



RESEARCH ARTICLE

OPEN ACCESS

DATA-DRIVEN ANALYSIS OF PRODUCT QUALITY DEFECTS IN MANUFACTURING SYSTEMS USING MACHINE LEARNING

Foysal Hossain^{1*}, S M Habibullah², Jahedul Islam Arif³, Maksud Miah⁴, Ariful Islam⁵, Fahim Rahman⁶, Al Akhir⁷

^{1,2,3}Industrial and Systems Engineering Lamar University, Beaumont, Texas, USA.

^{4,5,7}Business Analytics and Systems University of Bridgeport, Bridgeport, Connecticut 06604, USA

⁶University of Findlay, Findlay, Ohio 4584, USA

¹<https://orcid.org/0009-0009-9906-5376>, ²<https://orcid.org/0009-0008-8094-332X>, ³<https://orcid.org/0009-0009-0104-6565>,

⁴<https://orcid.org/0009-0003-0130-038X>, ⁵<https://orcid.org/0000-0001-6609-3324>,

⁶<https://orcid.org/0009-0002-5007-2187>, ⁷<https://orcid.org/0009-0001-3721-8439>

E-mail: mhossain71@lamar.edu, research.submission.quest@gmail.com, anikmail12@gmail.com, shabibullah@lamar.edu, jahedulislamarif01@gmail.com, marif6@lamar.edu, mamiah@my.bridgeport.edu, arislam@my.bridgeport.edu, rahmanf@findlay.edu, alakhir@my.bridgeport.edu

ARTICLE INFO

Article History

Received: February 08, 2026

Reviewed: March 10, 2026

Accepted: April 05, 2026

Published: May 29, 2026

Keywords:

Manufacturing quality defects,
Industrial engineering analytics,
Predictive maintenance,
Machine learning in
Manufacturing,
Quality improvement,
Process optimization.

ABSTRACT

The Product quality defects lead to substantial operational inefficiencies, financial losses, reputational costs, and customer dissatisfaction in contemporary manufacturing systems. Although Industry 4.0 environments generate extensive operational data, there remains limited empirical clarity regarding which factors most strongly drive defect occurrence and how such insights can support manufacturing decision-making. This study presents an empirical, data-driven assessment to evaluate operational, quality, and supply chain factors that influence product defects. The analysis is based on 3,240 manufacturing observations and 17 operational variables spanning production, maintenance, quality, efficiency, and supply-chain dimensions. Descriptive statistics and correlation analysis were conducted, followed by a Random Forest classifier to rank influential predictors of defect occurrence. Analysis of variance (ANOVA) was used to examine the effect of delivery delays. Model performance was evaluated using precision, recall, and F1-score. Maintenance hours and defect rate jointly accounted for approximately 44% of predictive contribution, with the model achieving precision and recall above 95% for defective items. Delivery delay showed no significant effect. The findings support the use of machine-learning (ML)-based analytics for data-driven quality improvement and process optimisation in manufacturing systems.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Product quality defects continue to be a significant and costly challenge in modern manufacturing systems, resulting in substantial financial losses, operational inefficiencies, and reduced customer satisfaction. Empirical evidence indicates that hidden quality costs, including rework, scrap, warranty claims, and process disruptions, can significantly erode organizational performance if left unaddressed [1]. In highly competitive industrial environments, product quality has therefore become a strategic imperative, closely linked to customer loyalty, brand reputation, and long-term sustainability [2]. As manufacturing systems grow increasingly complex, identifying and mitigating the drivers of quality defects has emerged as a critical managerial and analytical concern [3]. The advent of Industry 4.0 has transformed manufacturing by integrating cyber-physical systems, the Industrial Internet of Things (IoT), and real-time data analytics into production environments [4].

These developments enable the continuous monitoring of equipment, processes, and quality indicators, creating new opportunities for predictive maintenance and data-driven quality control [5]. However, the availability of high-volume and high-dimensional operational data also introduces analytical challenges, particularly in isolating the most influential defect drivers and translating analytical outputs into actionable insights for quality improvement [6]. Traditional quality management approaches, such as Statistical Process Control, Total Quality Management, and Six Sigma, remain valuable but are often reactive and constrained by assumptions of linearity and process stability that may not hold in digitally enabled manufacturing systems [7]. Recent advances in artificial intelligence (AI) and machine learning (ML) have expanded the analytical capabilities of quality assurance by enabling the modeling of nonlinear relationships, interaction effects, and dynamic process behavior [8]. ML-based quality prediction models have demonstrated strong potential in applications such as tool wear detection, defect classification, and failure prediction [9]. In parallel, Quality 4.0 frameworks emphasize the integration of analytics and automation to support zero-defect manufacturing objectives [10].

Despite these advances, much of the existing literature focuses on isolated technical use cases or individual process variables, which limits the generalizability of findings across broader manufacturing contexts. Importantly, prior studies have often examined quality determinants in isolation, such as maintenance practices, human factors, or supply chain responsiveness [11]. While these studies provide valuable insights, they rarely offer a comparative, multivariate assessment of operational, quality, and supply-chain factors within a unified analytical framework. Measurement challenges in empirical manufacturing research further complicate this issue, as heterogeneous data sources and inconsistent operational definitions can obscure the relative importance of competing predictors [12]. As a result, a gap remains in the literature regarding how multiple manufacturing variables jointly influence defect occurrence and how these relationships can be leveraged for data-driven quality improvement. From an information systems and industrial engineering perspective, there is a growing need to shift from descriptive quality reporting toward predictive and decision-oriented analytics that support proactive intervention. Predictive quality models based on machine-learning techniques, such as Random Forests and ensemble learning, have shown promise in enabling early detection of quality deterioration and supporting continuous improvement initiatives [13].

However, empirical evidence comparing the relative influence of internal operational factors and external supply-chain variables remains limited, particularly in studies that use secondary manufacturing data and have imbalanced defect distributions [14]. Against this backdrop, the present study conducts an empirical assessment of factors influencing product quality defects using secondary manufacturing data and predictive machine-learning techniques. Drawing on a dataset of 3,240 observations and 17 operational variables, the study applies descriptive analytics, Random Forest classification, and inferential testing to identify and prioritize influential predictors of defect occurrence. By addressing class imbalance and feature importance, the study adopts a data-driven approach to quality management that emphasizes interpretability and operational relevance. The contributions of this study are threefold. First, it provides a comparative evaluation of operational, quality, and supply-chain factors by quantifying their relative predictive importance in defect detection. Second, it demonstrates the applicability of ML-based analytics as a decision-support tool for manufacturing quality management in Industry 4.0 environments. Third, by leveraging secondary data, the study highlights the feasibility of cost-effective, data-driven quality analysis that can be adopted by manufacturing organizations seeking to enhance quality performance and reduce defect-related losses [15].

