Anomaly Detection in Ion Beam Etching Processes

Daniel Hamama

ID: 318652252

Holon Institute of Technology

Student of BSc Applied Mathematics

Introduction:

- **Context:**

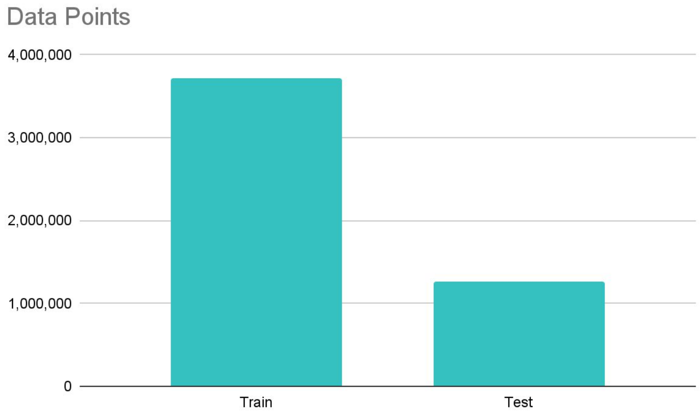
  Ion beam etching is a critical process in manufacturing high-precision components. Monitoring this process in real time helps avoid costly production failures.

- **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).
- **Target Variable: FLOWCOOLPRESSURE** from the domain knowledge.
- **Test Data:** 1.2 Million data points, 24 features.



Project Methodology

Data Preprocessing	EDA	Clustering	Anomaly Detection	Evaluation
• Data Loading • Data Cleaning • Scaling • Reducing RAM usage • Reshaping for LSTM • Label Encoding	• Distribution • Correlation • Behavior over Time	• Elbow Method • PCA • Cluster Understanding • Feature Importance • Comparative Analysis • Patterns	• Sample Dataset • Train-Test Split • LSTM Model • Model and Hyperparameter Tuning • Scaling to Full Dataset	• Train Data Performance • Applying on Test Data • Probability Thresholds • MSE Comparison

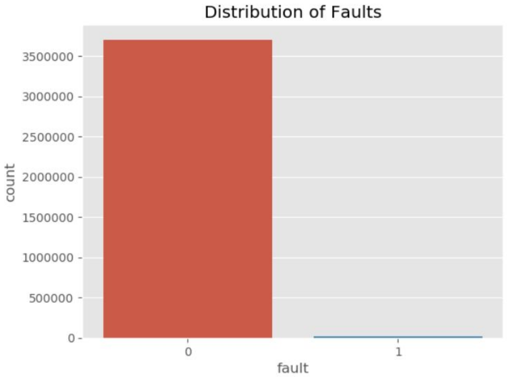
Data Preprocessing

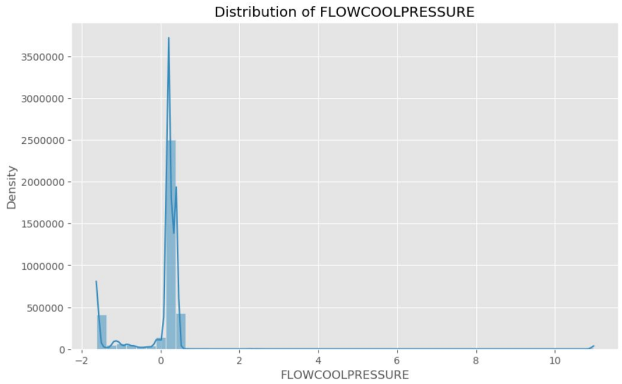
Objective: Cleaning the data as much as possible, cleaner data will yield better results, also preparing the data to the clustering and anomaly detection process.

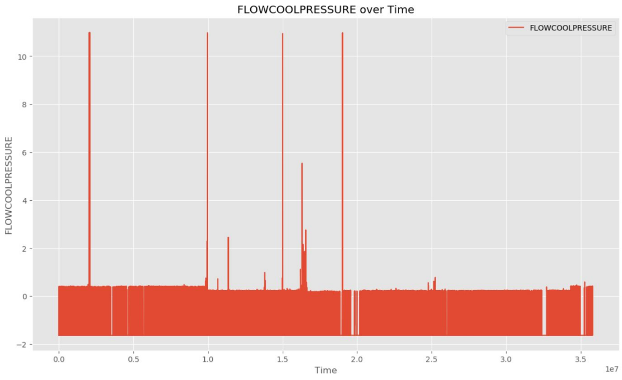
- Cleaning the data from irrelevant features.
- Checking for missing values and duplicate rows.
- Reducing RAM usage by converting numerical values to int16 bytes.
- Label Encoding categorical features.
- Scaling for Clustering and LSTM
- Reshaping data to 3D for LSTM
- Sequence Generation for LSTM (30 steps)

