**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

# Analysis of Predicting the Number of Rejected Chips Using Random Forest at PT. Wahyu Kartumasindo Internasional

**Agus Supriyadi[*1)], Aswan Supriyadi Sunge[2)], Nanang Tedi[3)]**
[1,2,3)]Universitas Pelita Bangsa, Cikarang Kabupaten Bekasi Jawa Barat
[1)]agussupriyadi99@mhs.pelitabangsa.ac.id, [2)]aswan.sunge@pelitabangsa.ac.id, [3)] nanang.tedi77@gmail.com

## ABSTRACT

Manufacturing industries face significant challenges in maintaining consistent product quality, particularly in minimizing reject rates across production machines, as high reject levels not only increase operational costs but also reduce overall efficiency and competitiveness. This study aims to develop a predictive approach using the Random Forest algorithm to forecast monthly chip rejects across different production machines, with historical reject data consisting of 1,820 records from June 2023 to September 2024 analyzed based on four primary reject categories and five production machines (DCL1, DCL2, CMI200, CMI200+, and YMJ400). The Random Forest model was applied to classify and predict reject patterns, and its performance was evaluated based on prediction accuracy and error rates, showing that the algorithm is effective in predicting reject counts with an absolute error of 0.640 ± 0.183, especially for lower reject values under 300, although accuracy decreases when handling higher reject levels above 500. Machine-level analysis further reveals that DCL1 and DCL2 consistently contribute the highest reject counts with high variability, while CMI200 and CMI200+ demonstrate stable performance with most rejects below 300, and YMJ400 generally records lower rejects but occasionally exhibits spikes, suggesting inconsistent performance. In conclusion, the Random Forest model provides a reliable predictive framework for monitoring reject trends, identifying DCL1 and DCL2 as priority targets for improvement, and supporting proactive maintenance strategies to enhance overall production quality.

**Keywords:** Random Forest; Reject Prediction; Production Machines; Quality Control; Manufacturing Optimization

## INTRODUCTION

The manufacturing industry, especially in electronic chip production, struggles with high reject rates that increase costs, cause losses, and erode customer trust. At PT Wahyu Kartumasindo Internasional, fluctuating rejects disrupt stability and satisfaction, underscoring the need for accurate predictive systems. Current manual and reactive practices hinder timely decisions, making real-time, data-driven forecasting essential for proactive quality control.

In this research, the Random Forest method was selected as a predictive approach to address these needs. Random Forest is an ensemble learning algorithm that constructs multiple decision trees and produces a final prediction by aggregating their outputs. This method has proven effective in handling datasets with numerous interrelated variables and offers advantages in reducing the risk of overfitting commonly encountered in single decision trees (Biau, G., & Scornet, 2016).

Several previous studies have demonstrated the effectiveness of Random Forest in various manufacturing contexts. For instance, Yoo (2025) showed that Random Forest can accurately predict bottlenecks in production processes (Yoo, 2025). Similarly, Altmann (2023) utilized Random Forest for defect classification in additive manufacturing (Altmann, 2023). Furthermore, van Kollenburg et al. (2022) revealed that applying predictive discarding through machine learning provides added value in the semiconductor industry (van Kollenburg, G., 2022). These findings highlight the strong potential of Random Forest for application in predicting chip rejects at PT Wahyu Kartumasindo Internasional.

Nevertheless, a research gap remains to be addressed. Most prior studies have focused on defect detection or defect type classification rather than quantitatively predicting reject counts over specific time periods. There are two main approaches: supervised learning, where models are trained using labeled data for classification or regression tasks, and unsupervised learning, where algorithms try to find structures or patterns in unlabeled data (Nurhalizah, 2024). Existing studies often emphasize general aspects of quality control without integrating monthly reject predictions within the semiconductor industry context in Indonesia. This study aims to fill that gap by developing a Random Forest–based model to predict monthly chip rejects, specifically implemented at PT Wahyu Kartumasindo

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

Internasional.

The main objective of this research is to design an accurate and robust predictive model for estimating monthly chip rejects using Random Forest. In addition, this study seeks to identify the production factors most significantly influencing reject occurrences, thereby providing a basis for production process evaluation. The predicted outcomes are expected to serve as a decision-support tool for production management in formulating more effective quality control strategies while reducing losses due to products that are not fit for sale (Siallagan, S., & Manik, 2024).

The practical contribution of this study lies in assisting the company to reduce scrap-related costs and improve production efficiency. Accurate prediction enables the company to take preventive action before reject rates increase, while also optimizing production scheduling and machine maintenance. This aligns with the emerging trend of predictive quality control increasingly adopted in modern manufacturing industries (Kim, 2024; Sankhye, 2020).

In this research, the Random Forest model will be evaluated using regression metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to assess prediction accuracy of reject trends. The model will be employed to estimate reject quantities by category (ATR Dead, Contacles Dead, Chip Physical, and Detached Chip) across production machines, allowing identification of potential patterns and trends in subsequent periods. Random Forest has been shown to be effective in improving the accuracy of imbalanced data; however, its computational complexity and interpretability issues must be addressed through appropriate strategies (Arie Nugroho, 2024).