II. METHODOLOGY

II.1 RESEARCH DESIGN

The current study employed a quantitative empirical research design, grounded in secondary data analysis, to investigate the operational and information-driven determinants of product quality defects in manufacturing systems (Figure 1). The methodological framework is explicitly oriented toward decision support and BI by integrating descriptive analytics, inferential statistics, and machine learning-based predictive modelling. The objective is not only to identify statistically significant drivers of defects, but also to develop an interpretable, data-driven model capable of supporting intelligent process optimization and managerial decision-making within Industry 4.0 environments.

II.2 DATASET DESCRIPTION

The data used were obtained from Kaggle (<https://www.kaggle.com/datasets/rabieelkharoua/predicting-manufacturing-defects-dataset>), and consisted of 3,240 manufacturing observations and 17 operational variables, spanning production, maintenance efficiency, quality, and supply-chain dimensions. The two target variables were binary (defect, where 0 indicated a non-defect and 1 indicated a defect), reflecting whether a manufactured item failed to meet quality standards. Exploratory analysis revealed a pronounced class imbalance, with approximately 84% of observations classified as defective. This imbalance was intentionally retained to reflect the underlying operational reality of the manufacturing context, where persistent quality failures represent a critical managerial risk. From a BI perspective, prioritising defect detection under such conditions is consistent with risk-averse quality management strategies, where false negatives are substantially more costly than false positives.

II.3 DATA PREPROCESSING

Data integrity checks confirmed that no values were missing across all variables. Descriptive statistics and exploratory visualizations were used to assess distributional properties, identify potential outliers, and evaluate pairwise correlations among predictors. Although Random Forest algorithms are robust to scale differences, variables were also reviewed for extreme values to ensure numerical stability. The dataset was partitioned into training and test sets using a reproducible random split, ensuring consistent results across model evaluations. This separation supports out-of-sample validation and enhances the reliability of the reported predictive performance.

III. ANALYTICAL METHODS

III.1 DESCRIPTIVE ANALYSIS AND CORRELATION ANALYSIS

Descriptive statistics, including means, standard deviations, and ranges, were used to summarize the continuous variables. Distributional properties, potential outliers, and data dispersion were examined using histograms, density plots, and boxplots. Pearson correlation coefficients were computed and visualized using heatmaps to assess linear associations among operational indicators and defect status. This analysis supports the identification of variables with potential relevance for subsequent predictive modeling and quality analysis.

III.2 RANDOM FOREST CLASSIFICATION

The primary predictive technique used for defect classification was a Random Forest classifier, an ensemble of decision trees. This method was selected due to its ability to capture nonlinear relationships, interaction effects, and class imbalance, which are commonly observed in manufacturing data [16]. In the Random Forest classifier, the final prediction is obtained by aggregating the outputs of multiple decision trees using majority voting, as expressed in Equation (1).

$$y = \text{majority}\{T_1(x), T_2(x), \dots, T_n(x)\} \quad (1)$$

where $\hat{y}$ is the predicted class label, x denotes the input feature vector, $T_i(x)$ represents the prediction of the i -th decision tree, and n is the total number of trees in the Random Forest. The final output is obtained through majority voting. Feature importance was assessed using the mean decrease in Gini impurity, which measures each predictor's relative contribution to classification performance. These importance values are interpreted as indicators of predictive relevance rather than causal effects. Feature importance was derived from the mean decrease in Gini impurity, calculated as in Equation (2).

$$\text{Gini} = 1 - \sum_{i=1}^C p_i^2 \quad (2)$$

where *Gini* is the impurity measure, C denotes the number of classes, and p_i represents the class proportion at a node.

Model performance was evaluated using precision, recall, and F1-score, with particular emphasis on recall for the defective class to reduce the risk of undetected quality failures. Model performance was evaluated using precision and recall, which quantify the accuracy of defect detection and the ability to identify defective items, respectively.

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

III.3 INFERENCE TESTING (ANOVA TEST)

To complement the predictive analysis, a one-way Analysis of Variance (ANOVA) was conducted to examine whether delivery delay has a statistically significant effect on defect rates. This inferential test provides supplementary evidence regarding the influence of supply-chain timing relative to internal operational factors, using a significance level of $\alpha = 0.05$.

III.4 TOOLS AND SOFTWARE

All analyses were conducted using the Python programming language. The libraries pandas and NumPy were used for data handling and preprocessing, Matplotlib and Seaborn for data visualization, scikit-learn for predictive modeling, and SciPy for statistical testing.

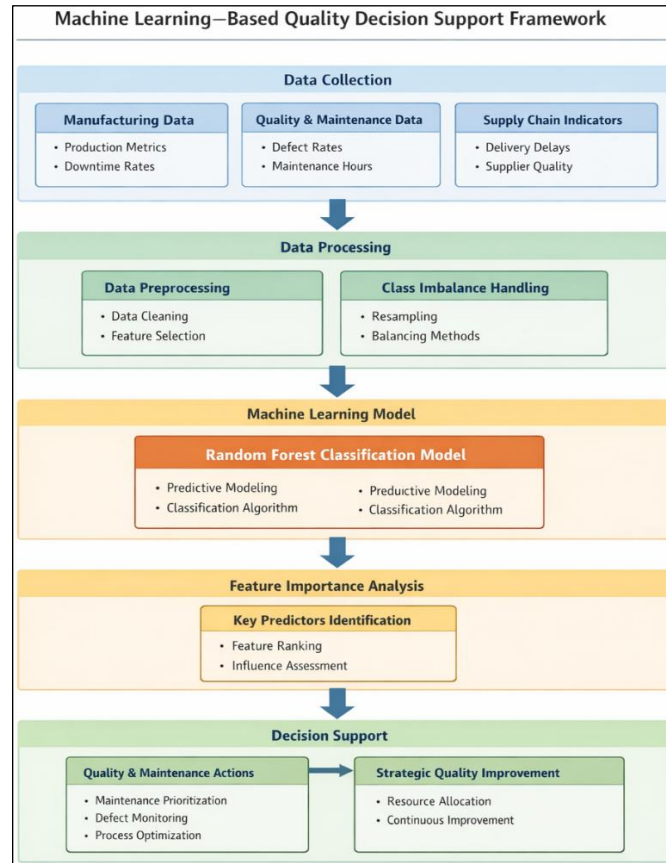


Figure 1: Illustrates the proposed machine learning–based quality decision support framework employed in this study.
Source: Authors, (2026).

