ACTIVITY RECOGNITION PROJECT

ASSIGNMENT 1

CSE 572: Introduction to Data Mining (SPRING 2019)

PROF. AYAN BANERJEE
I.R.A. FULTON SCHOOL OF ENGINEERING
ARIZONA STATE UNIVERSITY

GROUP – 25

PARV SETHI (1213445907)
VINIT GIRISHBHAI SHETH (1215128980)
SUMAN PARLAPALLI (1213698107)
THARUN CHINTHAM (1213185634)

INTRODUCTION :

This project involved 3 phases i.e. Data **Cleaning and organization,** Feature extraction and Feature Selection.

We are provided data for a given activity "eating action" which is mixed with other unknown activities. **Our main aim is** to identify the eating activities amongst all the noisy data. We will be working with real world wristband data which provides us with i) accelerometer, ii) gyroscope, iii) orientation, and iv) EMG sensors data.
In phase 2, the data from these sensors are extracted and analysed by selecting and implementing feature extraction methods based on which methods gave us maximum variation and plotted scatter plots to support our hypothesis. The aim is to use features that show clear distinction between the actions.
In phase 3, We reduce the feature space by keeping only those features which show maximum distance between the two classes (eating and non-eating). We will use Principal Component Analysis technique to obtain the informative features.

PHASE 1

Data :- The provided data is time series, gathered from myo arm band sensor which has Orientation, gyroscope, accelerometer and emg sensors.

The main challenge in dealing with this kind of data was that the sampling rate for all the sensors were not the same. So we tried following methods to organize the data.

We used following technique and methods to organize the data.

- From the ground truth frames which were synchronized to the myo sensors we calculated the time where the frames were for eating and labelled the corresponding records in EMG and IMU data as 1. We labelled the non eating movement as 0.
- The data is for multiple users, we clubbed the data in single feature matrix for future purposes.
- For better visualization we sampled the data into 201 records for eating and 201 records for not eating. The samples were created by taking the mean of the data using varying sample size for eating and non eating as the sample size for both is different.

PHASE 2 :

We selected and implemented these five feature extraction methods:

1) We calculated the Mean of the various samples taken in the x, y and z axes. The mean is given by the following formula:

$$\textbf{Mean(J)= } (\sum_{i=1}^{n} J) \textbf{ /n}$$

2) We calculated the Root Mean Square of the various samples taken in the x, y and z axes. The Root Mean Square is given by the following formula:

$$x_{rms}=\sqrt{((x1)^2+(x2)^2+\cdots.....+(xn)^2)}$$

3) We calculated the Minimum value of the various samples taken in the x, y and z axes. The minimum is given by the following formula:

$$\textbf{Minimum(X)= min } (x_1, x_2, x_3....., xn)$$

4) We calculated the Maximum value of the various samples taken in the x, y and z axes. The maximum is given by the following formula:

$$\textbf{Maximum(X)= max } (x_1, x_2, x_3....., xn)$$

5) We used a combination of mean and FFT since MATLAB's fft () gives both real and complex values. In order to obtain the real values and plot them, the following formula is used:

$$\textbf{Value=mean(abs((fft}(x_1,x_2,...,x_n))^2))$$

LIST OF FEATURES EXTRACTED

1. MEAN

INTUITION:

When taking mean of a series of data points we normalise the sample space into one value, hence getting rid of unnecessary noise and extraneous values. Hence mean was selected as a feature for differentiating between data points of the 2 activities, Eating and Non-Eating Activities