Exploratory Data Analysis (EDA) Pt.1

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

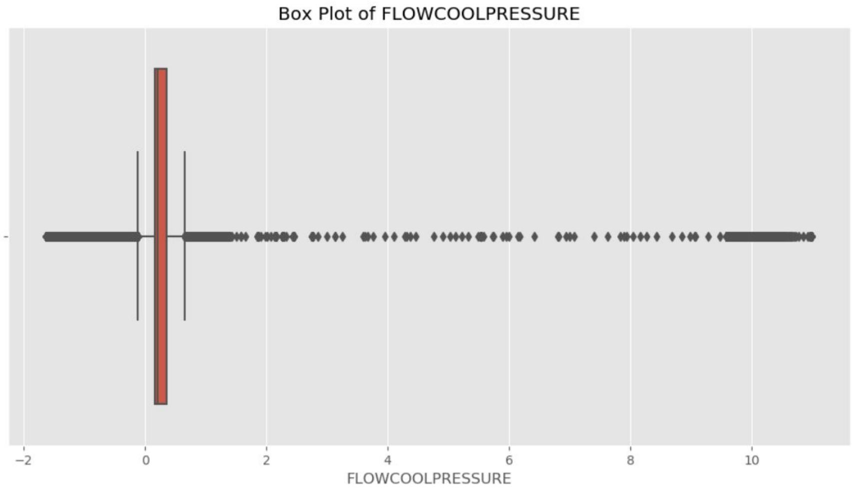


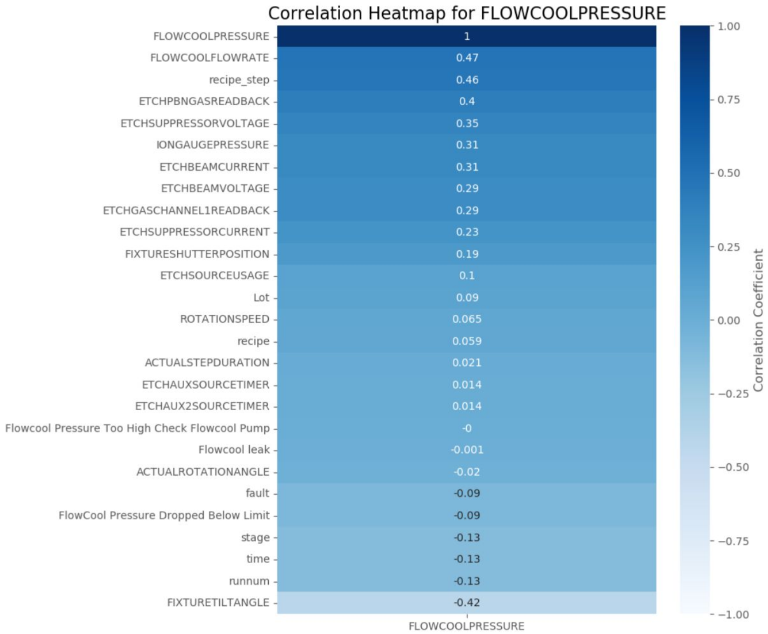




Exploratory Data Analysis (EDA) Pt.2

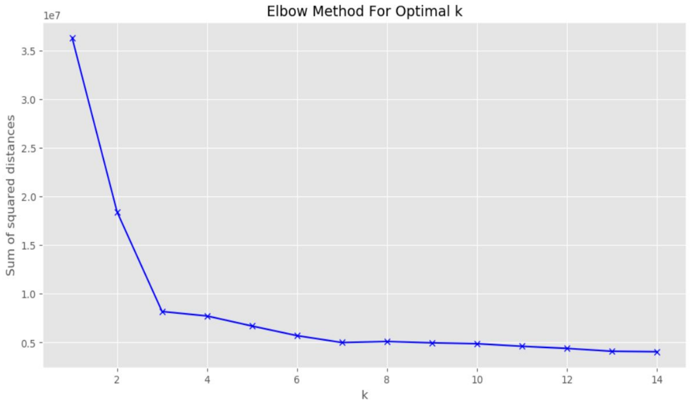
- **Outliers:** Using a Boxplot on the target variable to get a glimpse of it's behavior.
- **Correlation:** Using a heatmap to understand feature correlation with the target variable.

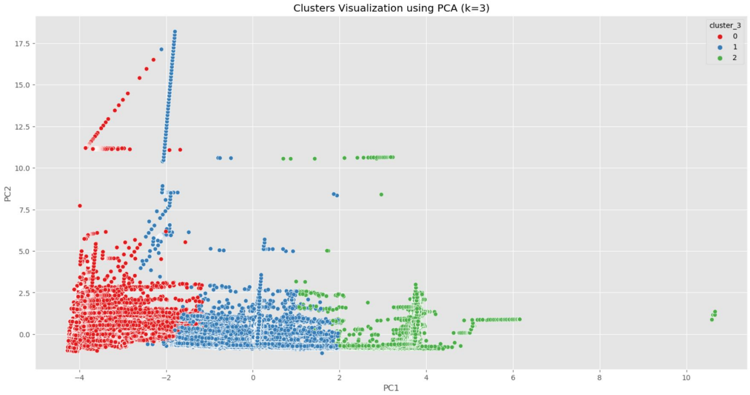




Clustering

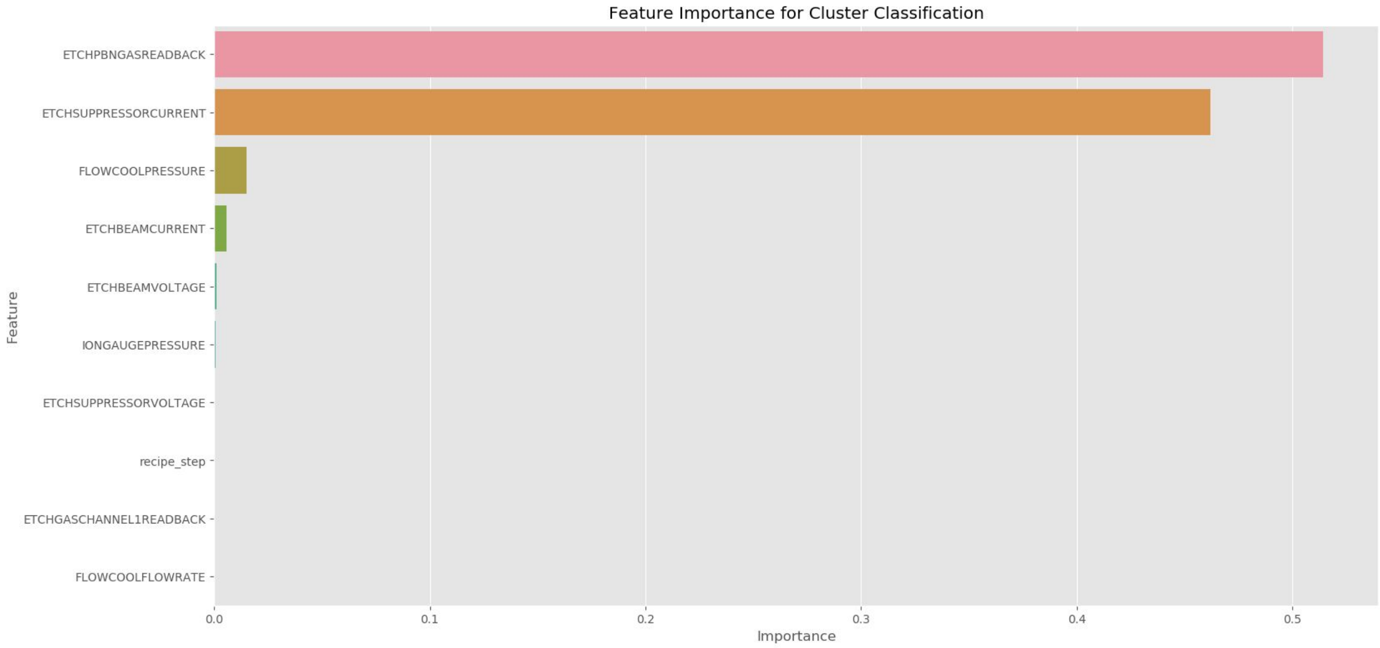
- **Clustering Method: MiniBatchKMeans** was applied to separate the data into clusters.
- **Elbow Method:** The Elbow method was used with **PCA** in order to determine the optimal number of clusters.
- **Feature Importance:** Deriving the features that had the most impact on the clustering process.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