IV. RESULTS

IV.1 DESCRIPTIVE CHARACTERISTICS OF THE DATASET

The manufacturing dataset comprises 3,240 observations across 17 variables, including production volume, cost structures, supplier quality, maintenance activity, workforce productivity, and energy usage. The key variables are presented in Table 1. The average production volume was 548.52 (SD: 262.40), with the highest volume ranging from 100 to 999 units, indicating significant variation. The average sample cost of production was \$12,423.02 (SD = \$4,308.05), with a range of \$5,000.17 to \$19,993.37, indicating varying production levels.

Table 1: Descriptive Statistics of Key Variables in the Manufacturing Dataset.

Variable	Mean	Std. Dev.	Min	Max
Production Volume	548.52	262.40	100	999
Production Cost	12423.02	4308.05	5000.17	19993.37
Supplier Quality	89.83	5.76	80.00	99.99
Delivery Delay	2.56	1.71	0	5
Defect Rate	2.75	1.31	0.50	4.99
Quality Score	80.13	11.61	60.01	99.99
Maintenance Hours	11.48	6.87	0	23
Downtime Percentage	2.50	1.44	0.00	5.00
Inventory Turnover	6.02	2.33	2.00	10.00
Stockout Rate	0.05	0.03	0.00	0.10
Worker Productivity	90.04	5.72	80.00	99.99
Safety Incidents	4.59	2.90	0	9
Energy Consumption	2988.49	1153.42	1000.72	4997.07
Energy Efficiency	0.30	0.12	0.10	0.50
Additive Process Time	5.47	2.60	1.00	10.00
Additive Material Cost	299.52	116.38	100.21	499.98
Defect Status	0.84	0.37	0	1

Source: Authors, (2026).

The dataset showed that supplier quality is high ($M = 89.83$, $SD = 5.76$, range = 80-99.99) and the defect rates are low ($M = 2.75$, $SD = 1.31$, range = 0.50-4.99). The overall product quality was 80.13 ($SD = 11.61$), and the average delivery delay was 2.56 days ($SD = 1.71$). Operational indicators included the maintenance hours ($M = 11.48$, $SD = 6.87$), downtime ($M = 2.50$, $SD = 1.44$), inventory turnover ($M = 6.02$, $SD = 2.33$), and worker productivity ($M = 90.04$, $SD = 5.72$). The average safety incident rate was 4.59 ($SD = 2.90$), and the average energy consumption was 2,984.49 units ($SD = 1,153.42$). The average energy efficiency stood at 0.30 ($SD = 0.12$). The time averaged 5.47 hours ($SD = 2.60$) in the additive process, and the cost of materials amounted to \$299.52 ($SD = \$116.38$). The average defect level ($M = 0.84$, $SD = 0.37$) indicated that 84% of the products were defective. From a BI standpoint, these descriptive patterns underscore the need for continuous monitoring systems capable of integrating cost, maintenance, and quality metrics to support early intervention

IV.2 DISTRIBUTION ANALYSIS AND CLASS IMBALANCE OF KEY VARIABLES

The distribution of Defect Status demonstrates a pronounced imbalance (Figure 2). Non-defect status products (Defect Status = 0) comprise approximately 16 per cent of the dataset (520 cases), and defect status products (Defect Status = 1) comprise approximately 84 per cent (2,720 cases). This pattern signals a systemic quality issue rather than sporadic failures, reinforcing the relevance of predictive analytics for proactive quality management. Rather than undermining model validity, this imbalance reflects a realistic decision environment in which defect detection is a dominant managerial concern.

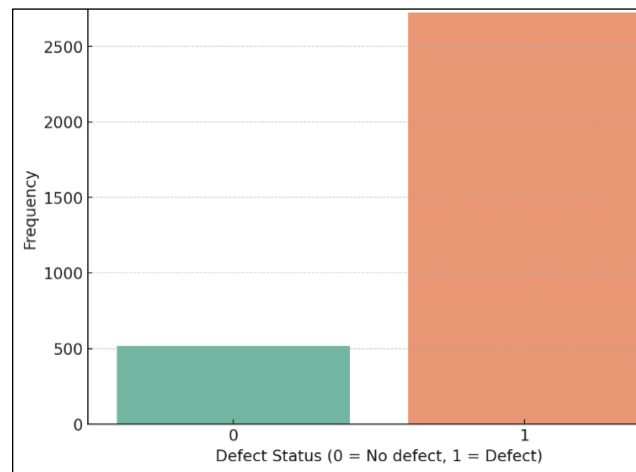


Figure 2: Distribution of defect status in the manufacturing dataset.
Source: Authors, (2026).

Figures 3–5 provide exploratory validation of key continuous variables used in subsequent modeling. Production volume (Figure 3) exhibits a broadly uniform distribution across operational scales, indicating the absence of artificial batching or scale-induced bias. The defect rate distribution (Figure 4) is centered around moderate defect levels with limited extreme values, supporting the use of probabilistic rather than threshold-based classification. The production cost (Figure 5) exhibits stable dispersion across the observed range, indicating consistent cost structures and justifying its inclusion as a control variable rather than a primary quality driver. Collectively, these distributions confirm the suitability of the dataset for predictive business intelligence modeling.

IV.3 RELATIONSHIP AND CORRELATION ANALYSIS

Correlation analysis reveals that defect status is positively associated with defect rate and maintenance hours, and negatively associated with overall quality score. Production volume and production cost are strongly correlated, reflecting scale effects in manufacturing operations. Maintenance hours and defect rates are also positively correlated with downtime, suggesting a compounding of operational inefficiencies. These relational patterns are visually reinforced in Figure 6, which illustrates the joint behaviour of defect status, defect rate, and maintenance hours, and in Figure 7, which summarises pairwise correlations across all operational and quality variables. Collectively, these relationships provide actionable insights by identifying variables that can function as leading indicators in quality dashboards and decision-support systems. These relationships provide actionable insights by identifying variables that can function as leading indicators in quality dashboards and decision-support systems.

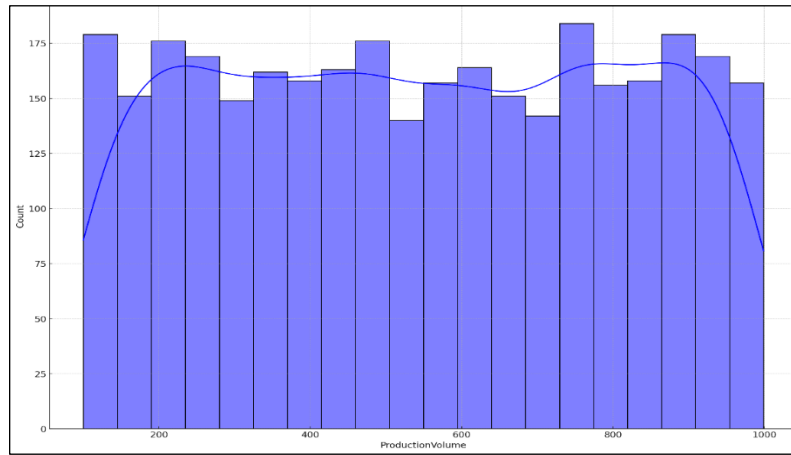


Figure 3: Distribution of production volume in the manufacturing dataset. The figure illustrates the dispersion of production volume across manufacturing observations, confirming broad operational variability and the absence of scale-specific clustering that could bias predictive modeling.

