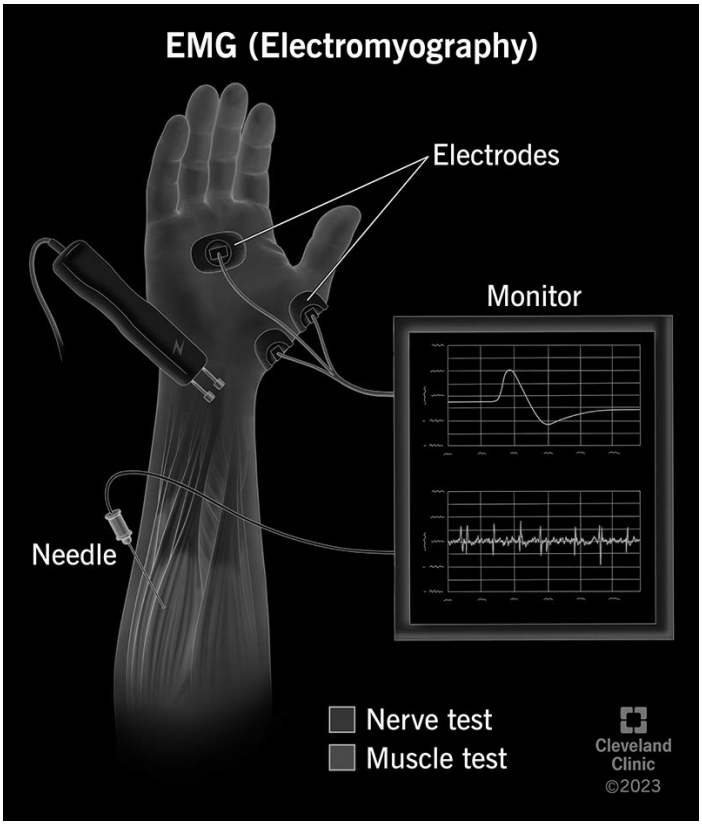
Amarildo Xhyra

Feature extraction and evaluation



Measurement of the differentiation power of features on 2 classes of EMG signals

Table of Contents

Introduction

The purpose of this practical work is to be familiar with feature extraction and analysis while using signal processing and statistical techniques. We are supplied with a database of 20 EMG signals during 5 seconds with a sampling frequency of 10 kHz, where the signals are separated in half into two different classes (EMG1-EMG10, EMG11-EMG20). The 20 EMG signals are provided as matlab files (EMG*.mat) where each signal has 50000 samples. In order to be able to properly handle the data we create a matrix called the emgMatrix which is a 20x50000 matrix that contains all of the provided signals.

With this data, we looked at different features, which are temporal, spectral, and statistical. For each feature, we evaluated it using boxplots, the Receiving Operator Characteristic (ROC) Curve, the Area Under the Curve (AUC), Max Youden's J Criterion, and other metrics like accuracy, precision, recall, and F1 score.

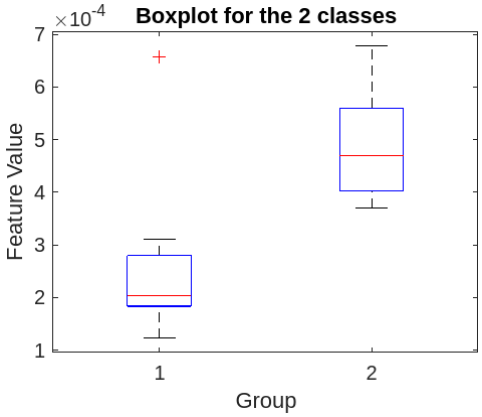
After getting the evaluations we analyze each feature and rank them based on the mentioned metrics, to see which features perform best with our data. This information helps us to understand our data better and to know which features to work on for future machine learning related tasks. The entire data handling, processing, feature calculation, and feature analysis is done in Matlab.

Discriminative Power Measurement

Boxplots

Boxplots, also known as whisker plots, are a graphical way of representing the distribution of data. To do this, certain important numbers are used: first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Visually this is represented as a rectangular box between the first and third quartile (the IQR), along with whiskers representing the variability of the data typically have a length of 1.5*the IQR. The median is represented as a line inside the rectangular box. All data points outside the rectangle and whiskers are considered outliers. The 5 most important metrics to look at when constructing a boxplot are the following ;

1. Minimum: The smallest data point, excluding outliers.
2. First Quartile (Q1): The median of the lower half of the data, representing the 25th percentile.
3. Median (Q2): The middle value of the dataset, indicating where half the values lie below and above.
4. Third Quartile (Q3): The median of the upper half of the data, representing the 75th percentile.
5. Maximum: The largest data point, excluding outliers.



IQR: The interquartile range is calculated as IQR = Q3 - Q1. This range provides insight into the variability of the central half of the data.
Outliers: Points that fall outside of 1.5 * IQR from the quartiles are considered outliers and are plotted as individual points.

Boxplots can be particularly useful for comparing distributions between two or more classes on a given feature. By visualizing the central tendency, spread, and potential overlap of distributions, boxplots provide a quick visual assessment of how distinct or similar classes are based on specific features. When comparing two classes A and B, we may assess these specific features of the boxplot :

- Median Values: The median values of the classes provide insight into the central tendency of each class.
- Spread and Overlap: The extent of the IQR and the presence of outliers reveal the variability and potential overlaps between classes. A significant overlap may suggest that the feature is not a good discriminator, while distinct boxes suggest strong discriminative potential.
- Outliers: The presence of outliers in one class and not in another can indicate anomalies or special cases that might inform the classification process.

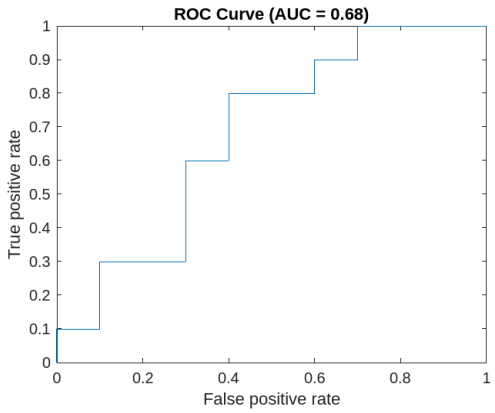
By analyzing boxplots, one can identify features that may be relevant for building predictive models and gain initial insights into the effectiveness of features for class discrimination.

Receiver Operating Characteristic (ROC) Curves

The ROC curve is a graphical representation of a binary classifier's performance across various threshold settings. The curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at different threshold levels.

$$\text{True Positive Rate (TPR)} \quad \textit{also called sensitivity/recall/hit rate} \quad = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR)} \quad \textit{also called fall out} \quad = \frac{FP}{N} = \frac{FP}{FP + TN}$$



Interpreting ROC Curve

- Curve Shape: An ROC curve closer to the top-left corner indicates a better performance, meaning higher true positive rates and lower false positive rates.

- Diagonal Line: A diagonal line from the bottom left to the top right represents random guessing. A model that performs better than random guessing will lie above this diagonal.

Area Under the Curve (AUC)

The area under the ROC curve (AUC) provides a single metric to evaluate the model's performance. It quantifies the overall ability of the model to discriminate between the positive and negative classes.

- AUC = 1: Perfect discrimination.
- AUC = 0.5: No discrimination (equivalent to random guessing).
- AUC < 0.5: Indicates that the model performs worse than random guessing.

The AUC is particularly useful because it summarizes the model performance across all possible thresholds, providing a comprehensive evaluation of its discriminative power.