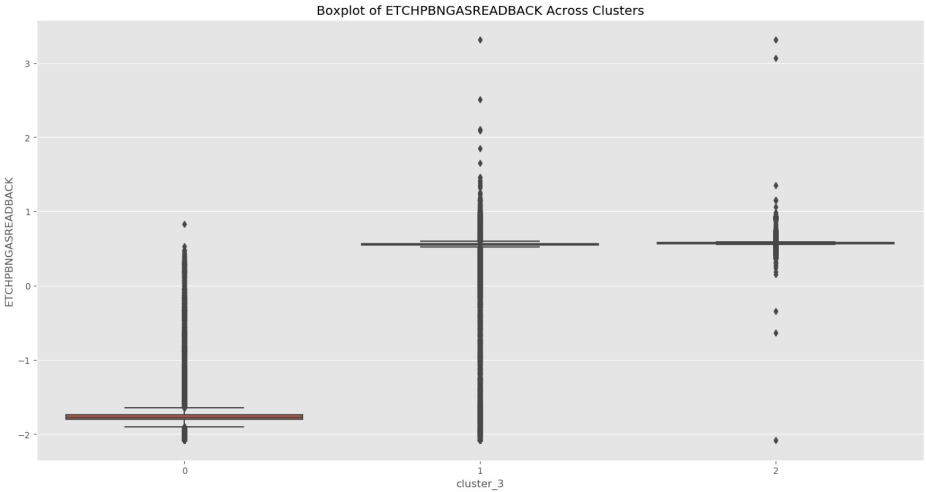
Clustering Plots 1:

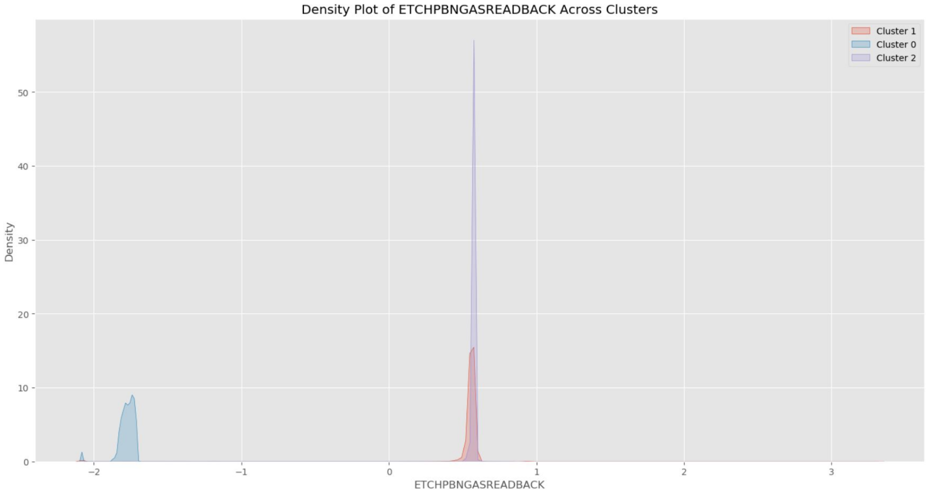
Feature Importance: These features were selected by a **Decision Trees Classifier** based on their correlation with the target variable.



Clustering Plots 2:

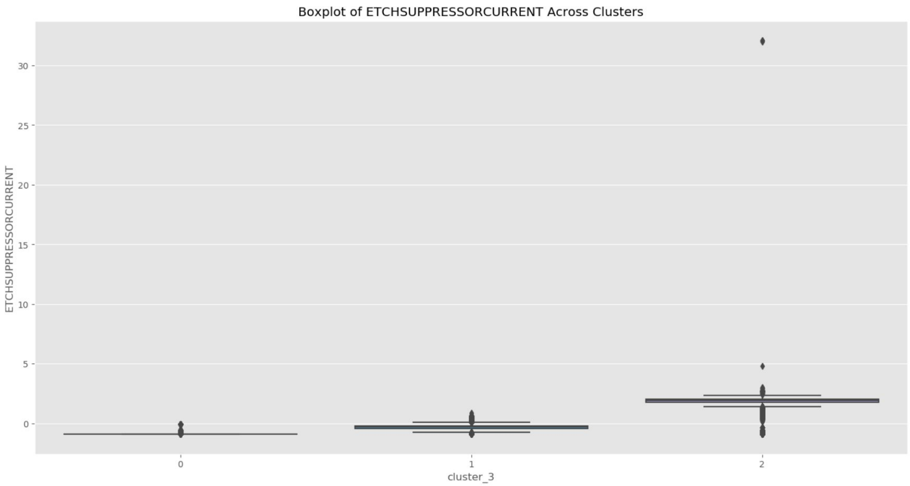
ETCHPBNGASREADBACK Comparison:

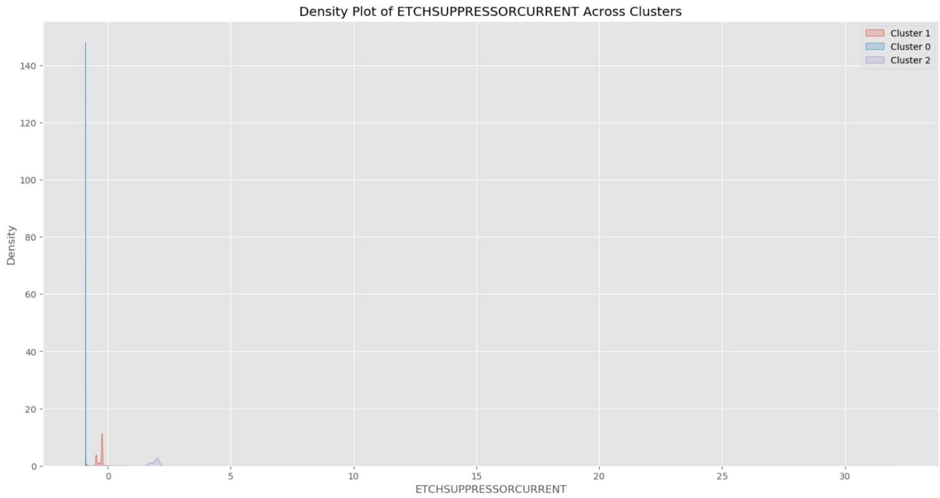




Clustering Plots 3:

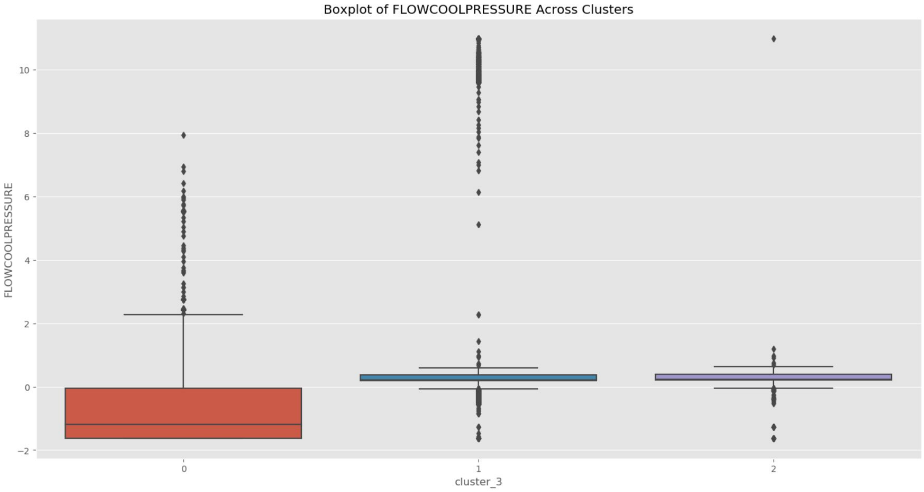
ETCHSUPPRESSORCURRENT Comparison:

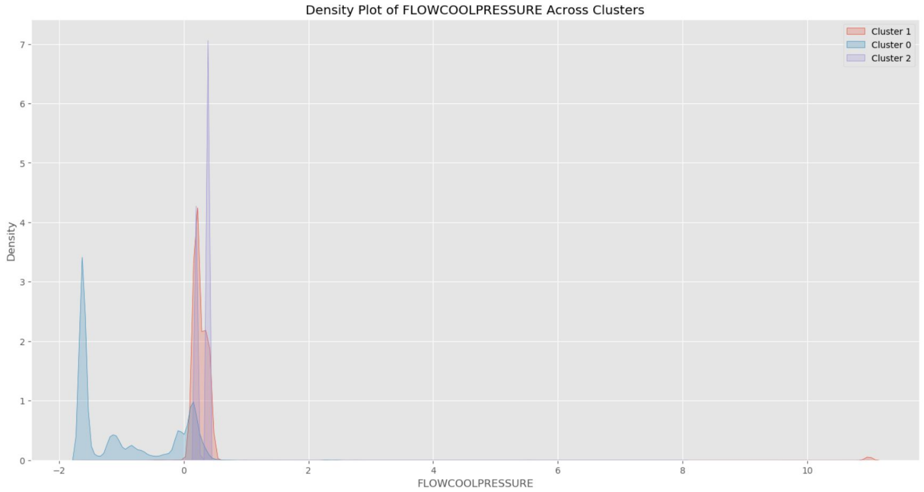




Clustering Plots 4:

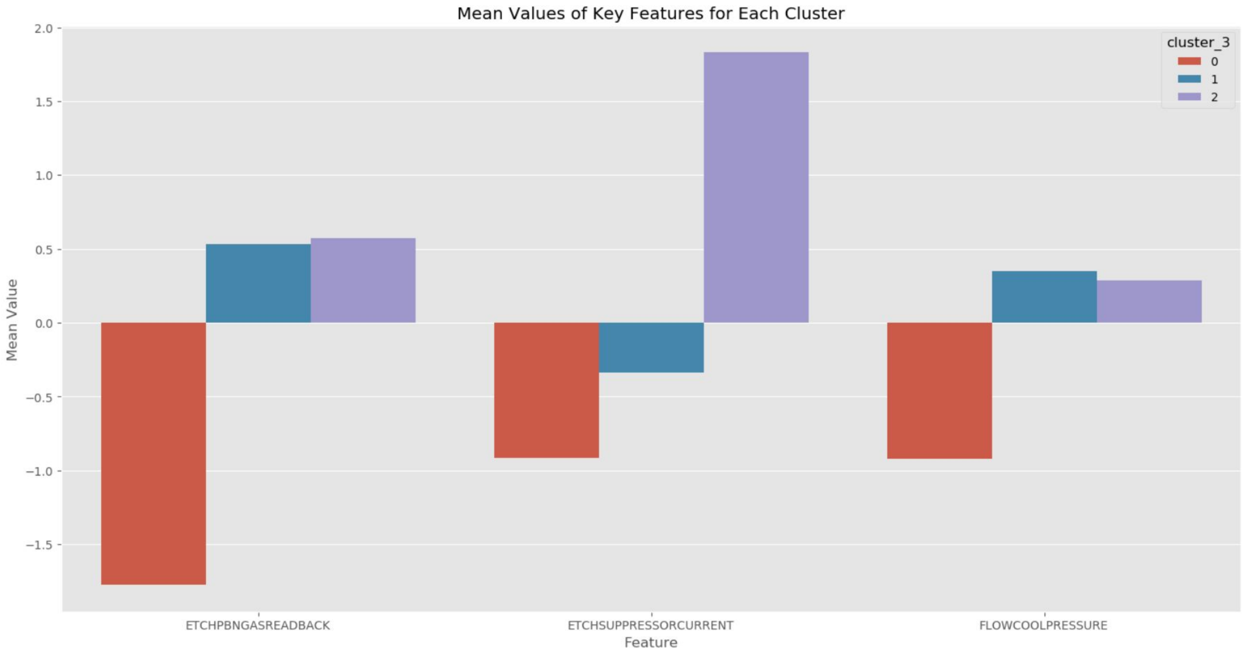
FLOWCOOLPRESSURE Comparison:





Clustering Plots 5:

Mean Value Across Clusters:



Analysis Summary:

Cluster 0:

- **ETCHPBNGASREADBACK**: Consistently low values, tightly clustered around -2.
- **ETCHSUPPRESSORCURRENT**: Low variability, grouped around -0.9.
- **FLOWCOOLPRESSURE**: Predominantly negative values, indicating lower pressures.

Cluster 1:

- **ETCHPBNGASREADBACK**: Tighter grouping around positive values with occasional outliers.
- **ETCHSUPPRESSORCURRENT**: Wider variability, ranging from -0.9 to 0.8.
- **FLOWCOOLPRESSURE**: Moderate range, values close to zero with some significant outliers.

Cluster 2:

- **ETCHPBNGASREADBACK**: High variability with extreme outliers, suggesting instability.
- **ETCHSUPPRESSORCURRENT**: Shows extreme outliers, with very high values.
- **FLOWCOOLPRESSURE**: Mostly positive values but with notable outliers.

