Machine Learning

Submission Assignment 02

von
Benjamin Ellmer
(S2210455012)

Mobile Computing Master
FH Hagenberg

November 26, 2023

1. Data Preprocessing

1.1. Preprocessing - Missing values

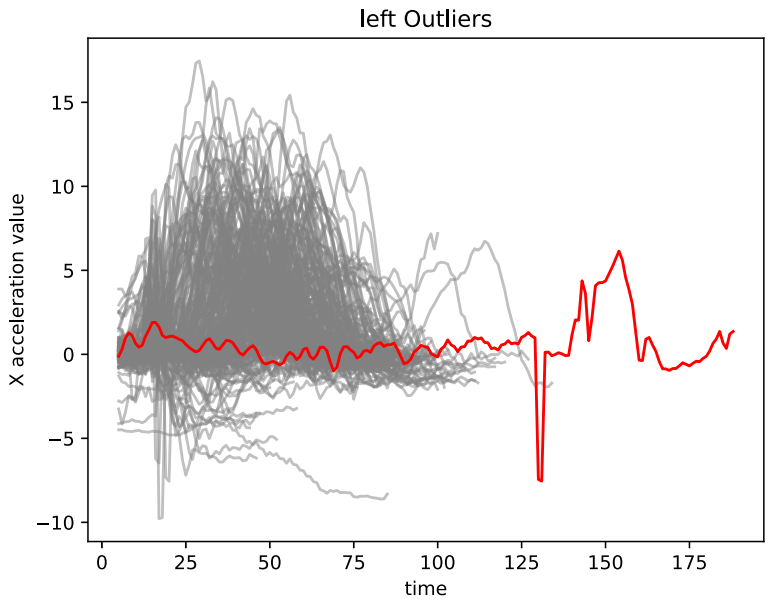
Are there any missing values which need to be taken care of?

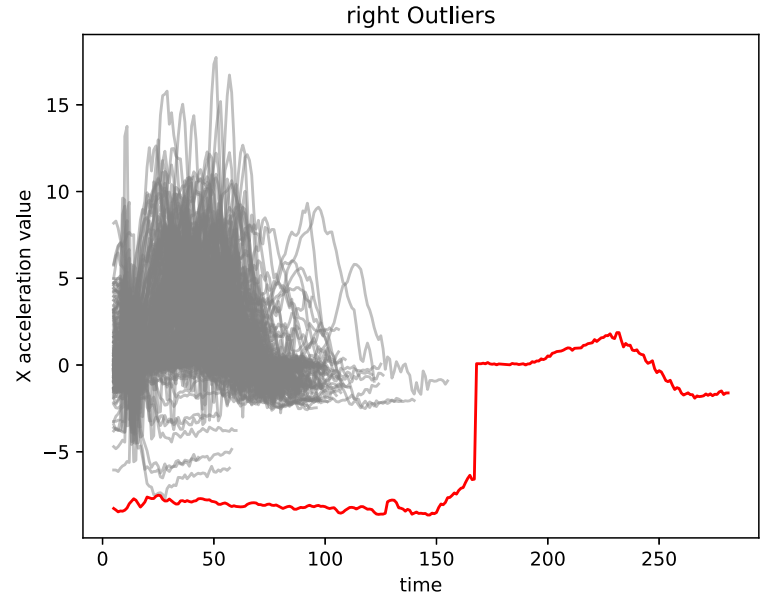
No, I did not find a sample that has missing values. This was done by counting the number of values that are not nan for each sample. Then I counted the number of nan values for the first values.

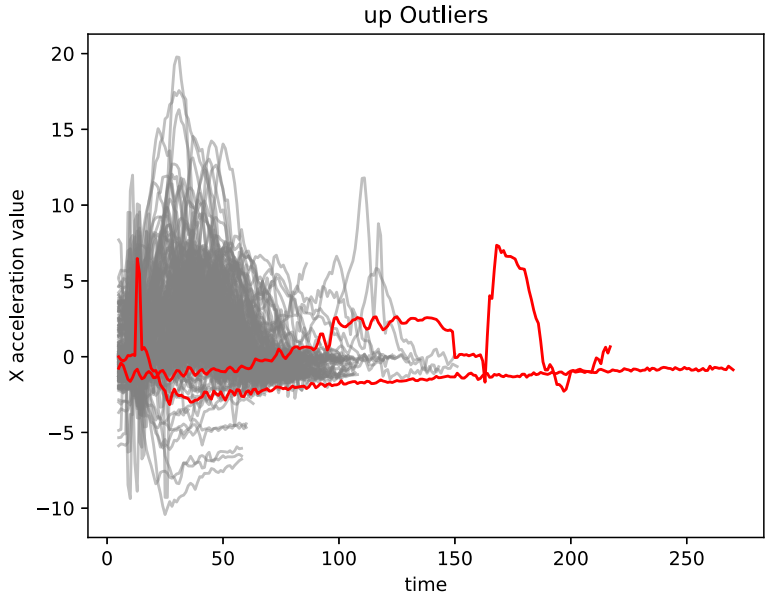
This means if the sample has 100 not nan values I took the first 100 values and counted the number of nan values. If the number of nan values in the first 100 values is 0, then there is no missing data in the sample.

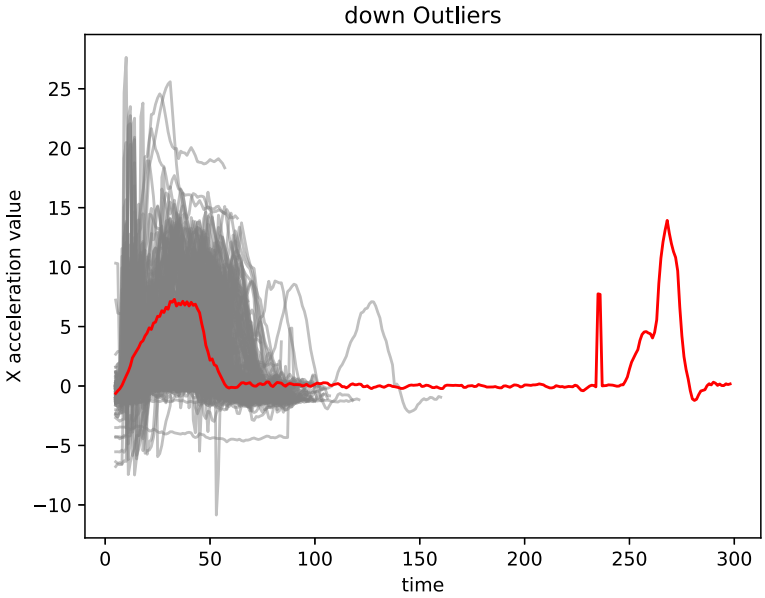
1.2. Preprocessing - Outliers

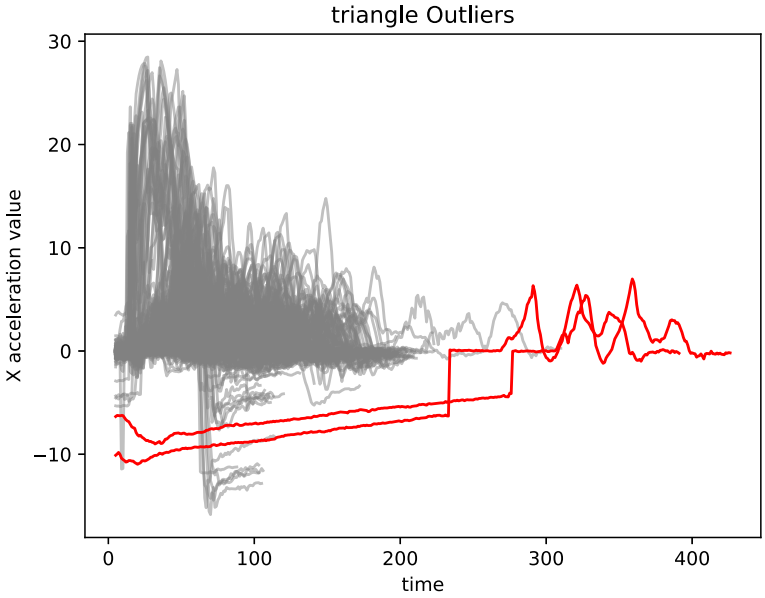
Looking at the mean, or the std it is hard to determine which samples are real outliers. In Figure 9 and Figure 8 we can see that there is a large amount of outliers std and mean. Since the data will also be filtered I decided to remove only the outliers that are visually recognizeable, especially by having an irregular length. Those outliers are visualized in the figures Figure 1, Figure 2 Figure 3, Figure 4, Figure 5, Figure 6 as the red samples.