Source: Authors, (2026).

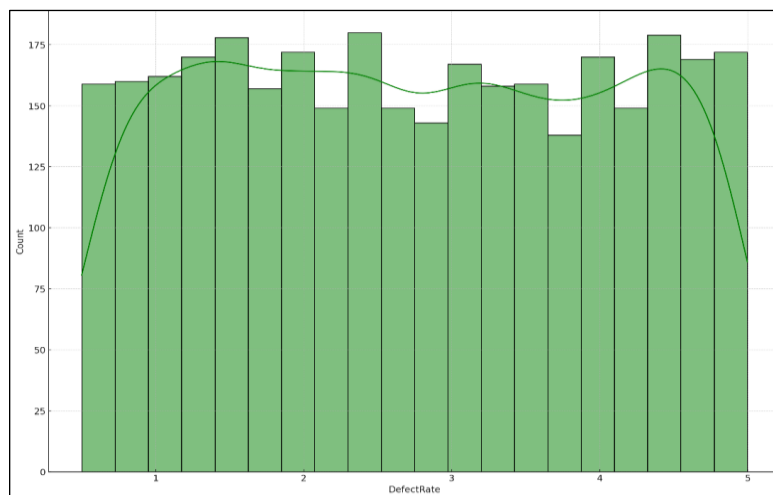


Figure 4: Distribution of production cost in the manufacturing dataset. The distribution exhibits a continuous spread of defect rates, with limited extreme values, supporting the use of probabilistic classification methods over rule-based thresholds for defect prediction.

Source: Authors, (2026).

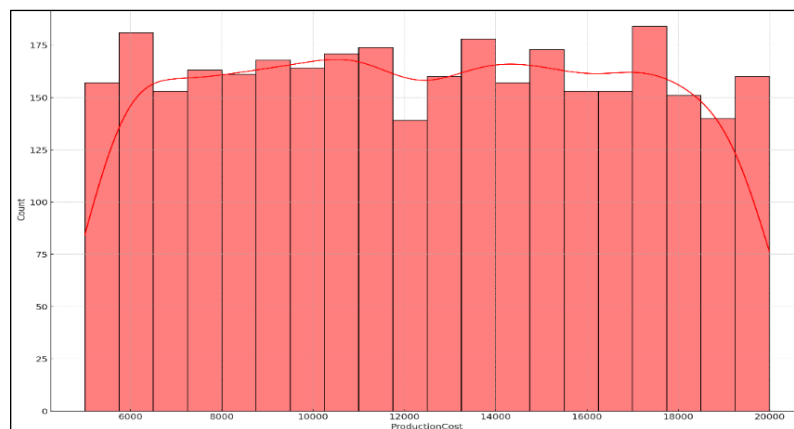


Figure 5: Distribution of production costs in the manufacturing dataset. This figure depicts the variability of production costs across observations, indicating stable dispersion consistent with output heterogeneity and justifying its inclusion as a control variable in the predictive analysis.

Source: Authors, (2026).

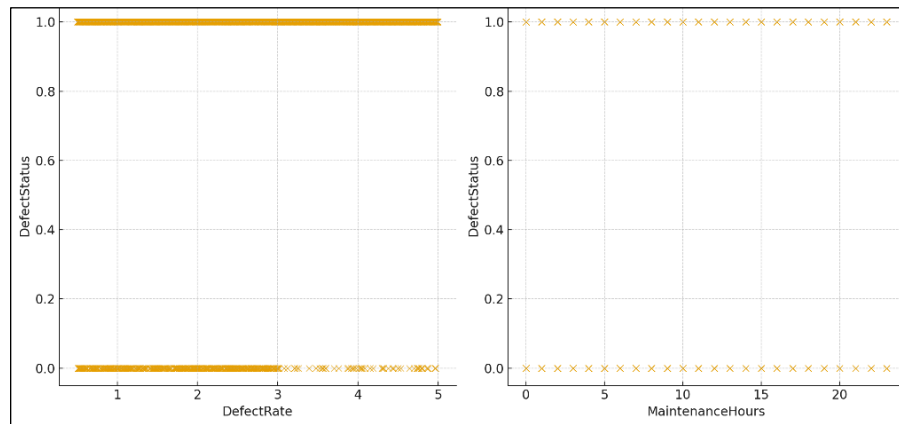


Figure 6: Relationship between defect status, defect rate, and maintenance hours. The figure visualizes how defect status varies with defect rate and maintenance hours, highlighting their strong discriminative patterns and supporting their prioritization as key predictors in the defect classification model.

Source: Authors, (2026).

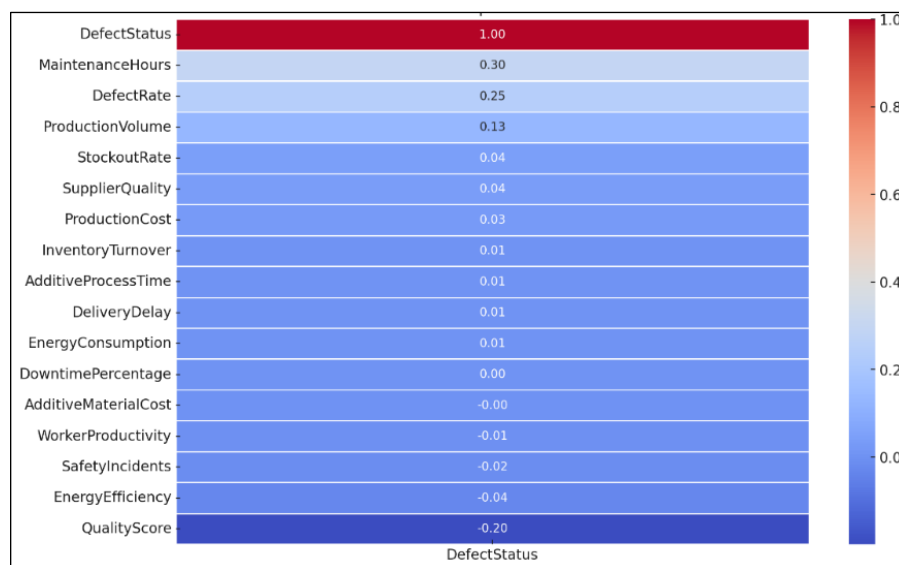


Figure 7: Correlation heatmap of operational and quality variables in the manufacturing dataset. The heatmap summarizes pairwise Pearson correlations among study variables, indicating the strongest associations with defect-related outcomes and informing the selection and interpretation of predictors for the predictive analytics model.