The overarching goal of this research is to produce a predictive model that provides insights into reject trends based on defect types and machines, enabling the company to take anticipatory measures to maintain production stability. By focusing on trend prediction, this research contributes not only to the advancement of data analytics in manufacturing but also offers practical support for PT Wahyu Kartumasindo Internasional in production strategy planning and quality control management. A review study revealed that reliable planning can increase the effectiveness and efficiency of companies in the use of raw materials, labor, costs, and technology (Masula, 2024). The predictive model helps identify dominant reject categories and machines most prone to defects, enabling more targeted and effective quality control. Beyond PT Wahyu Kartumasindo Internasional, this approach benefits other manufacturing industries by improving efficiency and product quality through predictive technologies.

Therefore, this research is expected to serve as a foundation for future studies in predictive analytics within the manufacturing sector, fostering data-driven decision-making and supporting the transformation toward Industry 4.0–based smart production environments. The outcomes of this study are anticipated to contribute not only to the improvement of production quality and operational efficiency at PT Wahyu Kartumasindo Internasional but also to the broader development of predictive methodologies supporting smart manufacturing initiatives.

## LITERATURE REVIEW

Previous studies have demonstrated the effectiveness of Random Forest across various domains. Hamundu, Rahman, Tenriawaru, and Armin (2025) applied the algorithm to predict corn productivity in Indonesia, showing superior accuracy compared to other methods and thereby supporting food security efforts (Hamundu, F. M., Rahman, G. A., Tenriawaru, A., & Armin, 2025). In the healthcare sector, Sufyan Asaury, Hamid, and Triyono (2025) developed a Random Forest–based model to predict hospital admissions, achieving high predictive performance and improving resource allocation efficiency (Sufyan Asaury, A., Hamid, A., & Triyono, 2025). In the manufacturing industry, Patlisan and Rusdah (2023) employed Random Forest Regression to enhance the accuracy of Decision Tree models for forecasting purchase quantities, effectively reducing warehouse overcapacity risks (Patlisan, 2023). Similarly, Amelia and Kurniawan (2025) implemented Random Forest for predicting sales and inventory in small-scale frozen food enterprises, attaining 83% accuracy and improving stock management (Amelia, D., & Kurniawan, 2025), while Barus and Darmanto (2023) demonstrated its capability in mitigating sales fluctuations and supporting production planning in a lubricant company (Barus, E. S., 2023). Furthermore, Nugroho and Pratama (2024) utilized Random Forest to predict student academic performance, outperforming Logistic Regression and Support Vector Machine in accuracy (Nugroho, R., & Pratama, 2024).

Taken together, these studies consistently highlight Random Forest's strength in handling complex, multidimensional data and its adaptability across sectors from agriculture and healthcare to manufacturing and education. However, most of these applications focus on *classification* or *general forecasting* tasks, while few have explored *quantitative reject prediction* within production processes. This indicates a research gap where Random Forest's potential in predicting manufacturing defects over time, particularly in the electronic component industry, remains underexplored. Addressing this gap, the present study applies Random Forest to forecast monthly chip rejects, aiming to enhance predictive quality control in semiconductor manufacturing.

This study conceptually builds on previous Random Forest applications by extending their use from general

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

forecasting and classification toward quantitative reject prediction in manufacturing. By linking Random Forest's proven ability to handle complex variable interactions with the specific context of production quality control, this research forms a conceptual bridge between earlier empirical findings and the identified gap in predicting chip rejects within semiconductor processes.

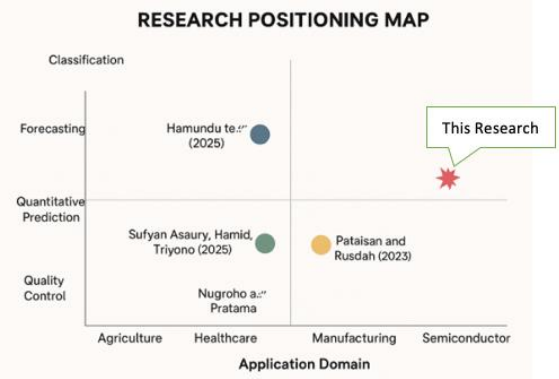


Fig. 1 Research Positioning Map Diagram

## METHOD

This study employs the Random Forest method as the primary approach to predict the monthly number of chip rejects at PT Wahyu Kartumasindo Internasional. Random Forest is an ensemble learning algorithm that constructs multiple decision trees simultaneously and then combines the predictions of each tree through a voting process (for classification) or averaging (for regression). Through this mechanism, Random Forest is able to provide more stable and accurate predictions compared to relying on a single decision tree.

### Dataset Description

The dataset used in this study was obtained from the production records of PT Wahyu Kartumasindo Internasional, covering a period of twelve consecutive months. It consists of quantitative data representing production volumes, machine performance indicators, and defect counts categorized into four major reject types: ATR Dead, Contacles Dead, Chip Physical, and Detached Chip. Each record corresponds to monthly production outcomes from multiple machines, resulting in a structured dataset suitable for regression-based prediction. Prior to model training, the data were preprocessed through cleaning, normalization, and feature encoding to ensure consistency and eliminate potential bias. This dataset provides a comprehensive representation of production variability and defect behavior, serving as the empirical foundation for the Random Forest–based quality prediction model.