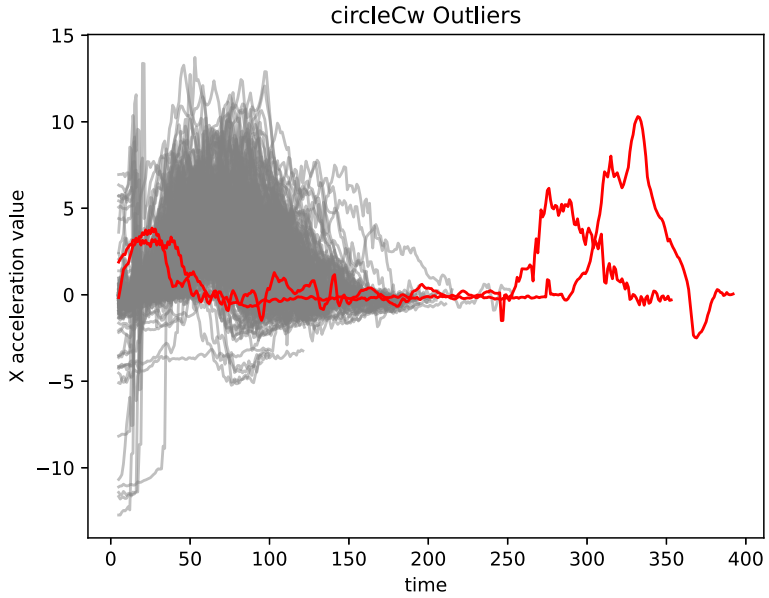



right Outliers
X acceleration value
15
10
5
0
−5
0
50
100
150
200
250
time


up Outliers
X acceleration value
20
15
10
5
0
−5
−10
0
50
100
150
200
250
time


down Outliers
X acceleration value
time
25
20
15
10
5
0
−5
−10
0
50
100
150
200
250
300


triangle Outliers
X acceleration value
time
30
20
10
0
−10
0
100
200
300
400

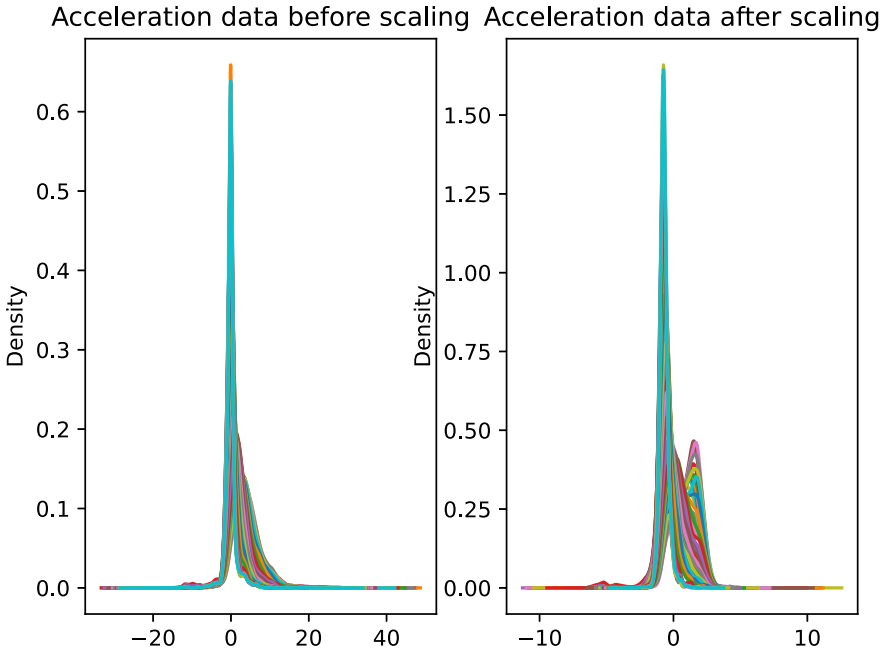


1.3. Preprocessing - Feature Reduction

To get an equal amount of acceleration values I interpolated the gesture data and took an equal amount of values for each sample. According to the description the sensor was recording with 100Hz and the max frequency that makes sense is 20Hz, this means 84.6 (423/5) values should be sufficient. Therefore, I visualized the comparison of the interpolated data for each gesture using 50, 100 and 200 values compared to the original data. As a result I decided to continue to work with only 50 values, because I think it still shows the mandatory information. The comparisons can be seen in Appendix Interpolations 4.1.

1.4. Preprocessing - Normalization

I decided to normalize the data using scaling. Figure 7 shows the distribution of the data before and after normalization. The normalization was done, because it afterwards helps the models to work with the data.



1.5. Preprocessing - Filtering

I decided to use filtering to reduce the noise in the data, as suggested in the exercise hints. I tried the suggested filters (running mean, running median and savgol filter).

In my opinion the savgol filter preservs the trends of the gestures best, therefore I continued with the data that was filtered using the savgol filter. The comparisons can be seen in Appendix Filters 4.2. Regarding the windowsize I tried some sizes and ended up with 8, but I think including theses plots here would be too much.

1.6. Preprocessing - Feature Addition

In the last exercise we already saw, that there is a correlation between the length of a sample and the gesture type. But, by processing the samples to get samples with equal lenghts, we lost this information. Therefore, I added it manually as an extra feature, describing the original length of the recording.

2. Feature Extraction

Yes I think it makes sense to derive more features besides the acceleration values, or at least try and look if there might be ones that make sense. I chose to extract the following features. The proof that it did make sense to derive more features is discussed in Section 3.

Based on the x-axis data (preprocessed):

- raw values
- min
- max
- mean
- median
- standard deviation
- innerquartile range
- median absolute deviation
- number of minimas
- number of maximas
- zero crossing rate
- median crossing rate
- frequency power
- frequency angle
- autocorrelation
- wavelet components

Based on the 1st derivative of the x-axis data:

- raw values
- min
- max
- mean
- median
- standard deviation
- innerquartile range
- median absolute deviation
- autocorrelation
- wavelet components

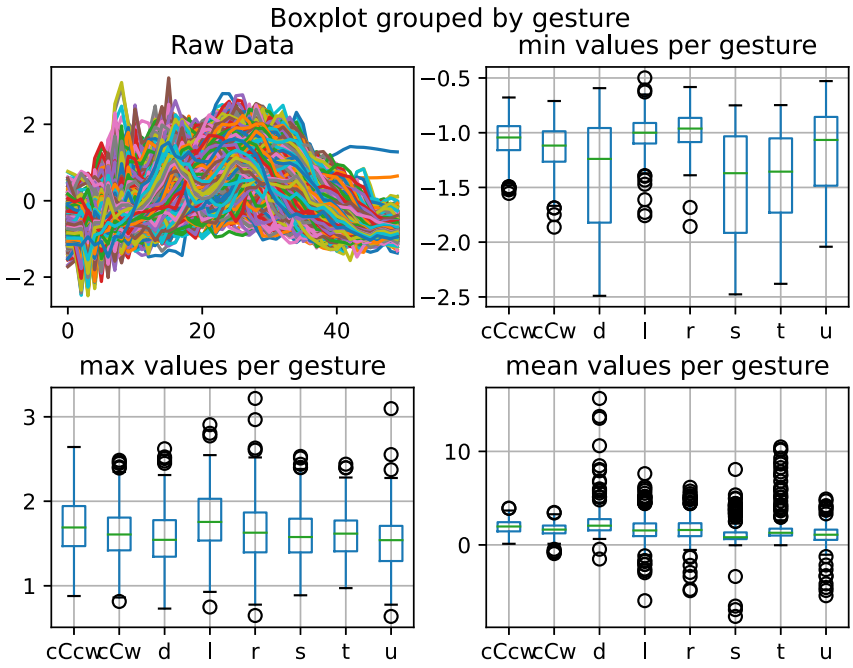
Based on the 2nd derivative of the x-axis data:

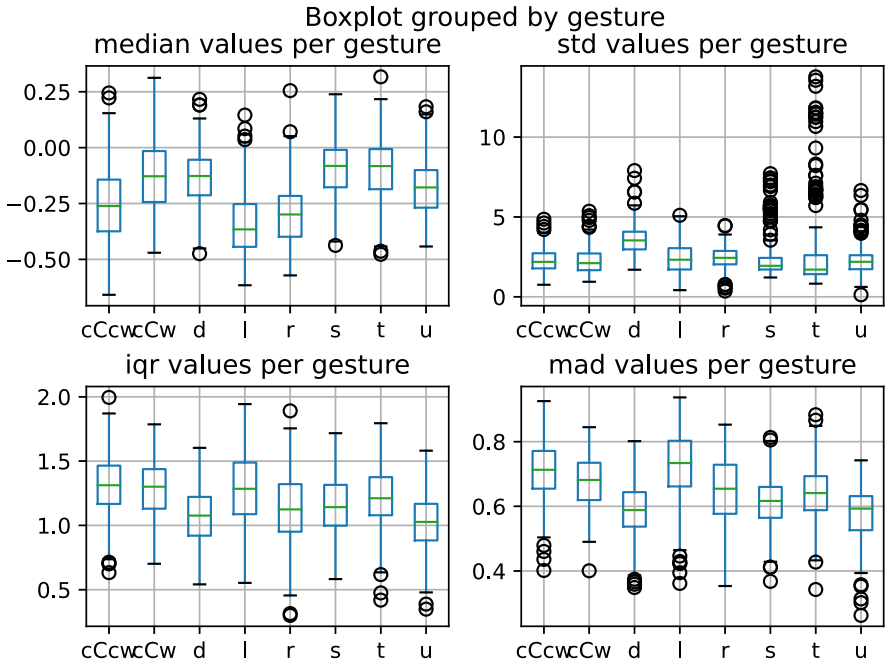
- raw values
- min
- max
- mean
- median
- standard deviation
- innerquartile range
- median absolute deviation
- wavelet components

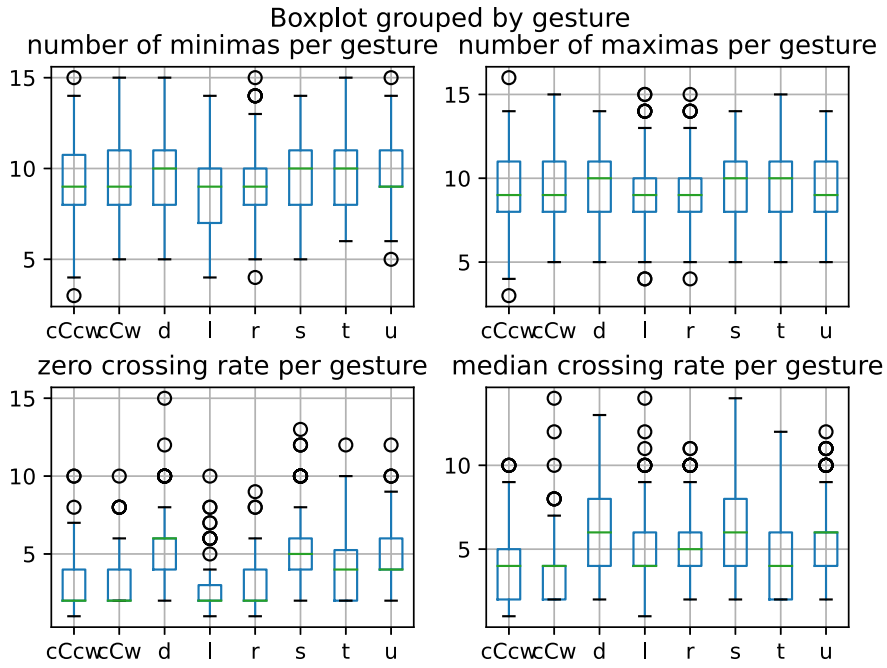
Afterwards I evaluated the features and chose the most promising ones, as described in Section 3.

