# Machine Learning-Based Process Improvement in Optical Module Manufacturing Through Predictive Analytics

**Nattapon Wongngern**
Industrial Statistics and Operational Research Unit (ISO-RU), Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Thailand
nattapon.wong@dome.tu.ac.th

**Busaba Phruksaphanrat**
Industrial Statistics and Operational Research Unit (ISO-RU), Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Thailand
lbusaba@engr.tu.ac.th (corresponding author)

*Received: 17 June 2025 | Revised: 27 August 2025 and 14 September 2025 | Accepted: 16 September 2025*

*DOI: https://doi.org/10.48084/etasr.12802*

## ABSTRACT

**This study aimed to enhance the optical module manufacturing process using Machine Learning (ML) by enabling early defect detection. The traditional step-by-step testing workflow (OPA, LOPT, OTSM, and OSET) results in excessive retesting cycles, increasing production time, and reducing overall efficiency. A structured method was employed, including data preparation, statistical hypothesis testing (t-test and F-test) for feature selection, and comparative evaluation across multiple ML models. Two prediction strategies were investigated: (i) a step-by-step model that predicts the outcome at each test stage, and (ii) a cross-step model that directly predicts the final OSET outcome from early-stage OPA data. The results show that the step-by-step prediction with XGBoost increased the total yield from 48.5% to 98.1%. Meanwhile, the cross-step prediction using Random Forest, leveraging early-stage OPA data, further improved the yield to 98.9%. In a real-world implementation involving 100 optical modules, the ML-based process reduced retesting from 53 units to only 3, saving approximately 795 minutes (13.25 hours) of testing time per batch while maintaining the same number of finished goods. These results demonstrate that ML-based predictive analytics can significantly improve process efficiency, minimize unnecessary retesting, reduce tester load, and accelerate production throughput without compromising product quality.**

***Keywords-machine learning; optical module manufacturing; process improvement; predictive analytics; industrial AI***

## I. INTRODUCTION

Fiber optics is essential in modern data communication in the digital age, as data transmission is vital to both the global economy and technological progress. Fiber optics serves as a critical infrastructure component in telecommunications and enterprise data networks because of its ability to transmit data at high speeds with increased capacity and minimal signal degradation over extended distances. Optical modules perform essential functions in fiber optic communication through the conversion of the electrical signal to optical for transmission. Data transmission capabilities improve through reduced latency, expanded bandwidth, and minimized signal loss [1]. Technical innovation leads to improved optical modules, which power 5G networks and AI applications by increasing data capacity and using power resources efficiently [2].

The four essential testing stages require extensive pre-production time because the Optical Precision Alignment (OPA) process starts module testing by establishing performance benchmarks, which lead to Live Optical Performance Tracker (LOPT) stability checks, followed by the Optical Thermal Stability Monitor (OTSM) and temperature endurance testing, before the Optical Stress Endurance Test (OSET) completes the sequence to verify long-term durability. The reliability phase of testing persists, but unexpected obstacles reduce testing efficiency and require additional resources and time, which negatively affect product manufacturing speed and delivery schedule management.

Defective optical modules need retesting to confirm failure before being sent to the debug station for in-depth analysis and troubleshooting. The total testing time increases as retesting defective units occupies testing time slots, delaying the evaluation of other units. The defective units impact production output, as new incoming units must wait for open test slots, leading to processing delays and reducing the number of deliverable units per day. The balance between achieving excellent product reliability and maintaining efficient production flow becomes evident through the necessity of reducing test failures to optimize throughput and decrease delays in high-volume manufacturing settings.

From a case study, the module testing process—comprising OPA, LOPT (yield rate: 80.6%), OTSM (yield rate: 76.2%), and OSET (yield rate: 78.9%)—suffers from a low overall yield rate of only 48.45%. The testing process becomes longer and more complex when units fail testing, which leads to increased workload and delayed production, and reduced product output. Identification of failure units remains ineffective. A predictive tool is necessary to detect potential defects and minimize unproductive testing, boosting the overall production.