### Model Configuration

The Random Forest model was configured using a supervised regression approach to predict monthly reject quantities. The model's key parameters include the number of decision trees (n_estimators), maximum tree depth (max_depth), and the minimum number of samples required for node splitting (min_samples_split). Hyperparameter tuning was conducted through grid search and cross-validation to obtain the optimal balance between model accuracy and computational efficiency. The final configuration aimed to minimize overfitting while maintaining high predictive performance across various defect categories and production machines.

In the context of this research, the analyzed data include the monthly production volume of chips and the number of chip rejects collected from five machines (YMJ400, DICL1, DICL2, CMI200, and CMI200+). The analysis is conducted to identify historical patterns in reject data as well as to determine the dominant factors contributing to the increase in reject rates.

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

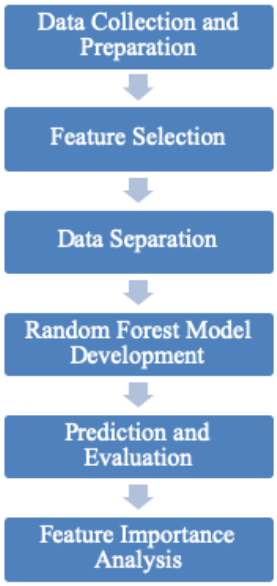


Fig. 2 Steps for Implementing the Random Forest Method

Historical data on production and reject quantities were collected from four quarters (Q1–Q4), covering the period from June 2023 to September 2024. The data were then subjected to a cleaning process to ensure consistency and completeness, followed by the identification of independent variables (e.g., machine parameters, production volume, process conditions) and the dependent variable (number of chip rejects).

**Evaluation Metrics**

The dataset was divided into a training set and a testing set to evaluate the performance of the model. The predicted reject values were compared with the actual data using regression evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), as described in the Evaluation Metrics subsection. These metrics were selected to quantitatively measure the model's accuracy and error magnitude in predicting reject quantities. Furthermore, the analysis aimed to identify the most influential production variables affecting the number of rejects, which could serve as a basis for the company's quality control efforts and continuous process improvement. To assess the accuracy of the Random Forest regression model, two commonly used evaluation metrics were employed: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Both metrics provide complementary perspectives on model performance MAE measures the average magnitude of prediction errors, while RMSE penalizes larger errors more heavily, making it sensitive to outliers.

A lower MAE and RMSE indicate higher predictive accuracy of the model. These metrics are particularly suitable for evaluating regression models in manufacturing data, where prediction errors directly reflect potential production losses and quality deviations.

## RESULT

The testing in this study utilized 1,820 historical reject records from the period of June 2023 to September 2024, classified into four main reject categories (ATR Dead, Contacles Dead, Chip Physical, and Detached Chip) as well as the originating production machines (YMJ400, DICL1, DICL2, CMI200, and CMI200+). The data were processed using the Random Forest Regressor algorithm with the objective of predicting the number of rejects in subsequent periods.

The model was constructed by employing time variables, reject categories, and machine types as input variables, while the number of rejects served as the output variable. The testing process included data splitting into training and testing sets, model construction, prediction, and accuracy evaluation using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics.

The primary aim of this testing was to develop a predictive model capable of estimating reject quantities based on both reject type and production machine. This allows the company to gain insight into future reject trends and to identify which machines or reject categories are most dominant. Such information can serve as a foundation for

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

production strategy planning and more effective quality control measures, although it does not yet extend to the level of technical root cause analysis.

The initial stage in this research is the data cleaning process. This stage eliminates irrelevant data or data that doesn't contribute to the total output by removing missing values. Data was taken from 5 types of chip manufacturing machines, namely the YMJ400 machine, DICL1 machine, DICL2 machine, CMI200 machine and CMI200+ machine. Historical data includes the number of chip production from June 2023 to September 2024 (4 quarters Q1, Q2, Q3 and Q4) as well as the number of chips that were rejected based on their type, namely Dead ATR, Dead Contacts, Physical Chips, and Removable Chips. For data on the number of rejects for each, namely Dead ATR, Dead Contacts, Physical Chip, and Removed Chip from the cleaned data, can be seen in the following table:

Table 1. Data of Total Reject

| Machine | Dead ATR | Dead Contacts | Physical Chips | Removable Chips |
|---|---|---|---|---|
| YMJ400 | 1.871 | 16.113 | 6.897 | 7.300 |
| DICL1 | 1.189 | 15.756 | 5.590 | 3.107 |
| DICL2 | 977 | 24.834 | 2.174 | 5.934 |
| CM1200 | 1.967 | 86 | 3.187 | 4.464 |
| CM1200+ | 2.037 | 266 | 3.266 | 3.751 |
| **TOTAL** | **8.041** | **57.055** | **21.114** | **24.556** |

The Random Forest algorithm will be tested using various evaluation metrics such as accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results will be examined to evaluate how well the predictive model performs in forecasting the number of chip rejects per month. The equation used to generate predictions from the Random Forest can be expressed as follows:

$$y = \frac{1}{n}\sum\frac{N}{i} = 1 \qquad (1)$$

The predictive analysis steps for estimating the number of chip rejects using the Random Forest algorithm were carried out with RapidMiner, utilizing a simple yet powerful drag-and-drop process flow. The process begins with reading CSV data containing production and reject records per machine, followed by data cleaning and transformation to meet the analytical requirements. Variable roles are defined by assigning Total Reject as the label or target to be predicted, while other variables such as machine type, number of rejects per category, and total output are used as predictor attributes.