2.1. Feature Plots based on raw x-axis data

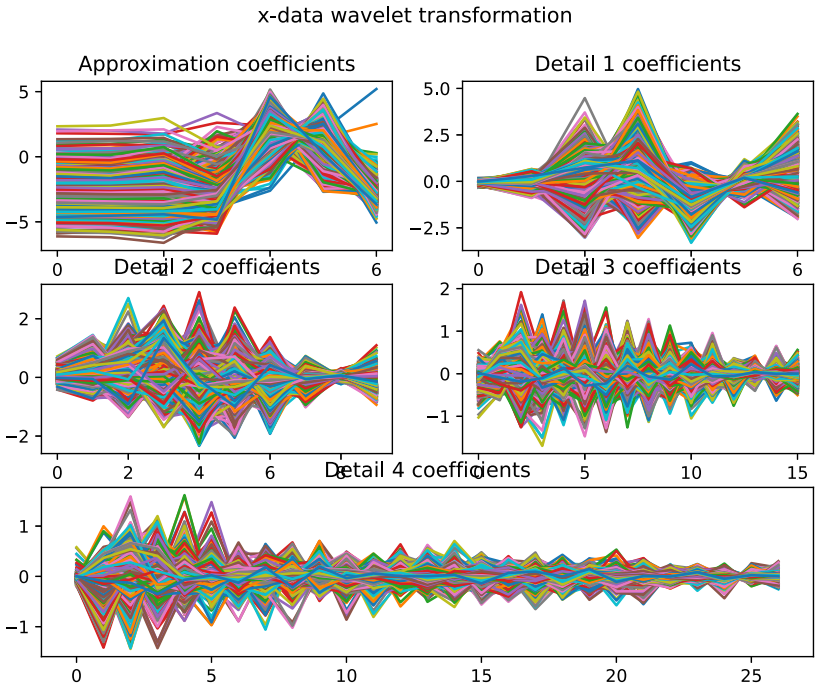
Note: The mean and the std are calculated based on the original data, because the samples were scaled during preprocessing.



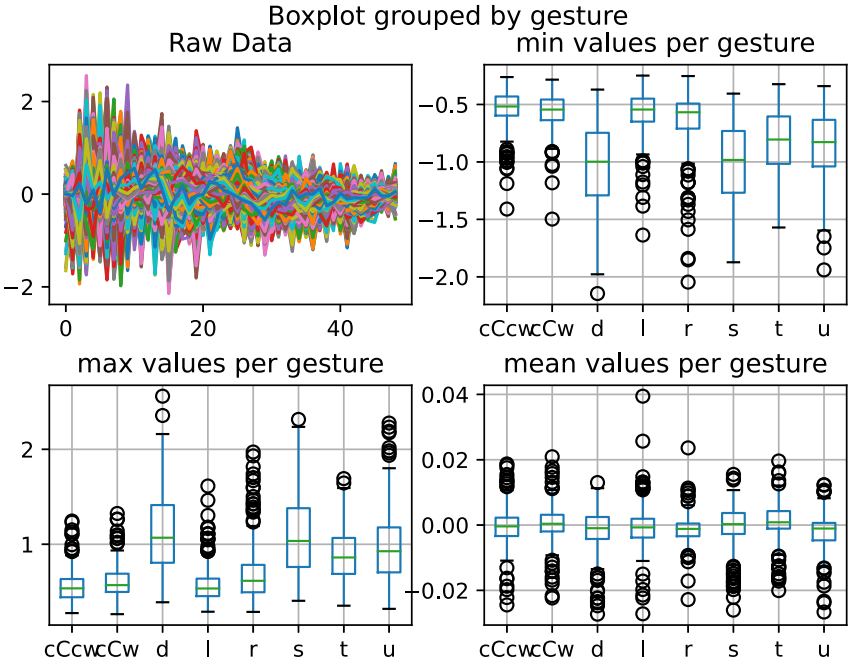


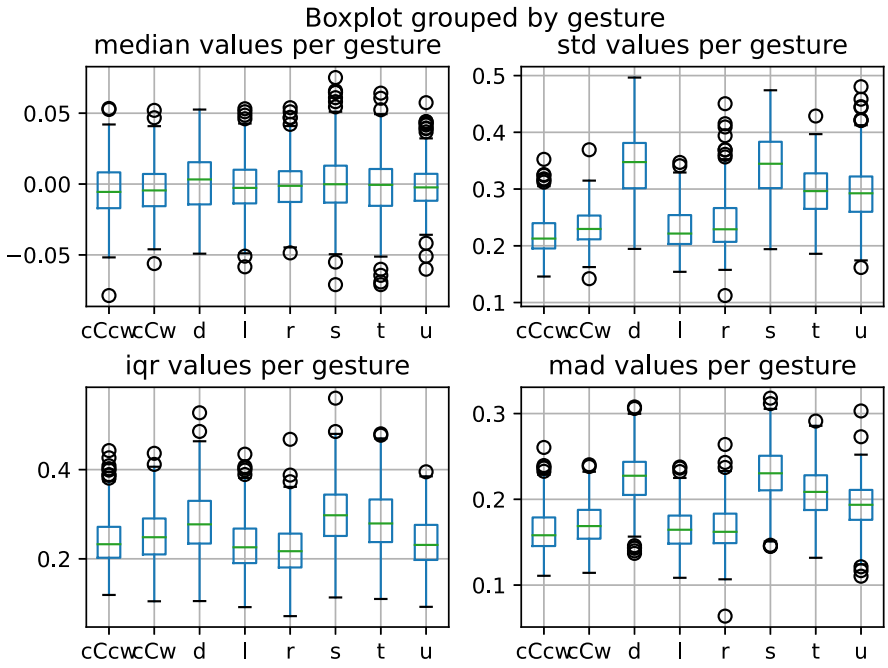


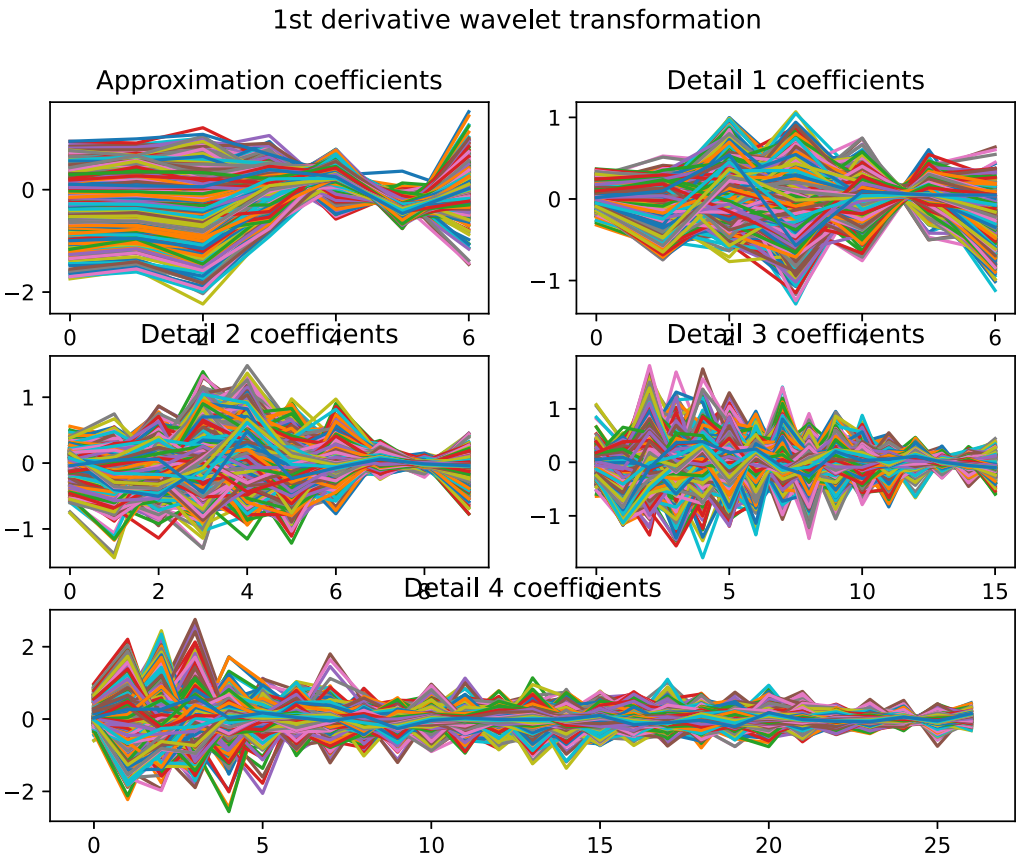
FFT power, FFT phase, ACF:
I created a plot that includes FFT power, phase and the ACF plots, but the created svg had almost 80MB, therefore I did not include it in the protocol.



