Analysis of Silicon Wafer defects in a production line

N. Heidaryan

I downloaded data on silicon wafer production from a semiconductor company's production line from Kaggle at the following link: https://www.kaggle.com/datasets/programmer3/semiconductor-sensor-data-for-predictive-quality

Data Description:

Data Card | Code (0) | Discussion (0) | Suggestions (0)

About Dataset

This dataset contains 4,219 real-time sensor readings simulated to reflect the advanced processes used in semiconductor fabrication — specifically during lithography, etching, and deposition.

Each row corresponds to a unique wafer processing instance, with measurements recorded from critical fabrication tools and environmental sensors. These sensor readings are used to determine the quality status of a wafer — whether it will be a Joining wafer (successfully proceeds to assembly) or a Non-Joining wafer (defective and rejected from the line).

Non-Joint Modelling: It is performed independently to predict quality control (fault prediction), without prior filtering or classification of faulty wafers.

Joint Modelling: A two-step model that first detects faulty wafers via classification, then uses prediction (including normal and faulty wafers) only on normal wafers to improve prediction accuracy and reduce error noise

Possible Questions:

1- Predict whether a wafer will fail or pass before process completion.
Key deliverable: “Predictive model to identify defective wafers early with 90% accuracy.”

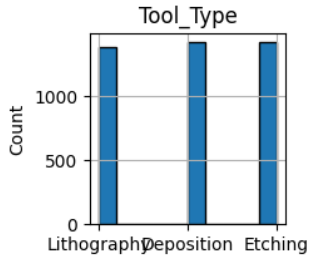
2- Which process parameters (temperature, pressure, gas flow, etching depth) most influence wafer failure?

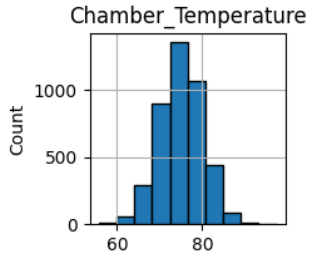
3- Train models to detect “abnormal” sensor values (Isolation Forest, Autoencoders).
Goal: find process runs that are unusual even if they haven’t failed yet.

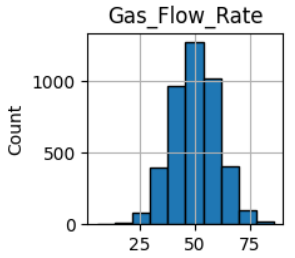
4- Cluster “good wafers” vs. “bad wafers” with K-Means or DBSCAN. Identify safe operating ranges for each parameter.
Deliverable: “Optimal process window: wafers most likely to pass when chamber temp = X–Y, vacuum = A–B, …”

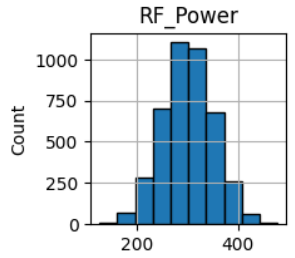
Features (parameters):

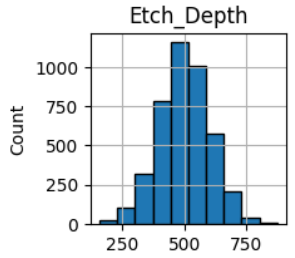
Process_ID
Timestamp
Tool_Type
Wafer_ID
Chamber_Temperature
Gas_Flow_Rate
RF_Power
Etch_Depth
Rotation_Speed
Vacuum_Pressure
Stage_Alignment_Error
Vibration_Level
UV_Exposure_Intensity
Particle_Count
Defect
Join_Status