The implementation of Machine Learning (ML) methods offers an effective approach to improve quality control analysis procedures and manufacturing system operations. Manufacturers use ML techniques for better defect detection along with failure prediction to achieve higher production efficiency. In [3], the effectiveness of ML-based anomaly detection was demonstrated in industrial IoT settings, highlighting the role of predictive models in improving system reliability and reducing unnecessary operational costs, which aligns with the improvement of process efficiency in manufacturing environments such as optical module production. ML enables the detection of failing units before initial tests, so that resources can be allocated more efficiently by reducing unnecessary retesting. Such methods facilitate the seamless addition of new units to the testing workflow.

This study presents an ML-based system for early-stage defect prediction in optical module manufacturing to improve the testing process. Instead of relying on defect detection after all testing stages, the proposed approach enables the early identification of potentially defective modules, allowing smarter routing decisions. This strategy reduces unnecessary retesting, shortens the overall test time, and alleviates the load on test stations and debug operations.

## II. MATERIALS AND METHODS

This research investigates the performance and reliability of optical modules designed for high-speed, long-distance data transmission using the Quad Small Form-factor Pluggable Double Density (QSFP-DD) standard [4]. ML methods use production line data to predict the results of the testing process. The workflow involves data collection, data cleaning, data preparation, and feature selection, leading to model training and validation. Finally, performance evaluation ensures accurate predictions.

### *A. Data Collection*

Performance testing data was collected from automated production-line testers under standardized manufacturing conditions between January 1, 2023, and December 31, 2023. A total of 1,050 optical modules were tested following a step-by-step approach consisting of four major stages: OPA, LOPT, OTSM, and OSET. Each module generated thousands of transactional records because all measurements were automatically logged by the tester machines, resulting in millions of entries overall. Therefore, the dataset accumulated in one year is sufficiently large, diverse, and representative to support reliable statistical analysis and robust model development. The details of each testing stage are summarized below.

- OPA: Conducted at a controlled room temperature of 25°C to calibrate and evaluate the optical module's initial performance. Each test lasted approximately 1 hour and 18 minutes, with 9 modules per batch. The dataset contained 49 features, including Wait for Demod Lock Time Result, Temperature – Laser, RXPD_D_FA, Temperature – PIC, TX_INIT_OUTPUT_POWER, RX90X Pre-/Post-Cal, Temperature – ASIC, and other optical/thermal parameters.
- LOPT: Performed at 25°C to evaluate module functionality under simulated real-use conditions, with a primary focus on optical performance. Each test cycle lasted 1 hour and 38 minutes, producing 12 modules per batch. A total of 96 parameters were collected, such as Client Lane 0 Total Bits_400ZR-OFEC, Media – FEC BER_-21.8dBm_400ZR-OFEC, Module Power_400ZR-OFEC, Optical Signal-to-Noise Ratio, Current consumption, and related optical/electrical measures.
- OTSM: Conducted at high temperature (45°C) and low temperature (-5°C) to assess thermal endurance. Each cycle lasted 3 hours and 10 minutes, with 12 modules per batch. A total of 214 features were recorded, including Wait for Demod Lock Time Result_-21.8dBm_400ZR-OFEC_45_Degree, OSNR – FEC Threshold_-12dBm_400ZR-OFEC_-5_Degree, Temperature – PCB_-12dBm_400ZR-OFEC_-5_Degree, Dark Current Variation, Thermal Drift, and other stability-related indicators.
- OSET: The OSET test subjects the module to repeated power cycling under fluctuating temperature conditions, focusing on assessing durability. Each test takes 6 hours and 49 minutes, with 40 modules tested per batch. This test does not include feature data, as it serves solely as the target variable for prediction.

### *B. Data Cleaning*

The data cleaning process eliminated rows containing missing values, ensuring that the analysis uses complete and accurate information.

### *C. Data Preparation*

The dataset was processed and divided into training and validation subsets using the train_test_split function in scikit-learn. After cleaning, a total of 6,783 complete records were available, with 5,426 (80%) used for training and 1,357 (20%) for validation. To evaluate model robustness, additional experiments were performed with different split ratios (60:40, 70:30, 80:20, and 90:10), following common practice in machine learning approaches: (i) 60:40 to test model behavior under limited training data, (ii) 70:30 as a balanced compromise often adopted in empirical studies, (iii) 80:20 as the most widely accepted standard providing sufficient training data while preserving a representative validation set, and (iv) 90:10 to examine performance when nearly all data are used for training. All splits were generated randomly with a fixed random_state of 42 to ensure reproducibility, while no stratification was applied so that pass/fail outcomes followed their natural distribution. Among these ratios, the 80:20 split consistently yielded the best predictive performance across key metrics. This finding aligns with prior research, such as [5], which reported that the 80:20 ratio delivered superior accuracy compared to other splits in multiple ML experiments.

