Statistical Quality Control and Machine-Learning–Based Anomaly Detection on the SECOM Semiconductor Manufacturing Dataset

Mobasshira Zaman

11/30/2023

Summary

This project integrates classical Statistical Quality Control (SQC) methods with modern anomaly detection and supervised learning techniques to analyze the SECOM semiconductor manufacturing dataset. The analysis includes univariate process monitoring using $\bar{X}$ and R charts, EWMA, and CUSUM for a selected sensor, followed by multivariate anomaly detection using both the original feature space and PCA-reduced space. In addition, supervised learning models (Logistic Regression and Random Forest) are trained using PCA-reduced features combined with SMOTE oversampling. The study contrasts interpretability and detection performance across these methods. Traditional control charts clearly highlight local instability in the selected sensor. Isolation Forest achieves 86.2% accuracy but detects only a small portion of faulty wafers. Supervised learning improves certain fault-detection metrics but faces challenges due to extreme class imbalance and noisy high-dimensional data.

1 Introduction

Semiconductor manufacturing systems involve numerous tightly controlled steps monitored by hundreds of sensors. Even small deviations in the process can significantly affect output quality and yield. Classical SQC charts such as $\bar{X}$, R, EWMA, and CUSUM provide interpretable, real-time process monitoring for individual variables. However, contemporary semiconductor datasets are high dimensional, noisy, and heavily imbalanced, making them suitable candidates for machine-learning–based anomaly detection. This project examines how both approaches perform on the SECOM dataset, comparing their strengths, limitations, and implications for yield monitoring.

2 Data Source

The SECOM Semiconductor Manufacturing Dataset, obtained from the UCI Machine Learning Repository,[1] contains 1,567 wafer production records and 590 continuous sensor readings for each record. Each observation is labeled as either normal (0) or faulty (1). The dataset is known for its severe class imbalance, with fewer than five percent of samples labeled as faulty, and for containing a large number of missing values. These characteristics make the dataset representative of real industrial monitoring challenges, where faults are both rare and difficult to distinguish from background noise.

[1] `https://archive.ics.uci.edu/ml/datasets/SECOM`

3 Data Preprocessing

Several preprocessing steps were required before applying SQC and machine-learning methods. Missing values were replaced using mean imputation for each feature. All features were standardized using Z-score normalization to ensure unit scale across the 590-dimensional space. Labels were mapped from the original $-1/+1$ convention to 0 and 1. A stratified 70/30 train–test split preserved the imbalance structure.

For dimensionality reduction, Principal Component Analysis (PCA) was applied to the standardized features. Retaining 95% of total variance reduced the dimensionality from the original 590 features to 163 principal components. PCA helped reduce noise and improved computational tractability for downstream models. The Synthetic Minority Over-sampling Technique (SMOTE) was applied only on the training portion of the PCA-reduced data to better balance minority faulty samples. Before SMOTE, the training set contained 1023 normal samples and 73 faulty samples; after SMOTE, both classes contained 1023 samples.

4 Methodology

The analysis proceeded in three main stages. First, univariate SQC charts were applied to a selected sensor measurement (Feature 0). $\bar{X}$ and R charts were constructed using subgroups of size five, and classical SPC constants were used to compute control limits. To detect more subtle patterns, EWMA and two-sided CUSUM charts were constructed using standard formulations. EWMA applied a smoothing parameter of $\lambda = 0.2$, while CUSUM used a reference value of $k = 0.5\sigma$ and decision interval $h = 5\sigma$.

The second stage involved applying Isolation Forest for unsupervised multivariate anomaly detection. Two versions were evaluated: one using the full 590-dimensional feature space and another using PCA-reduced features. Both models were trained exclusively on the training split and then evaluated on the held-out test set.

In the third stage, supervised learning methods were applied. Logistic Regression and Random Forest were trained on the SMOTE-balanced PCA-reduced data. Performance was evaluated on the original imbalanced test set to assess practical fault-detection capability.

5 Results

5.1 $\bar{X}$ Chart

The $\bar{X}$ chart for Feature 0 revealed one distinct out-of-control subgroup exceeding the upper control limit, suggesting a possible shift in the process mean. While the majority of subgroup means remained within control limits, several clusters displayed moderate fluctuations, which may reflect sensor noise inherent to semiconductor measurements.