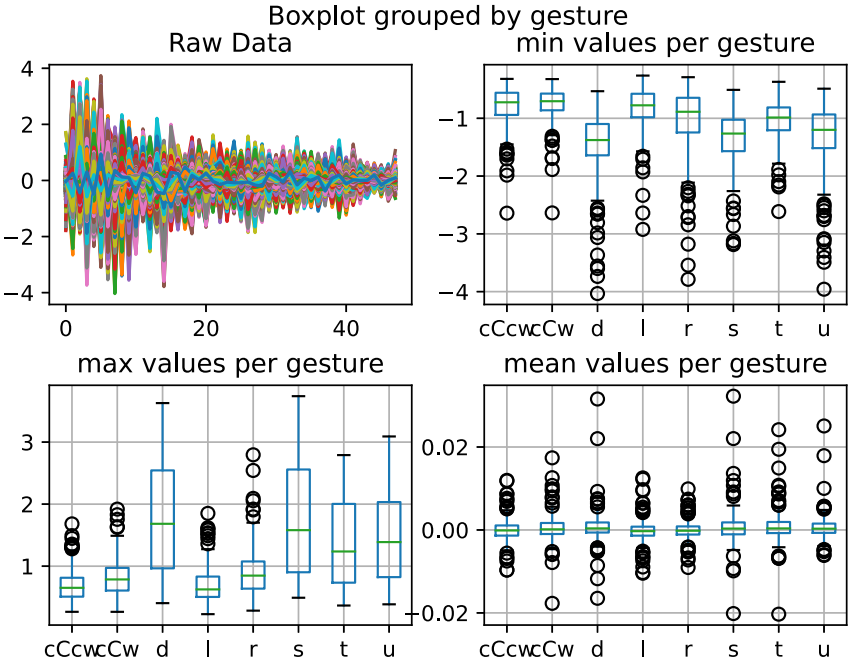
2.2. Feature plots based on 1st derivative

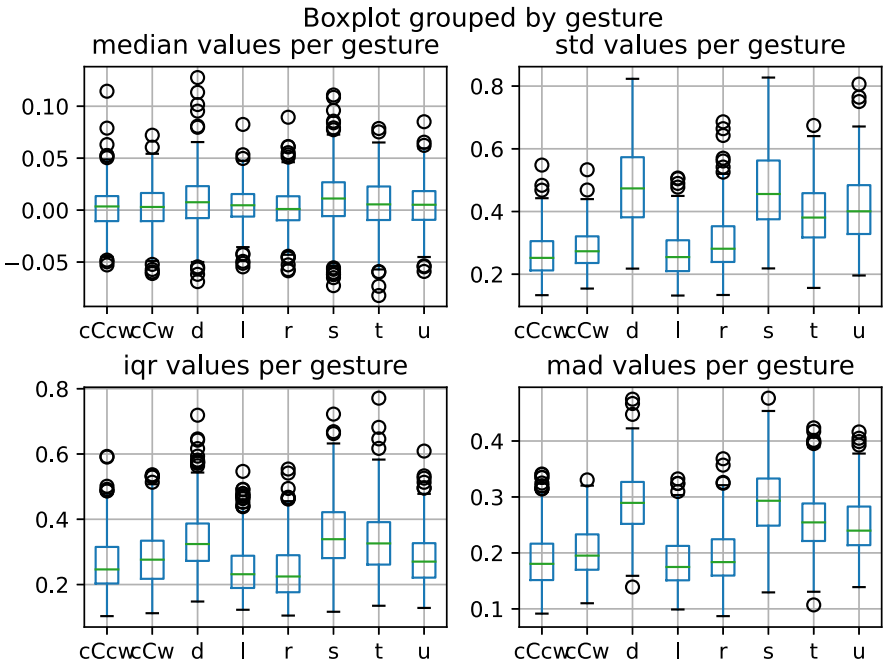


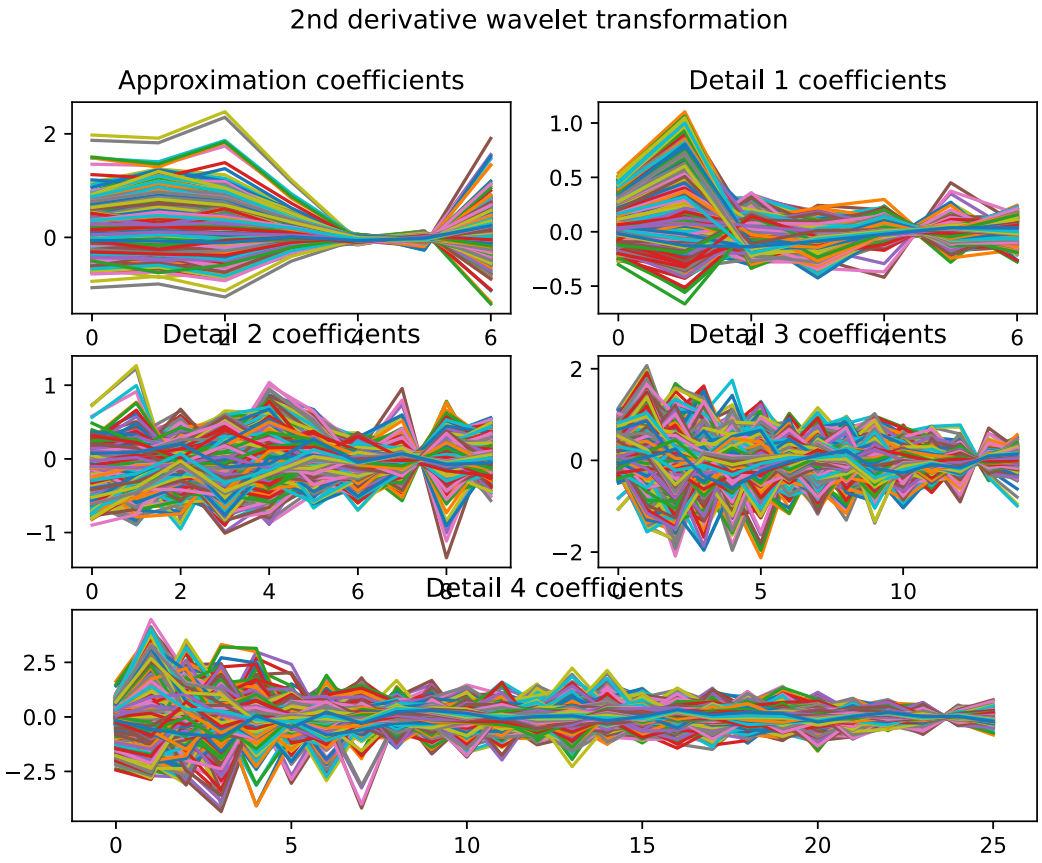



1st derivative wavelet transformation
Approximation coefficients
Detail 1 coefficients
Detail 2 coefficients
Detail 3 coefficients
Detail 4 coefficients

2.3. Feature plots based on 2nd derivative of x-axis data

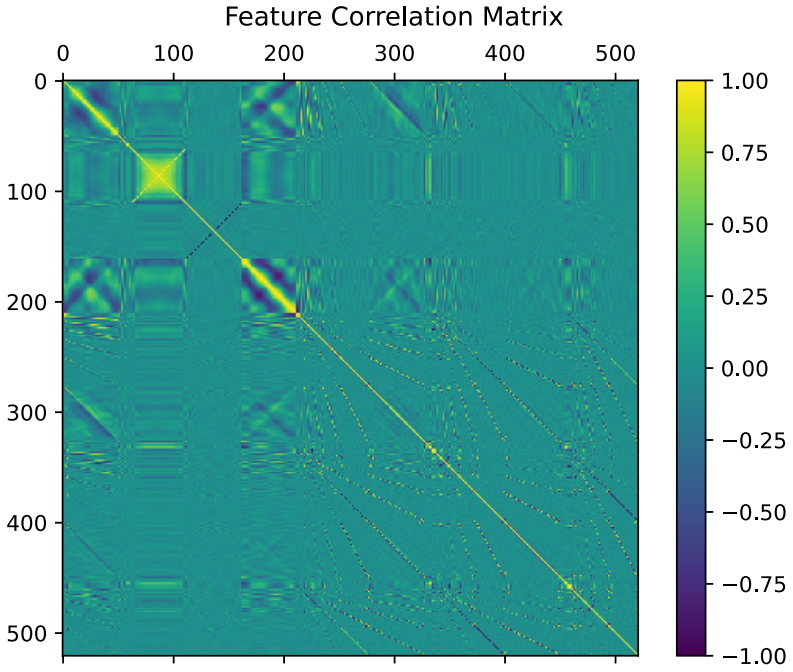





2nd derivative wavelet transformation
Approximation coefficients
Detail 1 coefficients
Detail 2 coefficients
Detail 3 coefficients
Detail 4 coefficients

