Anomaly Detection in Ion Beam Etching Processes

Daniel Hamama

ID: 318652252

Holon Institute of Technology

Student of BSc Applied Mathematics

Introduction:

- **Context:**

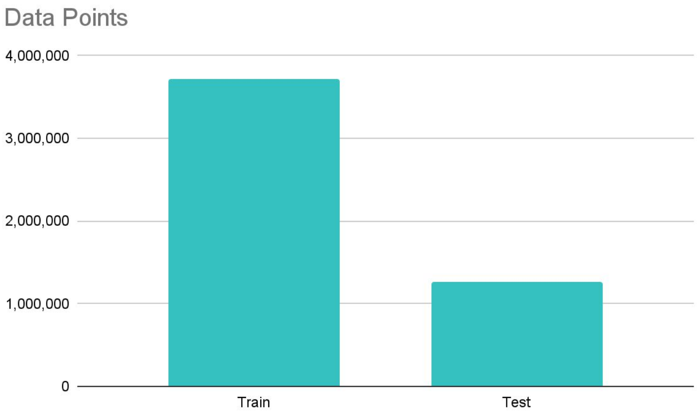
  Ion beam etching is a critical process in manufacturing high-precision components. Monitoring this process in real time helps avoid costly production failures.

- **Objective:**

  To develop an effective and scalable anomaly detection framework leveraging LSTM models, aimed at identifying rare faults in time-series dataset.

Dataset Overview

- **Data Source:** Sensor logs from an Ion Beam Etching machine.
- **Train Data:** 3.7 Million data points, 27 features.
- **Anomaly Variable:** ‘fault’ (Binary - 0: Normal, 1: Fault).
- **Test Data:** 1.2 Million data points, 24 features.

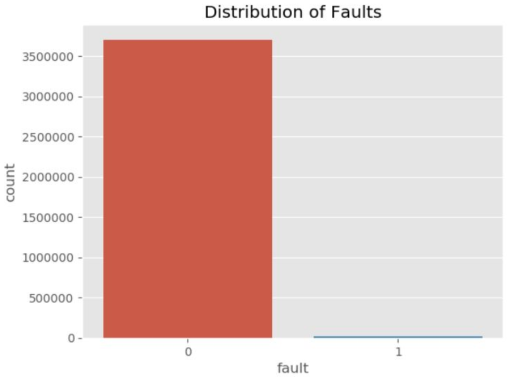


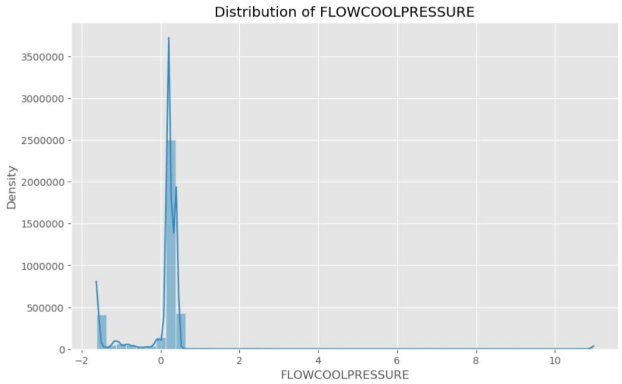
Project Methodology

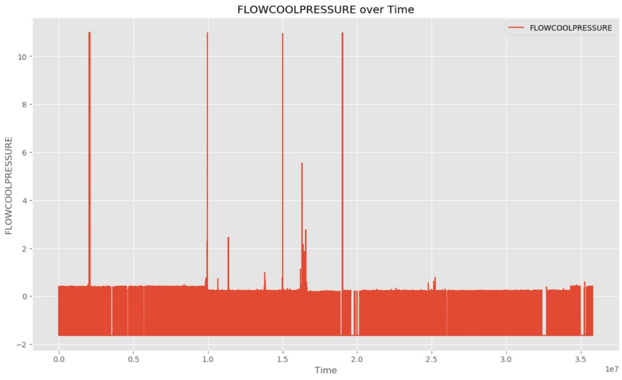
- **Data Preprocessing:** Cleaning, scaling, reshaping for LSTM, reducing RAM usage, label encoding.
- **EDA:** Understanding data distribution, correlation, feature importance and behavior over time.
- **Clustering:** Unsupervised clustering (MiniBatchKMeans) was applied in order to explore the data further, revealing underlying behaviors and patterns within the dataset.
- **Anomaly Detection Model:** LSTM-based model for time-series analysis.
- **Evaluation:** Classification metrics such as accuracy, precision and recall for the Train dataset and then using **MSE** on both Train and Test dataset to evaluate Test model performance.