Max Youden's J Criterion

Youden's J statistic is a measure used to assess the effectiveness of a diagnostic test or classifier. It is defined as:

$$J = \text{sensitivity} + \text{specificity} - 1 = \text{recall}_1 + \text{recall}_0 - 1$$

Youden's J statistic is a single value that takes into account both sensitivity and specificity, ranging from -1 to +1. A value of 0 signifies no discrimination, while a value of +1 indicates perfect discrimination. Identifying the maximum value of Youden's J across various thresholds helps in determining the optimal cutoff point for the classifier. This optimal threshold strikes a balance between sensitivity and specificity, enabling effective discrimination between classes.

Graphically on an ROC, (Receiver Operating Characteristic) curve, the Max Youden's J statistic represents the point where the sum of sensitivity and specificity is maximized. This point represents a balance of the optimal equilibrium between true positive rate (sensitivity) and false positive rate (1 - specificity). In essence, taking the graphical equivalent would be the threshold that maximizes the vertical distance from the diagonal (representing random chance), indicating the threshold with the best trade-off between true positives and false positives for the classifier.

Evaluating Classifier Performance

Once the optimal threshold is determined using Max Youden's J criterion, various performance metrics can be calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

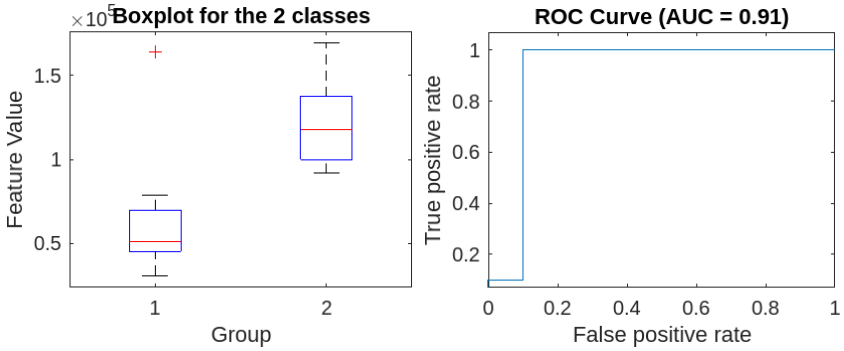
Feature Extraction

Temporal Features

1. Energy

Energy measures the total magnitude of the signal over time and provides insight into the overall activity level captured within the signal. In EMG analysis, energy helps quantify the total muscle activation during the recorded period. A higher energy value indicates greater overall muscle activity, while a lower energy value suggests reduced muscle activity. Energy can be calculated using the sum of the squared values of the signal, emphasizing larger signal amplitudes.

$$E = \sum_{n=-\infty}^{+\infty} |x(n)|^2$$



Boxplot Analysis: The boxplot illustrates notable differences in the distributions of values between the two groups. We can see that both groups are completely separate, apart from one outlier in class 1. This indicates that Energy is a good feature to discriminate over to obtain a good result.

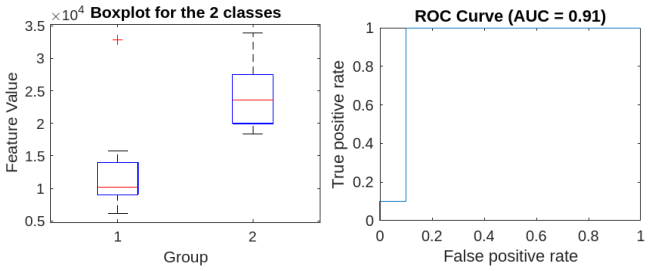
Performance Metrics: We observe very high performance metrics as expected (An AUC of 0.91 and Accuracy of 95%). This feature is very useful to separate both classes, the only mistake being the outlier in class 1.

Conclusion: Energy seems to be a promising discriminating feature. It allows separating both classes effectively, as indicated in the boxplot, thus bringing very high-performance metrics. The only reason the given performance is not perfect comes from an outlier energy-wise.

2. Power

Power in the context of signal analysis, especially in EMG (Electromyography), refers to the strength or intensity of the signal, providing insight into the energy carried by the signal per unit of time. It is a measure of the signal's magnitude squared, averaged over a given period. Power helps quantify the level of muscle activation, indicating how intense or forceful the muscle contractions are.

$$P = \lim_{N \to +\infty} \frac{1}{2N+1} \sum_{n=-\infty}^{+\infty} |x(n)|^2$$



Boxplot Analysis: Like for Energy, the boxplot illustrates notable differences in the distributions of values between the two groups. We can see that both groups are completely separate, apart from one outlier in class 1. This indicates that Power is a good feature to discriminate over to obtain a good result.

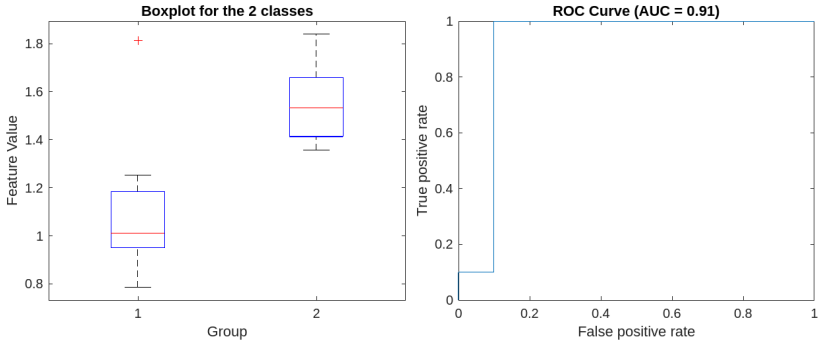
Performance Metrics: We observe very high performance metrics as expected (An AUC of 0.91 and Accuracy of 95%). This feature is very useful to separate both classes, the only mistake being the outlier in class 1.

Conclusion: Power seems to be a promising discriminating feature. It allows separating both classes effectively, as indicated in the boxplot, thus bringing very high performance metrics. The only reason the given performance is not perfect comes from an outlier power-wise.

3. Root Mean Square

The Root Mean square quantifies the intensity or power of signals like EMG signals which fluctuate in both positive and negative values. RMS provides a single positive value that reflects the overall strength of the signal over a given period. It is the square root of the signal's power at a specific time.

$$RMS(t_0) = \sqrt{P(t_0)}$$



Boxplot Analysis: Like for Energy and Power, from the boxplot we can see the notable differences in the distributions of values between the two groups. Both groups are separate, apart from one outlier in class 1. This indicates that just like Energy and Power, RMS is a good feature to discriminate over.

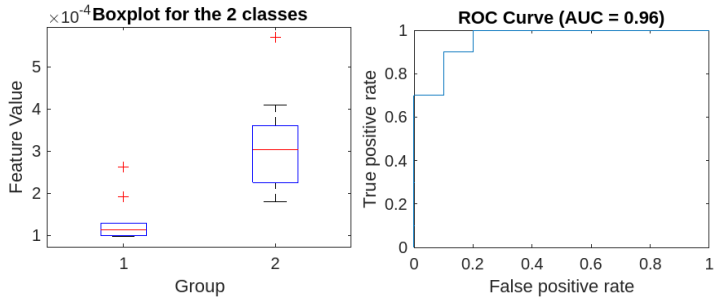
Performance Metrics: As expected from the graphs we notice very good performance metrics (AUC of 0.91 and Accuracy of 95%). This feature is very useful in separating both classes.