The dataset included temperature readings, electrical current measurements, and optical performance parameters collected from the OPA, LOPT, and OTSM stages, together with the final target variable (LOPT_RESULTS, OTSM_RESULTS, OSET_RESULTS), which indicates whether each module passed or failed the endurance test. This ensured that both environmental and optical parameters were incorporated into model training to provide robust prediction of yield outcomes.

### *D. Feature Selection*

Statistical hypothesis testing was applied to identify relevant features. Independent t-tests compared mean values, and F-tests assessed variance differences between passed and failed modules at a significance level of 0.05. Features with statistically significant differences were retained.

#### *1) Mean Hypothesis Test*

The t-test was employed to analyze average feature values between passed and failed modules using the method described in [6]:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the average values obtained from sample groups 1 and 2, $s_1$ and $s_2$ are the standard deviations of sample groups 1 and 2, and $n_1$ and $n_2$ are the sample sizes of group 1 and 2, respectively.

#### *2) Variance Hypothesis Test*

The F-test was used to evaluate variance feature values for passed and failed modules [7]:

$$F = \frac{s_1^2}{s_2^2} \tag{2}$$

where $s_1^2$ and $s_2^2$ are the variances of sample groups 1 and 2, respectively.

### *E. Machine Learning Models*

Several ML models were employed, with their respective parameter settings. Random Forest (RF), Gradient Boosting (GB), and Decision Tree (DT) were configured primarily for reproducibility. XGBoost (XGB, eval_metric=logloss) optimized log-loss during training. Neural Networks were implemented as Multi-Layer Perceptrons (NN, max_iter = 300) to ensure model convergence. Logistic Regression (LR, max_iter=200) was also constrained for convergence stability. All these models used a random state of 42. Naïve Bayes (NB, GaussianNB with default parameters) assumes a Gaussian distribution for continuous features. Finally, K-Nearest Neighbors (KNN, n_neighbors=5, metric=Euclidean distance, weights=uniform) was used [8-15].

### *F. Model Training and Validation*

The model training and data processing operations were carried out on an Intel Core i7-12700 (12th Gen) CPU with 16 GB RAM running Windows 10 Pro through Python 3.10.11. The dataset was divided into training (80%) and testing (20%) subsets, using 5-fold cross-validation techniques to achieve model generalization and prevent overfitting [16]. Hyperparameter optimization was performed using grid search, while early stopping was applied to prevent overfitting. The test data served to determine the complete performance of the models.

### *G. Performance Evaluation*

The evaluation of ML models relies on multiple essential performance metrics for assessment, including accuracy, precision, recall, and F1-score (F1) [17-20]:

$$\text{Accuracy (Ac)} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \tag{3}$$

$$\text{Precision (Pr)} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall (Rc)} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{F1} = 2 \times \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}} \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives, and FP is the number of false positives. AUC-ROC (Area Under the Receiver Operating Characteristic Curve, A-R) provides a measure of how well a model distinguishes between positive and negative instances [21].

## III. RESULTS AND DISCUSSION

The dataset was collected from real-world optical module manufacturing processes. It included various parameters that were recorded throughout the production and testing operations. These parameters served as key input for analyzing the factors influencing yield. The analysis process started with eliminating missing values, together with inconsistent data and potential outliers. Data preprocessing techniques were necessary to guarantee data validity, providing a solid base for both feature selection and model development procedures. The prediction model was optimized through effective feature selection using t-test and F-test, which are suitable for normally distributed data. These statistical tests help identify the parameters with the most significant influence on yield outcomes. This study employed two distinct prediction methods: a step-by-step analysis to evaluate feature contributions at each stage, and a cross-step evaluation to capture interactions across the full process flow from OPA to OSET.