Exploratory Data Analysis (EDA)

- **Purpose:** Understand Train dataset behavior, target variable (FLOWCOOLPRESSURE) distribution, correlation and patterns.
- **Insights:** Distribution, ‘fault’ count (13,693) , and behavior over time.

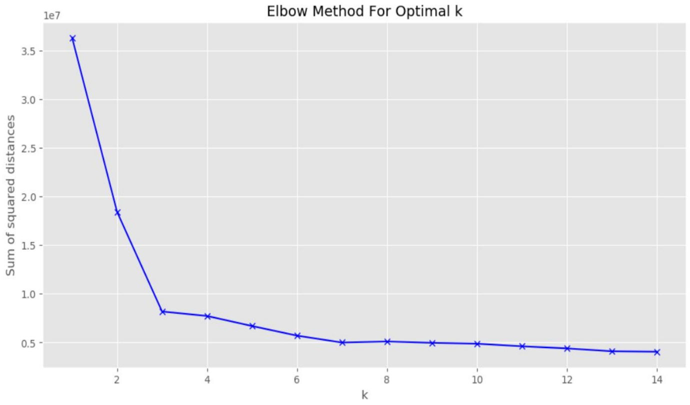


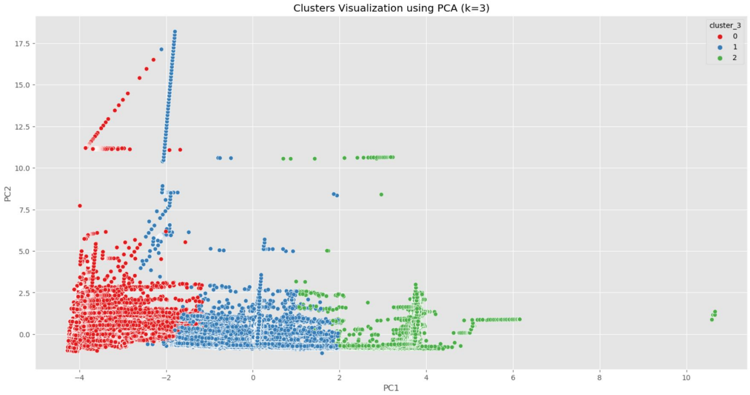




Clustering

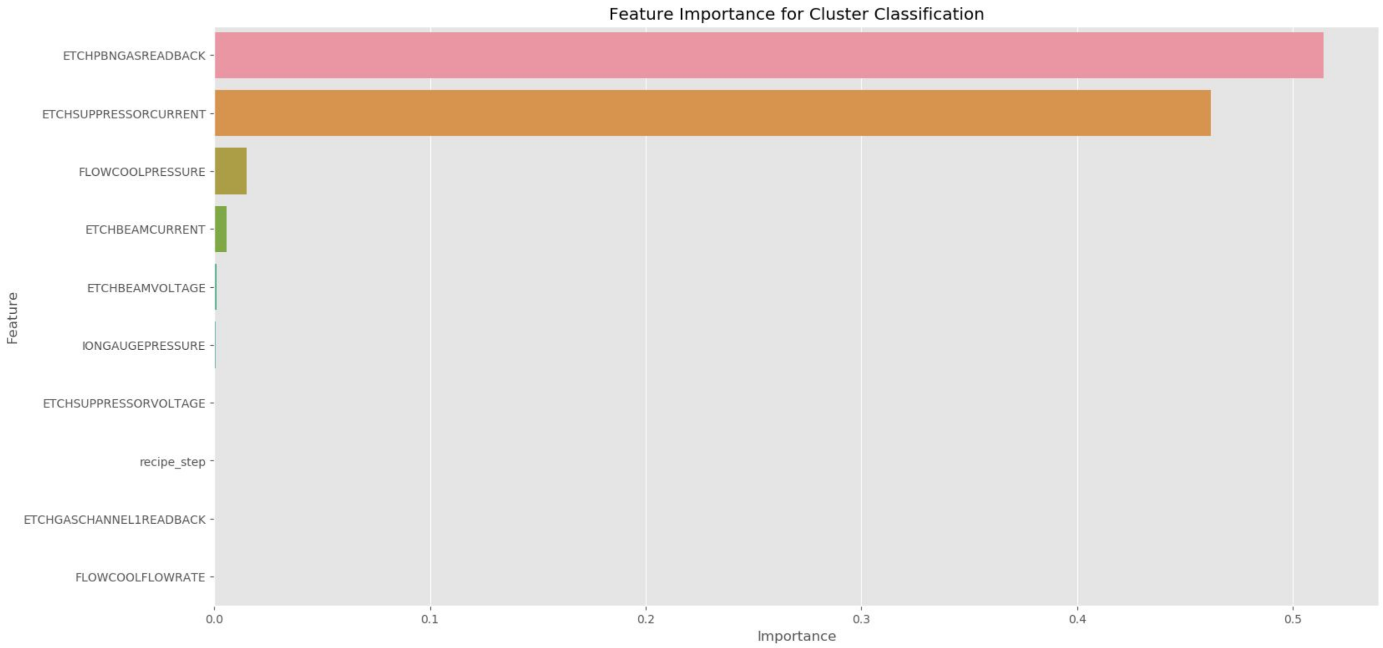
- **Clustering Method:** MiniBatchKMeans was applied to separate the data into clusters.
- **Elbow Method:** The Elbow method was used with PCA in order to determine the optimal number of clusters.
- **Feature Importance:** Deriving the features that had the most impact on the clustering process.
- **Comparative Analysis:** Comparing the top features in each of the clusters and how they behave.