Source: Authors, (2026).

IV.4 DIAGNOSTIC ASSESSMENT OF VARIABLE DISTRIBUTIONS AND MULTIVARIATE STRUCTURE

Figures 8 and 9 provide diagnostic support for the predictive analysis. Figure 8 illustrates the distributional properties and potential outliers of key continuous variables, confirming acceptable dispersion without extreme anomalies that could distort model estimation. There are limited extremes in production cost. The defect rate is concentrated around the median, with outliers at both ends. Maintenance hours are more varied and exhibit greater outliers, suggesting that some units need additional maintenance. Figure 9 presents multivariate pairwise relationships among production volume, production cost, defect rate, and maintenance hours, highlighting nonlinear patterns and interaction effects. Together, these diagnostics support the suitability of tree-based machine learning models for capturing complex relationships in manufacturing quality data. The correlation between volume and cost is strongly linear. On the contrary, the relationship between defect rate, volume, and cost is relatively weak and multifactorial. The maintenance hours appear to depend on volume and cost, suggesting that other variables may also influence them, such as equipment age or usage intensity.

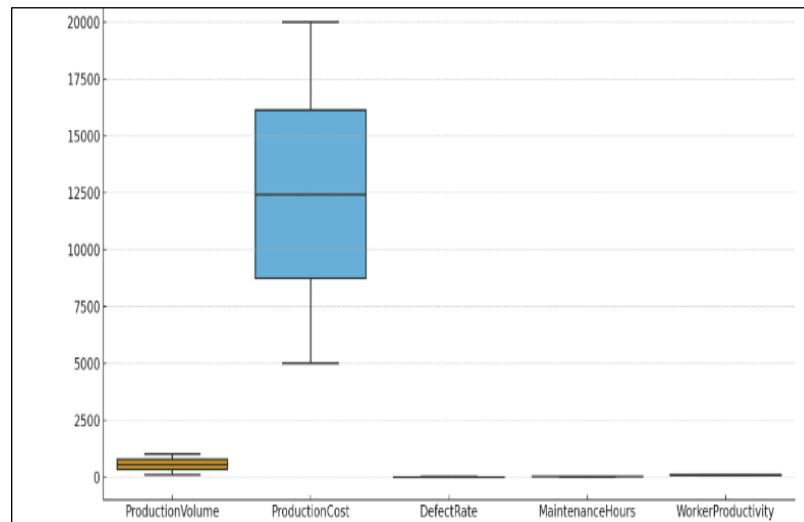


Figure 8: Boxplot of Important Variables in the Manufacturing Data. The boxplot is used to display the distributions of key variables, including production volume, production cost, defect rate, and maintenance hours. It allows for identifying potential outliers and delineating the spread of each variable, providing a view of each variable's variability within the dataset.

Source: Authors, (2026).

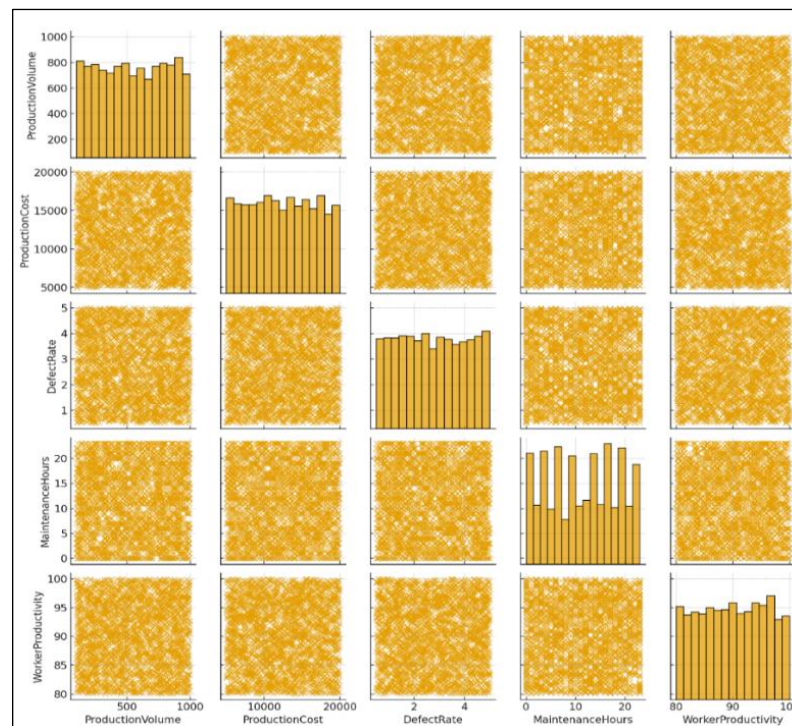


Figure 9: Key Variables Pair Plot in the Manufacturing Data Set. The pair plot provides a graphical representation of the inter-variable relationships, including production volume, production cost, defect rate, and maintenance time. It helps identify correlations between predictors and their interactions, which are essential for a detailed understanding of the determinants that affect the defect status.

Source: Authors, (2026).

IV.5 FEATURE IMPORTANCE FROM RANDOM FOREST MODELING

Feature importance rankings derived from the Random Forest model are presented in Table 2. Maintenance hours emerge as the most influential predictor (importance = 0.23), followed by defect rate (0.21), quality score (0.14), and production volume (0.11). Collectively, these variables account for approximately 44% of the model's predictive power. The remaining variables, including additive material cost, energy efficiency, supplier quality, stockout rate, downtime, additive process time, inventory turnover, worker productivity, production cost, and energy consumption, showed little significance (around 0.03), suggesting that internal operational conditions have a greater impact on quality outcomes in this dataset. Safety incidents and delivery delays had the least importance, with a score of 0.01 each, indicating they did not have a significant impact on defect status. These findings should be interpreted as predictive priorities for decision support, not as evidence of causal mechanisms.