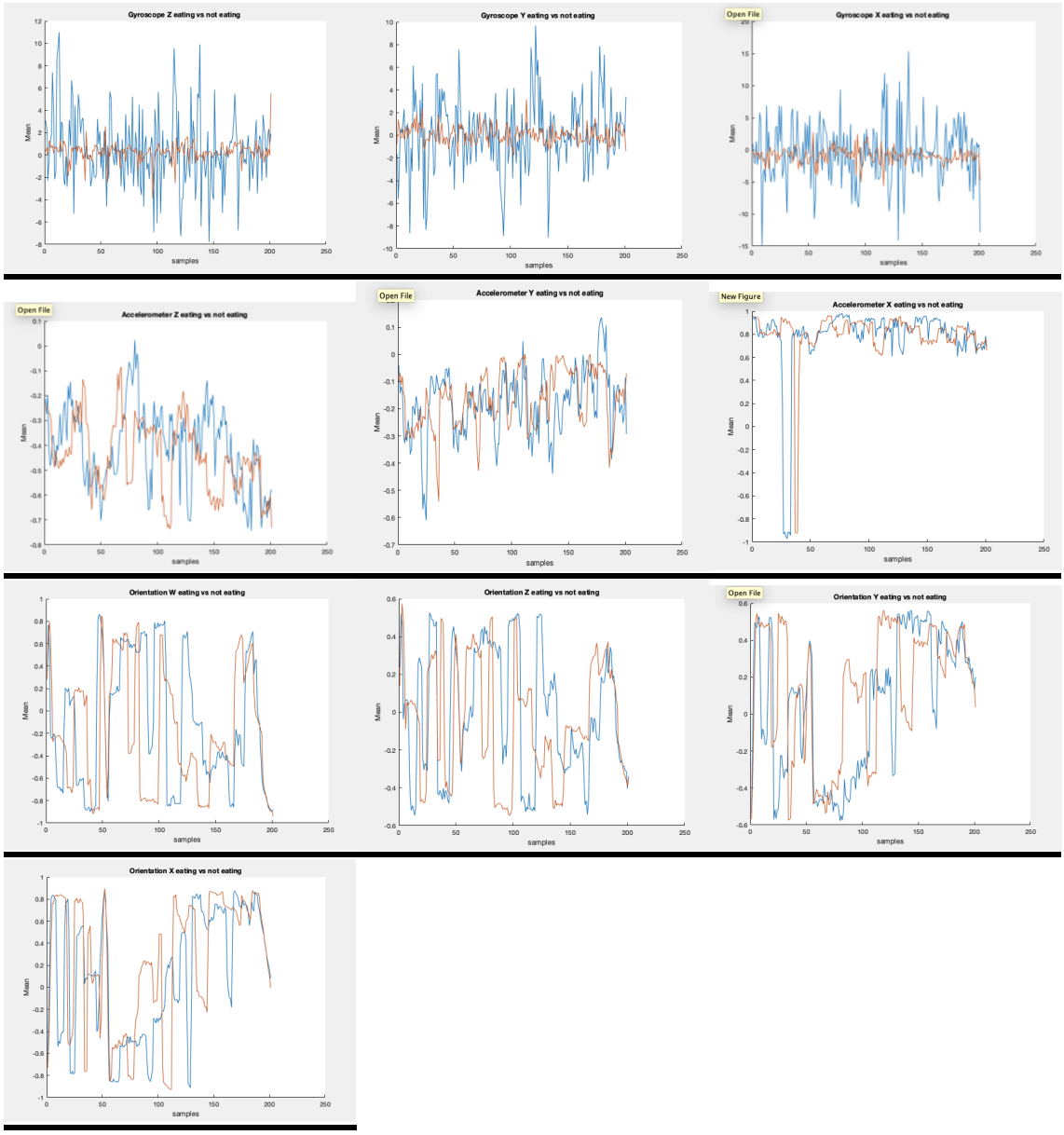


Fig.1: Comparison of mean plots between the two activities Eat Food vs Non Eating : IMU

