Abbreviations: FFT – Fast Fourier Transform, CWT – Continuous Wavelet Transform, STD – Standard Deviation, OOT – out of threshold

# Data cleansing and pre-processing

## Dataset overview

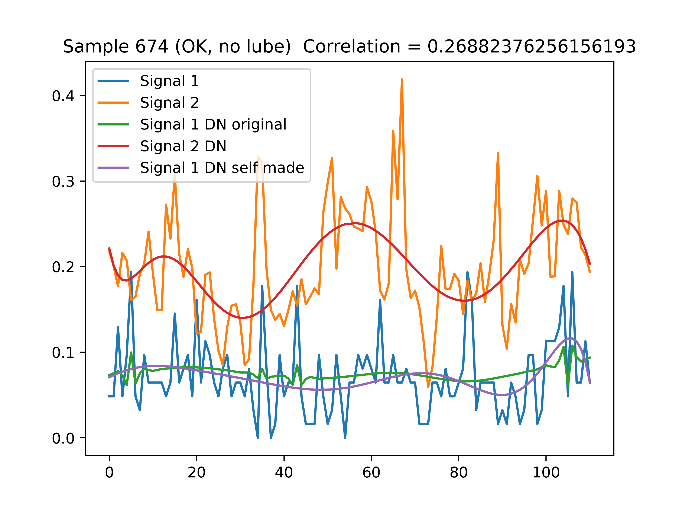
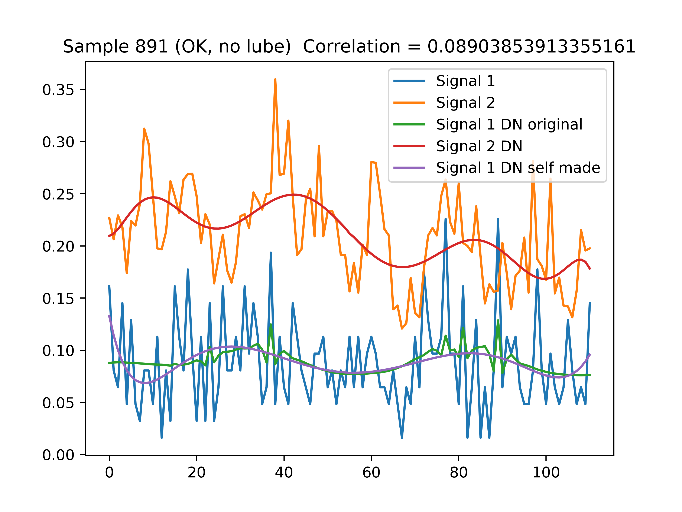
As mentioned in the task description the dataset contains measurements from two different photodiode sensors recorded back reflected laser radiation during laser beam welding. Additionally, denoised values for signal one were also given. Each welding seam was divided into five equidistant sections whose measurements were individually recorded as a time series in the dataset. Based on the assignment the welding parameters were assumed to be constant across the dataset.

Each signal log was categorized in four different, equally distributed binary categories: not OK, signal value exceeded, WD40 pollution and Gleitmo pollution. According to the assignment, labeling algorithm first detected if the signal value had been exceeded and if so, the OK/not OK classification was not carried out; thus, in the dataset all samples for which the signal value had been exceeded were uniformly marked as OK. In the second step the data was labeled according to the potentially present lubricant pollution. The labels of the lubricants were found to be exclusive, meaning that for no welding seam both WD40 and Gleitmo were present. In total, 77 (~6%) measurements were identified as “out of threshold”, 996 (~74%) as “OK” and 277 (~21%) as “not OK”. 450 (~33%) weld seams contained no pollution, 450 (~33%) were marked with WD40 and 450 (~33%) with Gleitmo.

## Pre-processing methodology

### Visualization

Aside from statistical inspection, visualizing the data was the first step in obtaining an overview about features of the dataset. The figures below show plotted signal values for randomly chosen samples with varying labels. As can be seen in the plots, attempts were made to reproduce the denoising procedure used for signal 1 for signal 2. Additionally, the automatically calculated Person correlation coefficients between the values for signal 1 and 2 were also displayed.