Table 2: Ranking of the Defect Status in the Manufacturing dataset based on order of importance

Feature	Importance
Maintenance Hours	0.23
Defect Rate	0.21
Quality Score	0.14
Production Volume	0.11
Additive Material Cost	0.03
Energy Efficiency	0.03
Supplier Quality	0.03
Stockout Rate	0.03
Downtime Percentage	0.03
Additive Process Time	0.03
Inventory Turnover	0.03
Worker Productivity	0.03
Production Cost	0.03
Energy Consumption	0.03
Safety Incidents	0.01
Delivery Delay	0.01

Source: Authors, (2026).

IV.6 PERFORMANCE OF THE PREDICTIVE MODELS

The Random Forest classifier demonstrates strong predictive performance, particularly for defective products. Precision (0.96), recall (0.99), and F1-score (0.98) indicate that the model is highly effective in identifying quality failures, which is critical in high-risk manufacturing environments (Table 3). In non-defective products, precision is 0.94, recall is 0.78, and F1-score is 0.85, which is less than the recall in class imbalance. Therefore, performance for non-defective products is comparatively lower, reflecting the inherent class imbalance and reinforcing the model's conservative bias toward defect detection. This bias is desirable, as the cost of undetected defects typically exceeds the cost of false alarms.

Table 3: Random Forest Model Evaluation.

Metric	Defect = 1 (Defective Products)	Defect = 0 (Non-Defective Products)
Precision	0.96	0.94
Recall	0.99	0.78
F1-Score	0.98	0.85

Source: Authors, (2026).

IV.7 STATISTICAL TESTING: EFFECT OF DELAY AT DELIVERY ON DEFECT RATE

Table 4 presents the ANOVA results for the impact of delivery delay on the defect rate. Between groups sum of squares = 3.2209 (df = 4); within groups sum of squares = 579.7014 (df = 3,235). The mean squares are 0.8052 and 0.1793, respectively, yielding an F-statistic of 1.2047 and a p-value of 0.3042. This difference is not statistically significant ($p = 0.05$), indicating that delivery delay does not affect the defect rate. This is also associated with its low feature importance, indicating that resources should be allocated to maintenance, defect rates, and quality scores rather than to delivery delays to reduce defect rates.

Table 4: ANOVA Table for Defect Rate across Delivery Delay groups.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	p-Value
Between Groups	3.2209	4	0.8052	1.2047	0.3042
Within Groups	579.7014	3235	0.1793		
Total	582.9223	3239			

Source: Authors, (2026).

V. DISCUSSION

This empirical study investigates the critical predictors of product-quality defects in manufacturing systems by integrating secondary data analysis with ML-based predictive modeling. The results demonstrate that equipment-related factors, particularly maintenance hours and observed defect rates, are the most influential determinants of defect occurrence, collectively accounting for approximately 44% of the model's predictive contribution.

This finding reinforces longstanding concerns in industrial engineering regarding measurement validity and operational prioritization, where inaccurate or delayed maintenance indicators can obscure true performance drivers and degrade quality outcomes [17].

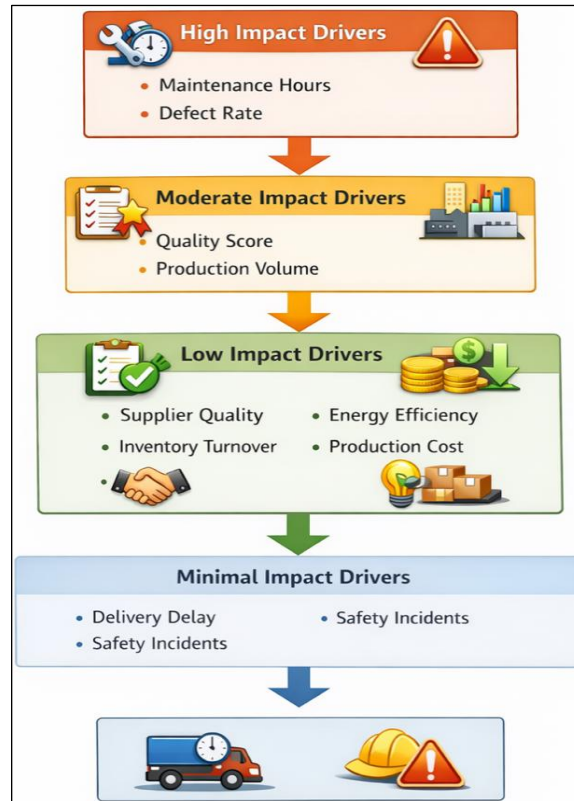


Figure 10: Feature-importance-based hierarchy of quality defect drivers identified using the Random Forest model.
Source: Authors, (2026).

By grouping the variables into hierarchical impact levels, Figure 10 highlights the dominance of maintenance hours and defect rate over other operational and supply-chain factors. This hierarchy supports the interpretation that internal operational conditions play a more decisive role in defect occurrence than external timing-related variables. The prominence of maintenance hours as the leading predictor (importance = 0.23) aligns strongly with the principles of TPM, reliability engineering, and human-centered maintenance frameworks, which emphasize proactive equipment upkeep as a mechanism for reducing process variability and preventing defect propagation [18]. Prior studies on human factors in maintenance and operational discipline suggest that inadequate maintenance planning and execution can amplify failure risks, particularly in complex, high-utilization manufacturing environments [18],[19]. From an operational perspective, increased maintenance activity contributes to improved process stability, reduced unplanned downtime, and enhanced equipment availability, thereby lowering the propensity for defect generation across production stages [18],[19]. The strong predictive contribution of defect rate (importance = 0.21) further underscores the reinforcing nature of quality failures, in which persistent defects signal deeper systemic inefficiencies rather than isolated process deviations. This observation is consistent with research on quality-cost dynamics, which demonstrates that hidden quality costs, such as rework, scrap, and warranty claims, often accumulate silently until defect levels reach critical thresholds [20]. Similar findings in quality culture research emphasize that organizations with weak feedback mechanisms tend to normalize recurring defects, allowing inefficiencies to persist over time [21]. Consequently, continuous defect monitoring serves not only as a quality control mechanism but also as an early warning indicator of declining operational health.

Consistent with prior empirical research, the findings confirm a robust association between preventive maintenance practices and improved product quality outcomes [22],[23]. Earlier work on predictive modeling in manufacturing systems shows that maintenance intensity and equipment condition are among the most reliable predictors of downstream performance, particularly when combined with operational efficiency indicators [23]. This study extends that literature by quantifying the relative importance of maintenance activities relative to a broader set of operational, quality, and supply-chain variables using ML-based feature-importance analysis, thereby offering a more integrated perspective than traditional univariate or siloed approaches [17]. In contrast to some supply-chain-oriented studies, delivery delay does not exhibit a statistically significant relationship with defect rate in the present analysis. This finding suggests that internal operational controls may outweigh external timing factors in determining quality outcomes within the examined context. Prior research indicates that buffering strategies, incoming quality inspections, and inventory policies can mitigate the quality risks associated with delivery variability, effectively decoupling logistics delays from production quality performance [24]. Studies on supply-chain responsiveness similarly argue that while delivery reliability affects customer satisfaction, its direct impact on defect formation may be secondary when robust internal quality systems are in place [25]. The Random Forest classifier demonstrates strong predictive performance for defective items, achieving high recall (0.96) and precision (0.99), which supports its suitability for quality-focused decision-support applications.