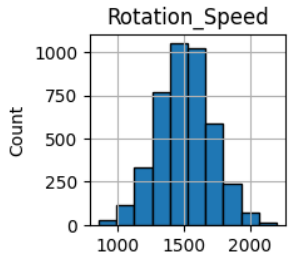


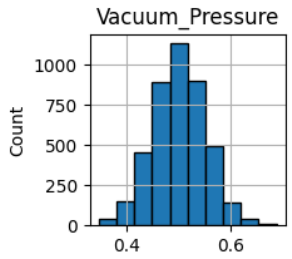


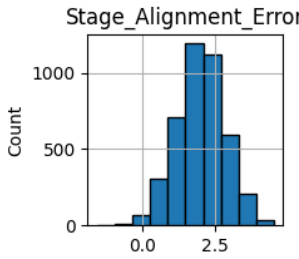


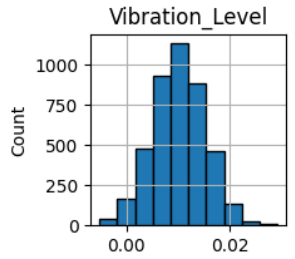


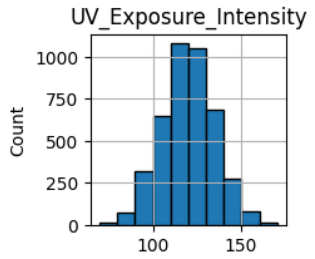