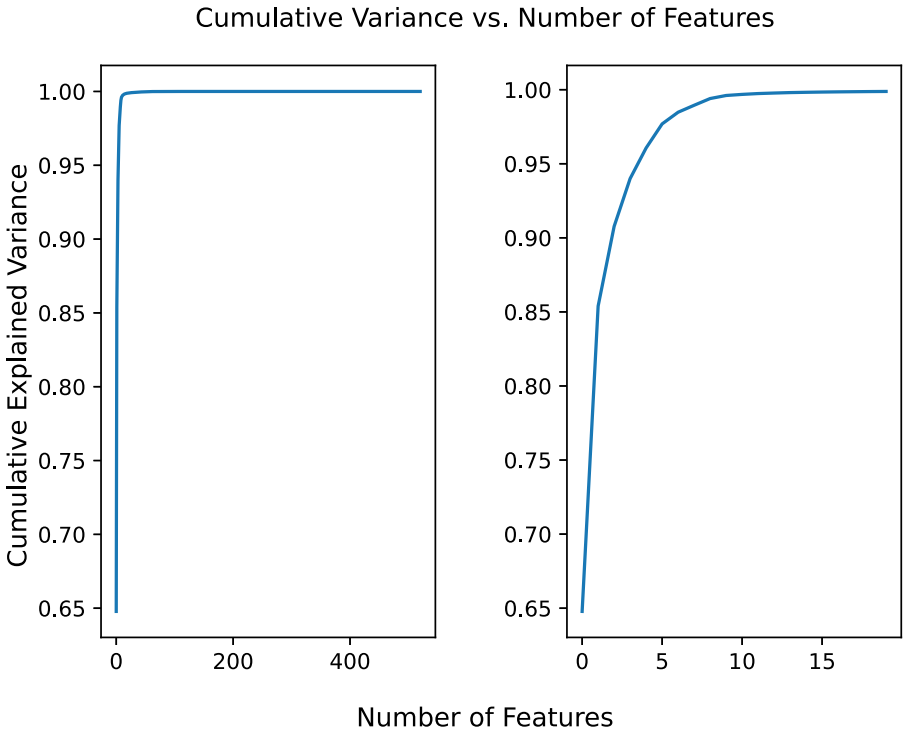
3. Feature Selection/Discussion

First of all, I plotted the correlation matrix to get a picture of the correlations beyond the features. Because of the huge amount of features, it is hard to get any useful information out of the plot. I removed the ticks and labels, because the texts were overlapping.



Still, we can see in Figure 18, that many features have a strong correlation and that we might be able to drop some of them.

Then I chose PCA to look at the variance of the features. In Figure 19 we can see that we should have about 99% variance with only 10 features. This is very interesting in the next assignment I will check which accuracy I can reach with only 10 features.



Before looking at the cross validation score of all extracted features I wanted to check the cross validation scores of only the acceleration data. If there would not be much difference, all the extracted features would be useless. Figure 20 shows the results with only the acceleration values.

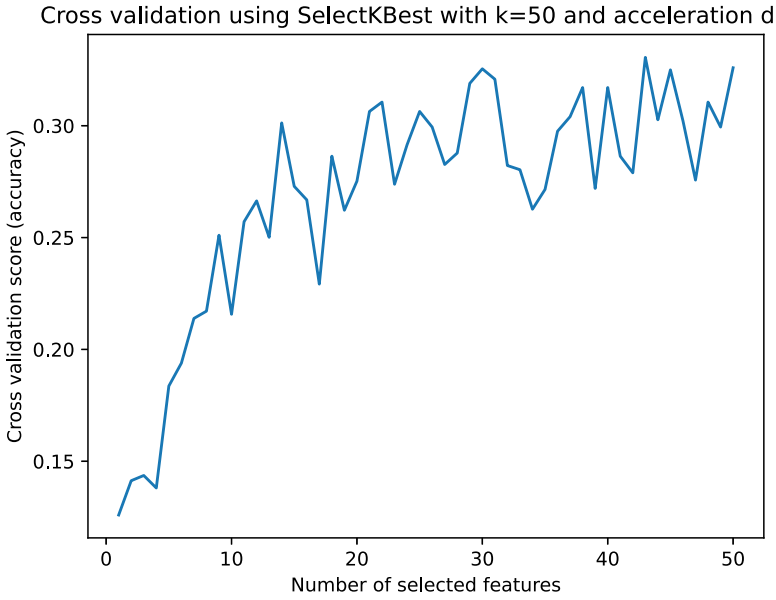
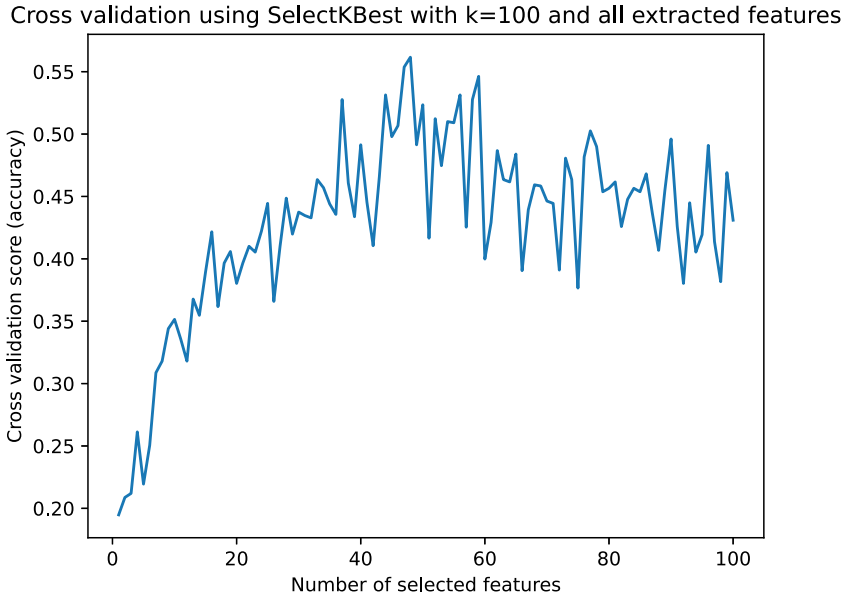


Figure 21 shows that it was worth extracting the features reaching almost 55%, which is about 25% better than using only the acceleration data.

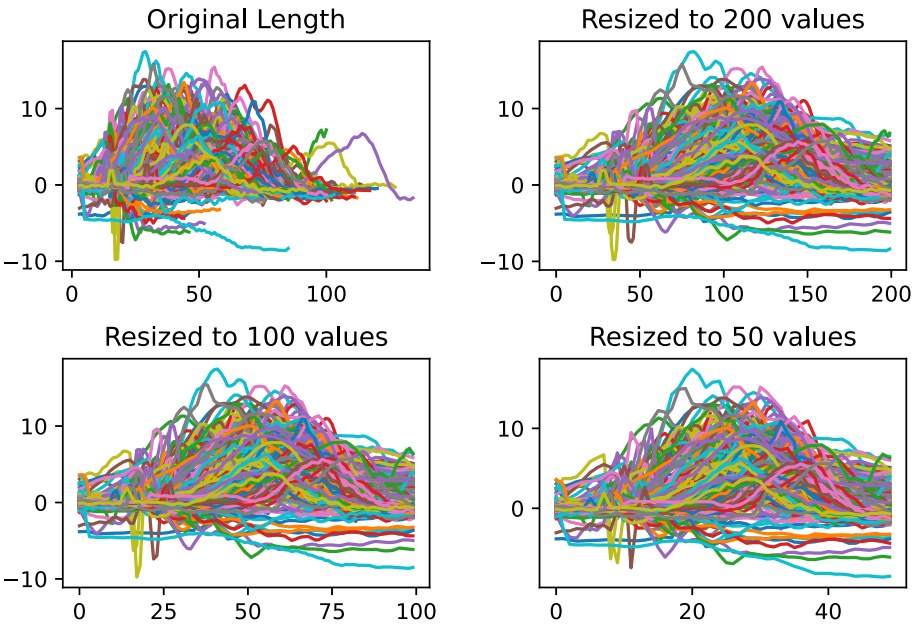


In the next assignment I will look at the features and choose the best ones this assignment showed only that it makes sense to derive more features