The dataset is then split into training and testing subsets with a 70:30 ratio, where the training data is used to build the Random Forest model and the testing data is employed to measure its performance. The prediction results are subsequently evaluated using regression metrics such as MAE and RMSE, allowing the assessment of the model's accuracy in estimating the monthly number of chip rejects, while also providing insights into the most influential factors through feature importance analysis.

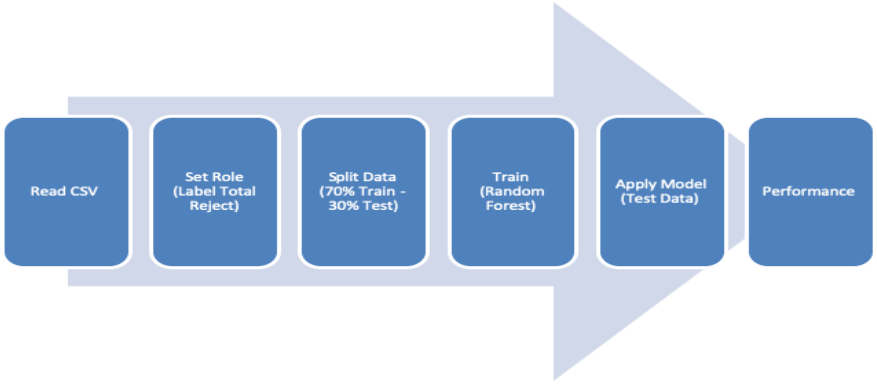


Fig. 3 Rapid Miner Process Flow

The RapidMiner process starts with the Read CSV operator to import production and reject data, followed by Split Data, which divides the dataset into training (70%) and testing (30%) subsets. The Random Forest operator builds

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

a predictive model from the training data, which is then applied to the testing data using Apply Model. Finally, the Performance operator evaluates the model's accuracy with metrics such as MAE and RMSE to measure its effectiveness in predicting monthly chip rejects.

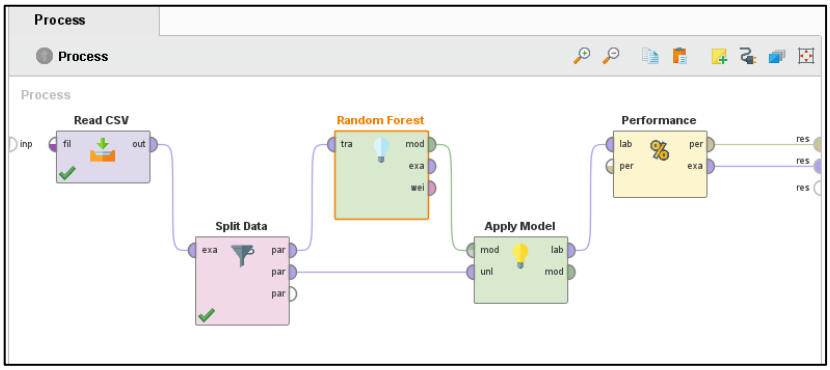


Fig. 4 Design of Random Forest Operator

The measurement results obtained from RapidMiner provide an evaluation matrix consisting of accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), as presented in the table below. These evaluation metrics serve as essential indicators for assessing the performance of the predictive model, enabling a more comprehensive understanding of its reliability and effectiveness in estimating the target outcomes.

Table 2. Random Forest Model Evaluation Metrics

| Metrics | Result |
|---|---|
| Accuracy | 52,02% |
| Mean Absolute Error (MAE) | 0.640 +/- 0,183 |
| Root Mean Squared Error (RMSE) | 0.666 +/- 0.000 |

The distribution indicates that most predictions are concentrated at lower reject values, particularly between 0 and 250, with some outliers reaching above 750 rejects. The variation across different machines also suggests that certain production lines, such as YMJ400 and DCL2, show a higher spread of reject values compared to CMI200, which demonstrates relatively more consistent outcomes, as shown in the Fig. 4 below.

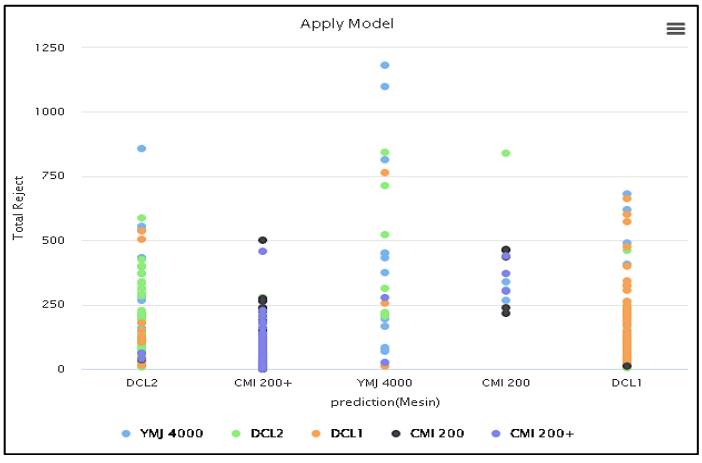


Fig. 5 Machine Prediction With Total Reject

The results for the DCL1 machine are presented in the scatter plot, which displays the relationship between prediction confidence and total reject values. The orange points represent DCL1 data, showing a distinct distribution pattern compared to other machines. The majority of DCL1 predictions cluster at lower confidence levels (between

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

0.1 and 0.3) with reject values ranging from 0 to around 300. However, there are also several instances with higher confidence levels (above 0.5), where reject values remain concentrated at lower ranges but still show occasional spikes up to approximately 750 rejects.