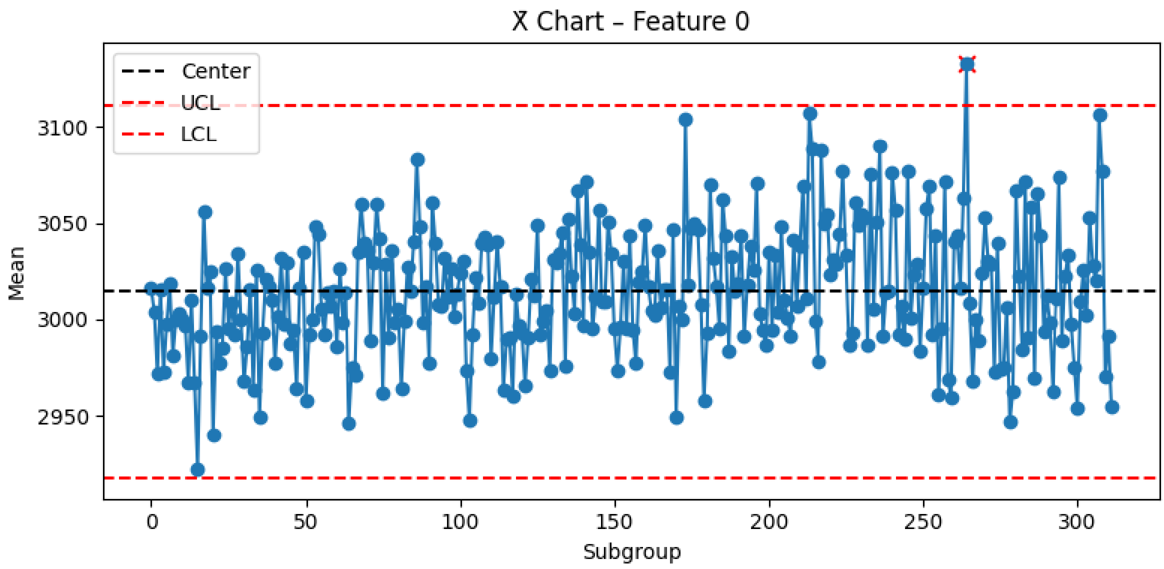

Figure 1: $\bar{X}$ Control Chart for Selected SECOM Sensor (Feature 0)

5.2 R Chart

The R chart showed multiple groups exceeding the upper control limit, indicating increased short-term variability. These excursions may correspond to abrupt operational disturbances or inconsistent tool performance. The presence of several high-range subgroups suggests unstable variability in the selected sensor reading.

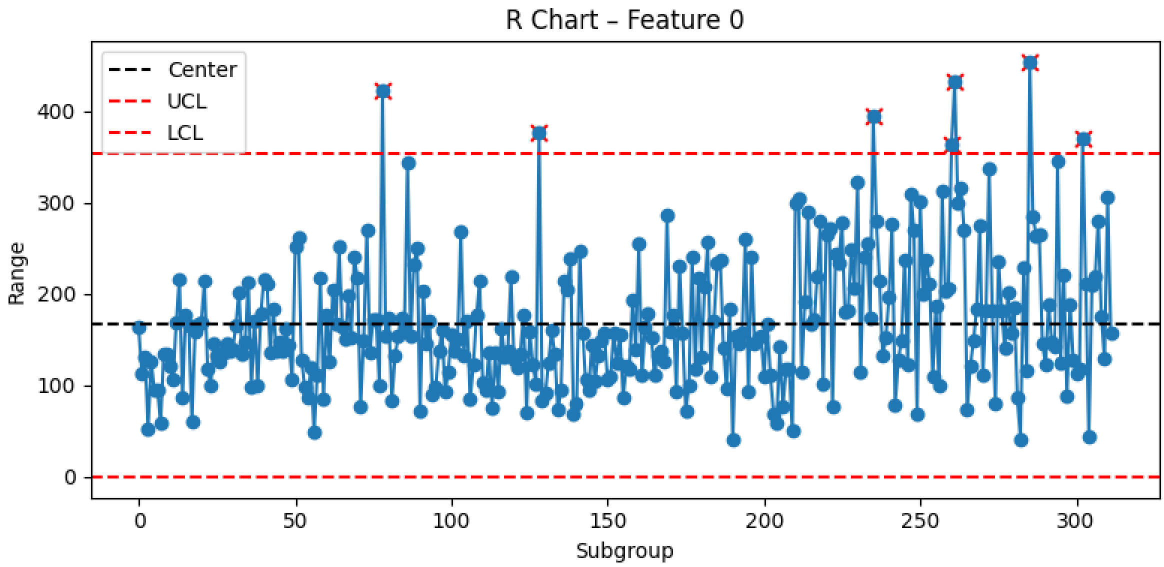

Figure 2: R Chart for Selected SECOM Sensor (Feature 0)

5.3 EWMA Chart

The EWMA chart for Feature 0 revealed several gradual deviations approaching the upper control limit. Compared to the $\bar{X}$ chart, the smoothed trajectory highlighted smaller yet persistent drifts that might otherwise remain hidden. These slow-moving deviations may reflect long-term tool degradation or environmental changes.