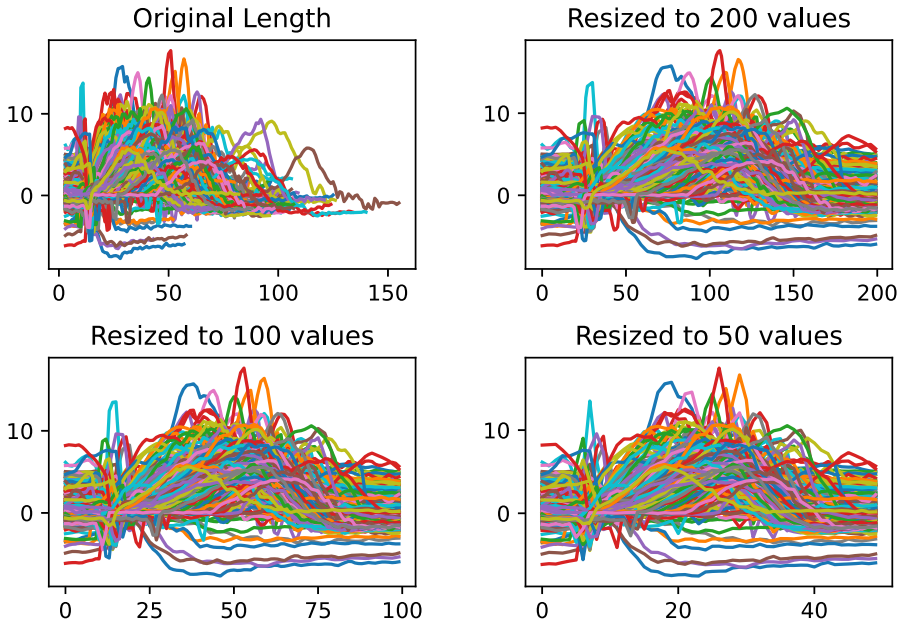
4. Appendix

4.1. Appendix Interpolations

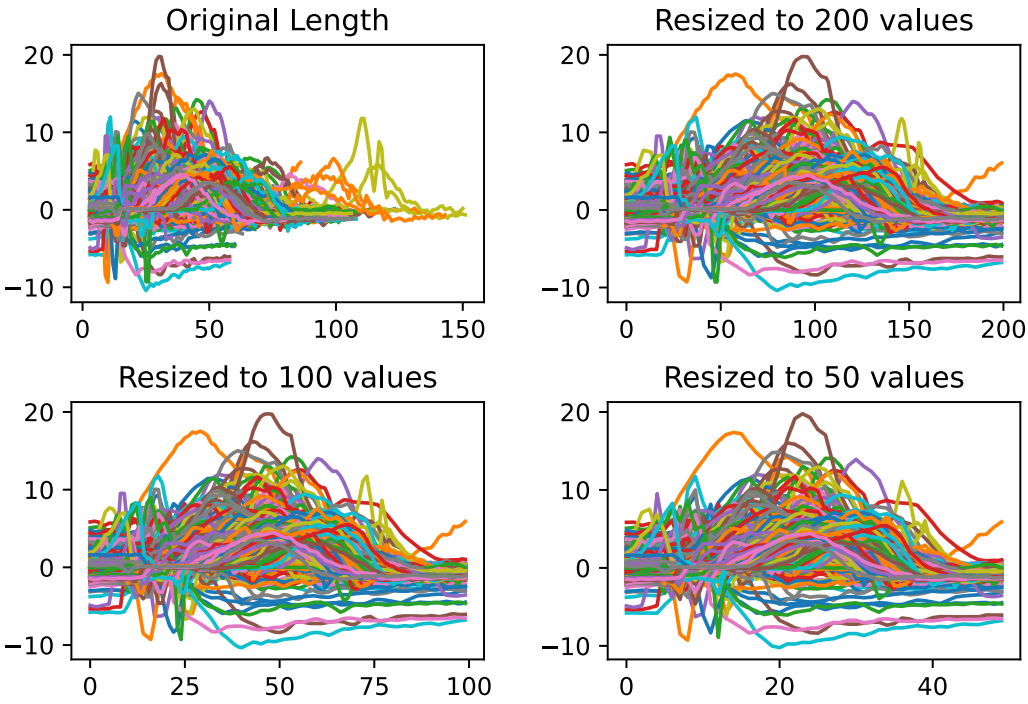
left gesture interpolation and resizing comparison



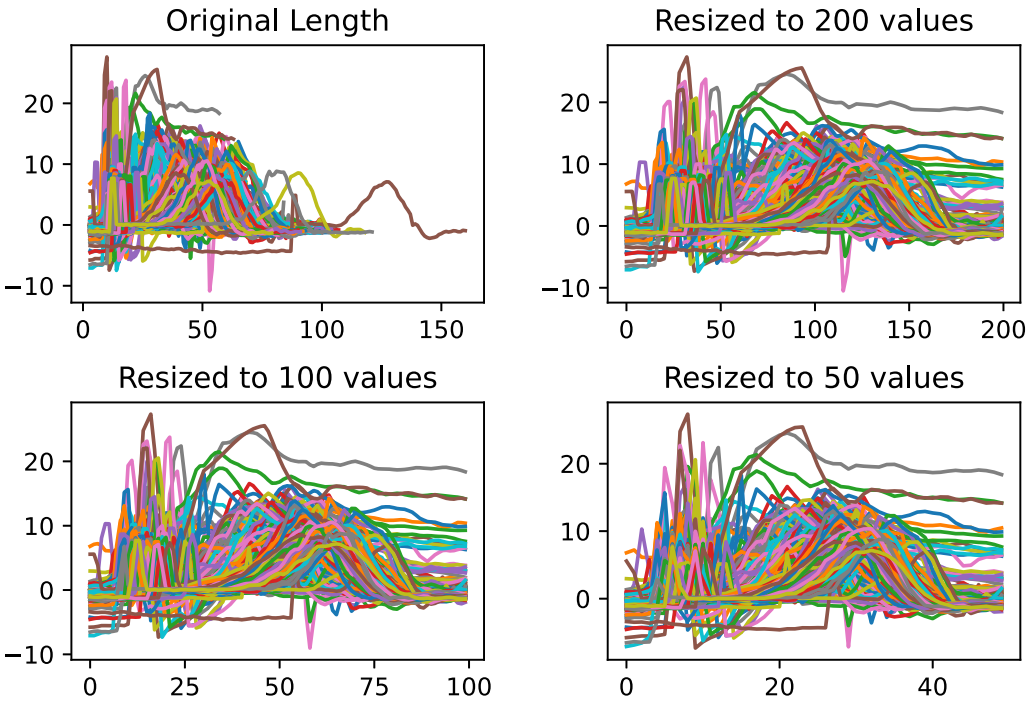
right gesture interpolation and resizing comparison

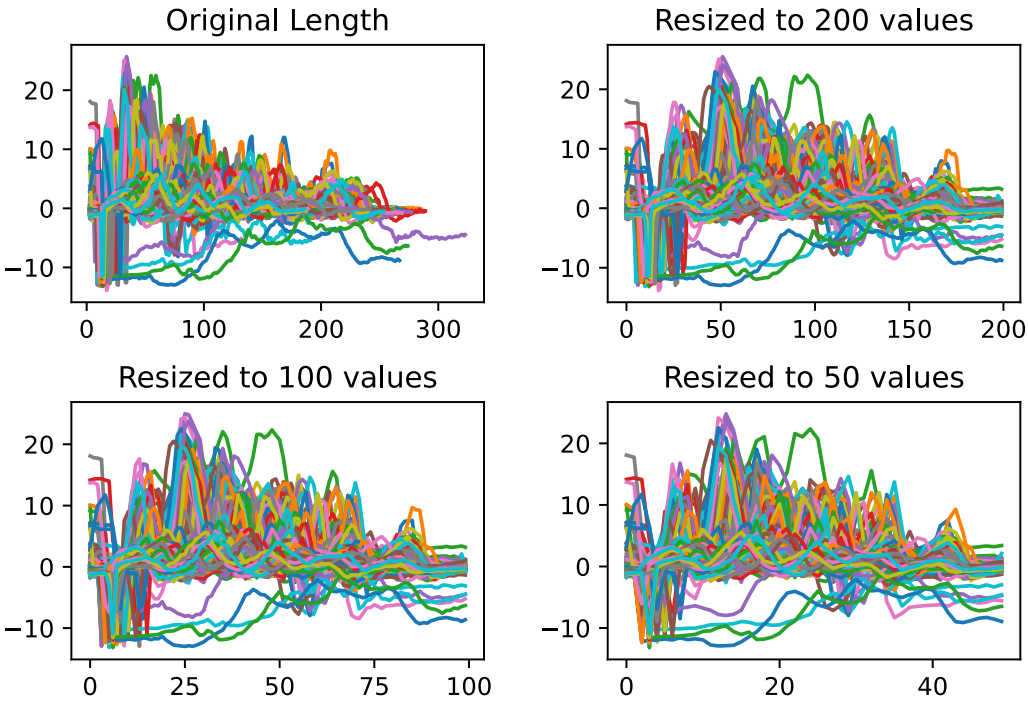


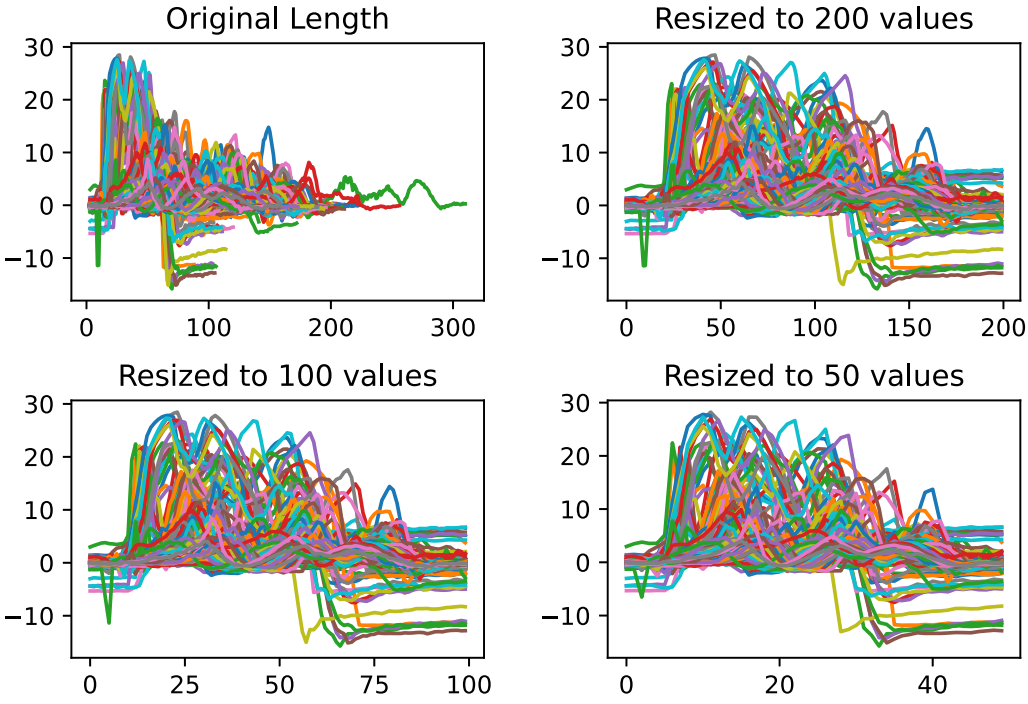
up gesture interpolation and resizing comparison