Conclusion: From the results we got RMS seems to be a promising discriminating feature. Both classes are separated effectively and the metrics show high performance.

4. Time Reversibility

Time Reversibility is a property where the statistical characteristics of the signal remain unchanged if the time axis is reversed. It is a measure of the asymmetry of a series under time reversal.So, if we reverse the time order of the signal, the resulting signal should have the same statistical behavior as the original.

$$Tr(\tau) = \frac{1}{N-\tau} \sum_{n=\tau+1}^{N} (S_n - S_{n-\tau})^3$$



Boxplot Analysis: Like for Energy, Power and RMS from the boxplot we can see the notable differences in the distributions of values between the two groups. Both groups are completely separated except the 2 outliers in the first group which overlap with the second. This indicates that Time Reversibility is a good feature.

Performance Metrics: As expected from the graphs we notice very good performance metrics (AUC of 0.96 and Accuracy of 90%). This feature is very useful in separating both classes.

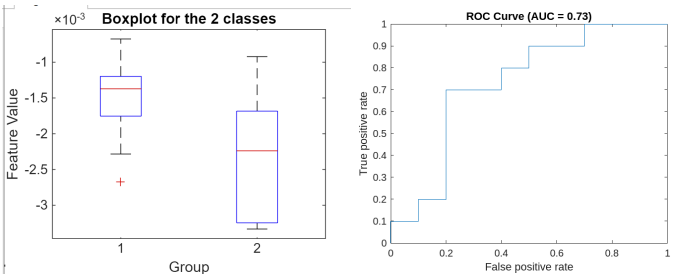
Conclusion: From the results we got Time Reversibility also seems to be a promising discriminating feature. Both classes are separated effectively and the metrics show high performance.

Statistical Features

1. Mean Values

The mean is a measure of central tendency, representing the average value of the EMG signal. It helps provide a baseline for the overall amplitude of the EMG data. A higher mean value indicates a shift towards higher muscle activity, while a lower mean suggests reduced muscle activity. To calculate the mean, we use the following formula:

$$\text{mean}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$$



Boxplot Analysis: Group 1 displays a more concentrated distribution and has less variability in its mean values while Group 2 reflects greater variability in mean values. The outlier below the lower whisker overlaps with the IQR of Group 2. When taking into account the box whiskers there is significant overlap, indicating they may be difficult to differentiate.

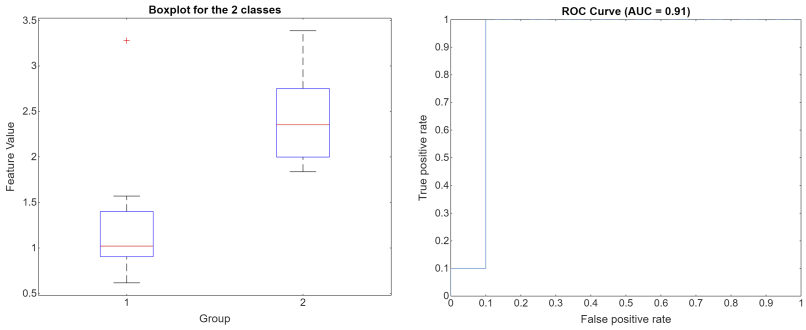
Performance Metrics: The AUC score of 0.73 is quite good. This is probably due to the higher concentration of the first class. Mean seems to be a satisfying classifier.

Conclusion: While the mean values show some differences between the two groups, particularly in terms of spread and central tendency, they do not provide enough discriminatory power for ideal classification. The good AUC score and accuracy highlights the mean values as a quite satisfactory feature.

2. Variance

Variance is a statistical measure that quantifies the degree of spread or dispersion in a set of data points, how each individual data point deviates from the mean of the dataset.Variance in a set can be calculated by the following formula:

$$\text{var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^2$$



Boxplot Analysis: Group 1's variance values are tightly clustered.There is one outlier above the upper whisker, suggesting an irregularity in Group 1's data. Group 2 reflects greater variation in its variance values and has a right-skewed distribution. Apart from the outlier, both boxes seem to be very separable, indicating the feature is very good for classification.

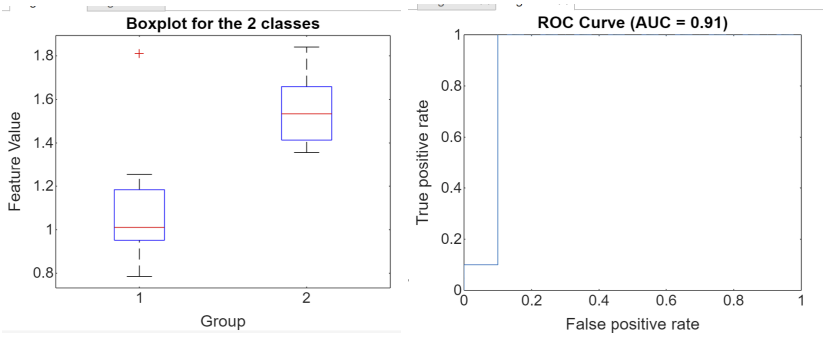
Performance Metrics: The model's accuracy,shows that 95% of the instances were correctly classified using variance as a feature. This indicates that the model is highly effective at distinguishing between the two groups based on variance. 90.91% of the positive predictions made by the model were correct, meaning it has very few false positive errors.

Conclusion: The variance feature is a highly effective discriminator between the two groups. The excellent classification metrics (high accuracy, precision, recall, and F1 score) suggest that the model performs very well when using variance to separate the groups.

3. Standard Deviation

Standard deviation measures the amount of variation or dispersion in the EMG signal and is derived from the variance. The standard deviation is calculated as :

$$\mathrm{std}(x) = \sqrt{\mathrm{var}(x)}$$



Boxplot Analysis: The boxplot reveals a clear distinction between the two groups based on their standard deviation values. Separation of the two groups, based on standard deviation, points to its effectiveness as a distinguishing feature between them.

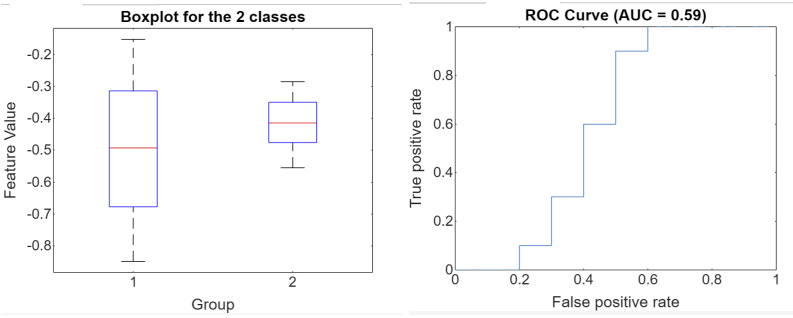
Performance Metrics: The model achieves excellent accuracy. 95% of cases are classified correctly using standard deviation. Of all positive predictions, 90.91% were correct, indicating that the model makes very few false positive errors when using standard deviation.

Conclusion: Standard deviation is an effective feature for distinguishing between the two groups. The high classification metrics (accuracy, precision, recall, and F1 score) confirm that standard deviation plays a significant role in separating these groups in the EMG signal analysis.

4. Skewness

Skewness is a measure of the asymmetry of a distribution. A positive skewness suggests that the tail of the distribution extends towards higher values, indicating a prevalence of higher muscle activity levels. On the other side, negative skewness indicates a predominance of lower muscle activity levels. To calculate skewness we consider the below formula:

$$\text{skew}(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^2\right)^{3/2}}$$



Boxplot Analysis: The boxplot reveals differences in the skewness values between groups. Group 1 indicates greater variability in the skewness of its data distribution. Group 2 exhibits a more compact range. While skewness might provide some discriminatory power, it is not as strong as other features like variance or standard deviation.