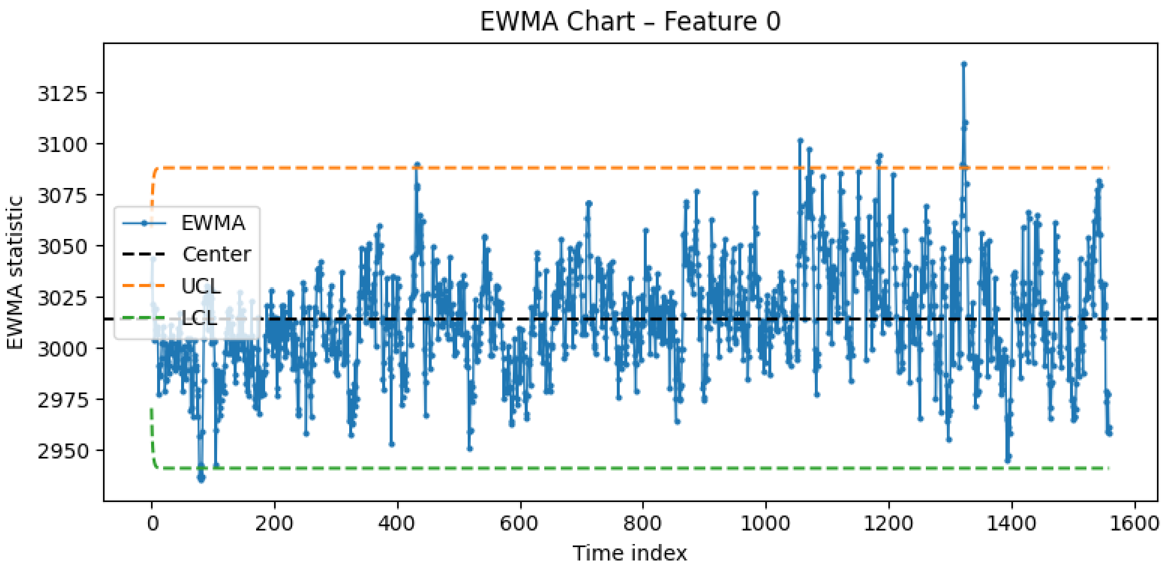

Figure 3: EWMA Chart – Feature 0

5.4 CUSUM Chart

The CUSUM chart demonstrated high sensitivity to small shifts in the process mean. Several segments of the positive and negative CUSUM traces came close to or exceeded the decision interval, indicating possible sustained deviations. The cumulative nature of CUSUM allowed it to detect multiple small changes that neither EWMA nor $\bar{X}$ charts captured as clearly.