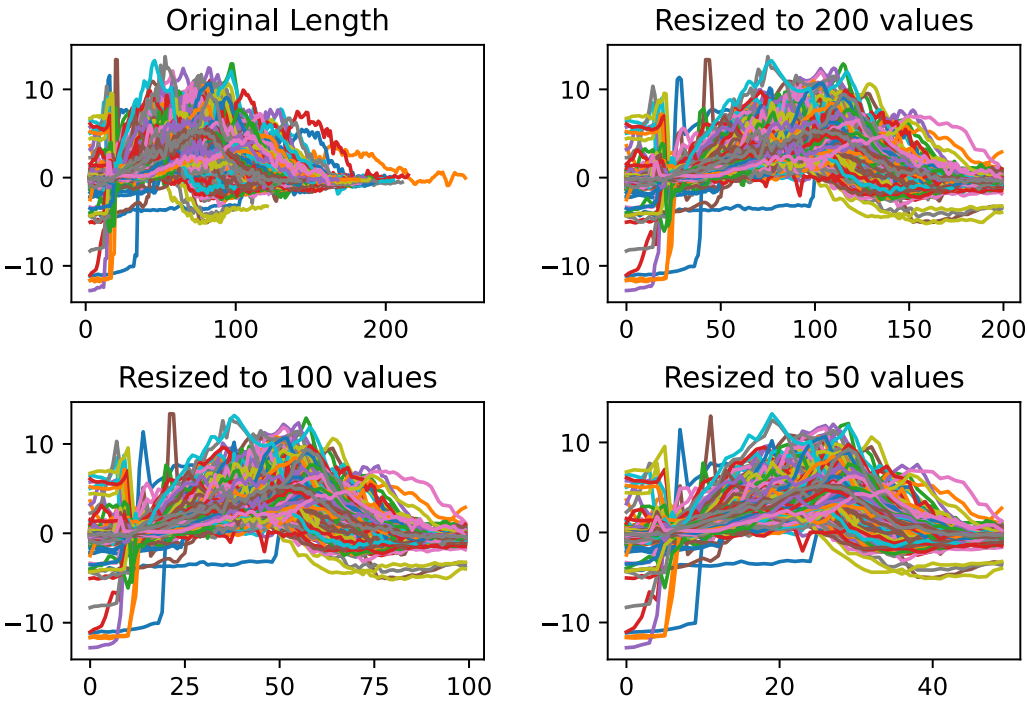
down gesture interpolation and resizing comparison



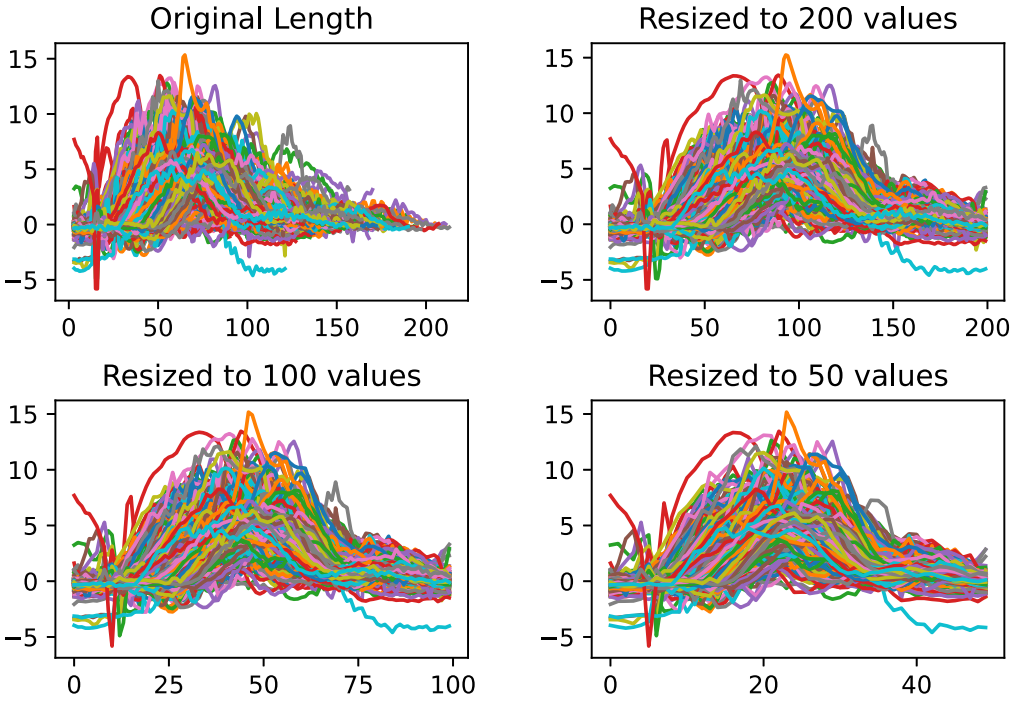

square gesture interpolation and resizing comparison
Original Length
Resized to 200 values
Resized to 100 values
Resized to 50 values


triangle gesture interpolation and resizing comparison
Original Length
Resized to 200 values
Resized to 100 values
Resized to 50 values

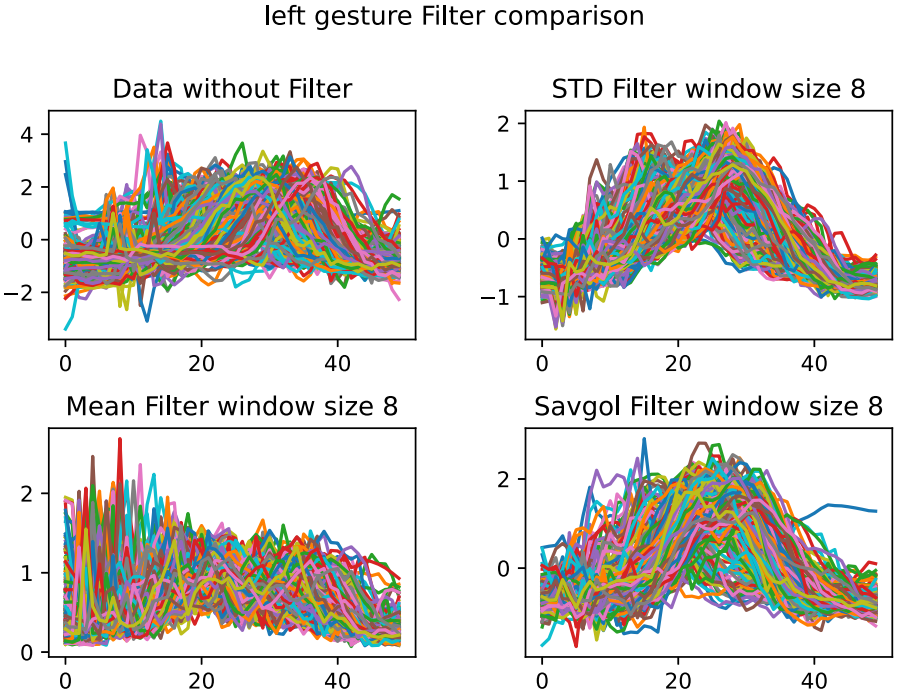
circleCw gesture interpolation and resizing comparison