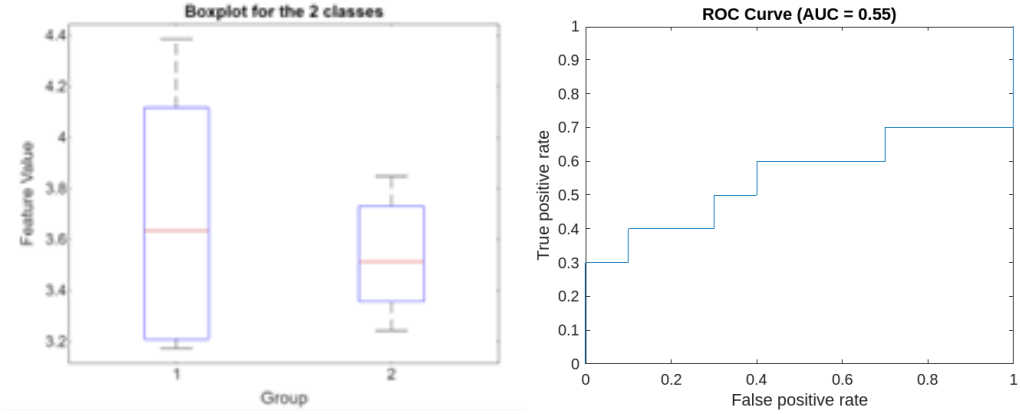
Performance Metrics: The model achieves 70% accuracy using skewness,Out of all positive predictions, 64.29% were correct, indicating that the model makes more false positive errors when using skewness.

Conclusion: The skewness feature provides moderate classification performance, offering some predictive value but not as strong as variance or standard deviation. The model has high recall, meaning it identifies most positive cases, but its lower precision suggests a tendency to misclassify negatives as positives.

5. Kurtosis

Kurtosis measures the occurrence of outliers in a distribution. In the context of EMG analysis, kurtosis can provide insights into the shape and nature of muscle contractions.To calculate we have considered the below formula:

$$\text{kurt}(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^2\right)^2}$$



Boxplot Analysis: Group 1 values indicate greater variability while Group 2 indicates less variability.. Although the distributions are slightly different, we don't have a clear distinction between the two groups. Kurtosis doesn't seem to be a useful feature.

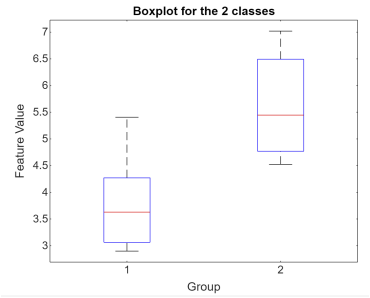
Performance Metrics: The AUC is not ideal and the model achieves 60% accuracy using kurtosis as a feature, which shows relatively bad discrimination power..

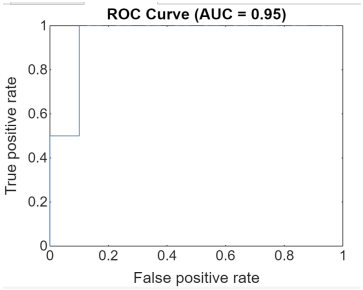
Conclusion: Kurtosis provides moderately bad classification performance with a high recall but lower precision. While kurtosis captures important characteristics related to the peakiness or tail behavior of the data, it is not as strong an individual predictor.

6. Peak Amplitude

The peak amplitude represents the maximum absolute value in each EMG signal. It is a useful metric for identifying the highest level of muscle activation within the signal.The peak amplitude is calculated by finding the maximum absolute value of the signal:

$$\text{peak\_amplitude}(x) = \max |x_i|$$





Boxplot Analysis: We have clear distinction in peak amplitude between the two groups. Group 1 values reflect less variability, peak amplitude values are tightly clustered around the median, suggesting a minor left skew. Group 2 values indicate a right-skewed distribution.

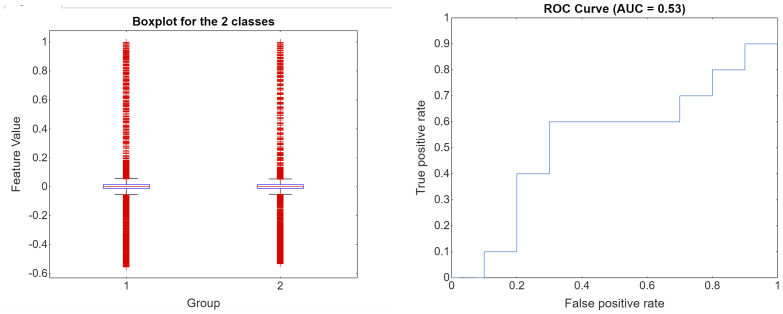
Performance Metrics: The model correctly classifies 95% of instances using peak amplitude as the feature. The precision of 90.91% indicates that when the model predicts a positive class, it is correct 90.91% of the time, demonstrating very few false positives.

Conclusion: The peak amplitude feature is a highly effective predictor for distinguishing between the two groups, evidenced by excellent performance metrics (high accuracy, precision, recall, and F1 score). Its strong ability to differentiate the groups suggests it is a valuable feature to include in the classification model.

7. Autocorrelation

Autocorrelation measures the similarity between a signal and a lagged version of itself. In EMG analysis, it helps identify repetitive patterns or periodicity in muscle activity. The autocorrelation function is defined as:

$$\text{Autocorr}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i x_{i+\tau}$$



Boxplot Analysis: Both Group 1 and Group 2 exhibit similar distributions of autocorrelation values, reflecting minimal differences in central tendency. The presence of numerous outliers, extending above and below the whiskers, highlights significant deviations in the autocorrelation values for both groups. Numerous outliers may indicate irregularities or noise in the data.

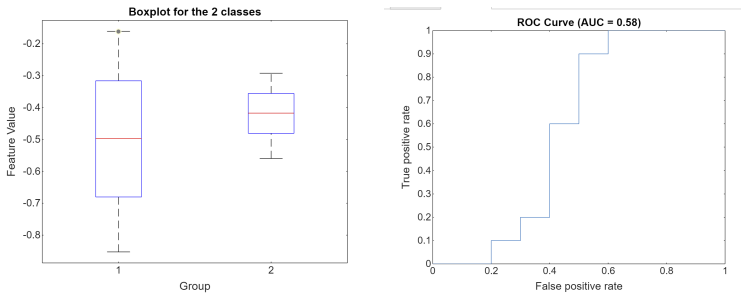
Performance Metrics: The model achieves a classification accuracy of 65% when using autocorrelation as a feature, which is significantly lower than the other features such as peak amplitude, variance, and standard deviation. With a precision of 66.67%, about two-thirds of the positive predictions are correct, indicating a notable number of false positives.

Conclusion: The autocorrelation feature demonstrates moderate classification performance but is less effective compared to other features. While it provides some insights into the data, its predictive power is limited on its own.

8. Moment Coefficient of Skewness

The moment coefficient of skewness is an alternative method to calculate skewness using moments of the distribution. This metric gives another perspective on the asymmetry of the EMG signal's distribution. It is calculated as:

$$\text{Moment Skewness} = \frac{E[X^3]}{(E[X^2])^{3/2}}$$



Boxplot Analysis: Group 1 exhibits more variability, values for this group are more dispersed and Group 2 demonstrates more consistent and tighter skewness values. The shorter whiskers reflect a smaller range of skewness values, highlighting a clear distinction between the groups.