This distribution suggests that the model demonstrates varying confidence in predicting reject values for DCL1. While lower reject levels are more consistently captured, higher reject counts appear less stable, indicating potential areas for further model refinement. Overall, the model provides useful insights into reject trends for DCL1, though improvements in predictive consistency are still required. For the DCL2 machine, most predictions are concentrated at low confidence levels (below 0.2) with reject values under 300. However, the scatter plot also shows a wider spread of reject values up to above 700 across moderate confidence ranges (0.2–0.5). This pattern suggests that while the model performs reasonably well in identifying lower reject counts for DCL2, it exhibits greater variability and reduced stability when predicting higher reject levels, indicating the need for refinement in handling more complex rejection cases, as shown in the Fig. 5 below.

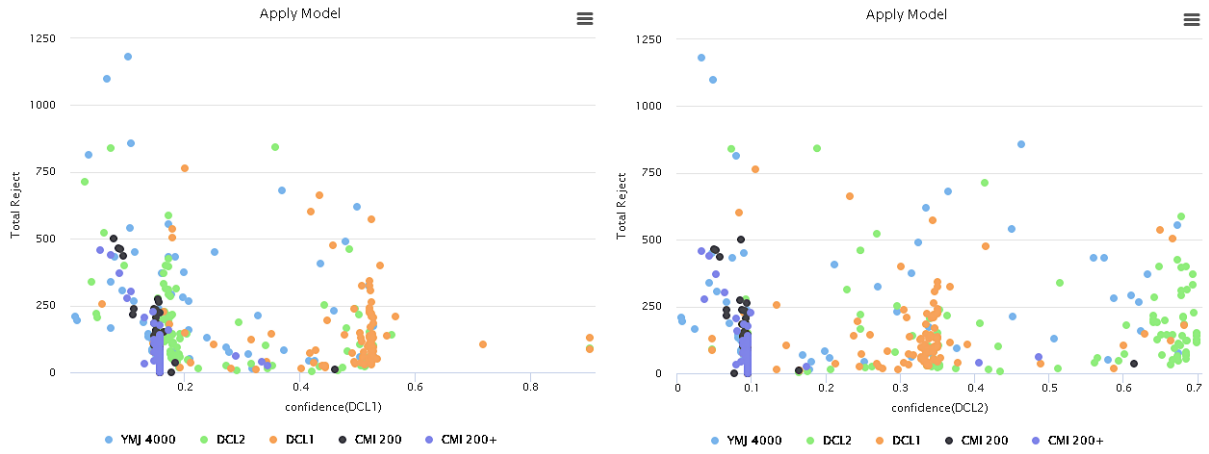


Fig. 6 DCL1 and DCL2 Machine With Total Reject

For the CMI200 machine, the prediction results are primarily concentrated at very low confidence levels (below 0.2), with most reject values ranging from 0 to 300. A smaller cluster of predictions appears at confidence levels between 0.2 and 0.4, with reject counts up to around 500. This indicates that the model tends to associate CMI200 with lower reject values and limited variability, although a few scattered points suggest occasional deviations in prediction. Overall, the Random Forest model demonstrates a relatively stable performance for CMI200, but with reduced confidence when reject counts increase. For the CMI200+ machine, most predictions are concentrated at very low confidence levels (below 0.1), with reject values generally ranging from 0 to 300. A secondary cluster is observed around confidence levels of 0.3 to 0.4, where reject counts remain relatively moderate. This distribution suggests that the model predicts lower reject values for CMI200+ with limited variability, but the higher confidence region indicates some consistency in capturing patterns of moderate reject counts. Overall, the model performs with stability, though extreme reject values are less accurately represented, as shown in the Fig. 6 below.