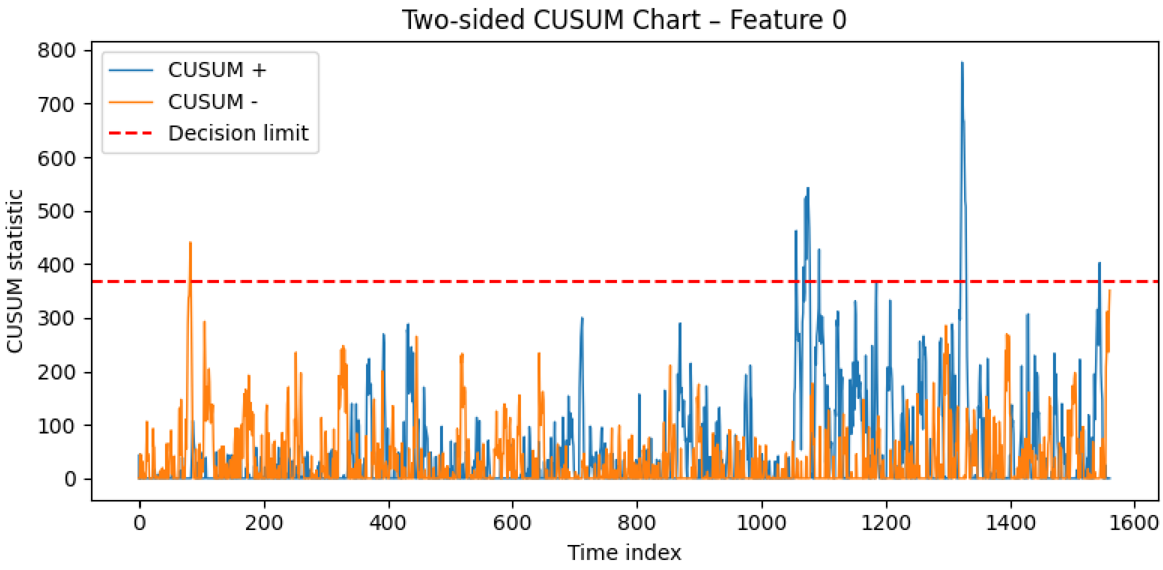

Figure 4: Two-sided CUSUM Chart – Feature 0

5.5 Isolation Forest (Full Feature Space)

The Isolation Forest applied to all 590 features achieved an overall accuracy of 86.2% on the test set. However, the model detected only eight of the 31 faulty wafers, leading to low recall for the minority class. The anomaly scores plotted across samples showed several low-valued points representing high anomaly likelihood, although many faulty samples were not isolated effectively.

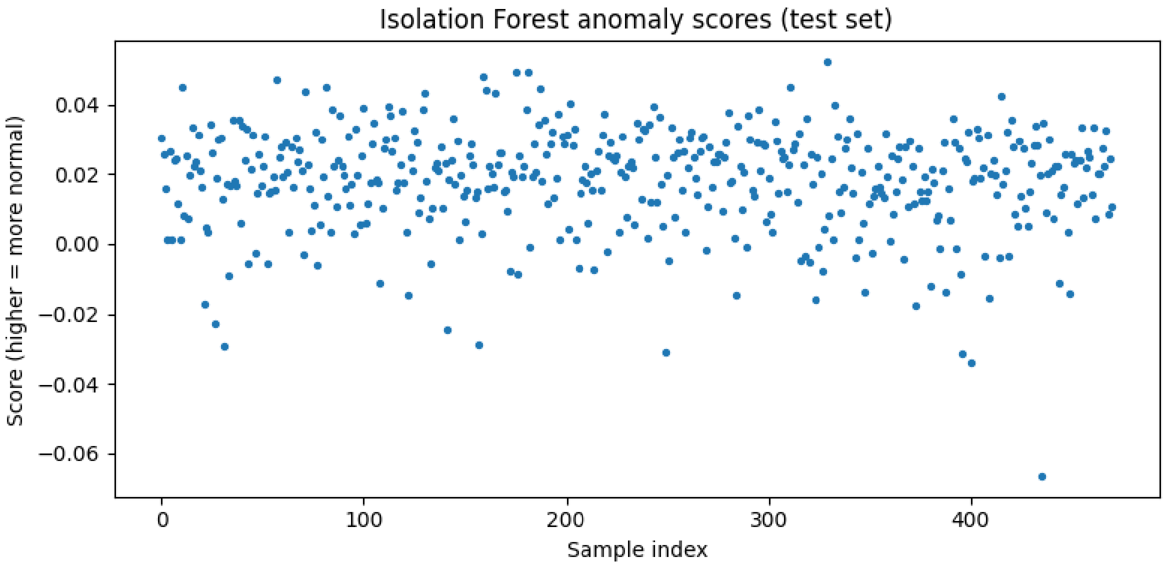

Figure 5: Isolation Forest Anomaly Scores (Full Feature Space)

The confusion matrix was:

$$\begin{bmatrix} 398 & 42 \\ 23 & 8 \end{bmatrix}$$

5.6 Isolation Forest (PCA Space)

Applying Isolation Forest to PCA-reduced data produced similar behavior. The PCA-based model slightly shifted the distribution of anomaly scores but did not significantly improve minority-class detection. The overall structure remained comparable to the full-space model, implying that dimensionality reduction alone does not overcome the extreme class imbalance.

