Abbreviations: FFT – Fast Fourier Transform, CWT – Continuous Wavelet Transform, STD – Standard Deviation

# 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 singe 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.

### 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 figure below shows the distribution of the mean values of signal one (unfiltered) across all measurements in the pre-processed dataset with the distinguishment between the OK/NOK labels.

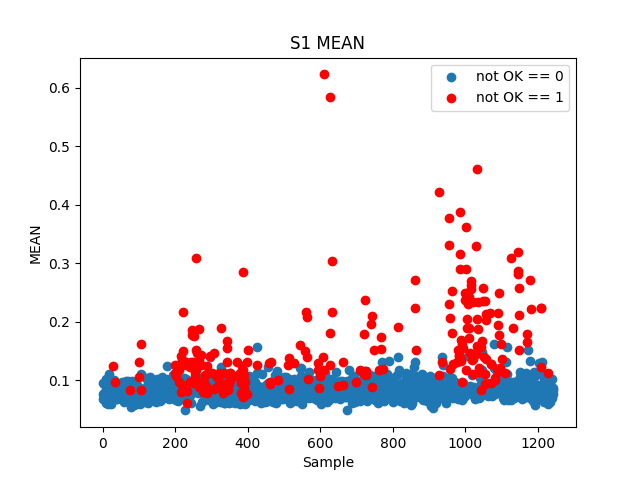


Figure # The mean values of signal 1 across samples.

As seen in the figure, 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.

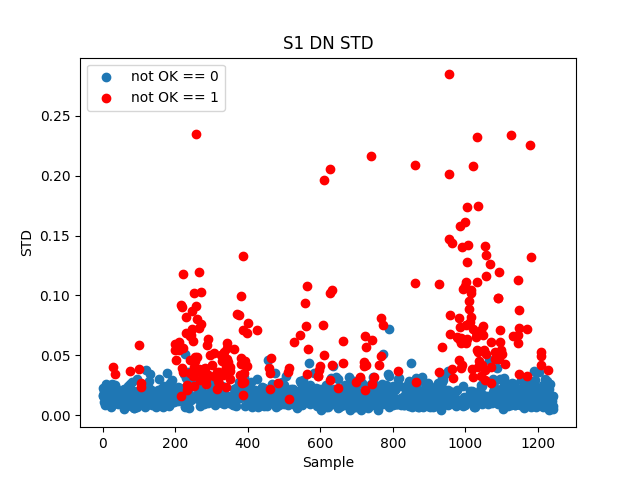


Figure # The mean values of signal 1 (denoised) across samples.

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 usefulness.

### Min/Max value 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.

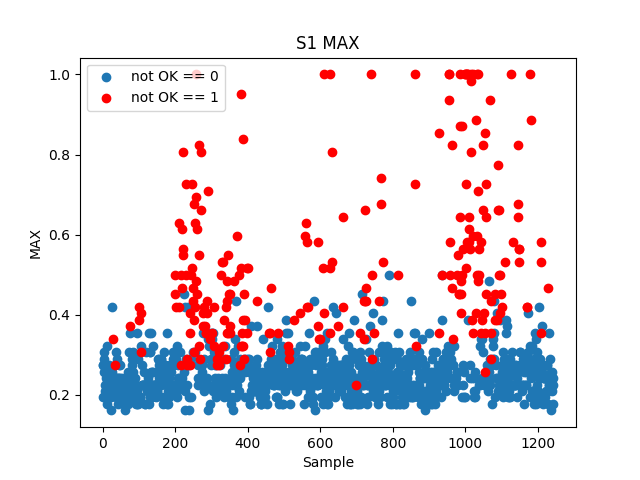


Figure # The maximum values of signal 1 across samples.

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.

## 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 **(2)**. 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.

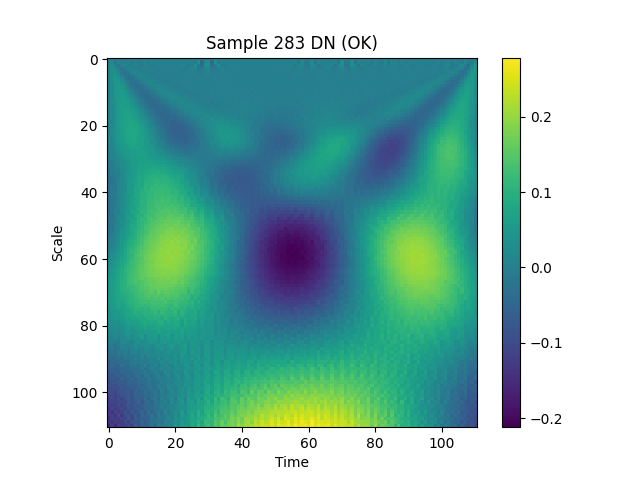


Figure # CWT visualization for sample 283 (Signal 1 denoised)

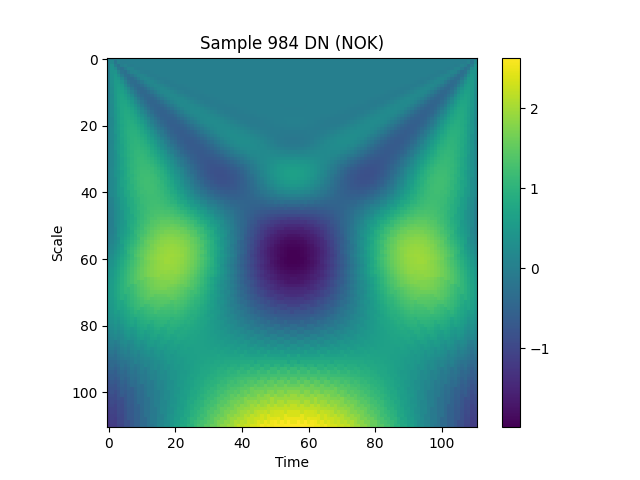


Figure # CWT visualization for sample 984 (Signal 1 denoised)

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 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.

# Bibliography

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

(2) 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/