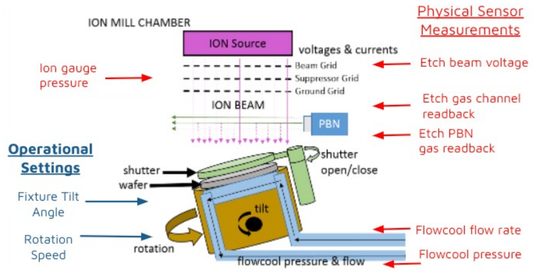
Fault Diagnosis for High-Precision Manufacturing Processes

Kayla Berman, Julia Brace, Olivia Johnian, Jordan Lian

Problem

Our project focuses on fault diagnosis for Ion Mill Etching (IME) manufacturing since that is the data provided to us by the 2018 PHM Data Challenge. This data includes both time series sensor data and fault data from the IME machine. Some of the sensor measurements and operational settings can be seen around the diagram of the IME machine below. Our goal was **to identify what type of failure is occurring each time there is a failure present within the data, i.e. Fault Diagnosis.**

This is an important issue in the manufacturing industry at large due to the cost impacts of faults and the ever-growing amount of data being supplied by various machines themselves. **The data from the IME machines serves as a type of case study for the larger issue of fault diagnosis.**

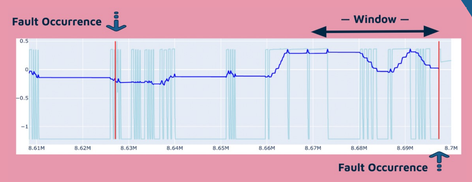


An ion beam is used to remove material from a workpiece that is placed on a rotating fixture which tilts at different angles. A particle beam neutralizer (or PBN) system controls the ion beam shape and distribution. The workpiece is cooled down by a flowcool system, which involves a flowcool liquid that runs behind the workpiece at a specified rate.

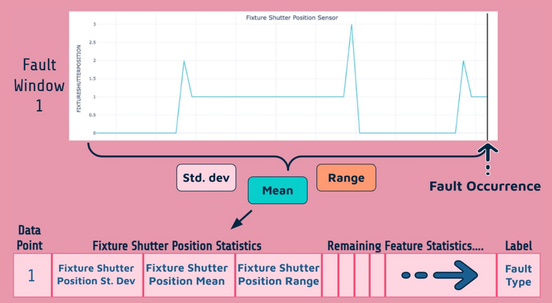
This project focuses on the flowcool system due to its importance for preserving the newly etched pattern and preventing thermal deformation. This system has 3 main failure mechanisms - leakage, high pressure, and low pressure. We will be analyzing all the sensor data for signs of degradation before faults.

Data Transformation

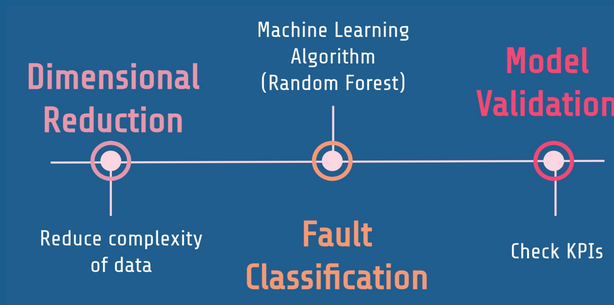
Time series data does not easily fit into typical classification algorithms, so the team had to transform the continuous sensor data into discrete data. To do this, the team needed to horizontally slice the data into the smaller instances of time called Fault Cycles of Fault Windows. An example of a Fault Window is shown to the right.



Once sliced into fault cycles, this data can become a **single sample point of data.** To do this, we extracted and stored summary statistics (mean, standard deviation, range, etc.) of selected sensors for each fault cycle. These statistics become the new features / columns of the model and the corresponding condition (normal or fault type) becomes the label. This process and an example row is shown below.



Solution



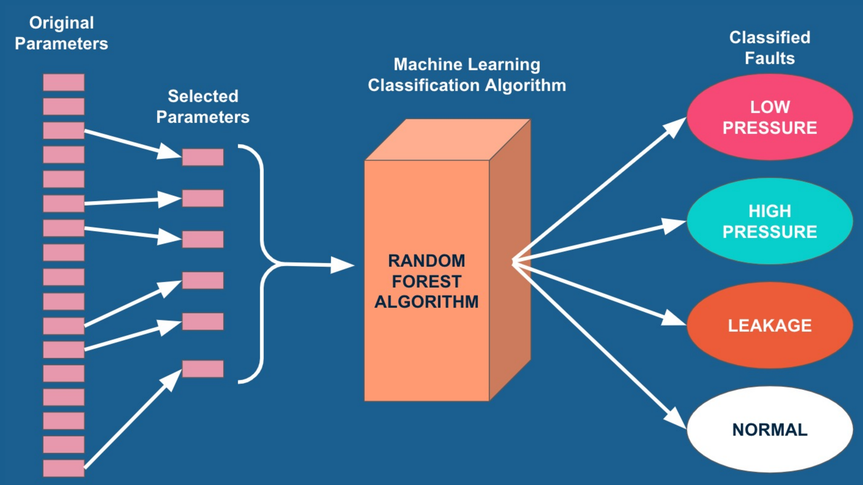
The above image gives an overview of our solution approach, each step is described below.