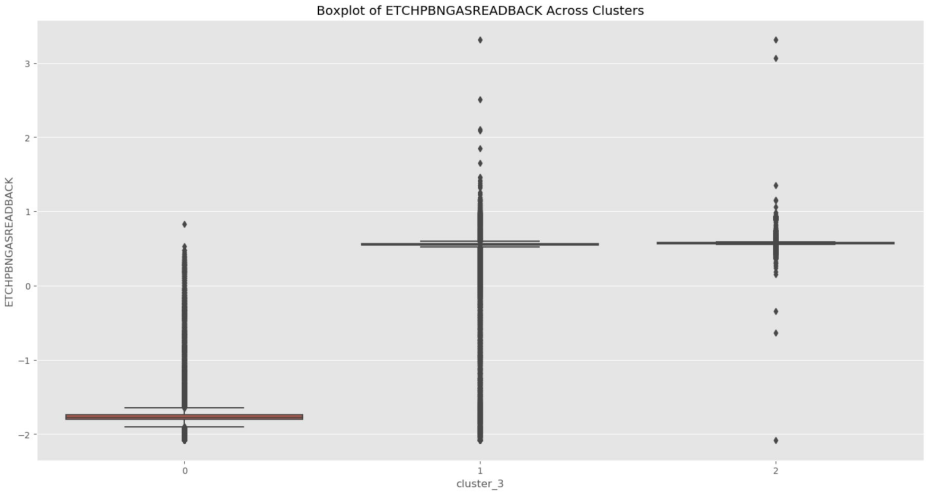
Clustering Plots:

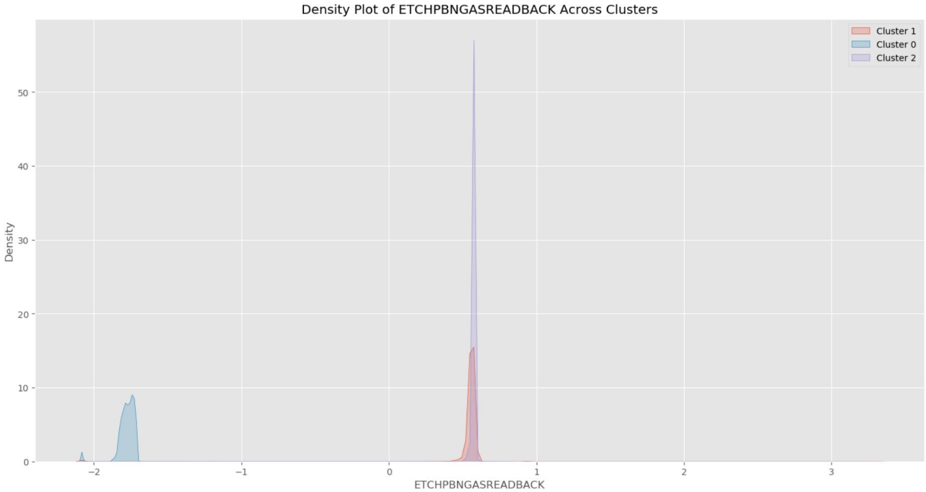
Feature Importance: These features were selected based on their correlation with the target variable.



Clustering Plots:

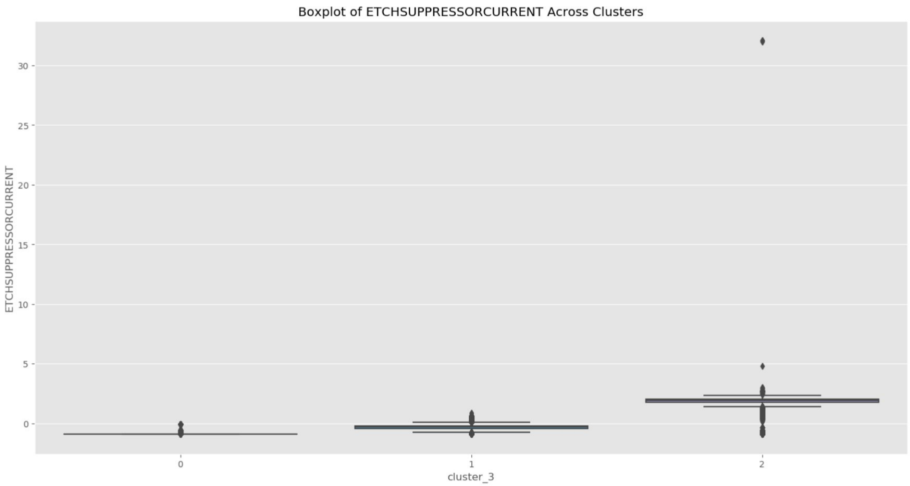
ETCHPBNGASREADBACK Comparison:

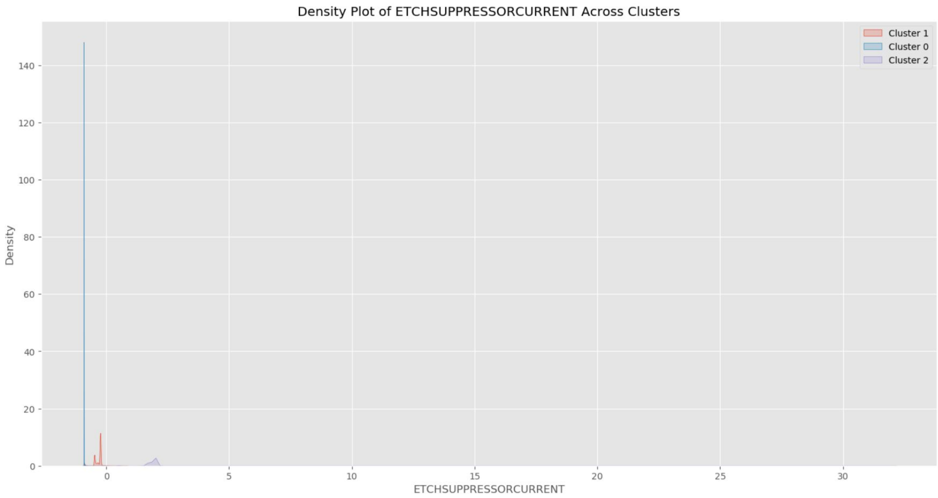




Clustering Plots:

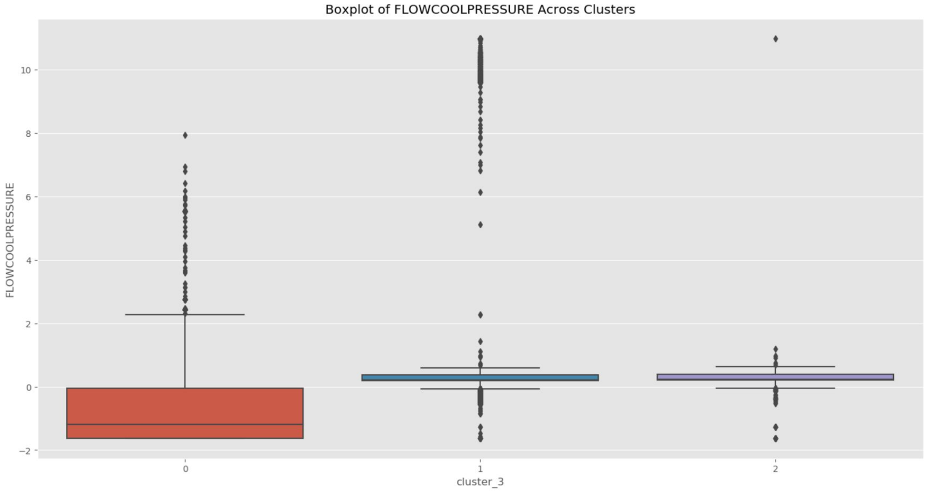
ETCHSUPPRESSORCURRENT Comparison:

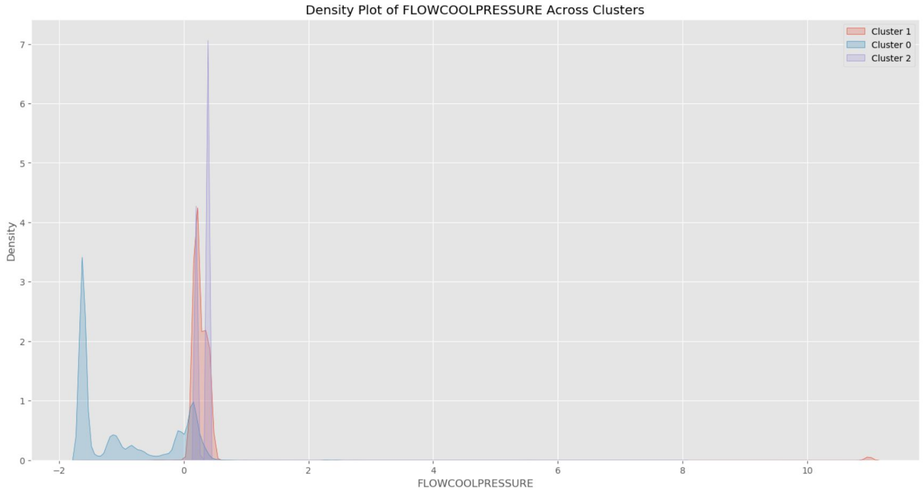




Clustering Plots:

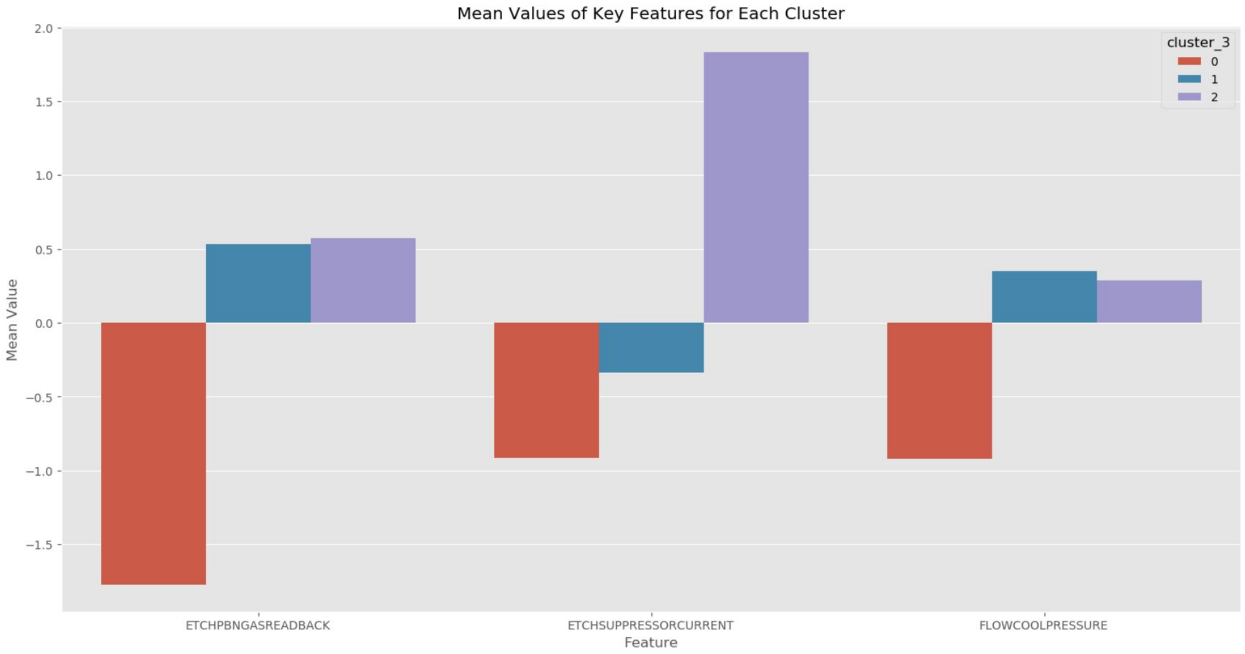
FLOWCOOLPRESSURE Comparison:





Clustering Plots:

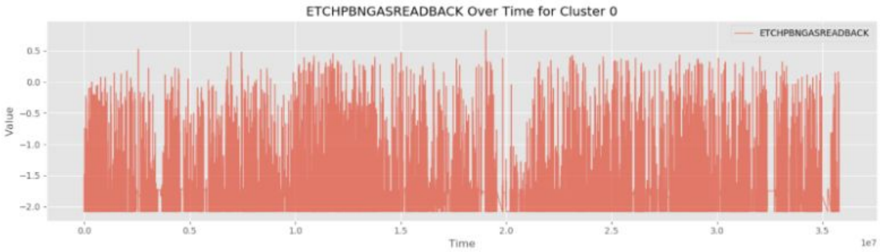
Mean Value Across Clusters:

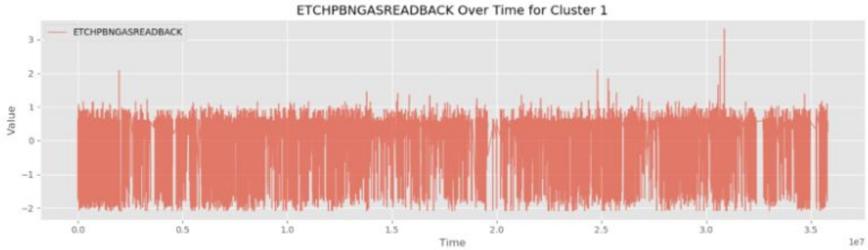


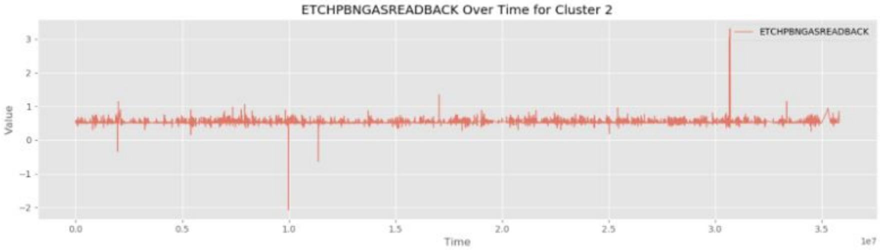
Clustering Plots:

Behavior across Time:

ETCHPBNGASREADBACK -



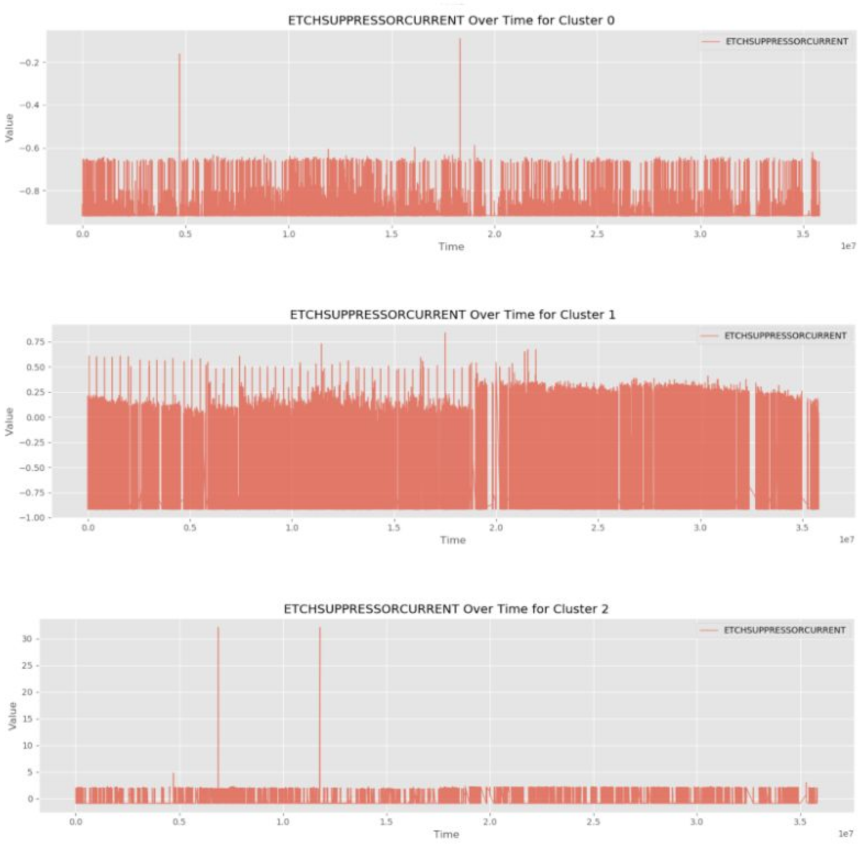




Clustering Plots:

Behavior across Time:

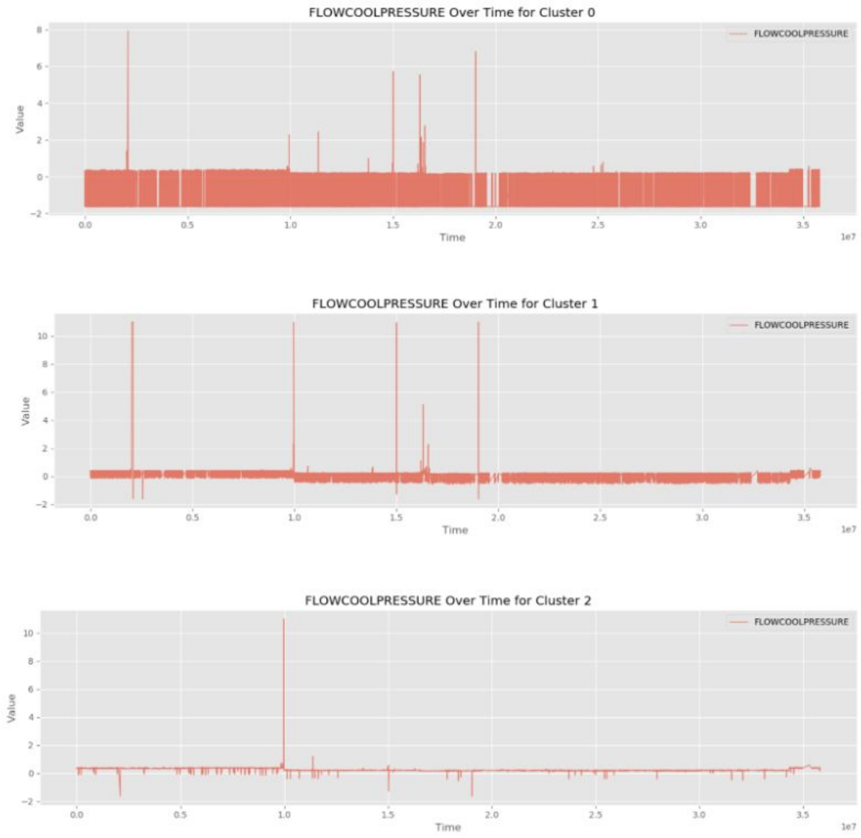
ETCHSUPPRESSORCURRENT -



Clustering Plots:

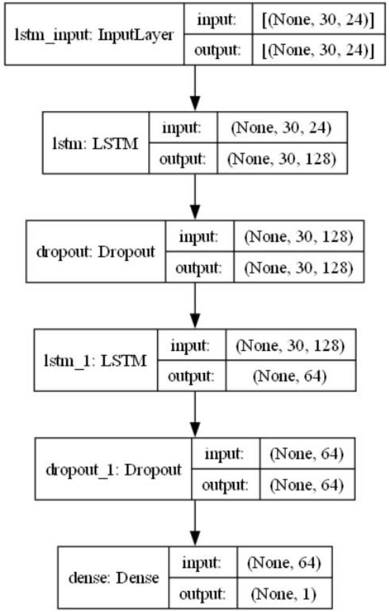
Behavior across Time:

FLOWCOOLPRESSURE -



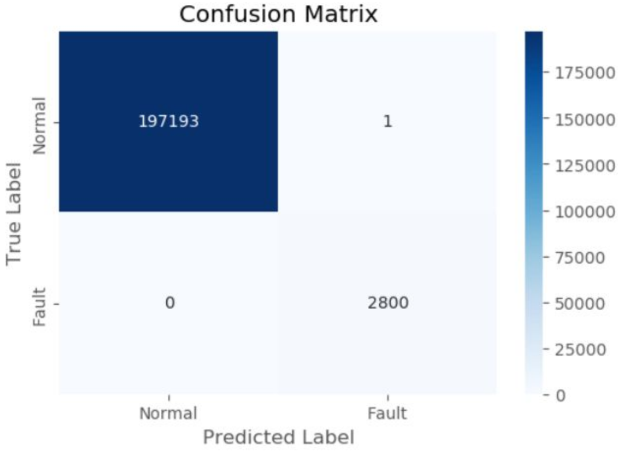
Anomaly Detection Model (LSTM)

- **Model Design:** Input Layer, 2 LSTM layers, Dense output layer, Loss function and an Optimizer.
- **Why LSTM? :** Long Short-Term Memory (LSTM) captures sequential time-series patterns effectively.



Experiment with Sample Dataset

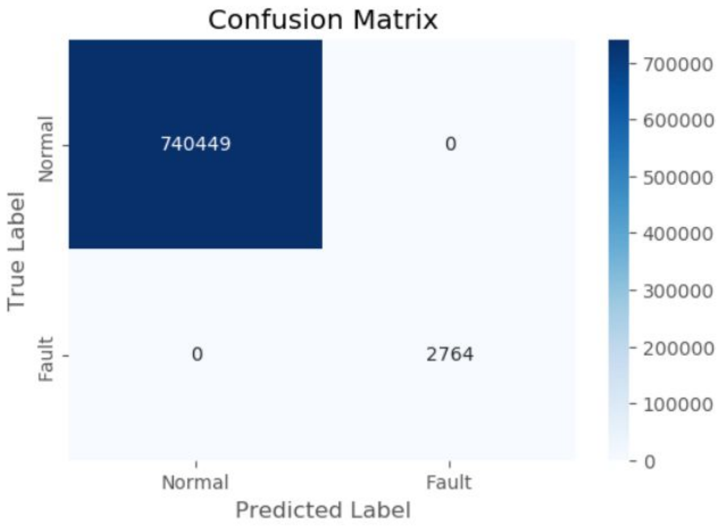
- **Sample Size:** 600K data points with 13,693 labeled anomalies.
- **Metrics Evaluated:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
- **Validation:** 20% of the data was used for validation set, Early Stopping was also implemented.
- **Results:**