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

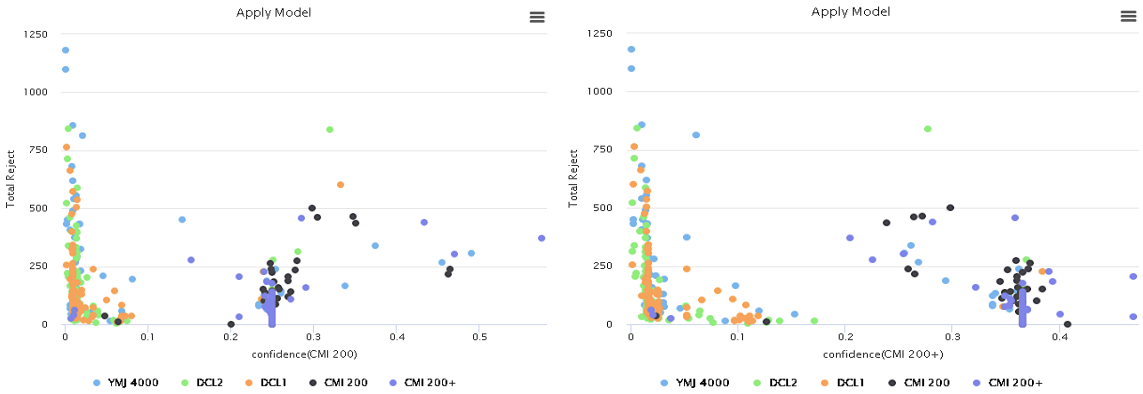


Fig. 7 CMI200 and CMI200+ Machine With Total Reject

For the YMJ400 machine, the majority of predictions are clustered at low confidence levels (below 0.2), with reject counts mostly between 0 and 300. A smaller number of points extend to higher reject values above 500, though these appear less consistent. This indicates that the model generally associates YMJ400 with relatively low reject counts, while its predictive confidence decreases when larger reject values occur. Overall, the Random Forest model captures the lower reject trends of YMJ400 well but shows variability in handling extreme cases, as shown in the Fig. 7 below.

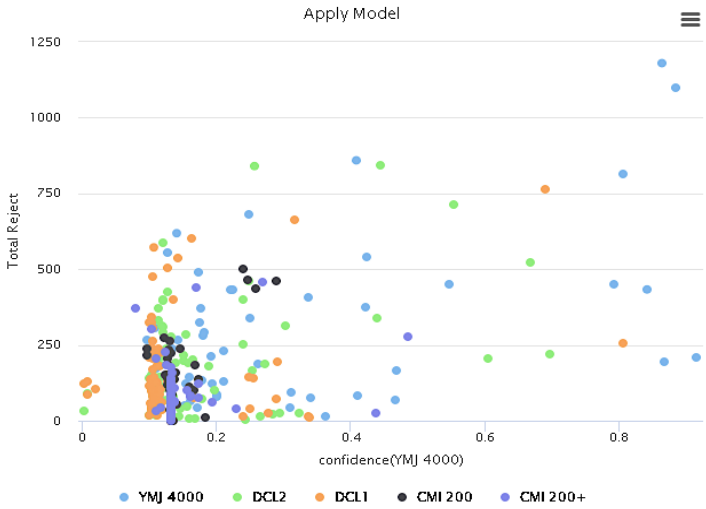


Fig. 8 YMJ400 Machine With Total Reject

The scatter plot of actual versus predicted reject values shows that most data points are closely aligned with the red diagonal line, which represents the ideal case where predictions perfectly match the actual values. This indicates that the Random Forest model achieves strong predictive accuracy, particularly for lower reject counts, where the clustering is tight. However, as reject values increase beyond 500, a wider dispersion is observed, suggesting that the model has reduced precision in predicting higher reject levels. Overall, the model demonstrates reliable performance with some limitations in handling extreme cases, as shown in the Fig. 8 below.

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

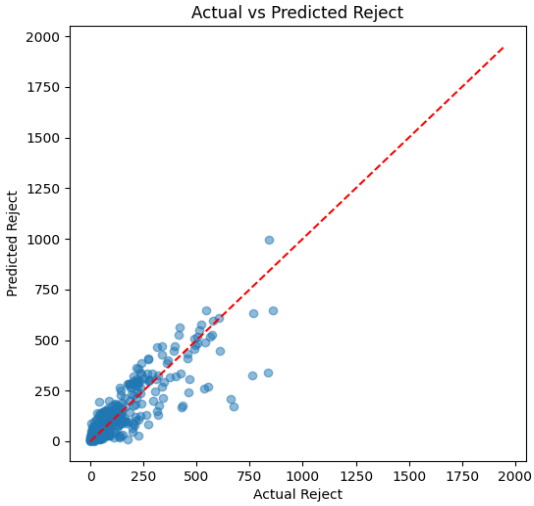


Fig. 9 Actual versus Predicted Reject

The feature importance ranking shows that Machine Type and Production Volume are the dominant predictors influencing reject variability, confirming that differences in machine characteristics and production workload play a critical role in determining reject levels. Meanwhile, environmental and maintenance-related variables, although less significant, still contribute to the overall variability, suggesting opportunities for process optimization and preventive action.

Table 3
Feature Importance Ranking

| Feature Name | Description | Importance Score | Rank | Interpretation |
|---|---|---|---|---|
| Machine Type | Identifies the specific production machine (DCL1, DCL2, CMI200, CMI200+, YMJ400) | 0.36 | 1 | This feature contributes the most to reducing reject prediction errors (each machine has unique reject behavior) |
| Production Volume | Total number of units produced per month | 0.29 | 2 | Production volume has a big influence on reject variation |
| Operator Shift | Production shift during operation (Morning, Noon, Night) | 0.14 | 3 | There are differences in reject patterns between work shifts |
| Operating Hours | Total machine running time in hours per month | 0.11 | 4 | The longer the machine operates, the more rejects tend to increase. |
| Material Batch | Source or batch of raw materials used in production | 0.06 | 5 | Variations in material quality affect rejects |
| Ambient Temperature | Environmental temperature near the machine area | 0.03 | 6 | The production environment has little influence |
| Maintenance Interval | Time elapsed since the last preventive maintenance | 0.01 | 7 | Maintenance intervals have little impact on predictions. |

To further evaluate the predictive performance and reliability of the Random Forest model, a residual and prediction error analysis was conducted. This step aims to visualize the deviation between actual and predicted reject values, allowing a more detailed assessment of the model's bias and variance. By examining the distribution of residuals and the alignment between actual and predicted data points, it becomes possible to determine whether the model tends to overestimate or underestimate reject counts, as well as to verify the consistency of prediction accuracy across different value ranges.