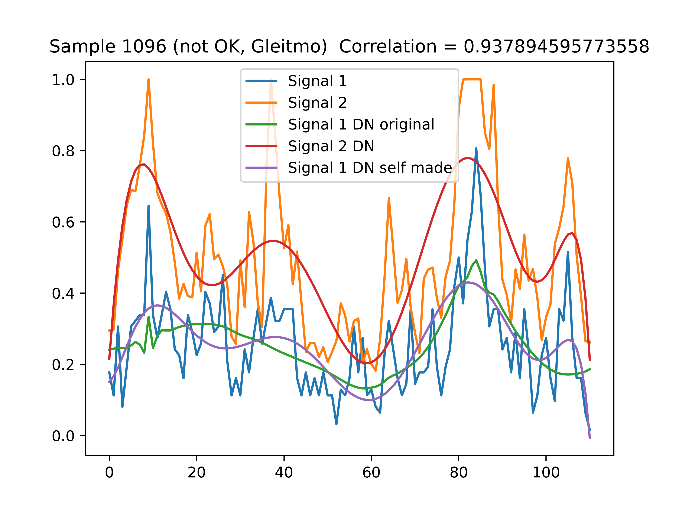
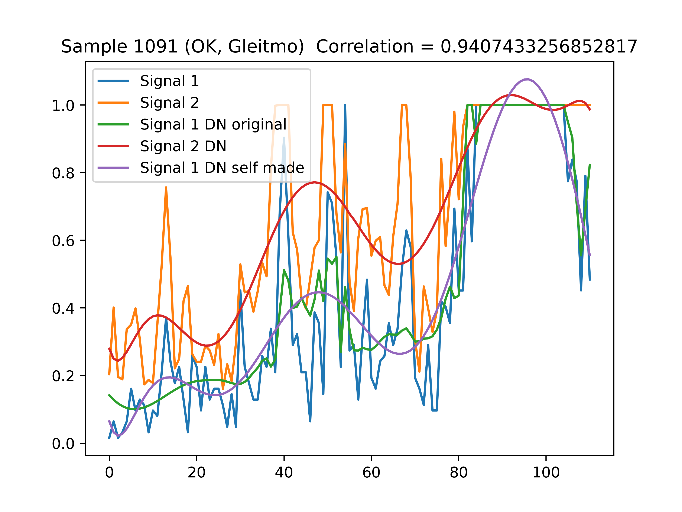


Figure # Visualization plots for four randomly chosen samples.

Based on the data inspection, a few important conclusions could be reached about the dataset: signal 2 was found to be typically much noisier than signal 1 and thus possibly worse for making predictions. Samples marked as “not OK” typically display higher values for all signals; however, no determining boundary distinguishing the two labels could be identified. Furthermore, the correlation coefficients between signals were found to carry little information about the weld quality. Despite reaching the mentioned conclusions, it was difficult to identify the decisive features only based on the visualization. Thus, further analysis including possible data transformation was needed.

### Data cleansing

After a careful examination of the dataset an inconsistency regarding the labelling of signals in the “out of threshold” category was identified: multiple samples across the dataset displayed only the value “1” for all signal readings of all the sensors. Although most of such samples were labeled as “out of threshold”, there was a significant number of logs where this was not the case. As the signal values were the only source of input feature for the prediction model, such inconsistency would pose a significant limit to the achieved accuracy. Thus, a new condition for the labeling of the signals was chosen based on the original set of OOT values and the data relabeled accordingly; the new condition being set at 13 occurrences of the value “1” in signal 1 logs, which was the minimum among the samples originally labeled as OOT.

# Data Transformation Methodology  
Data Transformation aims at extracting meaningful features from the dataset that can be used for training the classification model. This step is especially important while dealing with often noisy and unstructured time series data, such as provided in the given dataset **(1)**. There exist many possible approaches to the choice of features and the decision can greatly affect both the learning speed and the output quality of the model. For the given assignment both 1D and 2D transformation techniques were examined, each of them being briefly discussed in the following section. Visualization plots for selected representative samples were also provided for better evaluation of the statistical meaning of each examined feature.

## Statistical features

Extracting statistical features allows to transform the dataset of one signal into a single characteristic value that may be decisive about the classification of the signal. In the following section the applied statistical features and the results obtained from the dataset are briefly discussed. For the calculation of each feature, an affiliated method from the pandas, NumPy and SciPy modules were used. Since values obtained from signal 2 were typically noisier and thus yielded worse results for classification compared to signal 1 and signal 1 denoised, only visualizations for the two latter ones were provided.