Metric	Value
Accuracy	0.9999
Precision	0.9999
Recall	1.0
F1 Score	0.9998

Scaling to Full Dataset

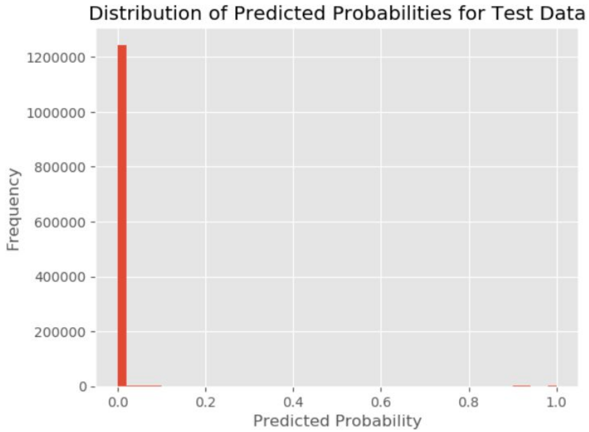
- **Dataset Size:** Expanded to 3.7M data points.
- **Challenges:** Memory management, class imbalance, long training time.
- **Results:**



Metric	Value
Accuracy	1.0
Precision	1.0
Recall	1.0
F1 Score	1.0

Applying the Model to Test Data

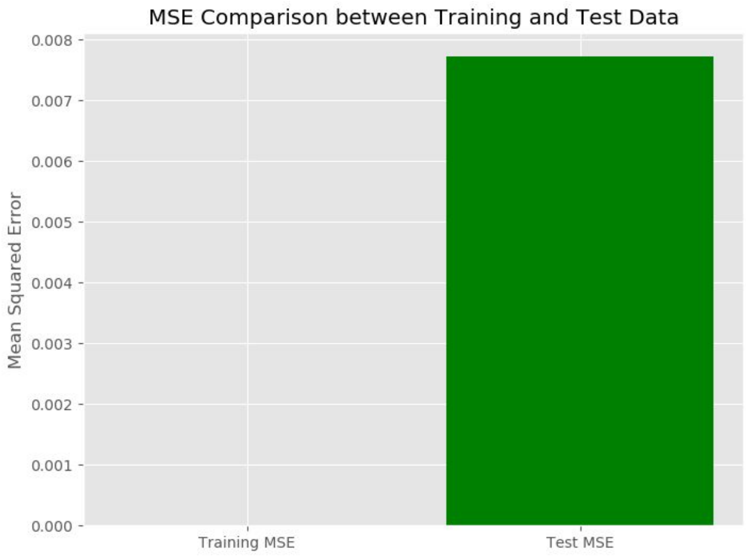
- **Test Data:** Unlabeled dataset with 1.27M samples.
- **Threshold-based Classification:** Probability threshold of 0.5 was implemented to determine anomalies.
- **Evaluation Method:** MSE-based difference between Train and Test data.
- **Findings:** Model performed well with low Test MSE of **7.72 x 10^-3** and 12,847 anomalies detected.



Evaluation and Key Results

- **Leveraging MSE for Model Evaluation:** The calculation of the MSE was applied to both the Training Data and the Testing Data while looking for the **minimal gap** between the two.
- **Findings:**

Dataset	MSE Value
Train Data	0.00000237 or **2.37e-06**
Test Data	0.007718233391642570 or **7.72 x 10^-3**



Limitations:

- **Unlabeled Test Data:** Since the test dataset lacks true labels for anomalies, the evaluation relies entirely on reconstruction error (MSE) and probability distributions.
- **Imbalanced Dataset:** The dataset is highly imbalanced, with anomalies being rare compared to normal data. The model could still have a bias toward normal data, missing subtle anomalies in the test data.
- **Limited Interpretability:** While the LSTM-based approach performs well at detecting anomalies, it lacks transparency regarding *why* certain instances are classified as anomalous.

Conclusion:

Summary of the Project:

- Successfully developed a robust anomaly detection system for ion beam etching using time-series sensor data.
- The LSTM-based model effectively detected faults, even when scaled from a 600k sample dataset to a full 3.7M-row dataset.

Key Achievements:

- **Model Scalability**: The model demonstrated consistent performance on datasets of varying sizes, highlighting its scalability and reliability.
- **Accurate Anomaly Detection**: The optimal MSE value of 2.37×10^{-6} (training) and 7.72×10^{-3} (test) confirmed strong generalization.

Thank You!

Github link of the project: https://github.com/Danielh2525/Anomaly-Detection-Ion-Beam-Etching

For further question you can reach me at:

Email: danielhofc@gmail.com

Daniel Hamama.