These performance levels are comparable to those reported in prior machine learning studies on manufacturing quality prediction, predictive maintenance, and operational risk assessment [26-28]. In particular, ensemble-based methods such as Random Forest and gradient boosting have been shown to outperform single-model approaches when dealing with nonlinear interactions and heterogeneous manufacturing data [29],[27]. Beyond classification accuracy, the model's ability to prioritize defect detection is especially valuable in high-risk production environments, where the cost of undetected defects significantly exceeds the cost of false alarms [30]. The limited influence of supply-chain timing variables further suggests that internal process optimization may yield greater quality improvements than external responsiveness initiatives alone. Prior research on quality management systems and Industry 4.0 performance measurement supports this conclusion, emphasizing that maintenance effectiveness, workforce discipline, and process monitoring exert stronger control over defect prevention than logistics acceleration strategies [31],[32]. Accordingly, quality-improvement initiatives may benefit from prioritizing predictive maintenance, continuous defect monitoring, and process control rather than focusing solely on supply-chain speed [18],[33],[34]. Although the dataset exhibits pronounced class imbalance, this characteristic reflects a realistic manufacturing environment characterized by persistent quality challenges rather than sporadic failures.

Prior studies on imbalanced learning in manufacturing and industrial analytics argue that such distributions necessitate risk-sensitive modeling approaches that emphasize recall for defective items [30]. This approach aligns with practical quality management objectives, where the operational and reputational costs of undetected defects outweigh the consequences of conservative classification decisions [31],[35],[36]. At a conceptual level, the study contributes to data-driven quality management by demonstrating how machine learning models can uncover nonlinear and interaction-driven relationships that traditional statistical methods may overlook [29]. The findings also challenge the assumption that supply-chain timing is a dominant determinant of quality outcomes, instead emphasizing the primary role of internal operational conditions—such as maintenance intensity, defect history, and quality scores—in shaping defect risk [21],[33]. Finally, the study affirms the viability of secondary data analysis as a cost-effective and scalable approach for generating actionable insights into manufacturing quality improvement within Industry 4.0 environments, particularly when supported by interpretable ML-based decision-support systems [37],[38]. In terms of industrial applicability, the proposed ML-based quality decision support framework is directly applicable to industrial manufacturing environments where operational, maintenance, and quality data are routinely collected. By identifying high-impact defect drivers, such as maintenance intensity and defect history, the framework enables practitioners to prioritize preventive maintenance, monitor defect risk, and allocate quality-improvement resources more effectively. The use of interpretable feature importance further supports practical adoption by engineers and managers without requiring advanced machine learning expertise.

VI. LIMITATIONS

This study has certain limitations. The analysis relies on secondary data from a single manufacturing context, which may limit generalizability across industries. In addition, feature importance reflects predictive relevance rather than causal relationships. Future research may extend the framework using multi-source industrial datasets, alternative machine learning models, or real-time deployment scenarios.

VII. CONCLUSION

This study employed secondary manufacturing data and ML-based analytics to identify the principal drivers of product quality defects in contemporary manufacturing systems. The findings indicate that equipment-related factors, particularly observed defect indicators and maintenance intensity, are the most influential predictors of defect occurrence, collectively accounting for nearly half of the model's predictive contribution. These results highlight the critical role of equipment condition and maintenance discipline in sustaining quality performance within Industry 4.0 environments. The Random Forest model demonstrated strong predictive capability, achieving precision and recall levels exceeding 95%, supporting its suitability for integration into quality monitoring and decision-support applications.

In contrast, delivery delay did not exhibit a statistically significant impact on defect rates, reinforcing the conclusion that internal operational conditions exert a greater influence on product quality outcomes than external supply-chain timing factors in the examined context. Overall, the study provides empirical support for data-driven quality management approaches that emphasize predictive maintenance, defect monitoring, and process optimization. By demonstrating the value of ML-enabled analytics in identifying high-risk operational conditions, the findings provide practical guidance for manufacturers seeking to enhance quality performance and reduce defect-related costs in digitally enabled production environments.

VIII. AUTHOR CONTRIBUTIONS

Conceptualization: Foysal Hossain, S M Habibullah, Jahedul Islam Arif, Maksud Miah, Ariful Islam, Fahim Rahman and Al Akhri.

Methodology: Foysal Hossain and S M Habibullah.

Investigation: Foysal Hossain and S M Habibullah.

Discussion of results: Foysal Hossain, S M Habibullah, Md Jahedul Islam Arif.

Writing – Original Draft: Foysal Hossain.

Writing – Review and Editing: Foysal Hossain and S M Habibullah.

Resources: S M Habibullah.

Supervision: S M Habibullah, Jahedul Islam Arif.

Approval of the final text: Foysal Hossain, S M Habibullah, Jahedul Islam Arif, Maksud Miah, Ariful Islam, Fahim Rahman and Al Akhri.