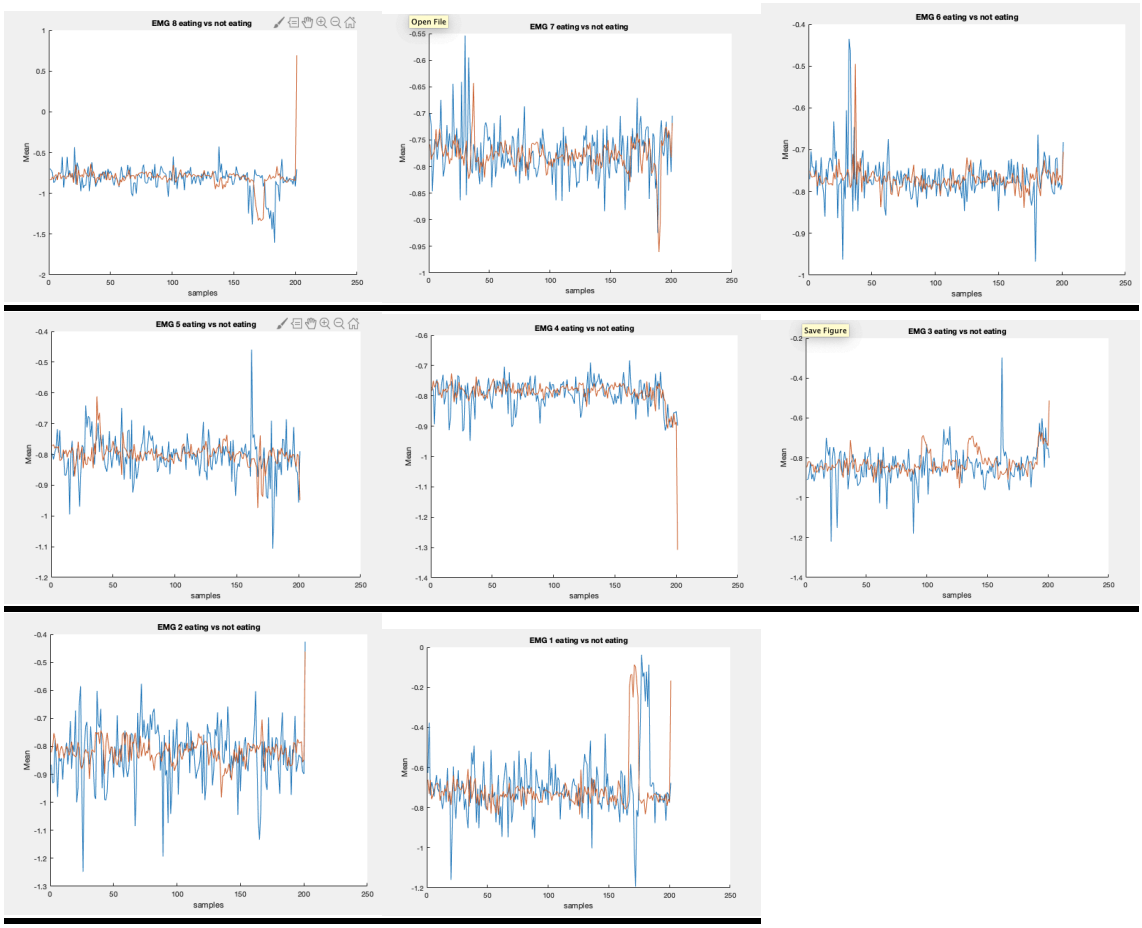


Fig.2: Comparison of mean plots between the two activities Eat Food vs Non Eating : EMG

2. ROOT MEAN SQUARE

INTUITION:

In calculating the Root Mean Square, we used of the samples of the mean that we had calculated earlier. Root Mean Square gives us the effective value in the series which best corresponds to the entire sample space. It was considered to be a feature which could be used in segregating the two activities.