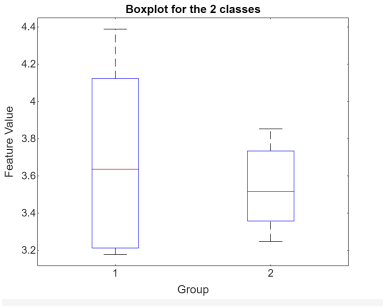
Performance Metrics: The model correctly classifies 70% of instances, consistent with the performance of skewness and kurtosis. Precision indicates that about 64.29% of positive predictions are correct, suggesting a notable number of false positives.

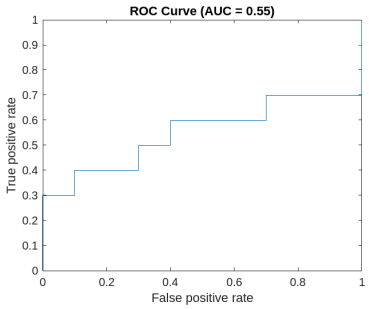
Conclusion: The feature provides moderate classification performance. While it effectively identifies a high percentage of true positive cases (high recall), it also exhibits lower precision, indicating the presence of a fair amount of false positives.

9. Moment Coefficient of Kurtosis

The moment coefficient of kurtosis is an alternative way to compute kurtosis using moments. This measure helps assess the "peakiness" and tail properties of the signal's distribution. It is calculated using the following formula:

$$\text{Moment Kurtosis} = \frac{E[X^4]}{(E[X^2])^2}$$





Boxplot Analysis: Group 2's kurtosis values are more concentrated, indicating less variability. Group 1 shows greater fluctuations in the peakness or tail behavior of the distribution, while Group 2 exhibits a more uniform distribution of values. Both groups have significant overlap.

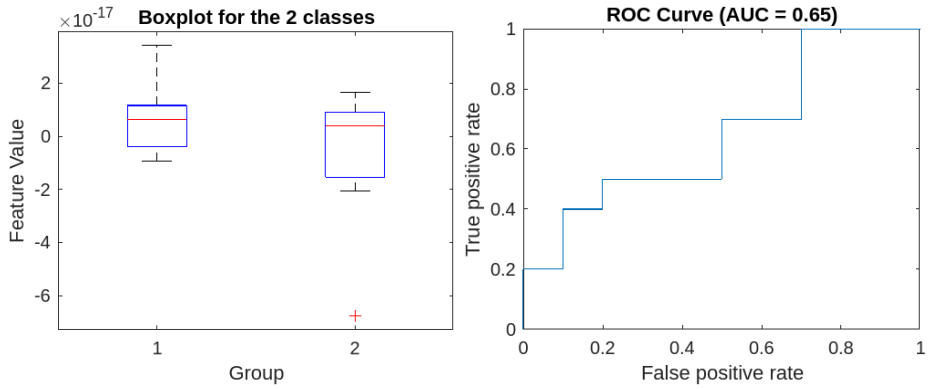
Performance Metrics: The model correctly classifies 60% of instances, which is low compared to stronger features like peak amplitude and standard deviation. 55.55% of the positive predictions are correct (precision). Generally, the performance of this feature is low.

Conclusion: The feature exhibits moderate classification performance. Its perfect recall indicates effectiveness in identifying all actual positive instances, but the low accuracy and precision suggest that it may produce many false positives.

10. Z-Score

The Z-Score is a statistical feature measured in terms of standard deviations from the mean, which we can see from the formula below. It is used to normalize the EMG signal by scaling it relative to its mean and standard deviation. This feature can be used to illustrate the nonlinearity of a signal.

$$\text{Z-Score} = \left| \frac{Tr_{org} - mean(Tr_{surr})}{std(Tr_{surr})} \right|$$



Boxplot Analysis: From the boxplot we cannot see notable differences in the distributions of values between the two groups the image we can see that the first group overlaps with the second, apart from one outlier in class 1. We can see that the two groups tend to overlap. This indicates that it is not the best feature to discriminate over.

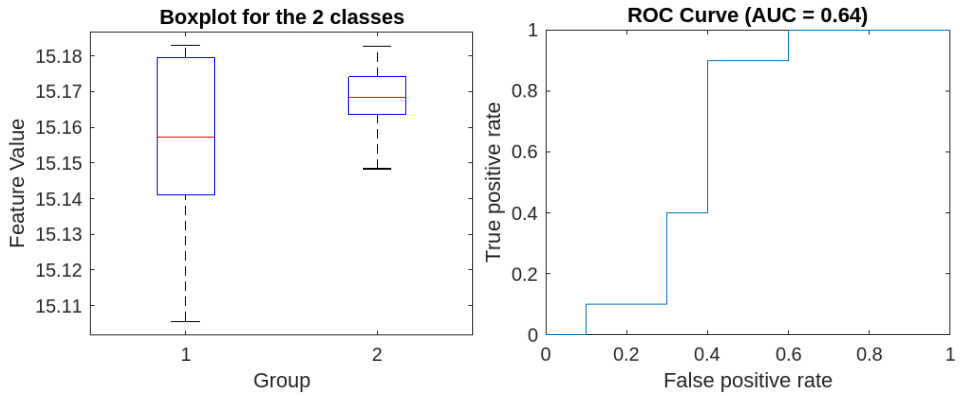
Performance Metrics: As expected from the graphs we notice not so great performance metrics (AUC of 0.65 and Accuracy of 60%). This feature acts just slightly better than random guessing.

Conclusion: From the results we got the Z-Score does perform a bit better than random guessing but it is still not the best feature we could have.

11. Entropy

Entropy is a measure that quantifies the level of uncertainty or information on average of a signal. This feature can provide information regarding underlying patterns and the complexity of the signal.

$$I_X = -\sum_{i=1}^{M} p_i \log p_i$$



Boxplot Analysis: For this feature, we can see from the boxplot that the two groups are not well separated, as group 2 overlaps with the first one. This shows that Entropy is not a good feature to discriminate over if we want to obtain a good result.

Performance Metrics: Based on the results from the boxplot we can expect the performance metrics to not be as high (AUC of 0.64 and Accuracy of 75%). With these metrics, the feature is closer to random guessing than a good separation of the groups.

Conclusion: What we were able to conclude from the boxplot and metrics is that Entropy is not a promising feature, as it does not allow for an effective separation of both classes.

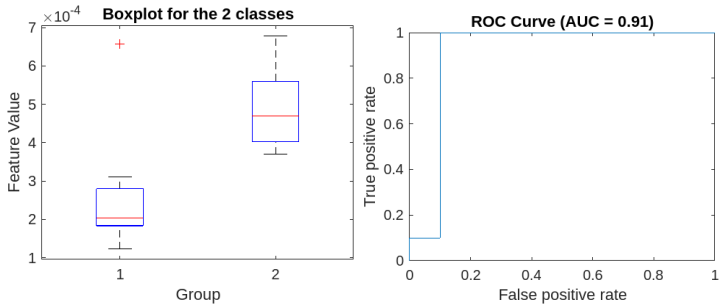
Spectral Features

Power Spectral Density (PSD)

1. Average periodogram PSD

The mean is a measure of central tendency, representing the average value of the EMG signal. It helps provide a baseline for the overall PSD power. A higher average value indicates a shift towards higher activity per frequency, while a lower average suggests reduced muscle activity. To calculate the average, we use the following formula:

$$\text{mean}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$



Boxplot Analysis: Similar to the power feature, we obtain two very distinct classes in our boxplot. We only observe an outlier in class 1 that will potentially be misclassified as class 2 due to its very high average PSD.