### Mean

One of the most prominent statistical approaches is the arithmetic mean. Although in many cases the method is very useful for characterizing the measurement, it also risks filtering out possibly meaningful outliers, which often makes it inadequate for highly varying data. The graphs below shows the distribution of the mean values of signal one (raw and denoised) across all measurements in the pre-processed dataset with the distinguishment between the OK/NOK labels.

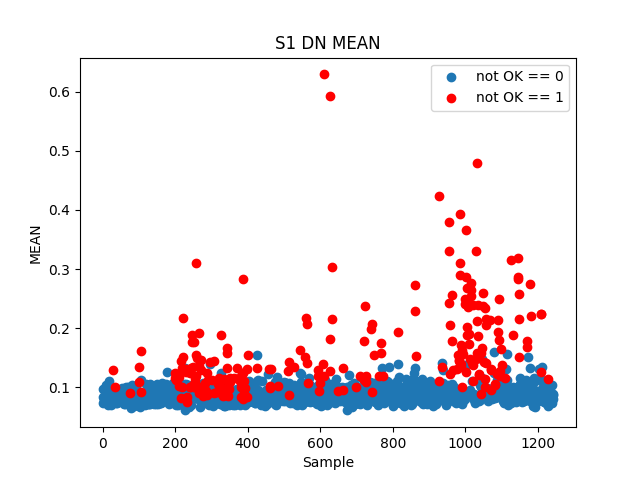
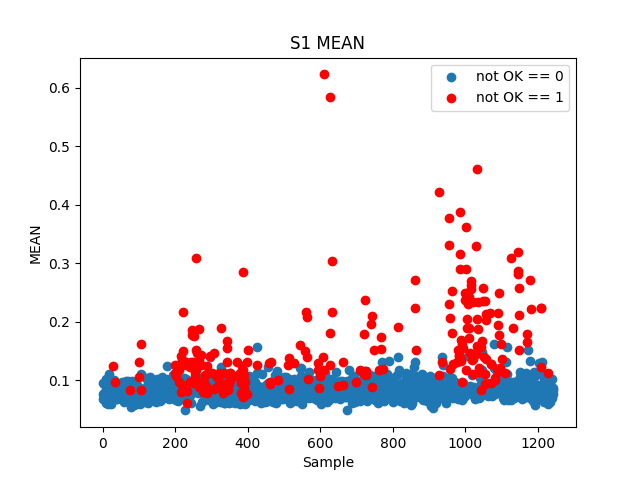


Figure # The mean values of signal 1 across samples.

As seen in the figure, for both signals the data seems to show a significant correlation between the mean value of the signal and the quality of the weld; poor quality welds tending to correspond to higher mean values. This is especially noticeable across samples affected by pollution with the lubricant Gleitmo (samples 200 through 401 and 950 through 1125). Despite that, there exists a vast region of signals with mean around 0.1 where both OK and NOK samples are found. This confusion region is likely to cause inaccuracies in any classification model using the metric as input which highly undermines its practicality for the model.

### STD

Standard deviation is a popular metric of spread of the values within a sample. Thus, it helps distinguish stable measurements from highly varying ones. In context of quality prediction based on time series data the feature can provide valuable information about possible changes detected in the weld seam. The figure below shows the distribution of standard deviation values across the dataset.

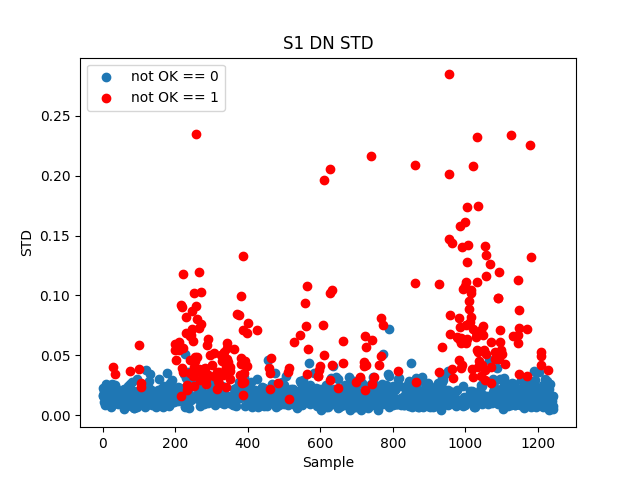
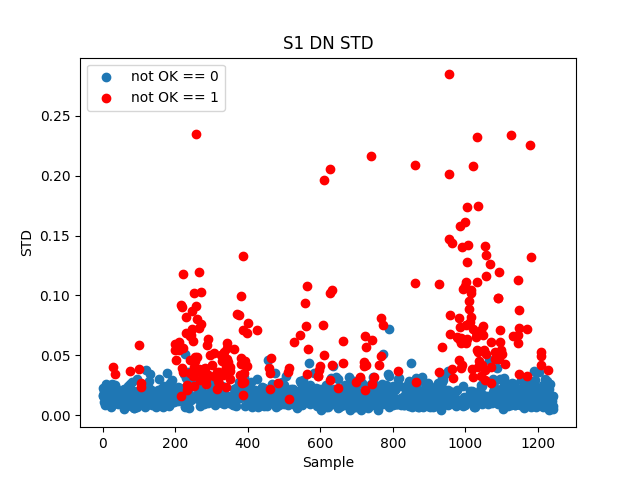