### *A. Statistical Feature Selection for Step-by-Step Prediction*

A statistical approach was utilized to determine the most relevant features for predicting whether a module would pass or fail. Specifically, mean and variance hypothesis testing (t-test and F-test) were conducted at a significance level of 0.05.

This ensured that only features exhibiting statistically significant differences between passed and failed modules were retained, thereby enhancing the predictive power of the model. Table I shows a summary of feature selection across stages. Only representative parameters are reported here due to space limitations; however, the same statistical approach was consistently applied to all features at each stage.

TABLE I. SUMMARY OF STATISTICAL FEATURE SELECTION

| Transition | Features available | Significant ($p < 0.05$) | Example parameters |
|---|---|---|---|
| OPA → LPOT | 49 | 24 | RXPD_D_FA , RFPD |
| LOPT → OTSM | 96 | 71 | Module Power_200ZR-OFEC |
| OTSM → OSET | 214 | 30 | Temperature – PCB_-12dbm_400ZR_OFEC_45_Degree |

### B. Statistical Feature Selection for Cross-Step Prediction

To identify the most relevant features for predicting OSET directly from OPA data, a statistical approach was utilized. Hypothesis tests for mean and variance were performed. Features demonstrating statistically significant correlations between OPA and OSET were included to detect long-range dependencies throughout the manufacturing process. The cross-step approach identified extended dependencies between initial parameters and final yield outcomes. Of the 49 features, 25 features met the selection criteria, all of which were identical step-by-step, with one exception (Temperature-ASIC feature).

To develop an effective predictive model for yield estimation, multiple ML algorithms were implemented and evaluated. To ensure consistency and reproducibility, all models were trained with a random state set to 42 (the best value of 0-100), allowing for fair comparisons across different approaches.

### C. Performance Evaluation

#### 1) Step-by-Step Prediction

This approach involved making predictions through sequential stages. The processes received the predicted output of their previous step, which served as input for the subsequent step. The predictive model started with OPA to generate LOPT, followed by LOPT to produce OTSM, and finally OTSM to create OSET. This approach used a controlled method to identify interdependent factors across workflows by utilizing the intermediate output predictions. Tables II-IV show that ensemble models (RF, XGB, GB) consistently outperformed other approaches in all transitions. Random Forest achieved 85% accuracy and 94% recall in the OPA→LOPT stage, confirming its high sensitivity for early defect detection. In contrast, models such as NN and NB performed poorly, with accuracy dropping below 55% in the LOPT→OTSM stage, highlighting their limited applicability in noisy, high-dimensional production data.

TABLE II. COMPARISON FROM OPA TO LOPT

| Model | Split validation (80:20) (%) | | | | | Test data new set (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ac | Pr | Re | F1 | A-R | Ac | Pr | Re | F1 | A-R |
| RF | 85 | 84 | 94 | 89 | 87 | 86 | 86 | 87 | 87 | 93 |
| XGB | 84 | 84 | 91 | 87 | 89 | 84 | 86 | 82 | 84 | 92 |
| GB | 84 | 83 | 93 | 88 | 88 | 84 | 85 | 83 | 84 | 93 |
| NN | 38 | 0 | 0 | 0 | 32 | 50 | 0 | 0 | 0 | 33 |
| LR | 83 | 83 | 93 | 87 | 86 | 86 | 85 | 87 | 86 | 92 |
| DT | 78 | 83 | 80 | 82 | 77 | 79 | 82 | 74 | 78 | 79 |
| NB | 78 | 81 | 85 | 83 | 80 | 79 | 76 | 83 | 80 | 86 |
| KNN | 72 | 73 | 85 | 79 | 71 | 64 | 61 | 81 | 69 | 69 |

TABLE III. COMPARISON FROM LOPT TO OTSM

| Model | Split validation (80:20) (%) | | | | | Test data new set (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ac | Pr | Re | F1 | A-R | Ac | Pr | Re | F1 | A-R |
| RF | 87 | 84 | 92 | 88 | 92 | 89 | 88 | 89 | 89 | 94 |
| XGB | 88 | 85 | 92 | 89 | 92 | 87 | 85 | 87 | 86 | 95 |
| GB | 87 | 84 | 93 | 88 | 92 | 88 | 88 | 88 | 88 | 95 |
| NN | 52 | 52 | 100 | 69 | 50 | 50 | 50 | 100 | 67 | 50 |
| LR | 52 | 52 | 100 | 69 | 56 | 50 | 50 | 100 | 67 | 60 |
| DT | 83 | 82 | 85 | 84 | 83 | 84 | 84 | 84 | 84 | 84 |
| NB | 51 | 58 | 21 | 30 | 49 | 50 | 52 | 12 | 19 | 42 |
| KNN | 53 | 54 | 62 | 58 | 52 | 53 | 53 | 50 | 51 | 53 |