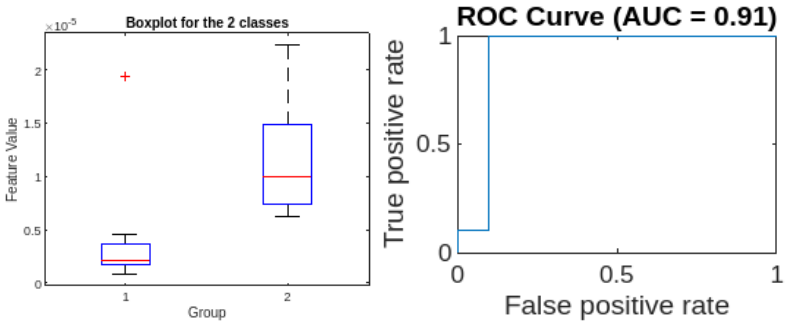
Performance Metrics: We observe very high performance metrics as expected (An AUC of 0.91 and Accuracy of 95%). This feature is very useful to separate both classes, the only mistake being the outlier in class 1.

Conclusion: Just like Power, Average PSD seems to be a promising discriminating feature. It allows separating both classes effectively, as indicated in the boxplot, thus bringing very high-performance metrics. The only reason the given performance is not perfect comes from an outlier.

2. Variance of PSD

As usual, variance is a statistical measure that quantifies the degree of spread or dispersion in a set of data points. Through it, we can tell how each data point deviates from the mean(average) of the dataset. We used this measure to show how far on average are the data points on a measurement from the mean of the PSD. Below is the formula for the calculation:

$$\mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mathrm{mean}(x))^2$$



Boxplot Analysis: Again, similar to before, we obtain two very distinct classes in our boxplot. We only observe an outlier in class 1 that will potentially be misclassified as class 2 due to its very high average PSD.

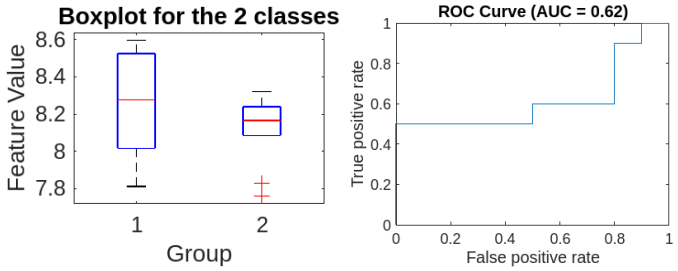
Performance Metrics: We observe very high performance metrics as expected (An AUC of 0.91 and Accuracy of 95%). This feature is very useful to separate both classes, the only mistake being the outlier in class 1.

Conclusion: As before, the variance of PSD seems to be a promising discriminating feature. It allows separating both classes effectively, as indicated in the boxplot, thus bringing very high performance metrics. The only reason the given performance is not perfect comes from an outlier.

3. Skewness of PSD

Skewness is a measure of the asymmetry of a distribution. Skewness measures the asymmetry of the EMG signal's probability distribution. A positive skewness suggests that the tail of the distribution extends towards higher values, indicating a prevalence of higher frequency activity. On the other side, negative skewness indicates a predominance of lower muscle activity frequency-wise. To calculate skewness we consider the below formula:

$$\text{skew}(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^2\right)^{3/2}}$$



Boxplot Analysis: Completely opposite to the other PSD measures, here we obtain two classes that aren't distinct at all. The first class has a much higher IQR, indicating that this measure is less precise in describing the first class. We observe that for this reason, the second class values are included inside the first class IQR interval, not allowing differentiation.

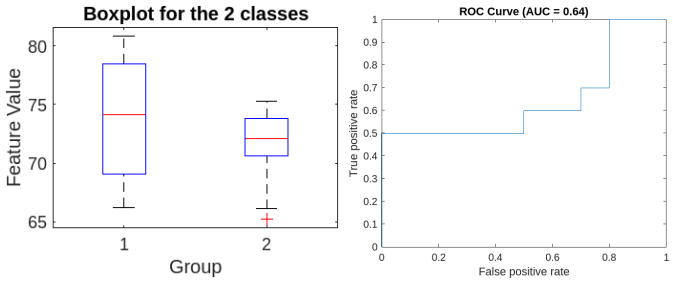
Performance Metrics: Difficult differentiation observed in the boxplots explains very bad performance results. An AUC of 0.62 indicates that our classifier is performing badly, although an accuracy of 0.7 is satisfying

Conclusion: Opposite to what we would like, using skewness as a discrimination feature leads to poor results as both classes are not different enough in this regard.

4. Kurtosis of PSD

Kurtosis measures the occurrence of outliers in a distribution. In the context of EMG analysis, kurtosis can provide insights into the shape and nature of muscle contractions. High kurtosis values suggest sharp, intense contractions, while low kurtosis values indicate smoother, more sustained contractions. To calculate we have considered the below formula:

$$\text{kurt}(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \text{mean}(x))^2\right)^2}$$



Boxplot Analysis: Again, here we obtain two classes that aren't really distinct. The first class has a much higher IQR, indicating that this measure is less precise in describing the first class. We observe that for this reason , the second class values are included inside the first class IQR interval, not allowing differentiation.

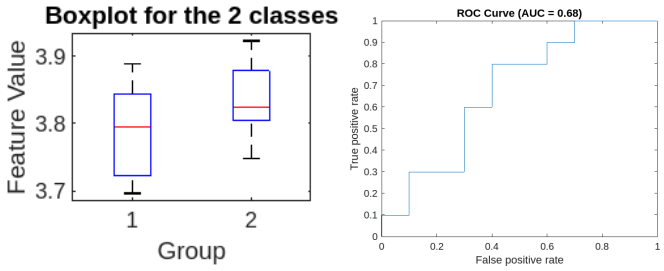
Performance Metrics: Difficult differentiation observed in the boxplots explains disappointing performance results. An AUC of 0.64 and accuracy of 0.7 indicates our classifier is not ideal.

Conclusion: Opposite to other PSD features, using kurtosis as an discrimination feature leads to very poor results as both classes are not different enough in this regard.

5. Entropy of PSD

Entropy is a statistical measure that measures the uncertainty or randomness in a distribution. In the context of EMG analysis, entropy can provide valuable insights into the complexity and variability of muscle contractions. Based on Shannon's information theory, higher entropy values suggest an uncertain function, indicating a possibly more complex task. Conversely, lower entropy values indicate more predictable patterns, which may suggest a more stable or controlled activity. To calculate entropy in this context, we use the following formula:

$$\begin{aligned} \mathrm{H}(X) &= -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \\ &= -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} \\ &= -\sum_{i=1}^{2} \frac{1}{2} \cdot (-1) = 1. \end{aligned}$$



Boxplot Analysis: Slightly differently here, we obtain two classes that have IQR interval overlap but not inclusion. This allows to differentiate slightly both classes, indicating we should obtain better performance but not excellent performance.