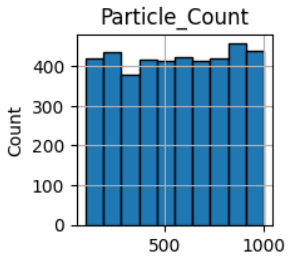


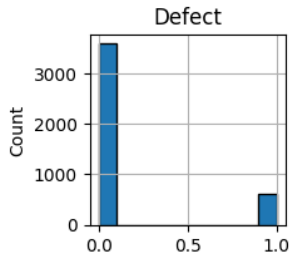


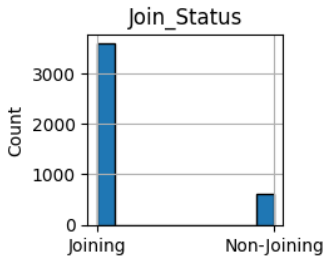


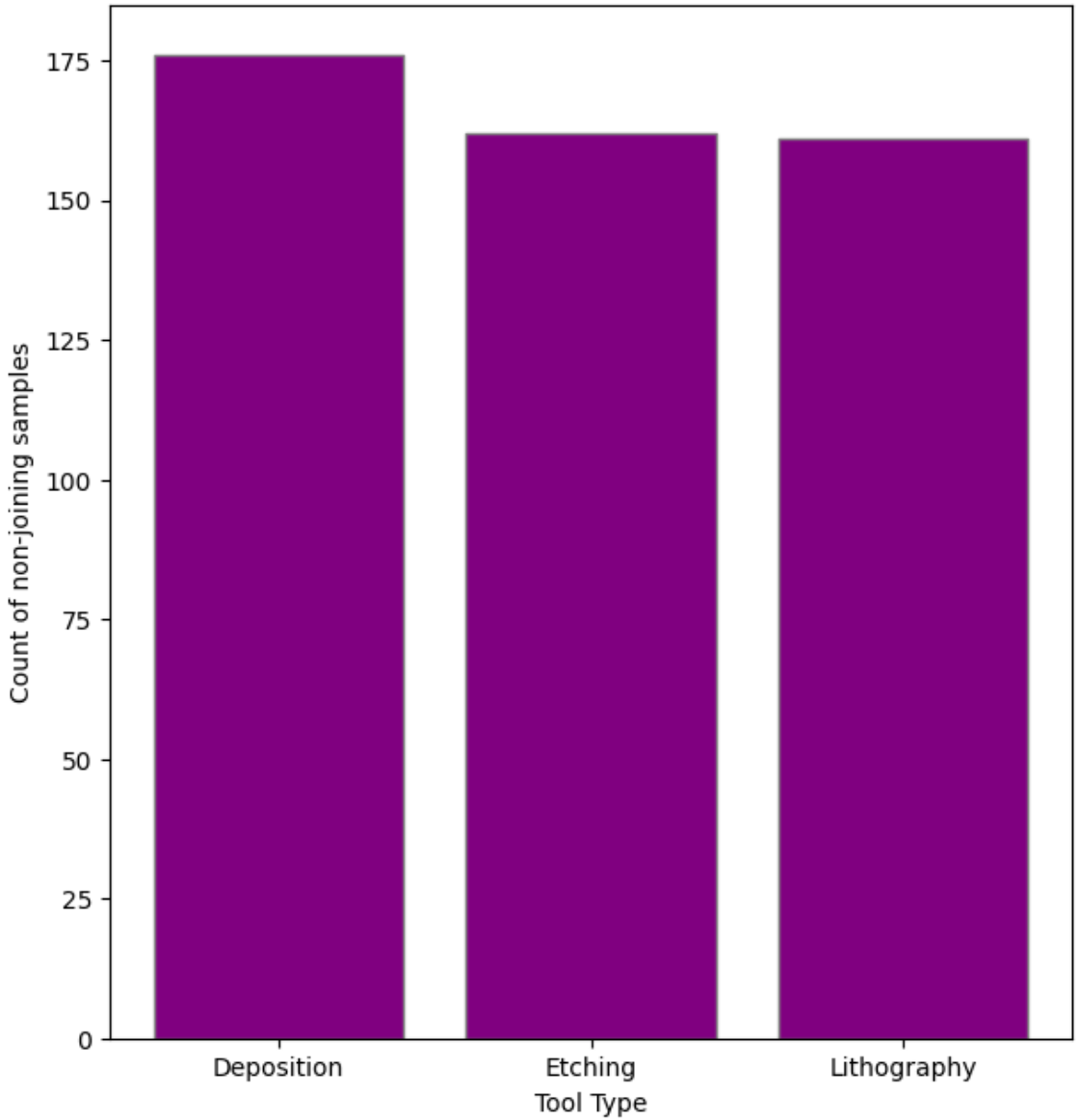


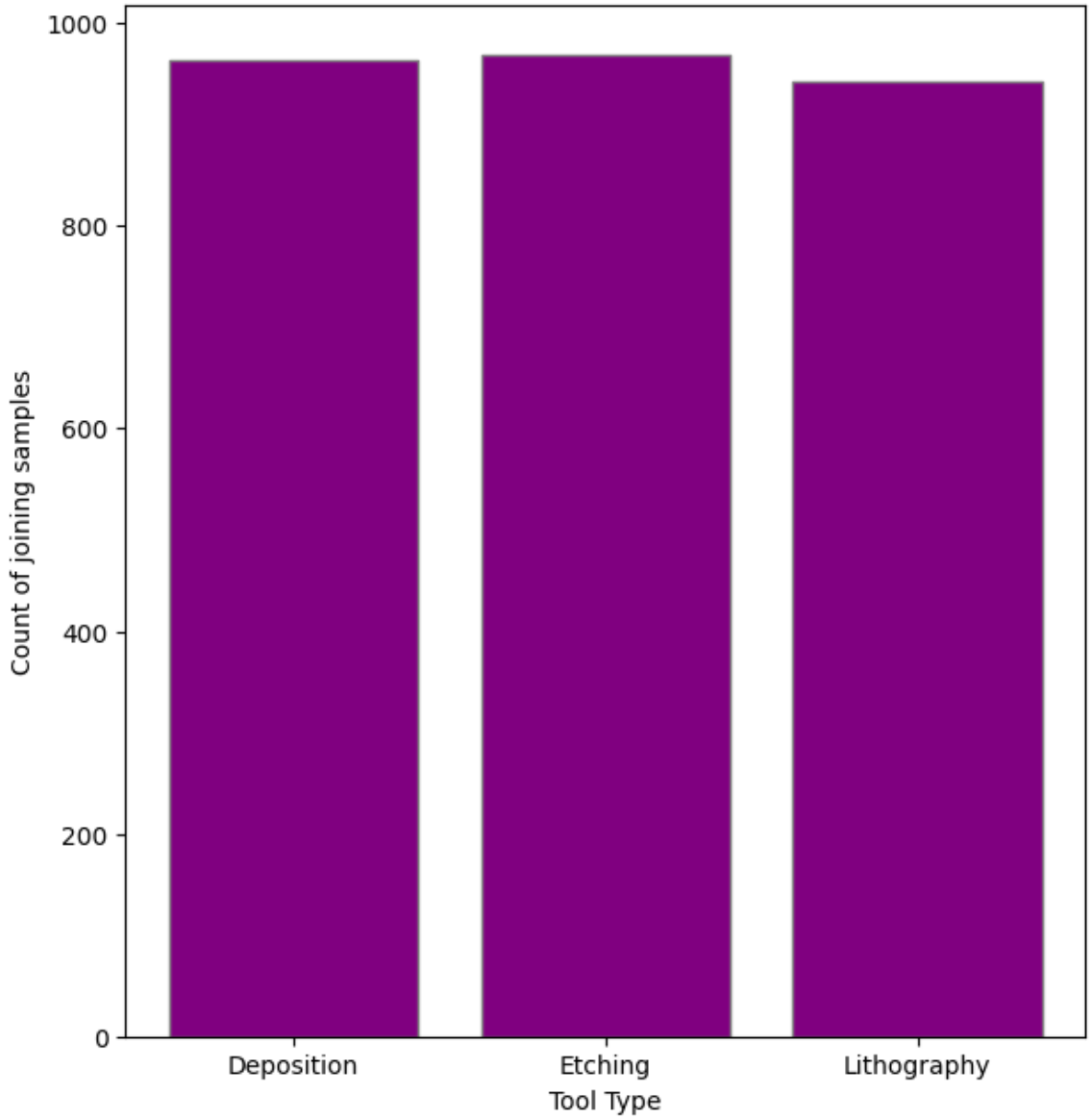