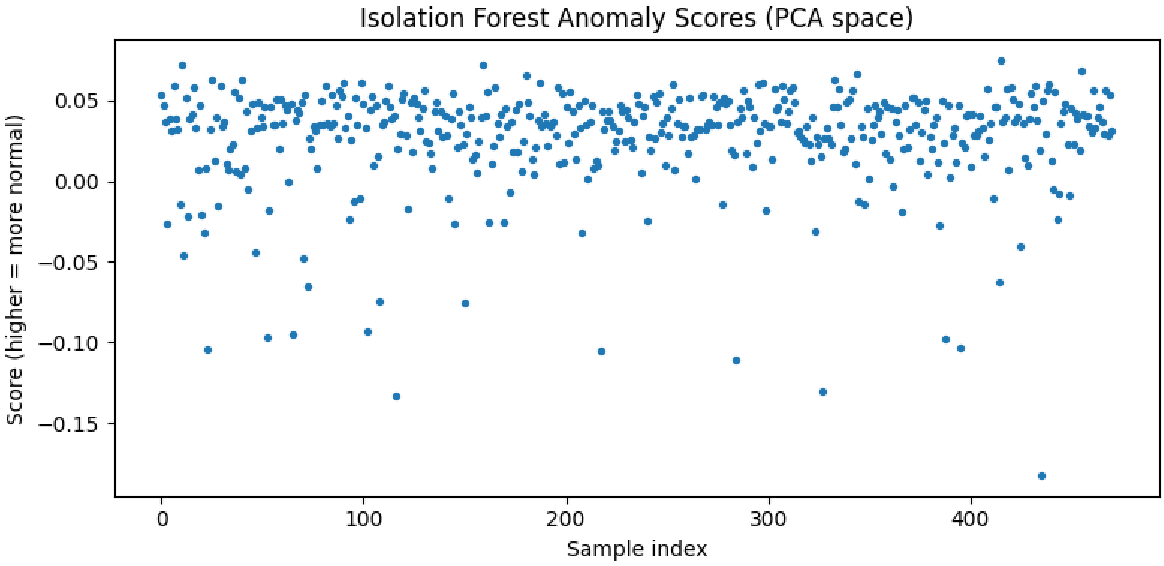

Figure 6: Isolation Forest Anomaly Scores (PCA Space)

5.7 Logistic Regression on PCA + SMOTE

Logistic Regression trained on the SMOTE-balanced PCA components achieved an overall accuracy of 81.1%. The model identified five faulty wafers, corresponding to a faulty-class recall of 16.1%. This is an improvement over Isolation Forest, but at the cost of more false alarms. The confusion matrix was:

$$\begin{bmatrix} 377 & 63 \\ 26 & 5 \end{bmatrix}$$

5.8 Random Forest on PCA + SMOTE

Random Forest achieved the highest overall accuracy of 93.2% but failed to identify any faulty samples. The model predicted almost all test samples as normal, demonstrating a typical failure mode in heavily imbalanced classification. While accuracy was high, the zero recall for faulty wafers limits its practical value. The confusion matrix was:

$$\begin{bmatrix} 439 & 1 \\ 31 & 0 \end{bmatrix}$$

6 Discussion

The results demonstrate that classical control charts remain highly interpretable and effective for analyzing individual process variables. The R chart indicated unstable variability, while EWMA and CUSUM detected gradual and cumulative shifts not visible in the $\bar{X}$ chart. However, SQC charts are inherently univariate and cannot utilize the high-dimensional structure of semiconductor data.

Isolation Forest offered a multivariate perspective but struggled to detect rare faults. Although PCA helped reduce noise and improved model stability, the unsupervised nature of Isolation Forest, combined with extreme class imbalance, prevented meaningful improvement in minority-class recall.

Supervised learning methods benefitted from SMOTE oversampling, particularly in enabling Logistic Regression to detect some faulty wafers. However, Random Forest illustrated the pitfalls of relying solely on accuracy as a performance metric, as it detected no faulty wafers despite excellent accuracy. The combination of PCA and SMOTE improved model tractability and class balance, but the complexity of the underlying process still posed challenges for fault detection.

7 Conclusion

This study illustrates the strengths and limitations of both traditional SQC and modern ML-based approaches for semiconductor process monitoring. Classical SQC charts provide interpretable diagnostics and can readily identify abnormal patterns in individual sensors. Machine-learning models offer richer multivariate detection but face challenges stemming from noise, high dimensionality, and severe imbalance. PCA reduces dimensionality and improves model efficiency, while SMOTE can partially alleviate imbalance in supervised settings. Nevertheless, no model achieved consistently strong recall for faulty wafers, emphasizing the difficulty of rare-event detection in industrial manufacturing.

Future work could explore cost-sensitive learning, anomaly scoring ensembles, Hotelling T^2 and multivariate EWMA charts, and deep-learning methods tailored for imbalanced industrial data.

References

Montgomery, D. C. *Introduction to Statistical Quality Control.* Wiley. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). "Isolation Forest." *Proceedings of ICDM.* UCI Machine Learning Repository. SECOM Dataset. `https://archive.ics.uci.edu/ml/datasets/SECOM`