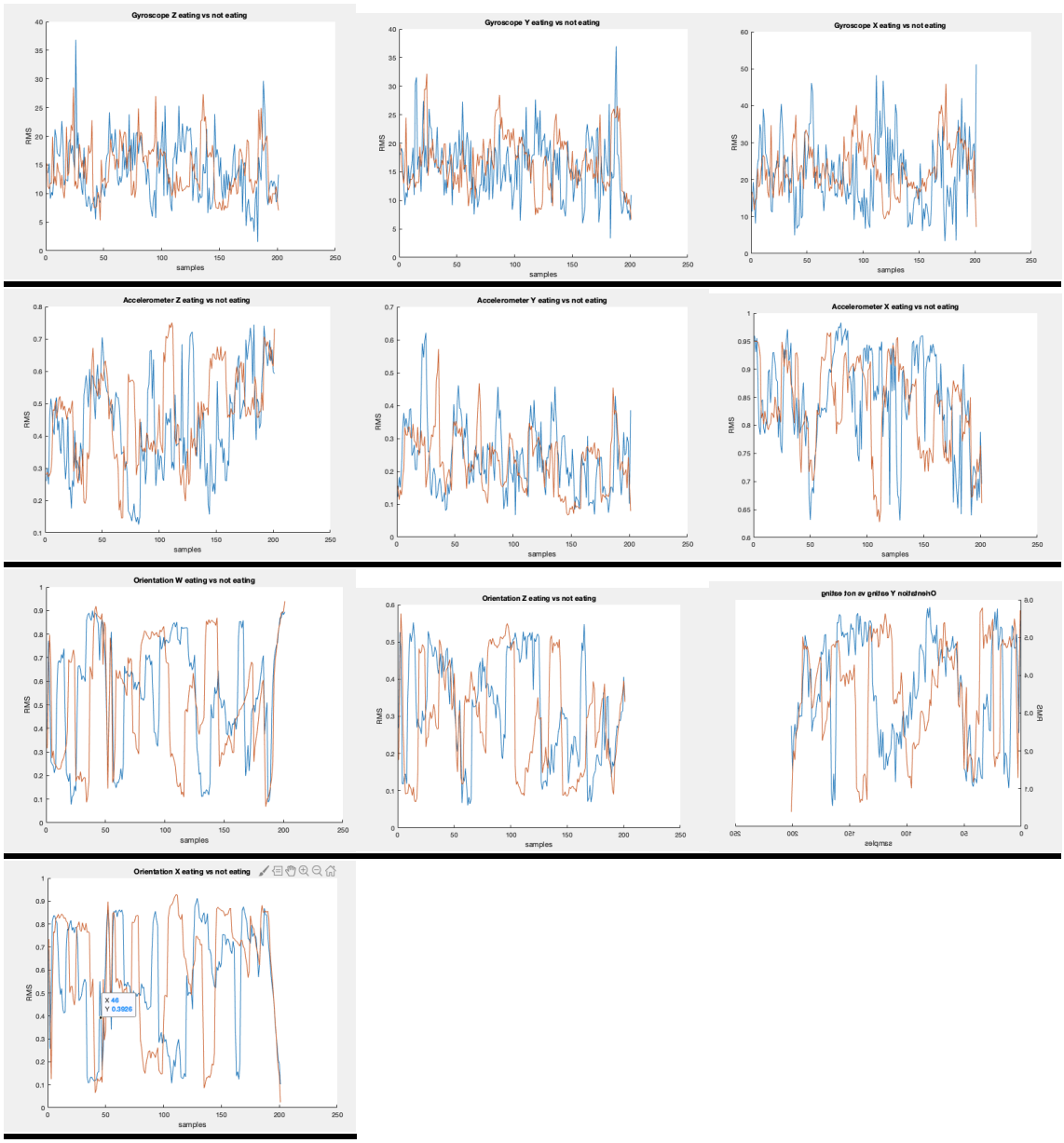


Fig.3: Comparison of RMS plots between the two activities Eat Food vs Non Eating : IMU