TABLE IV. COMPARISON FROM OTSM TO OSET

| Model | Split validation (80:20) (%) | | | | | Test data new set (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ac | Pr | Re | F1 | A-R | Ac | Pr | Re | F1 | A-R |
| RF | 79 | 70 | 94 | 81 | 90 | 84 | 79 | 92 | 85 | 93 |
| XGB | 79 | 72 | 87 | 79 | 89 | 86 | 82 | 93 | 87 | 92 |
| GB | 79 | 72 | 87 | 79 | 90 | 86 | 82 | 91 | 86 | 92 |
| NN | 66 | 64 | 55 | 59 | 66 | 71 | 78 | 57 | 66 | 82 |
| LR | 73 | 66 | 83 | 73 | 76 | 84 | 81 | 89 | 85 | 88 |
| DT | 77 | 70 | 85 | 77 | 78 | 76 | 74 | 81 | 77 | 76 |
| NB | 66 | 58 | 85 | 69 | 77 | 71 | 65 | 88 | 75 | 83 |
| KNN | 75 | 66 | 92 | 77 | 83 | 76 | 72 | 86 | 78 | 83 |

TABLE V. ACTUAL AND ML-PREDICTED YIELD (STEP-BY-STEP METHOD)

| Model/target prediction | LOPT | OTSM | OSET | Cumulative yield |
|---|---|---|---|---|
| Qty | 1030 | 1050 | 1040 | |
| Actual Yield | 80.6% | 76.1% | 48.9% | 48.5% |
| RF | 97.9% | 98.5% | 98.7% | 95.4% |
| XGB | 97.8% | 98.4% | 98.9% | 95.2% |
| GB | 92.9% | 95.2% | 92.1% | 81.6% |
| NN | 80.6% | 76.2% | 79.2% | 48.6% |
| LR | 83.1% | 76.2% | 83.0% | 52.6% |
| DT | 97.7% | 98.1% | 98.5% | 94.4% |
| NB | 86.2% | 71.4% | 80.2% | 49.3% |
| KNN | 86.1% | 79.5% | 88.2% | 60.3% |

Table V emphasizes the practical impact of ML-based prediction. Although the cumulative baseline yield at all stages was only 48.5%, RF and XGB nearly doubled this figure to ~95%. These improvements should be understood as effective filtering of defective modules before entering downstream tests. By identifying likely failures earlier, the prediction models prevent unnecessary units from consuming tester time, thereby reducing retesting operations and alleviating tester congestion. In practice, this means a more efficient workflow where good and bad units are separated earlier, test resources are better utilized, and overall process time is significantly shortened, all while maintaining customer quality expectations.

### 2) *Cross-Step Prediction*

This method bypasses intermediate steps by directly predicting the outcome. Instead of sequential predictions, OPA was used directly to predict OSET without using LOPT and OTSM as intermediate variables. This approach evaluates whether skipping intermediate steps affects accuracy and model performance.

Tables VI and VII evaluate the performance of ML models in predicting OSET outcomes using OPA data under a cross-step approach. Table VI presents both split validation and test results, demonstrating strong generalization performance. RF achieved 91% accuracy and 94% AUC-ROC in validation, increasing to 94% accuracy and 99% AUC-ROC in testing. XGB and GB also delivered high accuracy (93–95%) and AUC-ROC (99%) on test data. Table VI highlights the impact of adding Temperature–ASIC as an additional feature, increasing selected features from 24 to 25. This change improved prediction accuracy across models—RF increased from 96.77% to 98.9%, XGB from 97.8% to 98.6%, and NN from 88.3% to 95.7%. The Temperature–ASIC parameter, measured during OPA, showed a strong correlation with OSET outcomes, indicating it is an effective early-stage indicator of potential failure. As both OPA and OSET involve temperature-related conditions, early anomalies in Temperature–ASIC help detect unstable modules earlier, reduce retesting, and optimize downstream process efficiency.