Figure # The mean values of signal 1 across samples.

Based on the plot the signal’s standard deviation seems slightly more adequate for OK/NOK classification of the samples than the mean; the boundary between the two clusters being more protruding. The effect is again augmented by the presence of Gleitmo lubricant. However, the confusion region in this case is still very prominent which again questions the metric’s applicableness.

### Minimum, maximum and percentiles

Other easily identifiable characteristic features within a sample are the maximum, the minimum, as well as the value at a given percentile (most commonly used being Q25, Q75, and Q50 aka the median). The plot below shows the scattering of the first signal’s maximal value across samples which proved to be the most significant metric across all the others mentioned in this section.

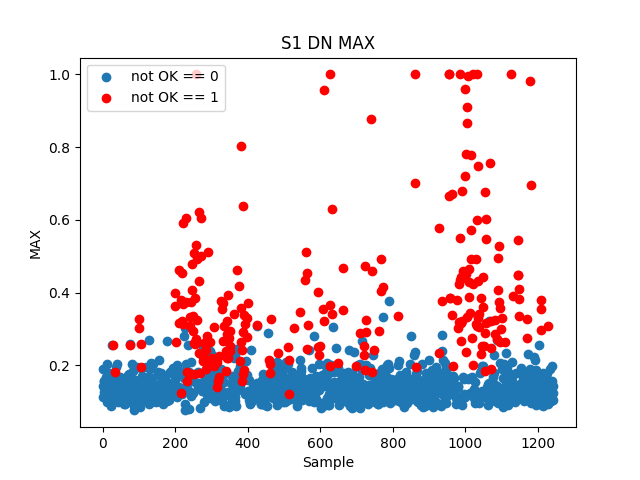
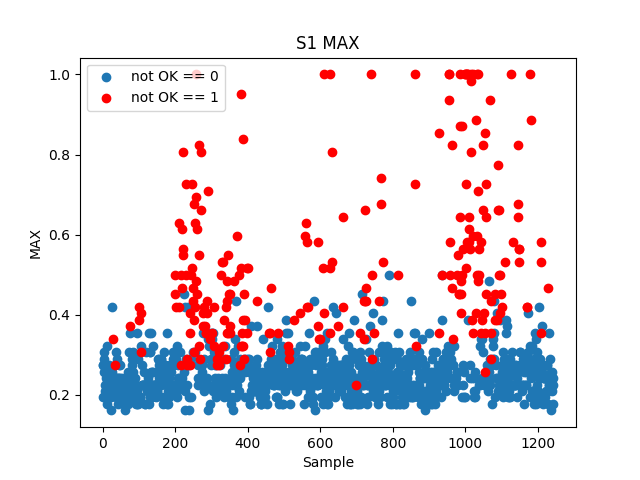
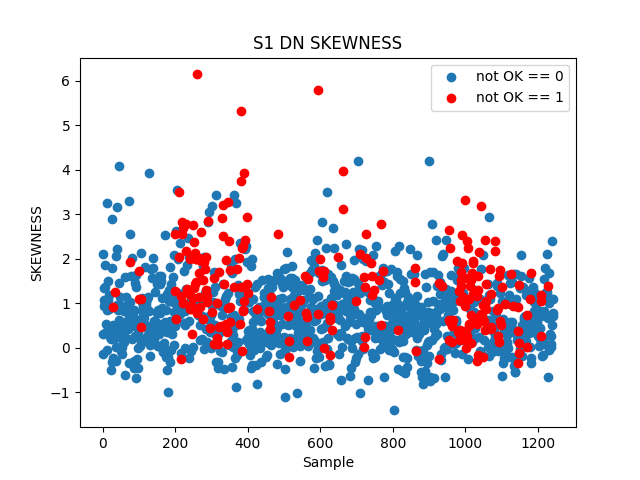
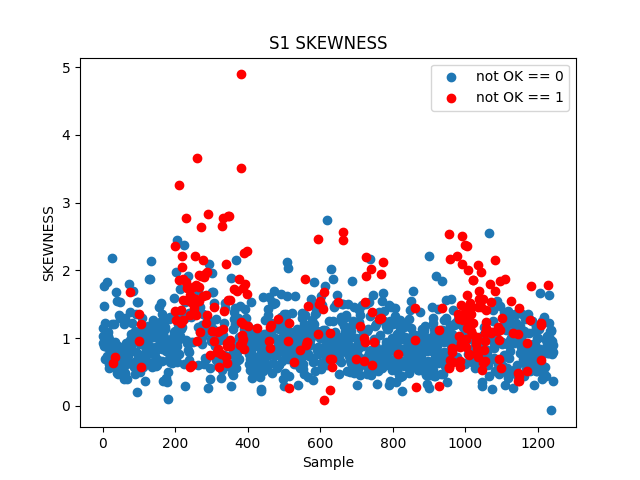