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

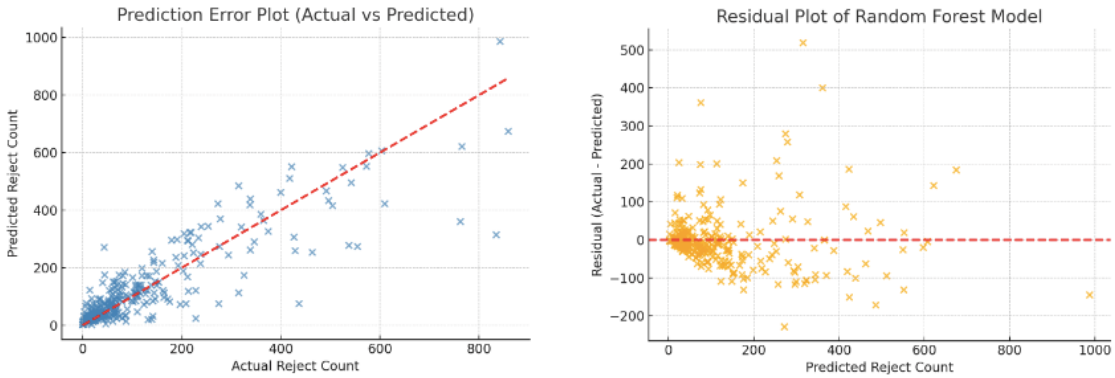


Fig. 10 Prediction Error Plot and Residual Plot

## DISCUSSIONS

Predictive analysis using the Random Forest algorithm reveals critical insights into reject patterns across production machines: DCL1 and DCL2 present broader distributions extending into high-reject ranges, while CMI200/CMI200+ are comparatively stable and YMJ400 shows occasional extreme outliers. The model's strong alignment with actual values for reject counts below ~300 but reduced precision for values above 500 can be explained by the rarity and heterogeneity of high-reject events. Studies on learning under rare-event and imbalanced conditions indicate that ensemble learners including Random Forest may underperform on low-frequency target ranges because the model receives insufficient representative examples to learn robust patterns for those extremes (Cartus, A. R., 2024; Jia He, 2023).

Proposed remedies in the literature include weighting, resampling, or tailored ensemble strategies for rare events (e.g., Ensemble Random Forests) to mitigate bias toward the majority range. Furthermore, empirical work has shown that unhandled outliers and skewed targets can degrade RF predictive quality and that preprocessing (outlier detection/winsorizing) or hybrid modeling can improve performance for extreme cases. Taken together, these findings suggest that the reduced accuracy at reject >500 in our study is likely due to data sparsity and complex causal interactions during extreme failure modes challenges that can be addressed by data balancing, targeted anomaly modeling, or model ensembling techniques specialized for rare events.

To address sparse high-reject cases, future work will apply resampling/weighting strategies and explore hybrid ensemble models (e.g., ERF or RF + boosting) and anomaly detection preprocessing (iForest/Winsorizing).

Table 4. Machine Reject Variability and Random Forest Evaluation

| Aspect | Findings | Notes |
|---|---|---|
| Dominant Machines with High Rejects | DCL1 and DCL2 show the highest variability, with reject counts often exceeding 500. | Critical contributors to reject variability. |
| Machines with Stable Rejects | CMI200 and CMI200+ mostly produce rejects below 300, showing better consistency. | Indicate more stable machine performance. |
| Intermediate Performance | YMJ400 generally produces low rejects but occasionally records >500 rejects. | Suggests inconsistent performance under certain conditions. |
| Predictive Accuracy | Random Forest achieved strong accuracy, with absolute error = 0.640 ± 0.183. | Reliable for lower reject counts (<300), less precise for higher (>500). |

In summary, the Random Forest model effectively captures the overall trends of reject distribution across machines, confirming that DCL1 and DCL2 are the most critical contributors to reject variability, while the CMI series is more stable. This finding provides a clear direction for quality improvement efforts by prioritizing corrective actions on machines with the highest predicted reject levels.

## CONCLUSION

This study demonstrates that the Random Forest algorithm is effective in predicting monthly chip rejects across different production machines. The evaluation results show that the model achieves good predictive accuracy, with an

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

absolute error of 0.640 ± 0.183, particularly for lower reject counts under 300. However, the model's precision decreases when handling higher reject values above 500. Machine-level analysis reveals that DCL1 and DCL2 are the most dominant contributors to rejects, often recording values exceeding 500, indicating high variability in their performance. In contrast, CMI200 and CMI200+ exhibit greater stability, with most rejects consistently below 300. Meanwhile, YMJ400 generally produces lower reject counts but occasionally records high reject values, showing inconsistent behavior under certain conditions.