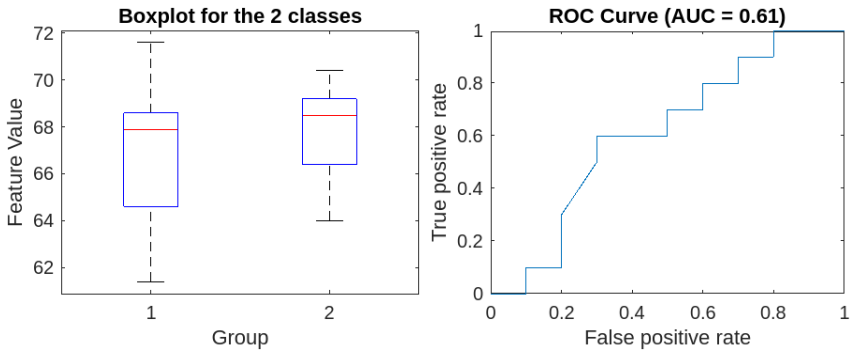
Performance Metrics: As supposed, we obtain acceptable performance but not great. Differentiation is difficult with many errors happening, thus resulting in an AUC of 0.68 and an accuracy of 70%.

Conclusion: This feature is slightly differentiating, but not enough to be our center of focus when choosing a way to classify our EMG signals.

6. Median Frequency

The Median Frequency as a feature is the frequency at which the cumulative power spectrum of an EMG signal equals 50%. It's the frequency that divides the power spectrum in half, with 50% of the total power located above and 50% below this frequency.

$$\int_{0}^{fmed} PSD(f)df = \int_{fmed}^{f\max} PSD(f)df$$



Boxplot Analysis: We can see from the boxplot that the two groups are not well separated, as there is some overlap between the 2 of them.

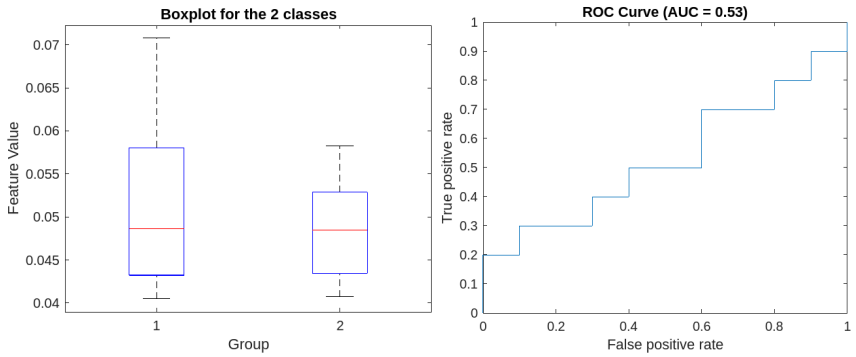
Performance Metrics: For the performance metrics the Median Frequency shows an AUC of 0.61 and Accuracy of 65%. With these metrics the feature is closer to random guessing than a good separation of the groups.

Conclusion: From the results obtained we can see that the Median Frequency is not a promising feature that allows for good separation, because of the low separation of the groups and the low performance metrics.

7. Relative Energy per Frequency Band

The Relative Energy per Frequency Band as a feature is a percentage of the EMG signal's overall energy that represents the amount of energy (or power) present in particular frequency bands of the signal. It facilitates the analysis of how much of the signal's energy is concentrated in specific frequency bands.

$$W_n = \frac{\int_{f_{n-1}}^{f_n} DSP(f)df}{M_0} \quad \text{with } f_n=(n/N)fmax \text{ and } N \geq n \geq 1$$



Boxplot Analysis: Same as for the Median Frequency we can see from the boxplot that the two groups are not well separated, as there is much overlap between the 2 groups. In this case group 2 fully overlaps with group 1.

Performance Metrics: For the performance metrics the REFB shows an AUC of 0.53 and Accuracy of 55%. With these metrics the feature is closer to random guessing than a good separation of the groups.

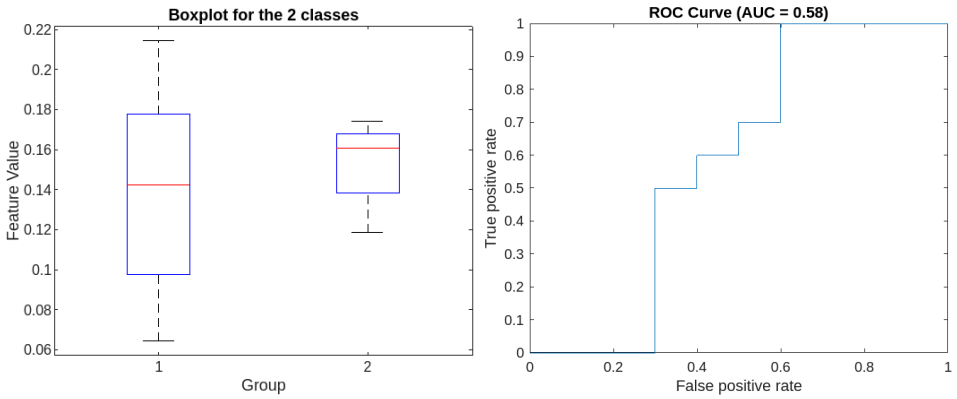
Conclusion: From the results obtained we can see that the REFB is not e great feature that does not allow for good separation, because of the low performance metrics.

8. High/Low Ratio

The H/L Ratio is a feature that provides a simple way to quantify the balance between the high-frequency and low-frequency components of an EMG signal. It compares the energy or power in high-frequency bands to that in low-frequency bands of the EMG signal.

$$\frac{H}{L} = \frac{\int_{fH1}^{fH2} PSD(f)df}{\int_{fL1}^{fL2} PSD(f)df}$$

with L = [L1,L2] and H = [H1,H2]



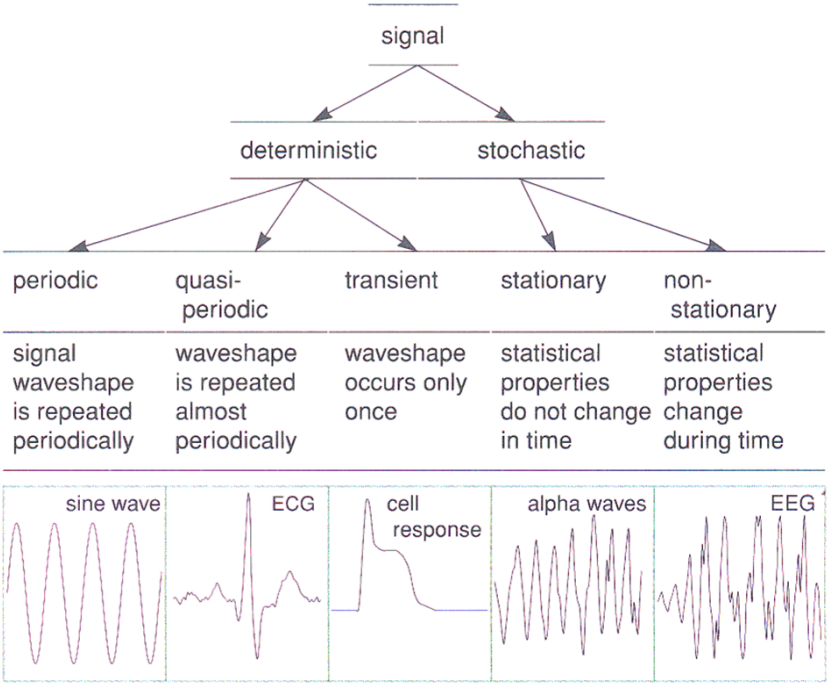
Boxplot Analysis: For this feature we can see from the boxplot that the two groups are not well separated, as group 2 overlaps with the first one. This shows that the H/L Ratio is not a good feature because of the low separability.