TABLE VI. COMPARISON OF MODELS FROM OPA TO OSET

| **Model** | **Split validation (80:20) (%)** | | | | | **Test data new set (%)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Ac** | **Pr** | **Re** | **F1** | **A-R** | **Ac** | **Pr** | **Re** | **F1** | **A-R** |
| RF | 91 | 92 | 91 | 92 | 94 | 94 | 99 | 89 | 94 | 99 |
| XGB | 87 | 87 | 88 | 88 | 94 | 93 | 99 | 88 | 93 | 99 |
| GB | 89 | 90 | 90 | 90 | 94 | 95 | 98 | 93 | 95 | 99 |
| NN | 75 | 72 | 88 | 79 | 82 | 93 | 96 | 90 | 93 | 96 |
| LR | 83 | 78 | 94 | 85 | 89 | 91 | 95 | 88 | 91 | 98 |
| DT | 80 | 78 | 87 | 82 | 79 | 91 | 93 | 89 | 91 | 91 |
| NB | 69 | 66 | 90 | 76 | 78 | 86 | 86 | 86 | 86 | 92 |
| KNN | 78 | 73 | 95 | 83 | 86 | 91 | 92 | 90 | 91 | 97 |

TABLE VII. YIELD PREDICTION: STEP-BY-STEP VS CROSS-STEP (OSET)

| **Model/target prediction** | **OSET step-by-step** | **OSET cross-step (24 features)** | **OSET cross-step 25 Features** |
|---|---|---|---|
| Qty | 1040 | 1040 | 1040 |
| Actual yield | 78.9% | 78.9% | 78.9% |
| RF | 97.1% | 96.77% | 98.9% |
| XGB | 98.1% | 97.8% | 98.6% |
| GB | 92.3% | 97.1% | 97.9% |
| NN | 80.2% | 88.3% | 95.7% |
| LR | 90.2% | 93.4% | 94.6% |
| DT | 86.6% | 91.6% | 92.6% |
| NB | 82.1% | 82.9% | 85.9% |
| KNN | 88.2% | 90.4% | 91.8% |

Comparable studies repeatedly demonstrate RF as a top-performing approach for predicting manufacturing quality. In semiconductor EDS-yield prediction, RF achieved the lowest error metrics (MAE ~0.520, RMSE ~0.648) among multiple models, enhancing production management [22]. In plastic injection molding quality classification, RF obtained approximately 95% accuracy on real production data [23].

Against this backdrop, the proposed RF-based step-by-step framework for optical modules achieved a cumulative yield of 95.4%, which is comparable to or higher than the results reported in other sectors. This confirms that RF consistently performs well in complex, high-dimensional industrial datasets, validating its application for predictive analytics in optical module manufacturing.

### D. *Practical Implementation and Process Efficiency Improvements*

To assess the practical benefits of the ML-based prediction model in actual production, a batch of 100 optical modules was evaluated under two conditions: Figure 1 illustrates the conventional process, where 53 units underwent retesting across LOPT, OTSM, and OSET stages, resulting in a cumulative retest time of approximately 697 minutes. Only 22 units were routed directly to the debug station, and the final yield remained at 67 units.

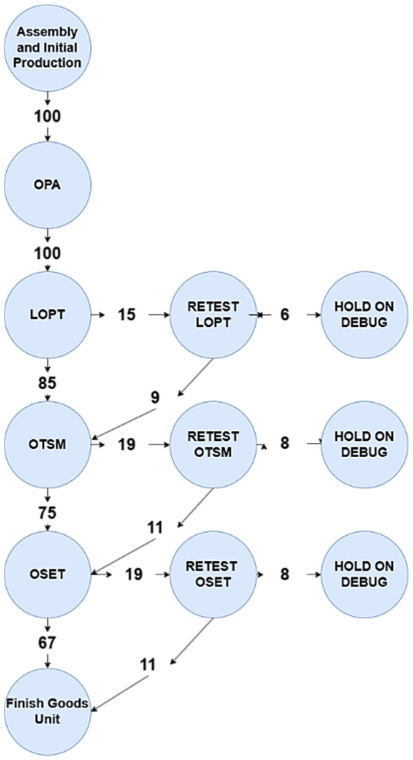

Fig. 1. Conventional optical module testing process.