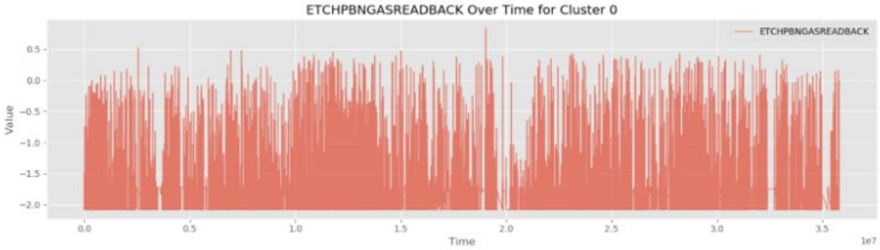
*Outliers:

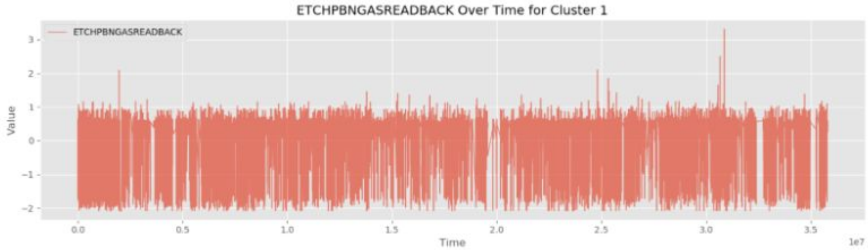
Clusters 1 and 2 exhibit more extreme outliers, particularly in **FLOWCOOLPRESSURE** and **ETCHSUPPRESSORCURRENT**, which could signify abnormal process conditions or potential anomalies.

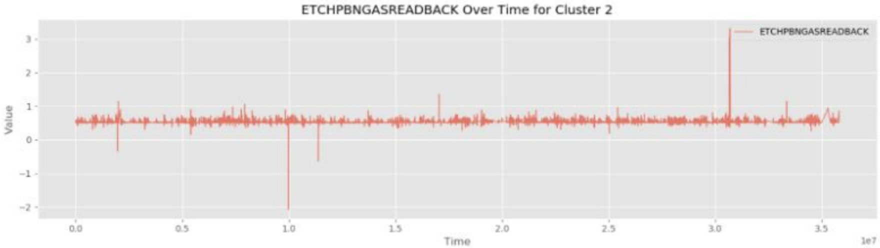
Clustering Plots 6:

Behavior across Time:

ETCHPBNGASREADBACK -



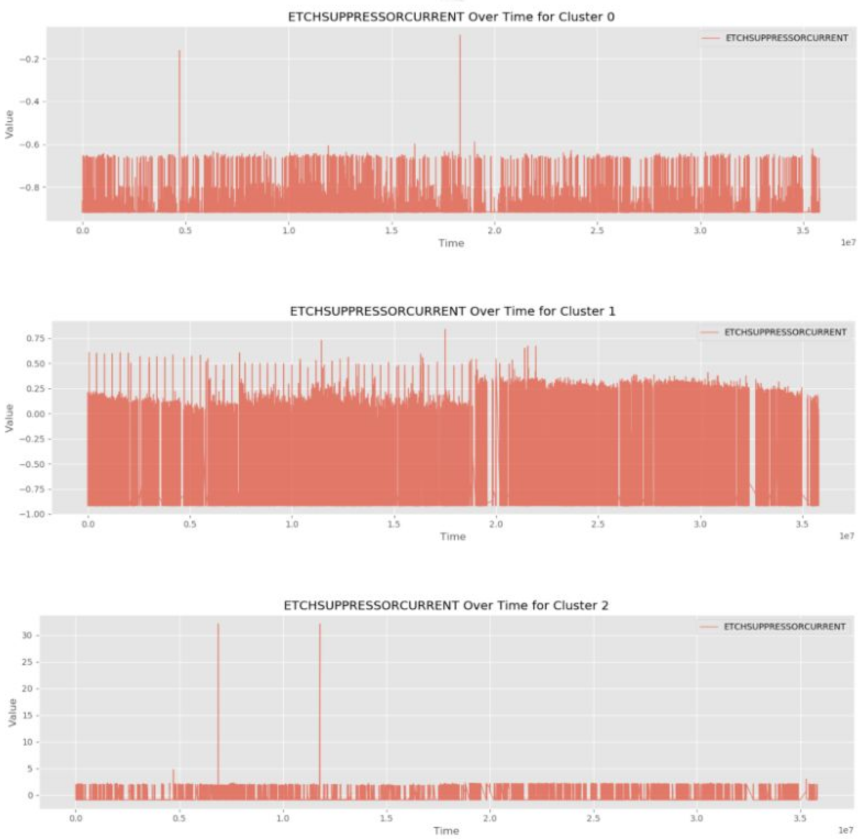




Clustering Plots 7:

Behavior across Time:

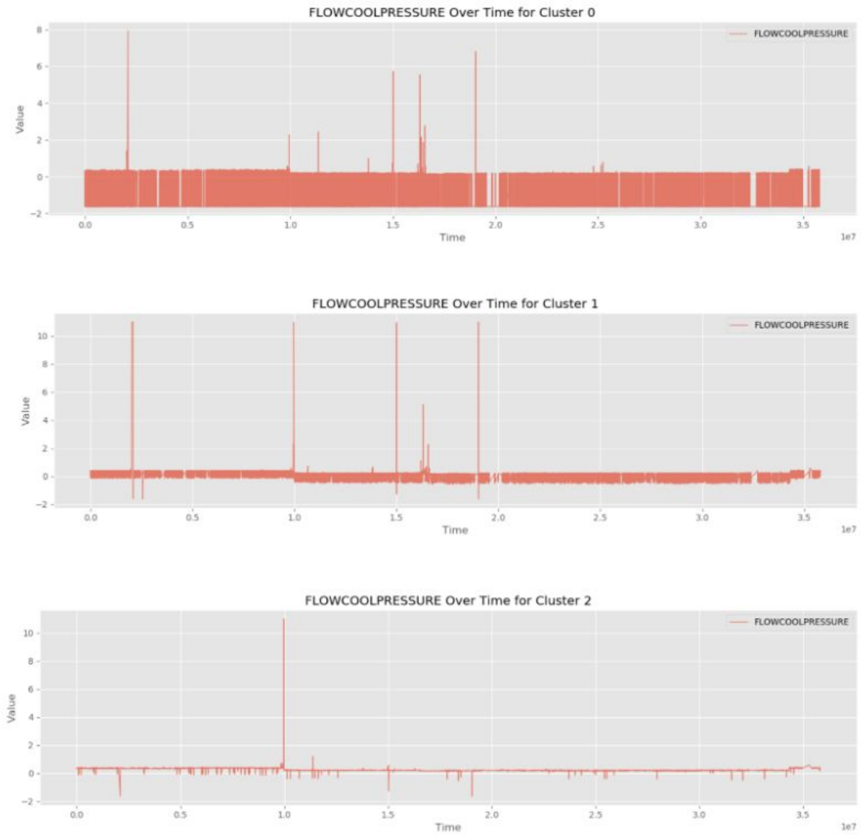
ETCHSUPPRESSORCURRENT -



Clustering Plots 8:

Behavior across Time:

FLOWCOOLPRESSURE -



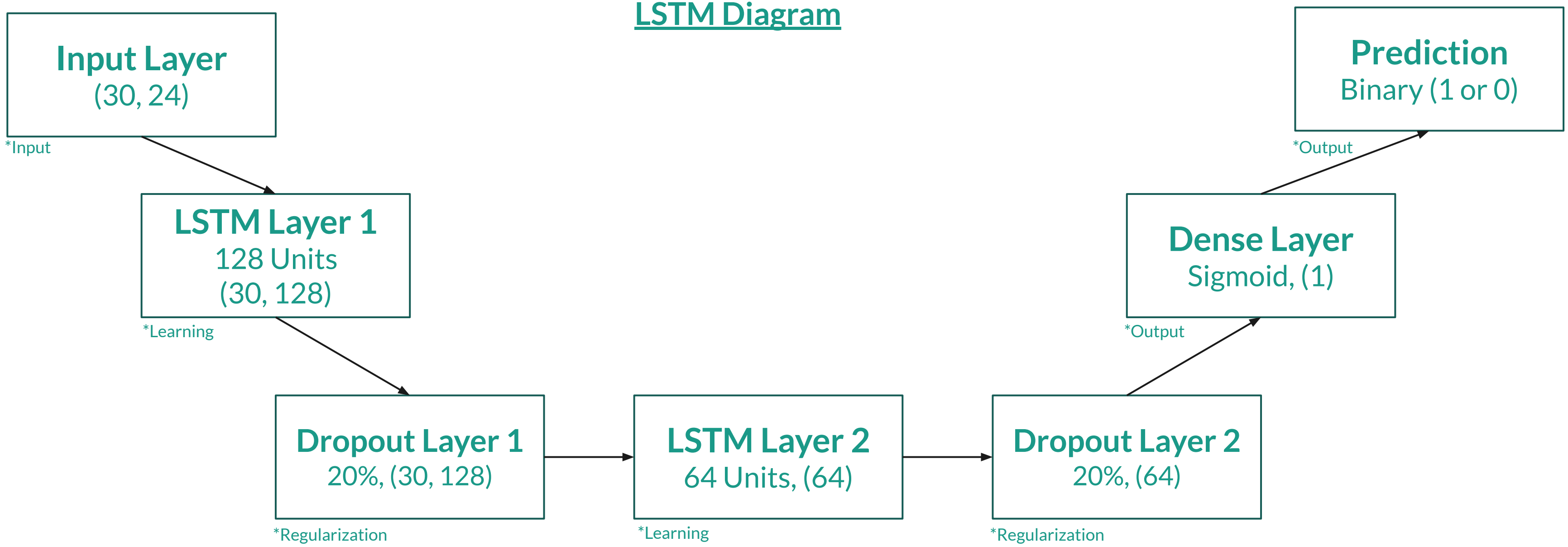
Time-Series Analysis Summary:

- **Cluster 0:** The Time series plots suggest consistent operation values over time, with fewer spikes or anomalies in the key features.
- **Cluster 1:** This cluster shows significant variability over time, with frequent spikes in **FLOWCOOLPRESSURE** and **ETCHSUPPRESSORCURRENT**, suggesting unstable behavior.
- **Cluster 2:** This cluster appears more stable overall but large spikes in **ETCHSUPPRESSORCURRENT** and **FLOWCOOLPRESSURE** could indicate critical operational issues during specific time windows.

Anomaly Detection Model (LSTM)

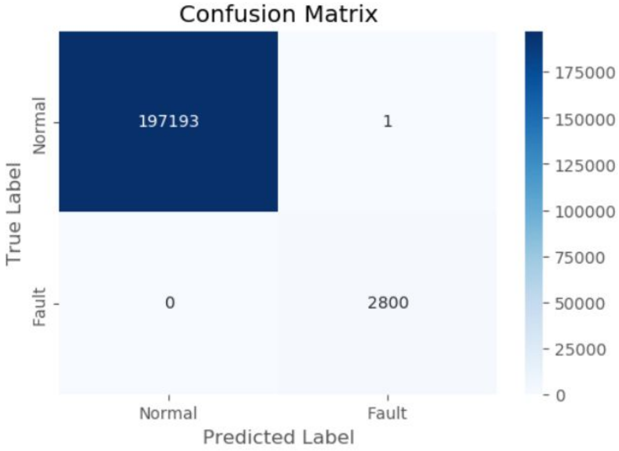
- **Model Design:** Input Layer, 2 LSTM layers, Dense output layer, Loss function and an Optimizer.
- **Why LSTM? :** Long Short-Term Memory (LSTM) captures sequential time-series patterns effectively.

LSTM Diagram



Experiment with Sample Dataset

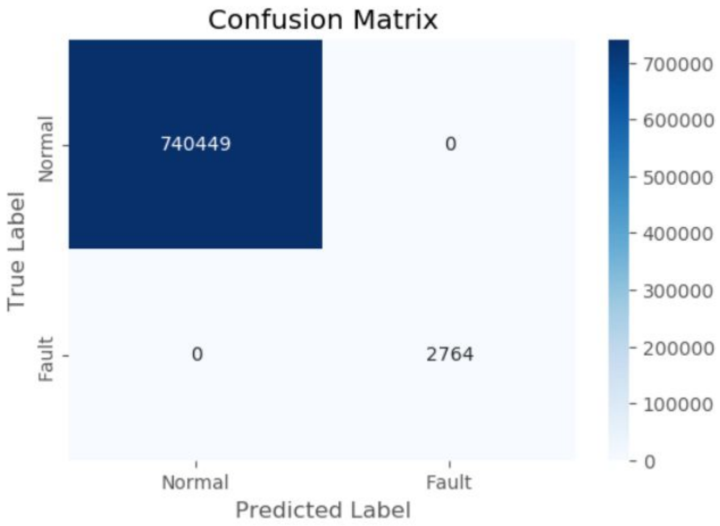
- **Sample Size:** 600K data points with 13,693 labeled anomalies.
- **Metrics Evaluated:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
- **Validation:** 20% of the data was used for validation set, Early Stopping was also implemented.
- **Results:**



