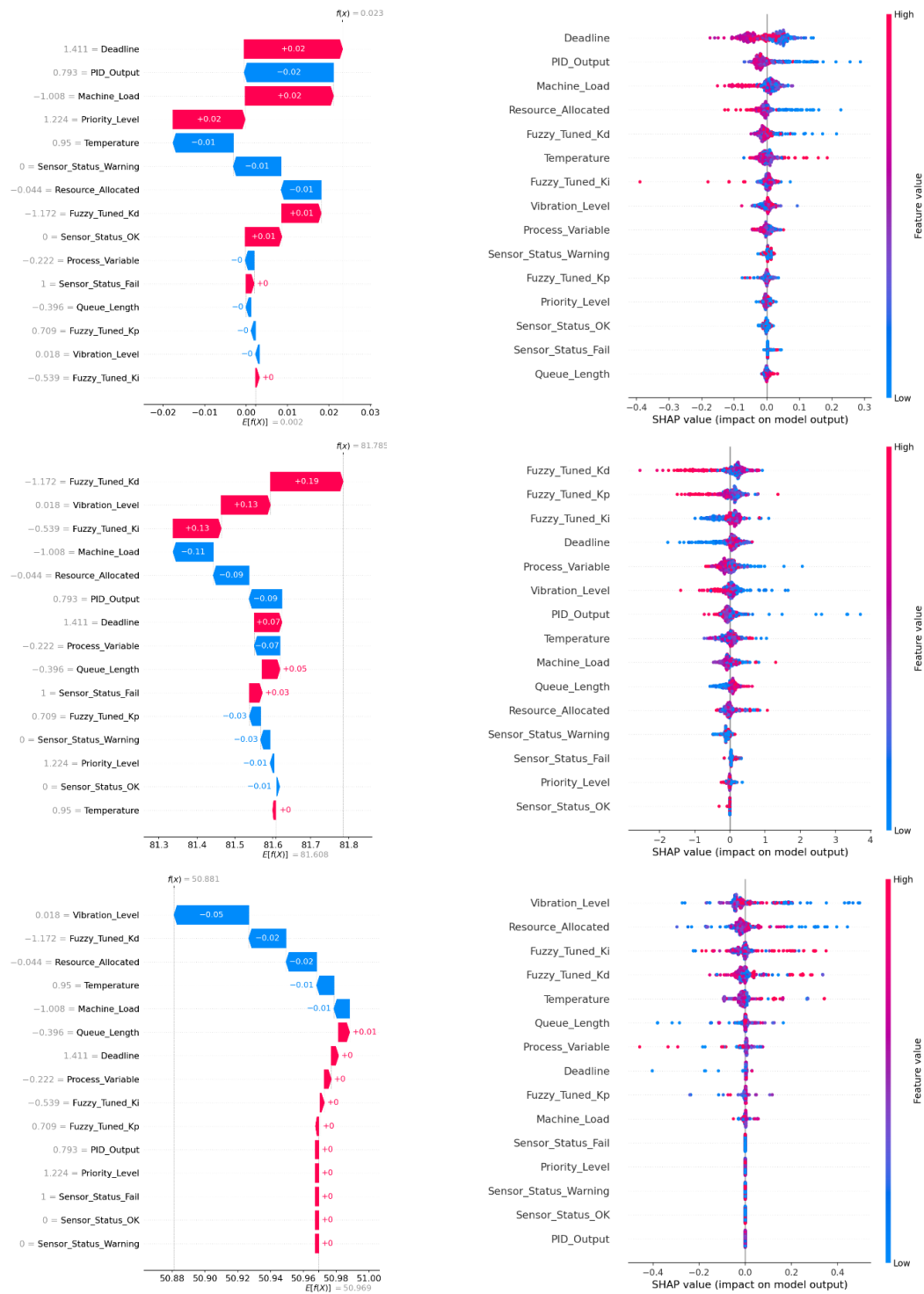


Figure 4. SHAP interpretability analysis of key operational drivers.

6. Conclusion

The black-box predictive models are still lagging behind in operational transparency while the field of smart manufacturing is rapidly growing. Traditional Industry 4.0 systems mostly prioritize accuracy at the expense of interpretability. This leads to failure in accounting for the inherent uncertainty in stochastic production environments. We addressed these gaps by proposing a framework that combines high-accuracy XGBoost model with explainable AI (XAI) and Uncertainty Quantification (UQ) to promote robustness in optimization.

The findings indicate that integrating uncertainty intervals significantly enhances the reliability of operational decision-making. The model demonstrated generalization capabilities on held-out data that achieved a Prediction

Interval (PI) coverage of approximately 90% across key performance indicators (KPIs). This reliability is an assurance for the fact that shop-floor managers are not merely presented with a single prediction, but are equipped with quantifiable confidence bounds necessary for risk-aware planning. The application of the decision optimization module also yielded visible improvements in production efficiency. The results show that the system successfully reduced completion time from 133.576 to 130.268. Simultaneously, energy usage was also decreased which clearly indicates that efficiency gains can be achieved with only negligible trade-offs in downtime. From a strategic perspective, the SHAP analysis yielded that process control variables are more significant drivers of performance than passive sensor status flags. This glass-box approach provides the stakeholders with the advantage of prioritizing flexible scheduling and aggressive resource allocation during high-demand periods with verified safety buffers.