Figure # The maximum values of signal 1 across samples.

Based on the visualization the maximal value seems to be the most promising feature for OK/NOK classification compared to the other examined ones. Nevertheless, a significant confusion region is still noticeable. This is likely due to the metric’s sensitivity to sudden changes which the signals exhibit.

### Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable around its mean value. For positive skewness more samples have a height value smaller than the mean; the opposite being true for negatively skew distribution **(1)**. The plots below show the distributions of sample skewness for signal 1 and signal 1 denoised across the measurements.



As visible in the graphs, the poor differentiation of skewness values between the labels for both signals make the feature unsuitable for the classification task. It suggests that there is hardly any correlation between the weld quality and the proportions of the signal’s distribution.

### Entropy

Entropy is another statistical feature often applied for time series data. It is the measure of impurity associated with a random variable. A sample containing more noise will have a higher entropy value **(2)**. The graph below shows the distribution of entropy values for signal 1 across the dataset.

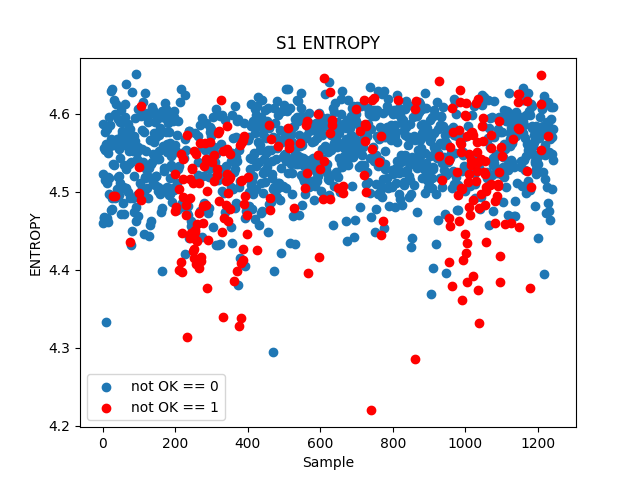
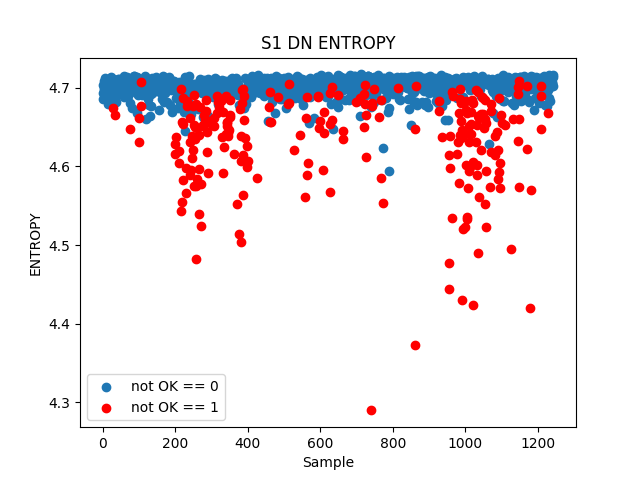


Figure # The entropy of signal 1 across samples.