First, dimension reduction helped to remove predictors that would not be useful in the model, either because they do not affect the response variable or because there are other variables that are better indicators of system degradation. The team utilized their physical understanding of the machine and correlation analysis to determine which sensors were important. Through this process, the dimensions were reduced from 17 to 6 final sensors.

The selected dimensions are...

ETCHBEAMVOLTAGE FLOWCOOLFLOWRATE FLOWCOOLPRESSURE ETCHPBNGASREADBACK FIXTURETILTANGLE FIXTURESHUTTERPOSITION

Once dimensions were reduced and data was transformed, the new data is entered into a machine learning classification algorithm in its adjusted form and the algorithm classifies each "window" of data as either a low pressure, high pressure, leakage, or normal window as seen in the diagram below.



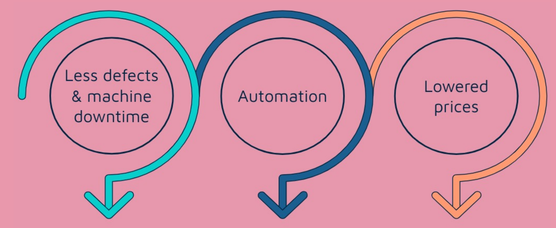
When it came to selecting a fault classification method, the team considered a few different algorithms. Each machine learning approach has its advantages and disadvantages depending on the type of data and what the expected output should look like. The team decided to use a Random Forest (RF) algorithm as the method for classification since it can also help select parameters / features that are most significant. The team used an open-source Python package (scikit-Learn) to develop the RF algorithm. The transformed data (i.e. the summary statistics of each window) was split into training and testing (80%, 20% split respectively) to give the model enough data to learn from without leaving too small of a sample to test. Iteration on the model came in the form of adjusting various parameters including the "window" size, presence of data smoothing techniques, number of data points, and the structure of the RF model itself until reaching an optimal accuracy within our constraints.

In terms of model validation, we set an initial accuracy KPI of 80% which we exceeded as shown in the "Results" section. Additionally, we met two key success factors:
1. Transparency for which sensor is showing degradation by analyzing the output of the model to see which features were most indicative of a failure occurring.
2. Understanding the capabilities and limitations of a Random Forest in fault diagnosis as shown by the Confusion Matrix in the "Results" section

Impact

A benefit to using intelligent fault diagnosis is that you can reduce defects and downtimes. This is an important goal in the manufacturing industry because these can be very costly. If an algorithm can pick up on when one of the sensors is starting to show symptoms of degradation, a correction can be made before a costly defect occurs.

This solution also increases automation of fault diagnosis rather than relying on traditional, more manual methods which requires experienced engineers. This has the potential to reduce labor needs and therefore costs. However, there are some ethical considerations with this since this could lead to layoffs if there is no program to keep people on and uptrain them to work with these algorithms.



Both previous cost saving benefits can lead to lowered prices for manufactured goods, which then increases the accessibility of the product to a wider consumer base.

This model can lead to increased profits for manufacturers and returned benefits to consumers.

Results

Below, on the right, is a graph of the model's accuracy across different test cases. The leftmost column represents the final multivariate model accuracy. The middle columns detail the accuracy of various binary classification test cases. The rightmost column displays the accuracy of the final model run on a different tool. This helps validate the model since this number is similar to the accuracy of the initial final model run.

On the left is a breakdown of the predicted condition versus the actual condition for a test run of the algorithm. This confusion matrix shows that the model has extremely high accuracy when classifying Normal and Low Pressure Fault points. However, the ability of the model to classify High Pressure Faults and Flowcool Leak Faults is very low.

These results indicate that a random forest algorithm can be used effectively in multi-class classification of multivariate time-series sensor data. Additionally, the lower accuracies found for certain fault types can educate the team and future researchers working with similar data on the drawbacks and limitations of the model.

Special thanks to all our supporters along the way, particularly our faculty advisor Prof. Xiaoning Jin and our design reviewer Dr. Huanyi Shui of Ford Motor Companies. Additionally, we would like to acknowledge Prof. Jaeger-Helton and Prof. McManus for all of their guidance and encouragement.