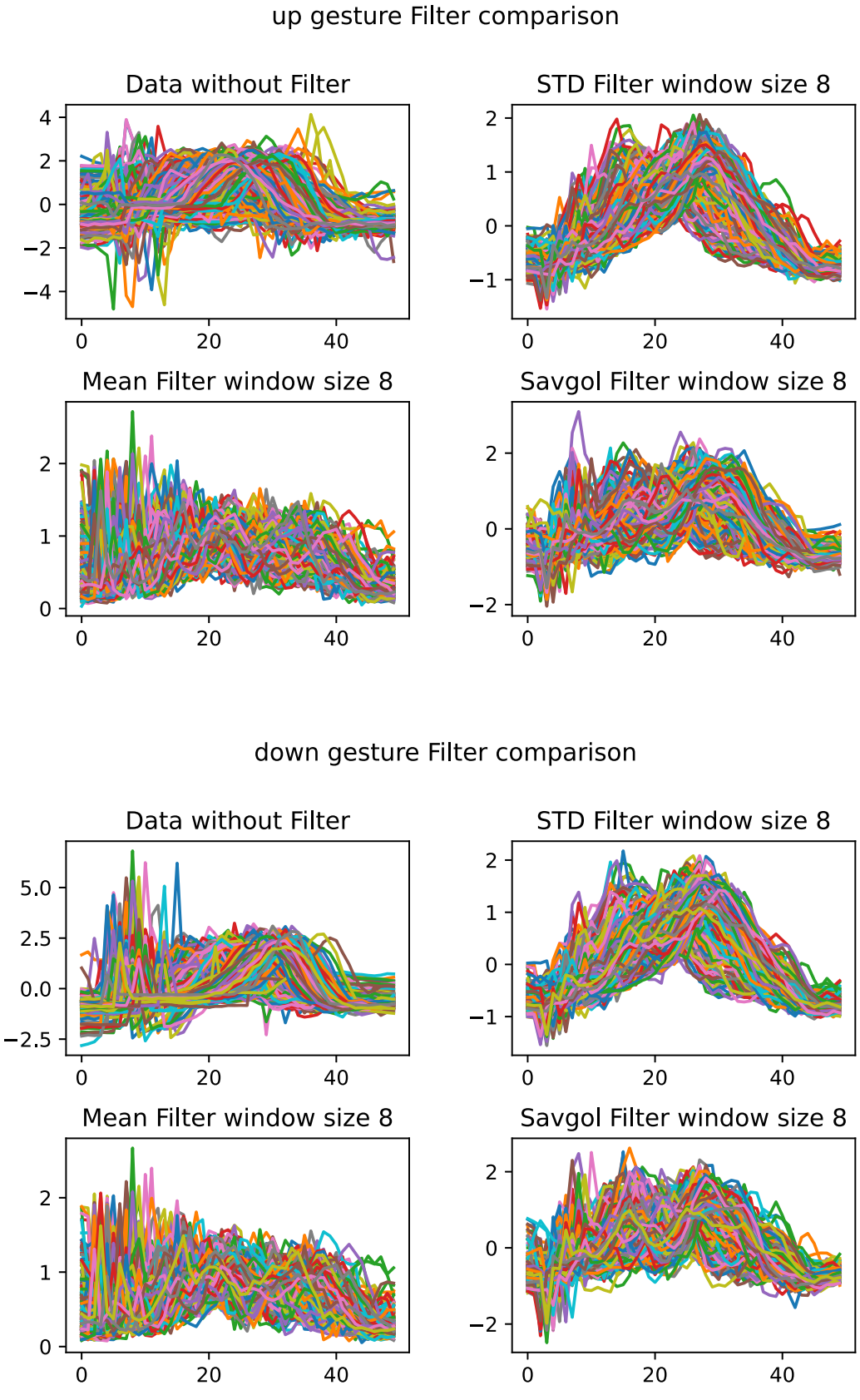


circleCcw gesture interpolation and resizing comparison

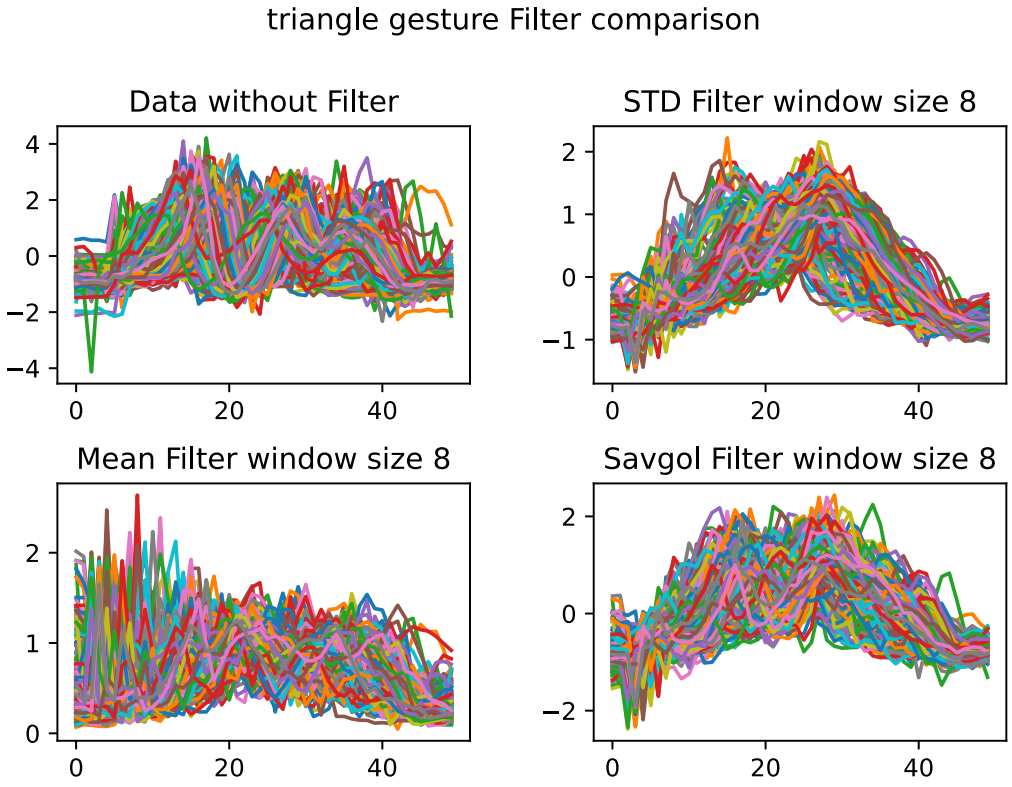


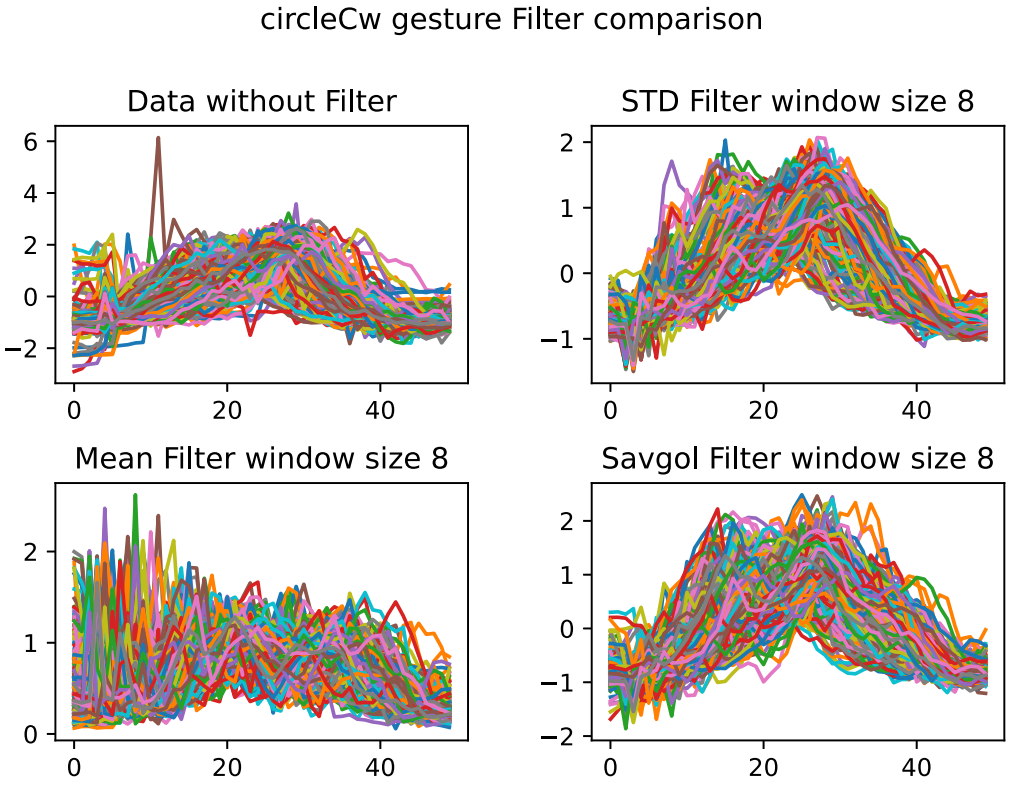
4.2. Appendix Filtering

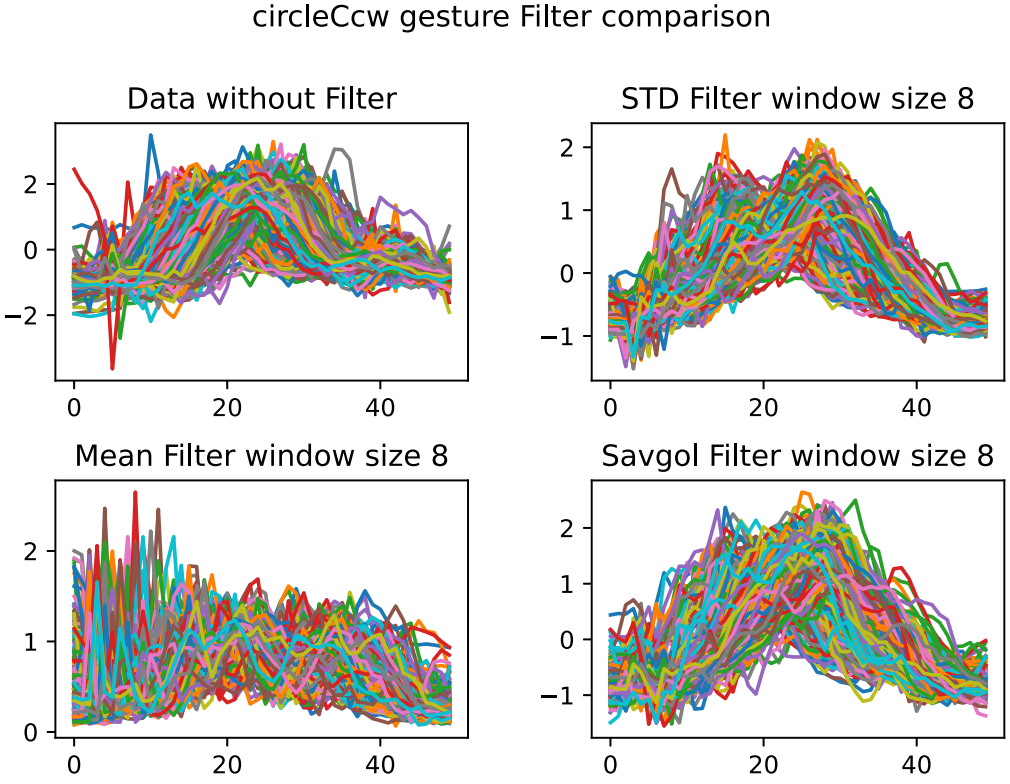



up gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8
down gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8


square gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8


triangle gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8


circleCw gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8


circleCcw gesture Filter comparison
Data without Filter
STD Filter window size 8
Mean Filter window size 8
Savgol Filter window size 8