Figure 2 shows the ML-assisted process based on the step-by-step prediction model. In this scenario, only 3 units were sent to the debug station after failing testing, while 30 units were correctly predicted as likely failures and routed out before unnecessary retesting. This shift led to a 94% reduction in retest operations, saving approximately 795 minutes (13.25 hours) per batch. The final output of 67 units was maintained, confirming that predictive analytics improved testing efficiency without compromising yield quality. These results highlight the real-world impact of predictive analytics not on yield increase, but on process optimization, by minimizing retesting, reducing tester load, and accelerating production throughput.

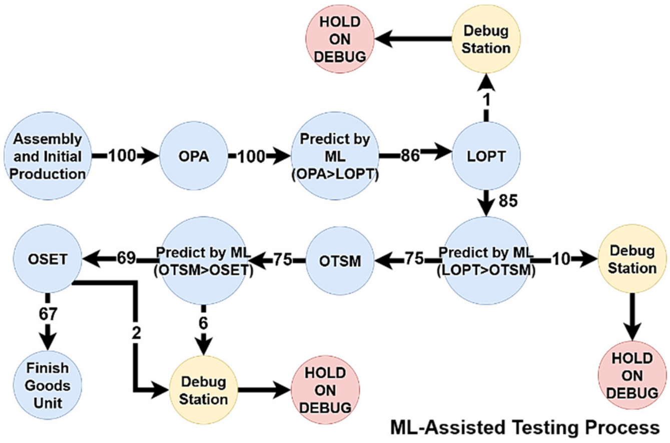

Fig. 2. Testing process with ML-based step-by-step prediction.

## IV. CONCLUSION

This study investigated the application of ML methods to optimize the testing process in the manufacturing of optical modules. By integrating predictive models across the standard testing sequence—OPA, LOPT, OTSM, and OSET—this study demonstrated how early-stage prediction can streamline production and reduce unnecessary retesting. A structured method was used, involving data preparation, statistical hypothesis testing (t-test and F-test), and multiple evaluations of the ML models. Among the models tested, RF and XGB showed the highest predictive accuracy.

Two prediction strategies were explored: step-by-step prediction, which used sequential test data, and cross-step prediction, which utilized only OPA data to predict the final OSET outcomes. The cross-step approach, particularly with the inclusion of the Temperature–ASIC feature, achieved up to 98.9% prediction accuracy using the RF model. Importantly, ML did not increase the number of good units, but it enabled early detection of likely failures. In practice, retesting was reduced from 53 units to just 3 per 100-unit batch, saving approximately 795 minutes of total testing time, without changing the final output quantity of 67 units. This confirms that ML-based prediction contributes to process optimization. The ability to bypass unnecessary stages, such as LOPT and OTSM, shortens the test cycle per unit from 13 hours to 8 hours, leading to higher throughput and less testing congestion. The findings emphasize the operational value of predictive analytics in high-volume manufacturing environments. By allowing faster routing decisions and early intervention, manufacturers can increase process efficiency while maintaining product quality.

Future work should focus on deploying models in real-time, integrating with automated control systems, and extending the approach to other manufacturing domains. Rather than reducing defects, ML enables smarter handling of them—transforming testing from a reactive process into a proactive strategy.

## ACKNOWLEDGMENT

This research was supported by the Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Thailand.

## REFERENCES

[1] A. Affandi and O. AL-Rusaini, "Implementation of Fiber Optic Communication System Using Developed Computer Program," *IOSR Journal of Electrical and Electronics Engineering*, vol. 11, no. 3, pp. 122–142, May 2016.

[2] U. J. Umoga *et al.*, "Exploring the potential of AI-driven optimization in enhancing network performance and efficiency," *Magna Scientia Advanced Research and Reviews*, vol. 10, no. 1, pp. 368–378, Feb. 2024, https://doi.org/10.30574/msarr.2024.10.1.0028.

