<https://www.mdpi.com/2075-1702/10/12/1233>

Two different datasets were used for this study. First of these datasets was UC Berkeley Milling Dataset, which contains 16 cases of example cuts in metal. Six parameters were utilised during its creation: the metal type, the feed rate and the depth cut, all three of them flitting between two values. Each case contains a combination of these values, the cases representing the tool when it is healthy, degraded or faulty in a gradual way.

During each cut, six types of signals were acquired from the sensors: acoustic emission and vibration from the spindle and the table alongside the AC and DC values of the spindle motor.

The other dataset used in this study was CNC Industrial Dataset, containing the data acquired during the process of metal machining of small ball-valves. The dataset has the manufacturing data of 5600 parts across a large array of the metal type and cutting parameters. Each of 35 cases in this dataset represents a unique roughing tool insert.

The current of the spindle motor was the principal signal collected from the CNC, and the data from its control system was acquired using a software from the equipment manufacturer. The current, the tool being used and when the tool started to cut the metal were recorded by this software at the frequency of 1000 Hz.

First, both of these datasets were pre-processed through labelling, classifying them into different health states. For the milling data, the health state (flank wear) at the end of the cut was used to label each case as healthy, degraded or failed. For the cuts that did not have any health state, an interpolation between the nearest cutes were utilised.

Afterwards, the stable cutting interval was chosen for each cut, based on the starting time of the cut. Sub-cuts from the data points were created alongside, and the problem was changed into a binary classification problem by merging the healthy and degraded labels into one healthy label.

For the CNC data, only the roughing tool was considered for pre-processed as it experienced the most frequent changes. Each sub-cut was extracted and given labels of either healthy or failed.

Next step was feature extraction, performed with the help of an open-source library named tsfresh. Unique feature methods such as FFT coefficients were generated through the help of this library. Afterwards, the data was split into training and testing sets and the features were scaled according to the maximum and minimum values of the training set.

As the data contained a large number of features, a feature selection method was necessary to choose viable features during the machine learning process. Two types of feature selection methods were used: randomly selecting several samples and using the selection method from the aforementioned open-source library.

As the datasets are highly imbalanced between the classes, oversampling and undersampling were preformed successively to acquire comparably more balanced datasets.

For the machine learning, eight models were tested: Gaussian naïve-Bayes classifier, logistic regression classifier, linear ridge regression classifier, linear SGD classifier, SVM classifier, KNN classifier, random forest classifier and gradient boosted machines classifier.

The experiment itself was conducted using Python, alongside many open-source libraries for each of these steps. During this experiment, it was concluded that the random search algorithm was the best model for these datasets. After the algorithms, both of these models were cross-validated for each dataset.

To evaluate the score of these ML models, the precision-recall area under curve (PR-AUC) was recognised as a viable measuring parameter for the binary classification. The model performance between two datasets differed, especially around 7% for the sensitivity. The reason may originate from the number of parameters that one can acquire, the complexity of the CNC dataset which also reflects the real-world manufacturing and all the noise arising from such an environment, the size difference between two datasets where the milling dataset had less cases than the CNC dataset. Finally, the smaller datasets can lead to overfitting during the machine learning.

Overall, it can be inferred that the CNC dataset’s quality is much larger than the milling dataset even though it also suffers from its small size.