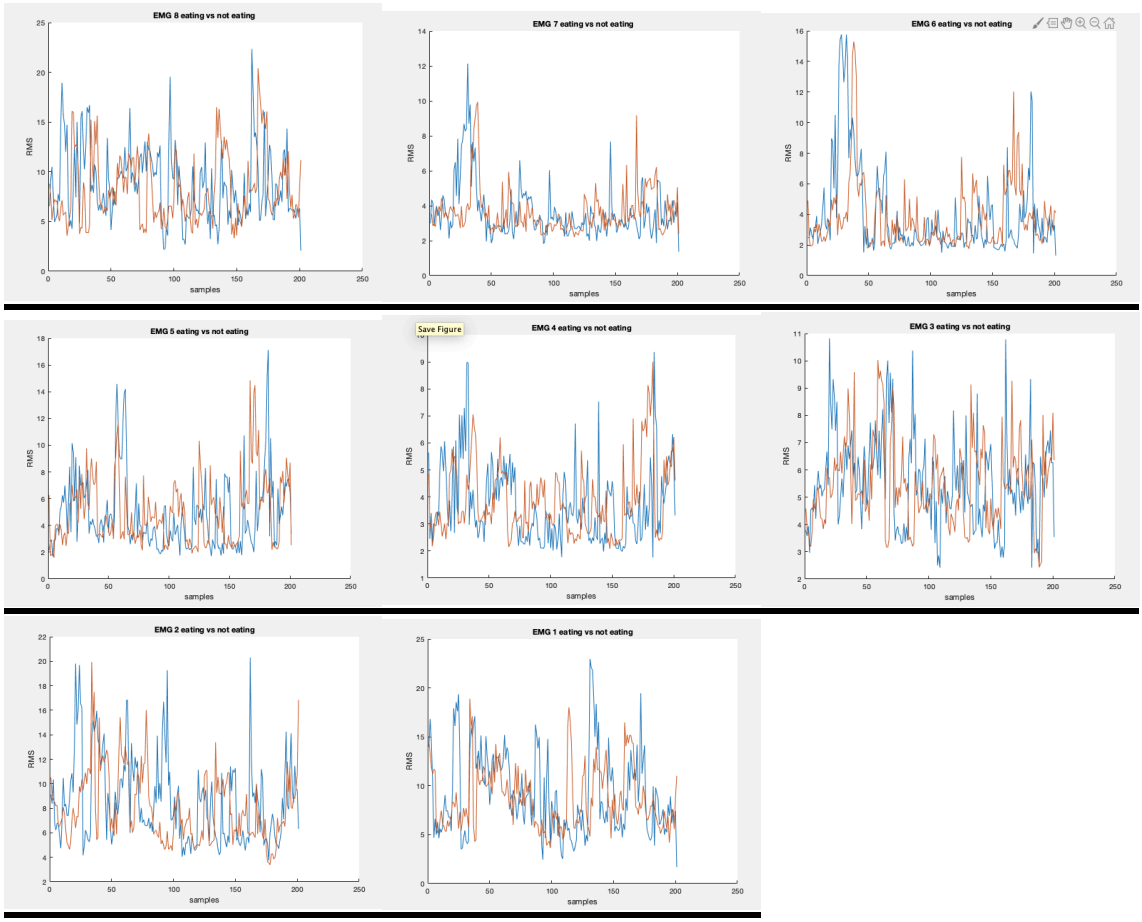


Fig.4: Comparison of RMS plots between the two activities Eat Food vs Non Eating : EMG

3. Minimum

INTUITION:

We found the minimum value for all the corresponding readings for samples (accelerometer, gyrometer, orientation) in all directions X, Y and Z. The intuition for finding the minimum value is that by calculating it, we obtain the data points across the samples having the least variance.

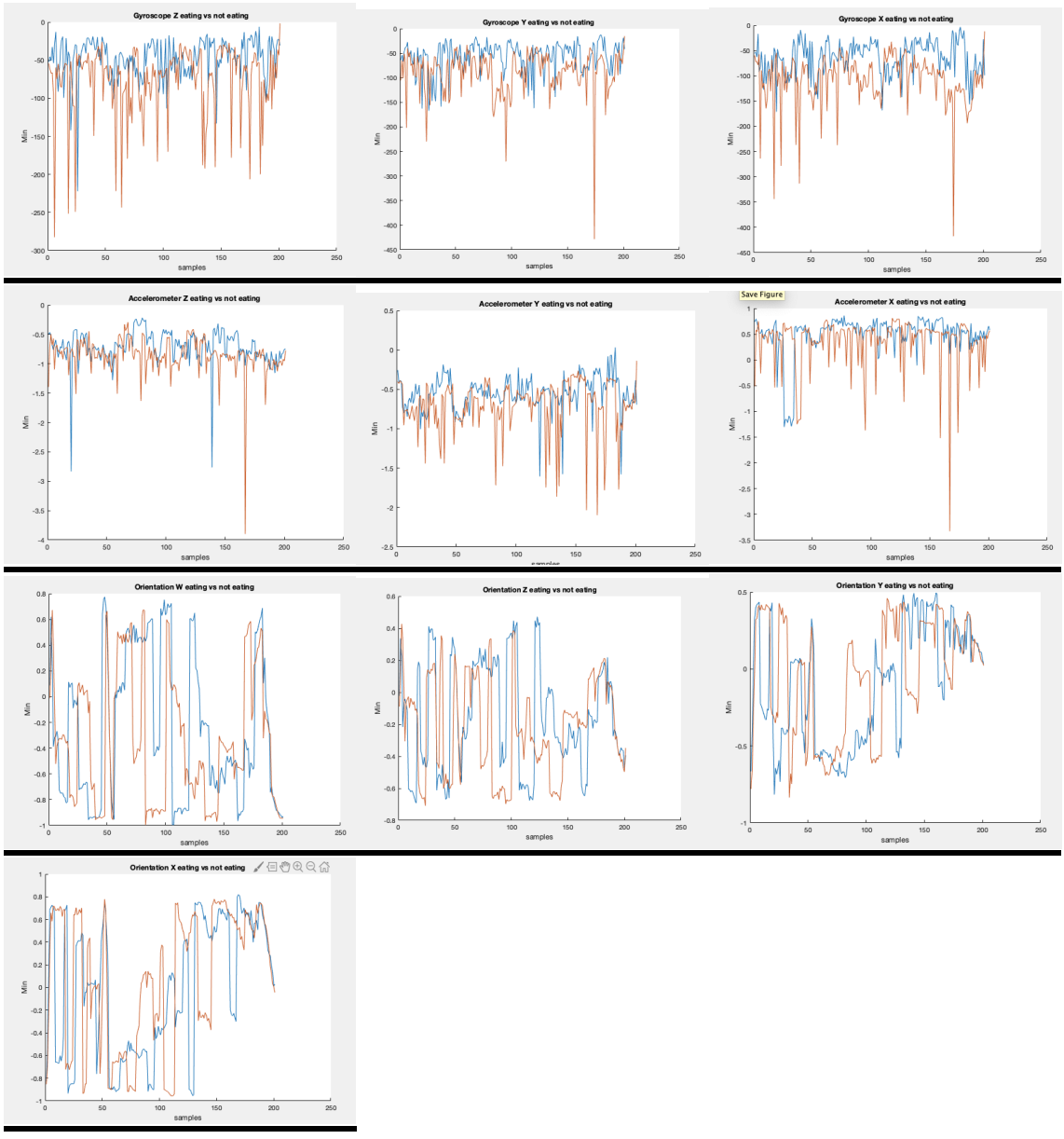


Fig.5: Comparison of Min plots between the two activities Eat Food vs Non Eating : IMU