Metric	Value
Accuracy	0.9999
Precision	0.9999
Recall	1.0
F1 Score	0.9998

Scaling to Full Dataset

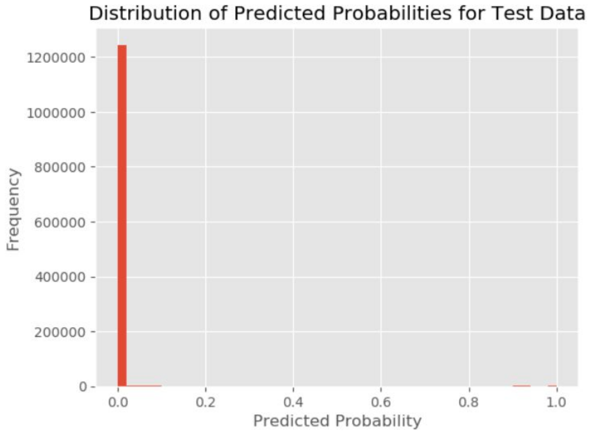
- **Dataset Size:** Expanded to 3.7M data points.
- **Challenges:** Memory management, class imbalance, long training time.
- **Results:**



Metric	Value
Accuracy	1.0
Precision	1.0
Recall	1.0
F1 Score	1.0

Applying the Model to Test Data

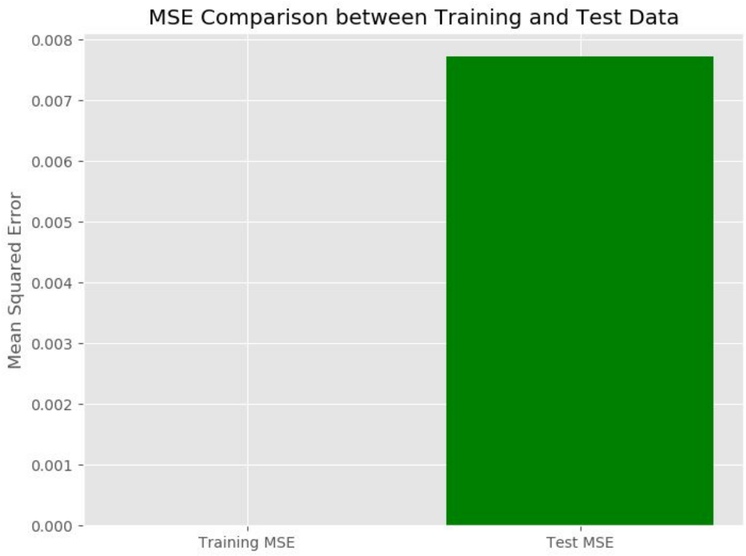
- **Test Data:** Unlabeled dataset with 1.27M samples.
- **Threshold-based Classification:** Probability threshold of 0.5 was implemented to determine anomalies.
- **Evaluation Method:** MSE-based difference between Train and Test data.
- **Findings:** Model performed well with low Test MSE of **7.72 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	0.0077182333916425705 or **7.72 x 10^-3**



Anomaly Investigation

We will compare anomalous data and normal data inside the following features:

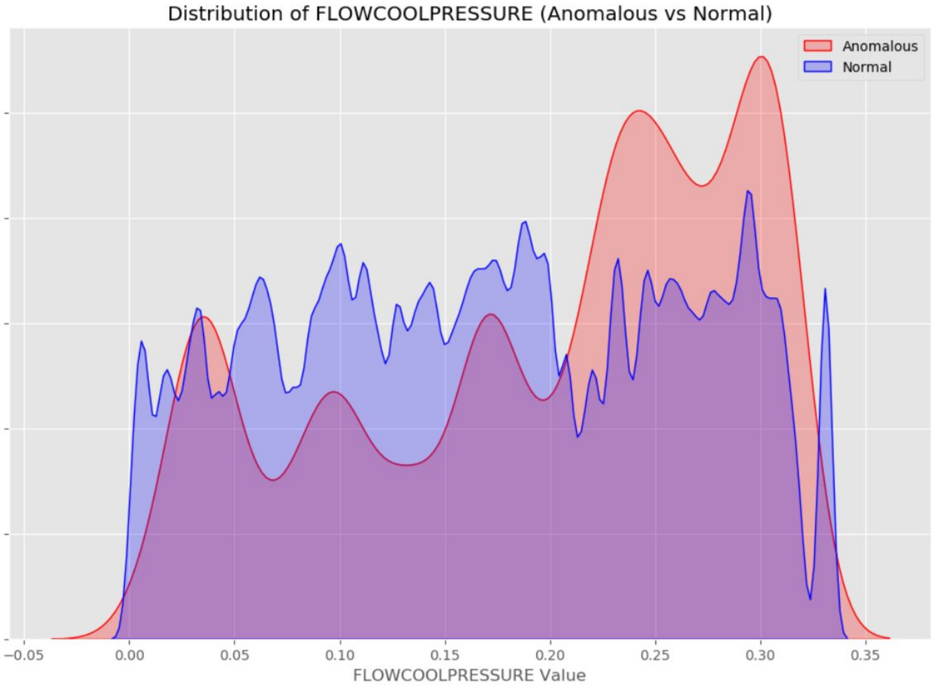
- **FLOWCOOLPRESSURE**
- **FLOWCOOLFLOWRATE**
- **ETCHSUPPRESSORCURRENT**
- **ETCHBEAMCURRENT**

Showcasing KDE Distribution plots for each of the features with a distribution of anomalous data and normal data overlapping each other and derive insights from that.

FLOWCOOLPRESSURE:

Key Observations:

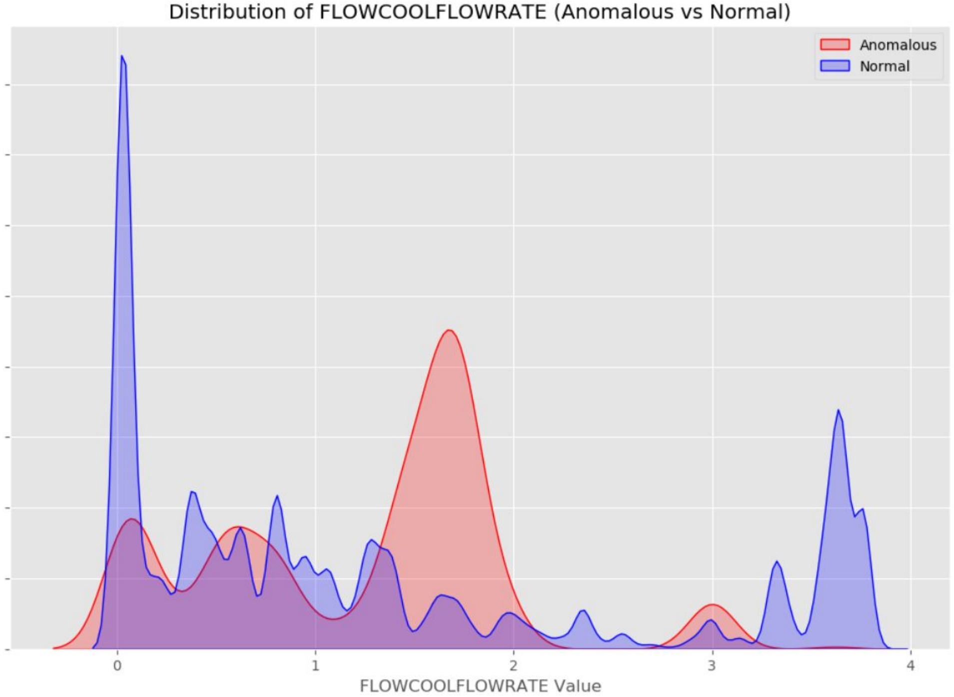
- Higher anomaly density peaks at elevated pressure levels, more notably between the values of **0.20** and **0.35**.
- Normal data has a narrower and smoother density curve.