Based on the graph for the denoised data yet again a characteristic pattern can be seen for samples of both categories; the NOK measurements having typically lower entropy values. Still, it is difficult to identify an unambiguous boundary between the two clusters which makes the metric unsuitable for effective classification. For the raw signal data the points are even more indistinguishable from each other. This is likely caused by the large amount of noise in the measurements.

## CWT

Similarly to Fourier Transform, the wavelet transform maps the signal from time to frequency domain; the key difference being that is also provides time resolution of the transformed signal which makes it more suitable for analyzing non-stationary systems **(3)**. The reasoning behind applying CWT upon the provided dataset was the assumption that certain changes in the signal’s frequency may be caused by particular defects in the welding seam. Furthermore, a wavelet-based algorithm was also mentioned as a part of the labelling process which suggested that the transform may yield valuable information about the weld seam quality.

The transform was carried out with help of the pywavelets module using the Morlet wavelet and varying ranges of scaling factors (up to 1000). The transformed data was subsequently visualized with the intention of identifying any possible underlying patterns. The following figures show exemplary plots for two randomly chosen samples with differing labels.

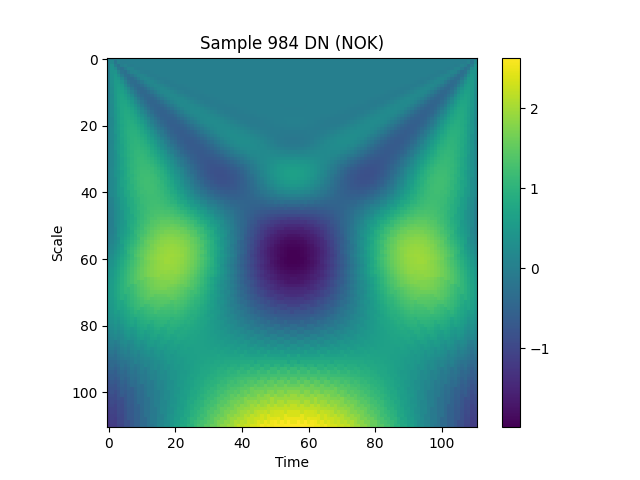
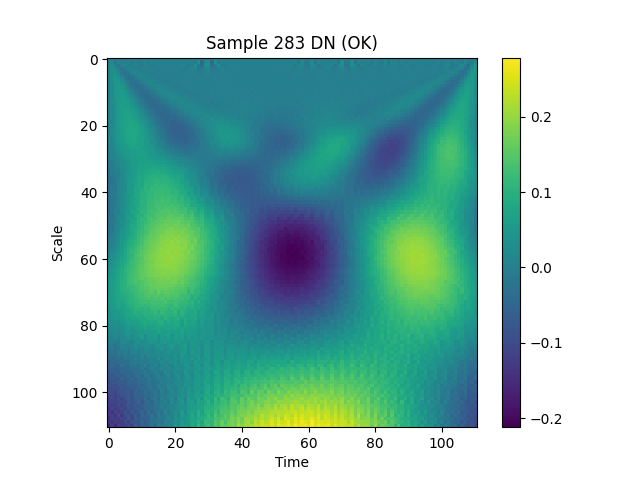


Figure # CWT visualization for samples 283 and 984 (Signal 1 denoised)

As demonstrated by the above figures, no compelling correlation between the labelling and the wavelet transform coefficients of the signal could be identified. Thus, it was decided not to include the transformed data in further consideration. It is possible that the seemingly poor performance of the approach resulted from the inappropriate choice of the transform parameters. However, their adjustment would require additional knowledge about the data collection and labelling process which was not provided.

**# Choice of the model**

**…**

# Bibliography

(1) Benker, M. (2022) Data Transformation [PowerPoint slides]. Technical University of Munich Institute for Machine Tools and Industrial Management

(2) Esmael B., Arnaout A., Fruhwirth R. K., Thonhauser G. (2013) A Statistical Feature-Based Approach for Operations Recognition in Drilling Time Series. *International Journal of Computer Information Systems and Industrial Management Applications.* <https://pure.unileoben.ac.at/portal/files/1073786/A_Statistical_Feature_Based_Approach_for_Operations_Recognition_in_Drilling_Time_Series.pdf>

(3) ML Fundamentals (2018, December 21). A Guide for Using the Wavelet Transform in Machine Learning [blog post]. Retrieved from: <https://ataspinar.com/2018/12/21/a-guide-for-using-the-wavelet-transform-in-machine-learning/>