IX. REFERENCES

- [1] A. Bousdekis, K. Lepenioti, D. Apostolou, and G. Mentzas, "Data analytics in Quality 4.0: Literature review and future research directions," *Int. J. Comput. Integr. Manuf.*, vol. 36, no. 5, pp. 678–701, 2023.
- [2] A. D. Pouliezios and G. S. Stavrakakis, *Real-Time Fault Monitoring of Industrial Processes*, vol. 12. Springer Science & Business Media, 2013.
- [3] A. K. Kalusivalingam, A. Sharma, N. Patel, and V. Singh, "Leveraging reinforcement learning and predictive analytics for continuous improvement in smart manufacturing," *Int. J. AI ML*, vol. 3, no. 9, 2022.
- [4] A. Keshar, "Advancing Industrial IoT and Industry 4.0 through digital twin technologies: A comprehensive framework for intelligent manufacturing, real-time analytics and predictive maintenance," *World J. Adv. Eng. Technol. Sci.*, vol. 14, pp. 228–240, 2025.
- [5] A. Q. Md et al., "A review on data-driven quality prediction in the production process with machine learning for industry 4.0," *Processes*, vol. 10, no. 10, p. 1966, 2022.
- [6] A. S. Moh'd Anwer, "An investigation of transportation logistics strategy on manufacturing supply chain responsiveness in developing countries: The mediating role of delivery reliability and delivery speed," *Heliyon*, vol. 8, no. 11, p. e11283, 2022.
- [7] B. P. Lovatti et al., "Use of random forest in the identification of important variables," *Microchem. J.*, vol. 145, pp. 1129–1134, 2019.
- [8] Chuang, C.T. (2006). A quality-based predictive methodology for manufacturing systems. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 13(3):287–298.
- [9] D. Kibira, K. C. Morris, and S. Kumaraguru, "Methods and tools for performance assurance of smart manufacturing systems," *J. Res. Natl. Inst. Stand. Technol.*, vol. 121, pp. 282, 2016.
- [10] D. Wu, C. Jennings, J. Terpenney, R. X. Gao, and S. Kumara, "Machine learning algorithms for tool wear prediction using random forests," *J. Manuf. Sci. Eng.*, vol. 139, no. 7, p. 071018, 2017.
- [11] F. Petropoulos et al., "Operations & supply chain management: principles and practice," *Int. J. Prod. Res.*, pp. 1–184, 2025.
- [12] F. Psarommatis and V. Azamfirei, "Zero defect manufacturing: A complete guide for advanced and sustainable quality management," *J. Manuf. Syst.*, vol. 77, pp. 764–779, 2024.
- [13] F. Psarommatis, G. May, and V. Azamfirei, "Zero defect manufacturing in 2024: A holistic literature review for bridging the gaps and forward outlook," *Int. J. Prod. Res.*, pp. 1–37, 2024.
- [14] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "Learning pattern classification tasks with imbalanced data sets," *Pattern Recognit.*, vol. 10, pp. 1322–1328, 2009.
- [15] Gonzalez Santacruz, E., Romero, D., Noguez, J., & Wuest, T, "Integrated quality 4.0 framework for quality improvement based on Six Sigma and machine learning techniques towards zero-defect manufacturing," *The TQM Journal*, 37(4), 1115-1155, (2025)
- [16] Guinet, A. (1993). Scheduling and planning in manufacturing systems: A decision support approach. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 1(2):115–124.
- [17] K. K. Boyer and M. Pagell, "Measurement issues in empirical research: improving measures of operations strategy and advanced manufacturing technology," *J. Oper. Manag.*, vol. 18, no. 3, pp. 361–374, 2000.
- [18] M. Sheikhalishahi, L. Pintelon, and A. Azadeh, "Human factors in maintenance: a review," *J. Qual. Maint. Eng.*, vol. 22, no. 3, pp. 218–237, 2016.
- [19] T. Zonta et al., "Industry 4.0 predictive maintenance review," *J. Manuf. Technol. Manag.*, 2021.
- [20] S. J. Cheah, A. Shah, F. Shahbudin, and Taib, "Tracking hidden quality costs in a manufacturing company: action research," *Int. J. Qual. Rel. Manag.*, vol. 28, no. 4, pp. 405–425, 2011.
- [21] M. Naor, S. M. Goldstein, K. W. Linderman, and R. G. Schroeder, "The role of culture as driver of quality management," *Decision Sci.*, vol. 39, no. 4, pp. 671–702, 2008.
- [22] Kim, K.C., Sun, J.U. and Lee, G.M. (2013). A hierarchical approaches to vehicle routing and scheduling with sequential services using the genetic algorithm. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 20(1):245–257.
- [23] Lee, G.M. and Kim, K.C. (2014). Predictive modeling approaches for performance improvement in manufacturing systems. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 21(2):134–146.
- [24] V. R. Kannan and K. C. Tan, "JIT, TQM, and SCM: understanding linkages," *Omega*, vol. 33, no. 2, pp. 153–162, 2005.
- [25] O. Kosasih, K. Hidayat, B. Hutahayan, and Sunarti, "Achieving sustainable customer loyalty in the petrochemical industry: The effect of service innovation, product quality, and corporate image with customer satisfaction as a mediator," *Sustainability*, vol. 16, no. 16, p. 7111, 2024.
- [26] M. Khosroniya, R. Hosnavi, and M. R. Zahedi, "Enhancing operational performance in Industry 4.0," *Int. J. Ind. Eng. Oper. Res.*, vol. 6, no. 1, pp. 96–122, 2024.
- [27] W. Qin, Z. Zhuang, L. Guo, and Y. Sun, "A hybrid multi-class imbalanced learning method for predicting the quality level of diesel engines," *J. Manuf. Syst.*, vol. 62, pp. 846–856, 2022.
- [28] Y. Yang and H. Wang, "Random forest-based machine failure prediction: A performance comparison," *Appl. Sci.*, vol. 15, no. 16, p. 8841, 2025.

- [29] S. Wang and Y. Chen, "Improved yield prediction and failure analysis in semiconductor manufacturing with XGBoost and SHAP models," in IEEE IPFA, pp. 01–08, Jul. 2024.
- [30] P. Tambare et al., "Performance measurement system and quality management in data-driven Industry 4.0: A review," *Sensors*, vol. 22, no. 1, p. 224, 2021.
- [31] S. M. Lee, D. Lee, and Y. S. Kim, "The quality management ecosystem for predictive maintenance in the industry 4.0 era," *Int. J. Qual. Innov.*, vol. 5, no. 1, p. 4, 2019.
- [32] O. B. Johnson et al., "Developing real-time monitoring models," *Int. J. Eng. Res. Dev.*, vol. 20, no. 11, pp. 1296–1304, 2024.
- [33] M. U. Mojumder, "Impact of lean six sigma on manufacturing efficiency," *ASRC Procedia*, vol. 1, no. 1, pp. 343–375, 2025.
- [34] Mansour, Hamdia, Hanaa Abohashima, Heba Elkhoully, and Nermine Harraz, "Smart quality control: integrating six sigma, machine learning and real-time defect prediction in manufacturing," *International Journal of Lean Six Sigma*, (2025).
- [35] N. Faisal et al., "Defect types," *Structural Health Monitoring Damage Detection Systems for Aerospace*, 2021, p. 15.
- [36] K. Papageorgiou et al., "A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing," *Front. Manuf. Technol.*, vol. 2, p. 972712, 2022.
- [37] M. Yc and J. Kumar, "Artificial intelligence in quality assurance: A new paradigm for total quality management," in *Operations Research Forum*, vol. 6, no. 2, pp. 1–23, Jun. 2025.
- [38] O. C. Abikoye et al., "Application of internet of things and cyber physical system in Industry 4.0 smart manufacturing," in *Emergence of Cyber Physical System and IoT in Smart Automation and Robotics*, Cham: Springer, pp. 203–217, 2021.