[3] M. I. H. Okfie and S. Mishra, "Anomaly Detection in IIoT Transactions using Machine Learning: A Lightweight Blockchain-based Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14645–14653, Jun. 2024, https://doi.org/10.48084/etasr.7384.

[4] M. Zeng and C. Chen, "Design of High-speed Data Transmission System Based on Fiber-PCIE," *Frontiers in Business, Economics and Management*, vol. 7, no. 3, pp. 84–89, Feb. 2023, https://doi.org/10.54097/fbem.v7i3.5397.

[5] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, 2024, https://doi.org/10.14569/IJACSA.2024.0150235.

[6] B. Gerald, "A Brief Review of Independent, Dependent and One Sample t-test," International Journal of Applied Mathematics and Theoretical Physics, vol. 4, no. 2, p. 50, 2018, https://doi.org/10.11648/j.ijamtp.20180402.13.

[7] R. Odek and G. Opuodho, "F-test and P-values: A synopsis," *Journal of Management and Science*, vol. 13, no. 2, pp. 59–61, June 2023, https://doi.org/10.26524/jms.13.22.

[8] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, June 2024, https://doi.org/10.58496/BJML/2024/007.

[9] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, no. 21, pp. 1–21, 2013, https://doi.org/10.3389/fnbot.2013.00021.

[10] Z. A. Ali, Z. H. Abduljabbar, H. A. Tahir, A. B. Sallow, and S. M. Almufti, "eXtreme Gradient Boosting Algorithm with Machine Learning: a Review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, https://doi.org/10.25007/ajnu.v12n2a1612.

[11] R. Qamar and B. A. Zardari, "Artificial Neural Networks: An Overview," *Mesopotamian Journal of Computer Science*, vol. 2023, pp. 124–133, Aug. 2023, https://doi.org/10.58496/MJCSC/2023/015.

[12] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, Sept. 2002, https://doi.org/10.1080/00220670209598786.

[13] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74–78, Oct. 2018, https://doi.org/10.26438/ijcse/v6i10.7478.

[14] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, no. 2, June 2019, Art. no. 91, https://doi.org/10.3991/ijes.v7i2.10659.

[15] P. Cunningham and S. J. Delany, "K-Nearest Neighbour Classifiers - A Tutorial," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–25, July 2022, https://doi.org/10.1145/3459665.

[16] M. M. Abdulrazzaq, I. F. Yaseen, S. A. Noah, and M. A. Fadhil, "Multi-Level of Feature Extraction and Classification for X-Ray Medical Image," *Indonesian Journal of Electrical Engineering and Computer*

*Science*, vol. 10, no. 1, pp. 154–167, Apr. 2018, https://doi.org/10.11591/ijeecs.v10.i1.pp154-167.

[17] H. Dong, R. Liu, and A. W. Tham, "Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation," *Journal of Risk and Financial Management*, vol. 17, no. 2, Jan. 2024, Art. no. 5, https://doi.org/10.3390/jrfm17020050.

[18] E. J. Michaud, Z. Liu, and M. Tegmark, "Precision Machine Learning," *Entropy*, vol. 25, no. 1, Jan. 2023, Art. no. 175, https://doi.org/10.3390/e25010175.

[19] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 2229–3981, 2011.

[20] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in *AI 2006: Advances in Artificial Intelligence*, vol. 4304, A. Sattar and B. Kang, Eds. Springer Berlin Heidelberg, 2006, pp. 1015–1021

[21] M. Majnik and Z. Bosnić, "ROC analysis of classifiers in machine learning: A survey," *Intelligent Data Analysis*, vol. 17, no. 3, pp. 531–558, May 2013, https://doi.org/10.3233/IDA-130592.

[22] Y. Lee and Y. Roh, "An Expandable Yield Prediction Framework Using Explainable Artificial Intelligence for Semiconductor Manufacturing," *Applied Sciences*, vol. 13, no. 4, Feb. 2023, Art. no. 2660, https://doi.org/10.3390/app13042660.

[23] A. Polenta, S. Tomassini, N. Falcionelli, P. Contardo, A. F. Dragoni, and P. Sernani, "A Comparison of Machine Learning Techniques for the Quality Classification of Molded Products," *Information*, vol. 13, no. 6, May 2022, Art. no. 272, https://doi.org/10.3390/info13060272.