Performance Metrics: For the performance metrics the H/L Ratio shows an AUC of 0.58 and Accuracy of 70%. With these metrics the feature is closer to random guessing than a good separation of the groups.

Conclusion: We can see that the H/L Ratio is not a promising feature that allows for good separation, because of the low performance metrics.

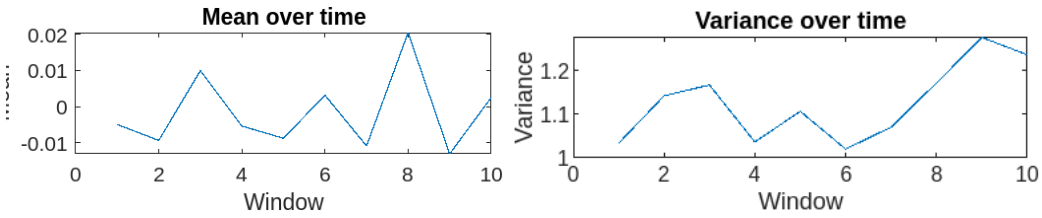
Potential advancements

An important aspect that was taken into account in this study was the properties of the signal.The properties of a signal guide the study into which features might be most representative, and thus best at differentiating it from other signals.



Properties of a signal

Through careful analysis of the signal and its evolution in time, we've found early on that our signals are not deterministic, i.e presents some randomness, by plotting them through time. Another important differentiation point would be the stationarity or not of our signal, indicating whether or not time is an important factor to consider when studying our signals. For that reason, we studied the evolution of different statistical properties (mean, variance,skewness, kurtosis etc.) over different windows of time and have come to the conclusion that our signal is non-stationary.



For this reason, some spectral features may be more adapted than others. We've decided to apply Power Spectral Density (PSD) which is purely a frequency based representation of our signal. Power Spectral Density (PSD) is more adapted to stationary signals. The PSD represents how the power of a signal is distributed across different frequencies. It assumes that the signal's statistical properties, such as its mean and variance remain constant over time, which is characteristic of stationary signals.

The fact that PSD is not the most adapted to explaining the signals prompted us to examine the wavelet transform. Through seeing the evolution of the statistical properties as represented above, we found that the wavelet transform is very well adapted for non-stationary signals because it gives a way to analyze both time and frequency components of the signal simultaneously, something that can't easily be done with classical Fourier transforms. The usefulness of this property in analyzing non-stationary signals, such as EMG, where the frequency content can vary over time due to changing contractions of different muscles and their respective movements, is very important.

This is an important advantage in the use of the wavelet transform. It can decompose the EMG signal into different frequency bands at various time intervals. For instance, higher frequency parts may dominate during a rapid muscle contraction, indicating a spike of muscle activity, while importance of lower frequencies indicate sustained or slower contractions.Wavelet transform is, in a way, a multi frequency time analysis of the signal, allowing for the examination of both precise and general changes in the EMG signal. This is particularly useful when dealing with different speeds of changes in signal activity, where other techniques might not be adapted. For this reason, it facilitates the extraction of features relevant for classification tasks such as distinguishing between different types of muscle contractions. Just as we've done with PSD, by utilizing the time-frequency representation obtained from the wavelet transform, one can also obtain classical characteristics (energy, entropy, and frequency bandwidth) that serve as informative features for machine learning classifiers.

Unfortunately, we weren't able to pursue this path as the wavelet transform is a very complex and expensive transformation, augmenting drastically our dimension of representation (as a new dimension is added for each wavelet). This complex operation requires the Wavelet toolbox library which is unfortunately inaccessible in the free version of Matlab. Similarly, one is required to build more complex classifier models (such as KNN) to differentiate our signals in this larger space. Satisfying results had already been obtained and these complex classifiers were not yet seen in our coursework, but wavelet transforms would be a path we would pursue.

There exists many other features and transformations to analyzing signals, including techniques like bilinear transforms, empirical mode decomposition, and the short-time Fourier transform. As

researchers worldwide explore new methods of signal analysis it was incredibly engaging to learn about the field during the ISCE course, with it offering only a glimpse into a much more fascinating area of study.

Conclusion

The conclusion of this report highlights the effectiveness of various feature extraction techniques and their evaluation in distinguishing between two classes of EMG signals.The analysis of temporal, statistical as well as spectral features made it possible to highlight a number of features that significantly contribute to classification accuracy, which in practice improved the classification. Parameters such as peak amplitude, energy or power demonstrated high discriminative power, achieving impressive performances at close to 95 % ranges in accuracy. These characteristics provided clear separations, with precision that was reliably supported by high values of AUC and precision scores.

On the other hand, features such as skewness, kurtosis, or even entropy but to a lesser extent relative to the other measures registering lower performances on accuracy and AUC metrics. The analysis also pointed out that there are indeed other spectral features, such as Power Spectral Density, which are useful depending on the specific measure used, such as average PSD versus skewness of PSD..

We believe that these results are strongly indicative of the nature of EMGs and what they measure. Electromyographies (EMGs) measure muscle response to a nerve's stimulation of the muscle. As we may imagine, it is essential for a muscle to respond quickly and with significant power. This idea justifies why most differentiating features are related to measures of quick and significant increases in a signal (Peak Amplitude, Power, Average PSD etc.)

The study demonstrates the importance of choosing the correct features for signal processing when doing classification, showing that not all features contribute equally to a model’s performance. For future advancements we could potentially explore more complex methods that were beyond the scope of this project due to resource constraints.

In conclusion, this practical work has provided a solid discovery of feature extraction and evaluation in signal processing, demonstrating the critical role of feature selection in enhancing the quality of model classification.

Annex

This table resumes the performance results of all presented features. It is ordered in descending accuracy.

Feature	Accuracy	Precision	Recall	F1 Score	AUC
Peak Amplitude	0.95	0.90909	1	0.95238	0.95
Energy	0.95	0.90909	1	0.95238	0.91
Power	0.95	0.90909	1	0.95238	0.91
Variance	0.95	0.90909	1	0.95238	0.91
Standard Deviation	0.95	0.90909	1	0.95238	0.91
Root Mean Square	0.95	0.90909	1	0.95238	0.91
Average PSD	0.95	0.90909	1	0.95238	0.91
Variance PSD	0.95	0.90909	1	0.95238	0.91
Time Reversibility	0.9	0.9	0.9	0.9	0.96
Entropy	0.75	0.69231	0.9	0.78261	0.64
H/L Ratio	0.70	0.625	1	0.76923	0.58
Mean	0.7	0.66667	0.8	0.72727	0.73
Kurtosis PSD	0.7	0.625	1	0.76923	0.64
Skewness	0.7	0.64286	0.9	0.75	0.59
Moment Coefficient of Skewness	0.7	0.64286	0.9	0.75	0.58
Skewness PSD	0.7	0.625	1	0.76923	0.62
Entropy PSD	0.7	0.66667	0.8	0.72727	0.68
Moment Coefficient of Kurtosis	0.6	0.55556	1	0.71429	0.55
Autocorrelation	0.65	0.66667	0.6	0.63158	0.53

Median Frequency	0.65	0.66667	0.6	0.63158	0.61
Kurtosis	0.6	0.55556	1	0.71429	0.55
Z-Score Mean	0.6	0.5625	0.9	0.69231	0.65
Relative energy per frequency band	0.55	0.52632	1	0.68966	0.53