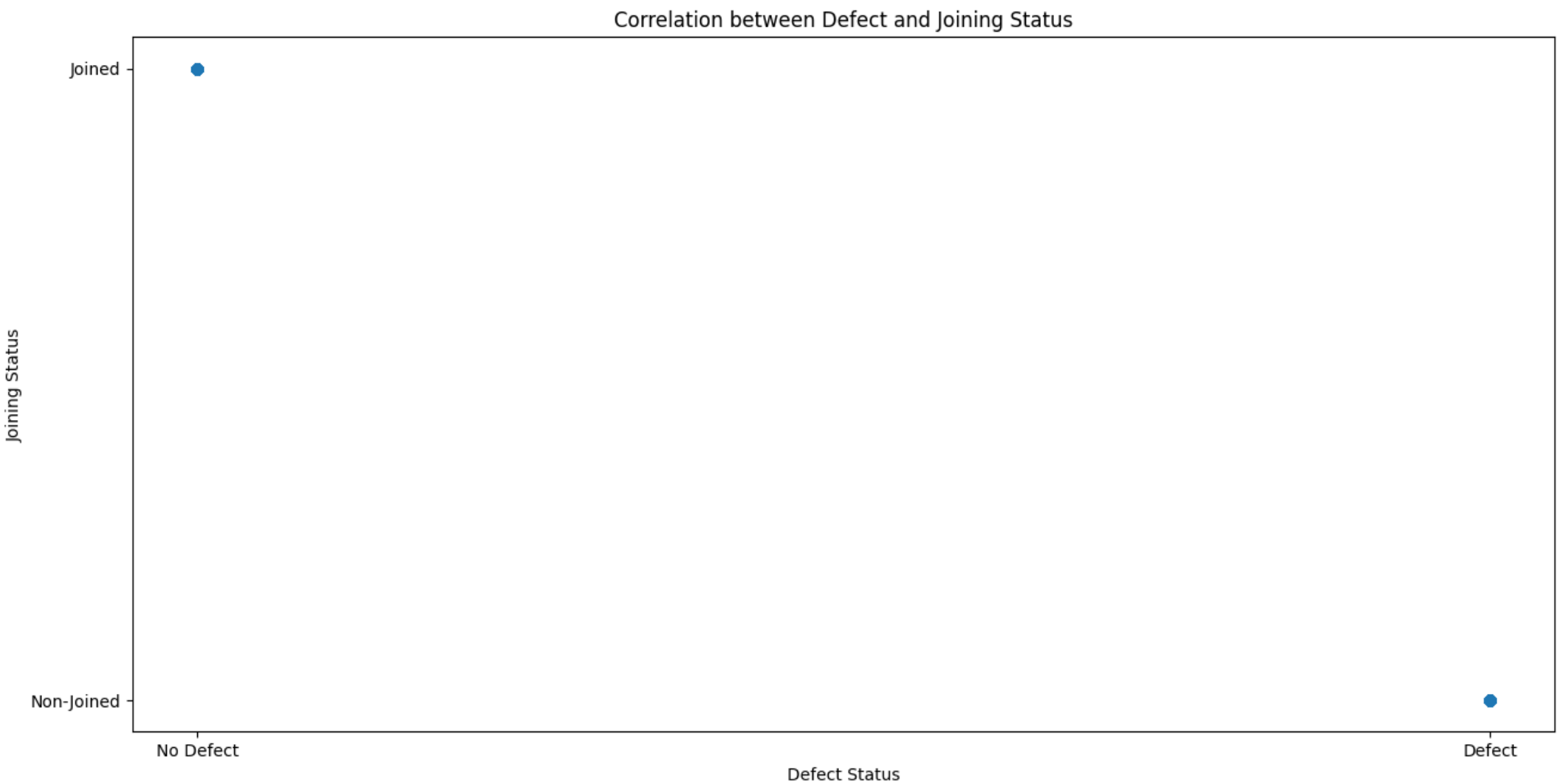







Count of non-joining samples
0
25
50
75
100
125
150
175
Deposition
Etching
Lithography
Tool Type


Count of joining samples
0
200
400
600
800
1000
Deposition
Etching
Lithography
Tool Type


Correlation between Defect and Joining Status
Joined
Non-Joined
Joining Status
No Defect
Defect
Defect Status

- There is no correlation between number of particles and other features:

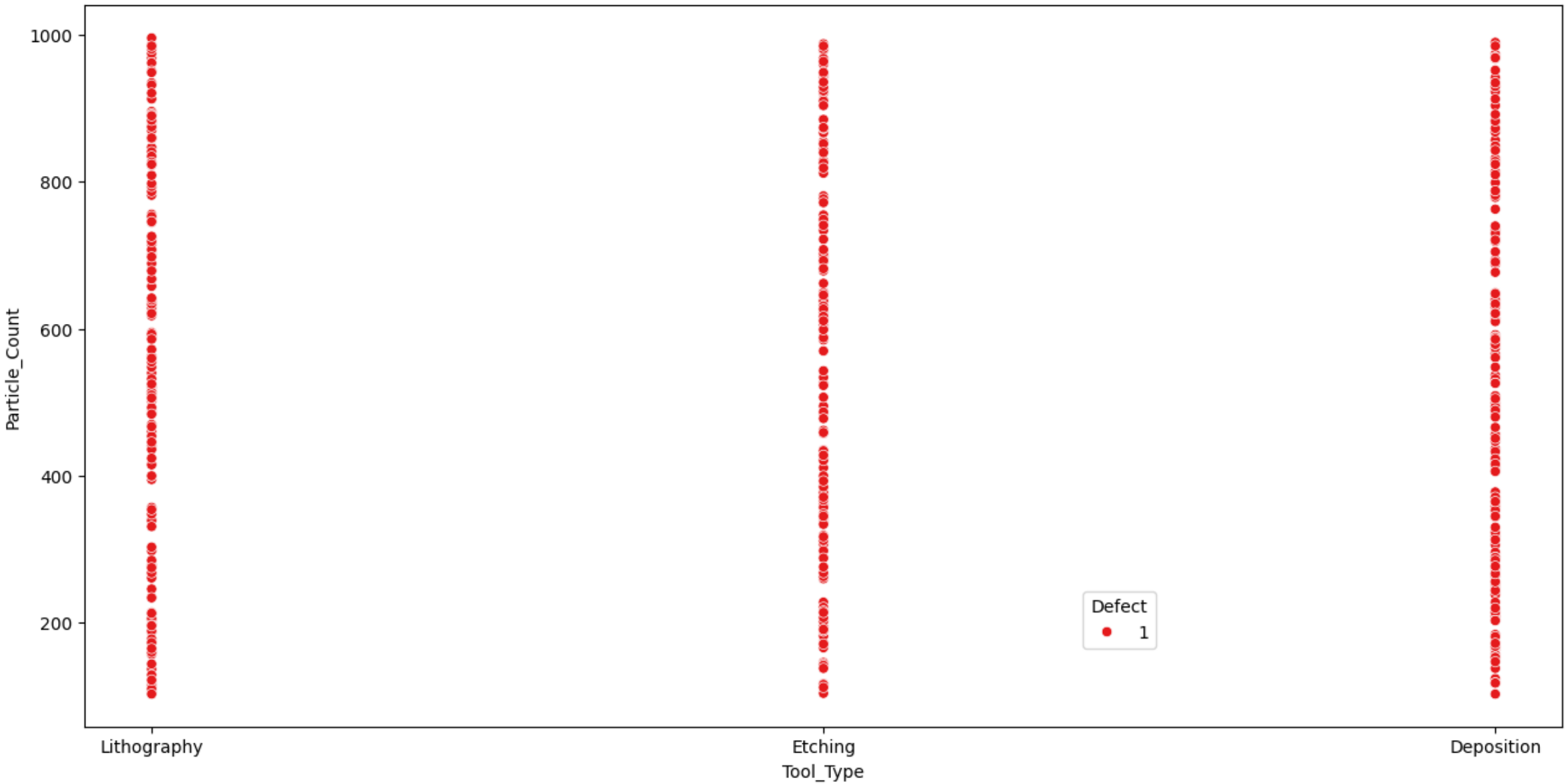
	Target	Feature	Pearson_coef	p_value
0	Particle_Count	Chamber_Temperature	-0.009082	0.555342
1	Particle_Count	Gas_Flow_Rate	-0.006903	0.653994
2	Particle_Count	RF_Power	-0.008946	0.561310
3	Particle_Count	Etch_Depth	-0.009274	0.547041
4	Particle_Count	Rotation_Speed	0.000126	0.993482
5	Particle_Count	Vacuum_Pressure	0.003901	0.800007
6	Particle_Count	Stage_Alignment_Error	-0.021193	0.168721
7	Particle_Count	Vibration_Level	-0.015467	0.315189
8	Particle_Count	UV_Exposure_Intensity	0.008222	0.593407
9	Particle_Count	Defect	-0.028406	0.065052
10	Particle_Count	join_helper	0.028406	0.065052


Chamber Temperature(°C)
90
80
70
60
No Defect
Defect
Defect Status


RF power (w)
450
400
350
300
250
200
150
No Defect
Defect
Defect Status


Rotation Speed (rpm)
2200
2000
1800
1600
1400
1200
1000
800
No Defect
Defect
Defect Status


Vibration Level
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−0.005
No Defect
Defect
Defect Status


1000
800
600
400
200
Particle_Count
Lithography
Etching
Deposition
Tool_Type
Defect
1

Conclusion:

- The analysis of the semiconductor production dataset shows that particle count, the key defect indicator, does not show a meaningful linear correlation with any single process variable. This suggests that wafer defects are not driven by one dominant parameter and may instead result from complex multivariate interactions, equipment behavior, or temporal effects not captured in the available data.
- However, the absence of strong correlations may also reflect limitations in the dataset itself. The data appears clean and stable, but it may lack sufficient variability, contain hidden bias (e.g., only “normal-operation” samples), or exclude other influential factors such as equipment aging, maintenance cycles, contamination history, or operator-dependent steps. Because of these factors, simple statistical analysis is not enough to explain defect formation.
- Overall, the results indicate that more advanced methods - multivariate modelling, anomaly detection, clustering, or machine-learning approaches - are required to uncover defect patterns. Additional process variables or a more diverse dataset would further improve the ability to predict wafer failures and identify their root causes.