This study contributes to the existing literature by integrating uncertainty quantification directly into the optimization loop. Such integration aids manufacturers to make decision regarding complex trade-offs between speed, energy, and risk. Future research may extend this methodology to real-time closed-loop control systems. This will further validate the scalability of interpretability-driven optimization in diverse industrial settings. Ultimately, this research facilitates the trade-off between model accuracy and operational transparency. Hence accuracy is not a necessary compromise in industrial applications. Since reliable efficiency gains can be secured without sacrificing safety, uncertainty-aware interpretability is not an optional feature, but a key requirement for the sustainable development of digital manufacturing.

References

- Ben-Tal, A., Ghaoui, L. E., and Nemirovski, A., *Robust Optimization*, Princeton University Press, 2009. <https://doi.org/10.1515/9781400831050>
- Lee, S., Jo, W., Koo, H., Kwak, J., Kim, H., Koo, J., and Kim, D., "A Framework for Tool Wear and Remaining Useful Life Prediction in Machining Process Using Monitoring Data and Machine Learning," *The International Journal of Advanced Manufacturing Technology*, vol. 138, no. 2, pp. 491–509, 2025. <https://doi.org/10.1007/s00170-025-15534-7>
- Levi, D., Gispan, L., Giladi, N., and Fetaya, E., "Evaluating and Calibrating Uncertainty Prediction in Regression Tasks," *Sensors*, vol. 22, no. 15, 5540, 2022. <https://doi.org/10.3390/s22155540>
- Lundberg, S. M. and Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," *arXiv preprint arXiv:1705.07874*, 2017. <https://doi.org/10.48550/arXiv.1705.07874>
- Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., Wang, Y., Zhang, X., and Hu, C., "Uncertainty Quantification in Machine Learning for Engineering Design and Health Prognostics: A Tutorial," *Mechanical Systems and Signal Processing*, vol. 205, 110796, 2023. <https://doi.org/10.1016/j.ymssp.2023.110796>
- Rudin, C., "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *arXiv preprint arXiv:1811.10154*, 2019. <https://doi.org/10.48550/arXiv.1811.10154>
- Tao, F., Qi, Q., Liu, A., and Kusiak, A., "Data-Driven Smart Manufacturing," *Journal of Manufacturing Systems*, vol. 48, pp. 157–169, 2018. <https://doi.org/10.1016/j.jmsy.2018.01.006>

Biographies

Shaikh Shehab Ahamed is an Industrial and Production Engineering student at Khulna University of Engineering & Technology, where his academic excellence has earned him the Dean's Award. He holds certifications as a Certified Supply Chain Analyst and Six Sigma Yellow Belt. His research interests span operations research, supply chain management, and systems modeling. He also serves as a peer reviewer for the Q1 journal Expert Systems with Applications.

Rashid Abrar Tanzil is a Bangladeshi engineering enthusiast currently pursuing his undergraduate studies in the Department of Industrial and Production Engineering at Khulna University of Engineering & Technology (KUET). Alongside his academic work, he has built a strong interest in optimization, production systems and data-driven decision-making. He has also explored modern technologies such as machine learning and explainable AI (XAI), applying them to problem-solving in areas like routing and operations research. His technical curiosity spans industrial engineering, manufacturing systems and computational optimization, with growing research interest in intelligent algorithms, supply chain analytics and sustainable production.

Md. Mahmudul Hasan Ratul is an undergraduate student in the Department of Industrial Engineering and Management at Khulna University of Engineering & Technology in Bangladesh. Alongside his academic coursework, he has developed a strong interest in machine learning, quality control and IoT. He is currently working on AI-integrated mechanical systems after a recent work in bearing fault diagnosis. Beyond academics

he is passionate about learning and using technology in real life to solve critical problems and has previously served in leadership roles for student organizations.

Prosenjit Sarkar is pursuing a Bachelor of Science degree in Department of Industrial Engineering and Management from Khulna University of Engineering and Technology (KUET), Bangladesh, where he is currently in the 3rd year, 1st semester. His research interests include operations research, optimization and simulation, and machine learning application in industrial system. He is skilled in programming languages such as C and Python and computer-aided design using SolidWorks as well. He works remotely on a part-time basis for a dispatch service agency in the United States. He aims to pursue a Ph.D. in a related field in the future.

Sawon Mukherjee is a student of the Department of Industrial Engineering and Management, Khulna University of Engineering & Technology, Khulna, currently pursuing his degree. He is a Certified SOLIDWORKS Professional (CSWP) with expertise in various engineering design applications and computer-based digital manufacturing tools. This experience has helped him gain higher research experience, including the implementation of Digital Twin technology in health infrastructure, focusing on efficiency improvement and decision-making. Machine learning, intelligent systems, and applications are his primary research interests. He dreams of pursuing higher education outside of his home country, contributing positively to advanced levels of research works.

Fawjia Islam is presently enrolled in B.Sc. Engineering from Khulna University of Engineering & Technology under the department of Industrial Engineering and Management. Her research interests include supply chain management, operations management, operations research.