FLOWCOOLFLOWRATE:

Key Observations:

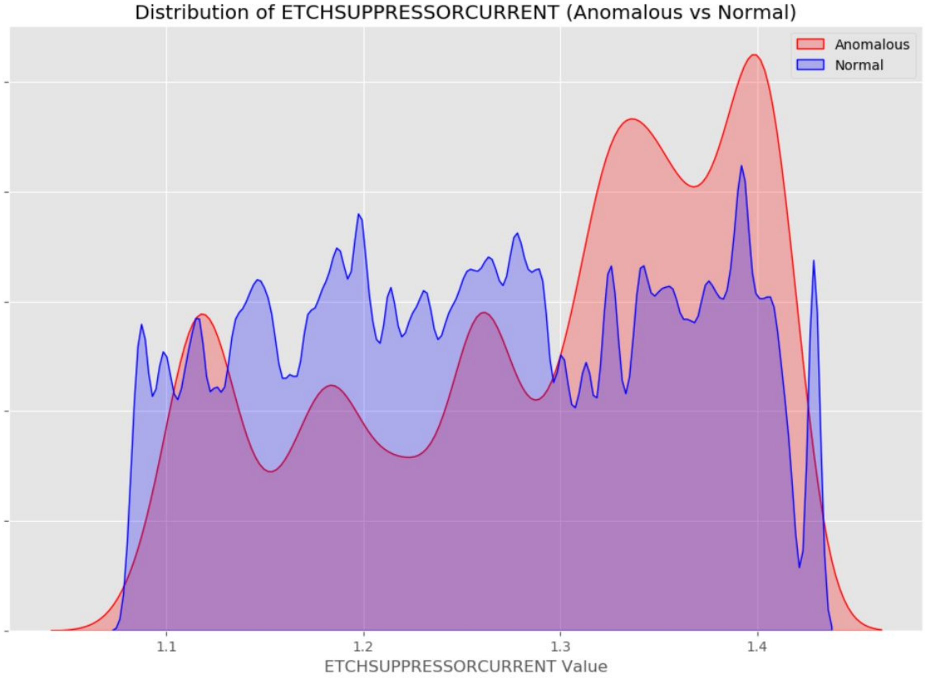
- Anomalous data shows distinct peaks at specific higher flow rate values, specifically between the values of **1 and 2**, diverging significantly from normal data.
- The normal distribution highlights a steady, consistent flow rate across most observations with peaks around the 0 value and 3 to 4 values.



ETCHSUPPRESSORCURRENT:

Key Observations:

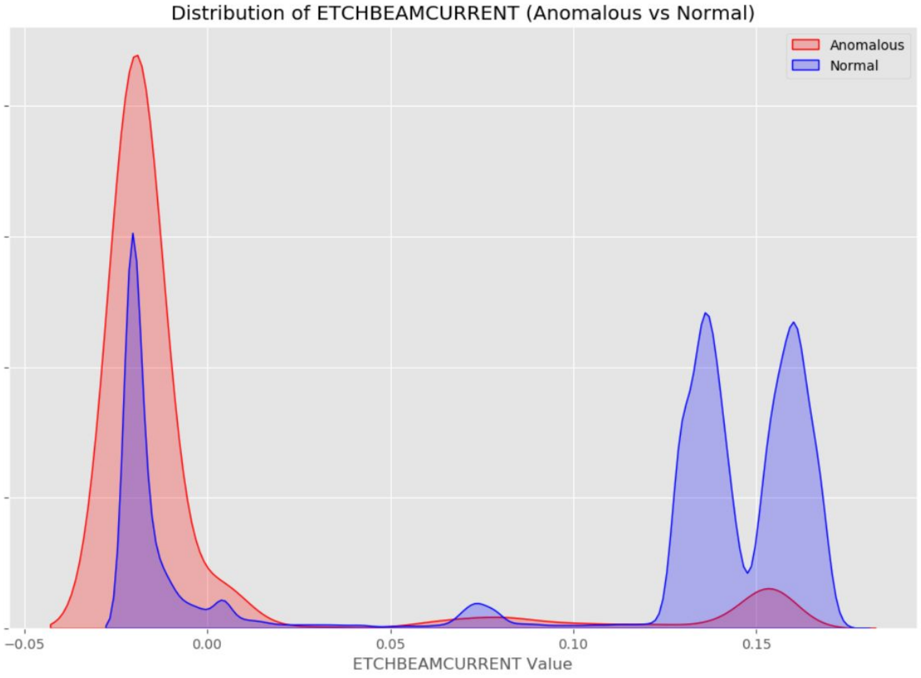
- Anomalous data shows a noticeable shift towards higher current values, very similar behavior to FLOWCOOLPRESSURE. We can also observe a peak between the values of **1.3 to 1.4.**
- The normal distribution is more tightly centered, indicating stable current levels during normal operations.



ETCHBEAMCURRENT:

Key Observations:

- Anomalous data shows a significant peak at the lower values, more specifically between **-0.05 and 0.**
- Normal data shows a variety of peaks, while also being present inside anomalous data there might be overlapping conditions between the two.



Anomaly Investigation Conclusion

- **FLOWCOOLPRESSURE** and **FLOWCOOLFLOWRATE** indicates notable deviations in their distributions for anomalous data compared to normal data, suggesting significant differences in these features under anomalous conditions.

- **ETCHSUPPRESSORCURRENT** shows elevated values during anomalies, which likely contribute to distinguishing anomalous behavior.
- Similarly, **ETCHBEAMCURRENT** exhibits a clear shift in its distribution, while overlapping with normal data it can still be a strong indicator of anomalies.

- These features collectively demonstrate their importance in identifying and characterizing anomalous behavior in the dataset.

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. To counter that we performed distribution analysis of both anomalous data and normal data.

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.