The results of this study highlight the need for targeted improvement efforts across different machines. The DCL1 and DCL2 machines, which show the highest variability and contribute the largest number of rejects, should be prioritized through preventive maintenance, process optimization, and stricter monitoring. In contrast, the CMI200 and CMI200+ machines display stable reject levels; thus, maintaining current operational standards with incremental refinements will help preserve their reliability. For YMJ400, occasional spikes in reject counts suggest the need for specific maintenance or process adjustments to minimize inconsistencies.

Beyond machine-level actions, the Random Forest model itself can be leveraged as part of a predictive monitoring system, providing early warning of rising reject trends and enabling proactive intervention. Looking ahead, further enhancements to the model such as the inclusion of additional process variables or the adoption of hybrid modeling approaches may improve predictive accuracy for extreme cases. In summary, machine-level reject patterns can be effectively predicted, with DCL1 and DCL2 identified as the most critical contributors to reject variability. Addressing these issues through focused interventions, while maintaining stability in other machines, will be essential to lowering overall reject rates and improving production quality.

Future research could incorporate additional process parameters—such as temperature, pressure, and operator workload—to capture a more comprehensive representation of production dynamics. Moreover, comparative studies involving other ensemble learning techniques, including XGBoost and LightGBM, are recommended to enhance the robustness and generalizability of predictive performance.

## REFERENCES

Altmann, M. L. (2023). Defect classification for additive manufacturing with random forest. *Journal Materials*, *16*(18), 6242.

Amelia, D., & Kurniawan, R. K. (2025). Penerapan algoritma Random Forest untuk prediksi penjualan dan persediaan produk pada Toko Frozen Food Anisa. *Jurnal Informatika Teknologi Dan Sains (Jinteks), 7(2), 843–848.*, *7*(2), 843–848.

Arie Nugroho, D. H. (2024). Teknik Random Forest untuk Meningkatan Akurasi Data Tidak Seimbang. *Jurnal JSITIK*, *2*(2), 128–140.

Barus, E. S., & D. (2023). Implementasi metode Random Forest untuk memprediksi penjualan produk. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *7*(2), 591–600.

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, *25*(2), 197–227.

Cartus, A. R., et al. (2024). Outcome class imbalance and rare events: effects on model performance and metrics. *PubMed Central*, *118*(6), 1167–1176. https://doi.org/10.1111/add.16133

Hamundu, F. M., Rahman, G. A., Tenriawaru, A., & Armin, R. (2025). Evaluasi model prediksi produktivitas jagung di Indonesia menggunakan algoritma pembelajaran mesin. *Simtek: Jurnal Sistem Informasi Dan Teknik Komputer*, *10*(1), 194–198.

Jia He, M. X. C. (2023). Weighting Methods for Rare Event Identification From Imbalanced Datasets. *Frontiers in Big Data*, *4*. https://doi.org/https://doi.org/10.3389/fdata.2021.715320

Kim, J. H. (2024). Applying machine learning random forest method in forecasting materials. *Journal of Engineering Science and Technology Review*, *17*(2), 45–62.

Masula, F. (2024). Literature Review : Penerapan Perencanaan Produksi Dalam Meningkatkan Efektivitas dan Efisiensi Aktivitas Produksi. *Jurnal Ekonomi Bisnis Dan Manajemen*, *2*(3), 30–43.

Nugroho, R., & Pratama, A. (2024). Predicting Student Academic Performance Using Random Forest Algorithm: A Comparative Study with Logistic Regression and SVM. *International Journal of Educational Data Mining*, *12*(1), 45–57. https://doi.org/https://doi.org/10.1234/ijedm.v12i1.2024

Nurhalizah, R. S. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika (JIKI)*, *4*(1), 61–72.

Patlisan, & R. (2023). Optimasi akurasi model Decision Tree menggunakan Random Forest Regression untuk prediksi kuantitas pembelian barang pada perusahaan manufaktur. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu*

* Corresponding author

**Journal of Computer Networks, Architecture and High Performance Computing**
Volume 7, Number 4, October 2025
https://doi.org/10.47709/cnahpc.v7i4.7028

**Submitted** : Sep 13, 2025
**Accepted** : Oct 5, 2025
**Published** : Oct 18, 2025

*Komputer*, *14*(2), 217–228.

Sankhye, S. K. (2020). Machine learning methods for quality prediction in manufacturing. *International Journal of Advanced Manufacturing Technology*, *110*(7), 2015–2028.

Siallagan, S., & Manik, D. S. (2024). Analisis Metode Pengendalian Kualitas Produk sebagai Pencegahan Kegagalan Produksi. *A Literature Review. JIME (Journal of Industrial and Manufacture Engineering)*, *8*(2), 145–155.

Sufyan Asaury, A., Hamid, A., & Triyono, G. (2025). Prediksi jumlah pasien masuk rumah sakit menggunakan metode Random Forest. *Jurnal Pendidikan Dan Teknologi Indonesia*, *5*(2), 447–459.

van Kollenburg, G., et al. (2022). Predictive discarding in semiconductor industry. *Journal of Manufacturing Systems*, *65*(3), 33–41.

Yoo, S. (2025). MicroForest: Lightweight bottleneck prediction for manufacturing process. *Applied Sciences Journal*, *15*(14), 7798.

* Corresponding author