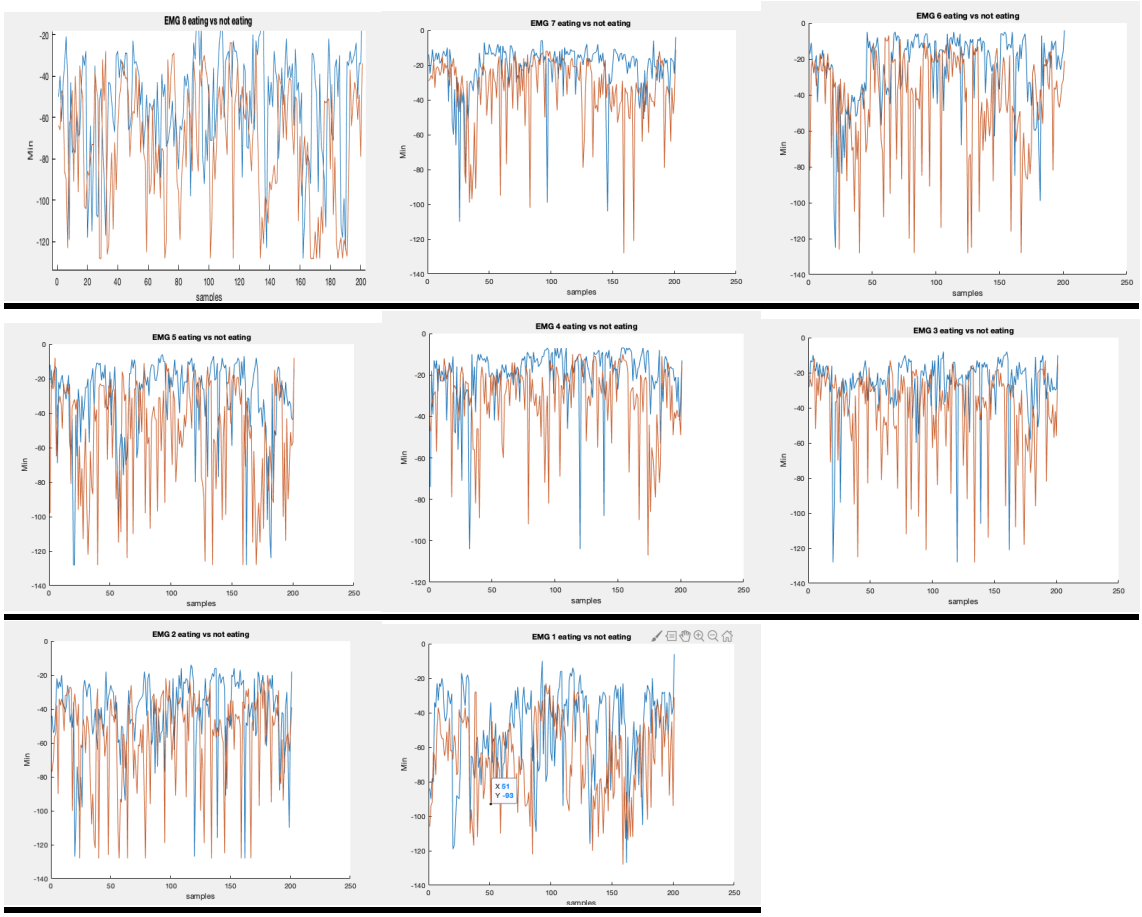


Fig.6: Comparison of Min plots between the two activities Eat Food vs Non Eating : EMG

4. Maximum

INTUITION :

The intuition behind finding the maximum value is that by calculating it, we obtain the data points across the samples having the highest variance. Thus , Obtaining minimum (previous feature) and maximum now we tend to have a better comparison of the features for each sample considered.

Fig.7: Comparison of Max plots between the two activities Eat Food vs Non Eating : IMU

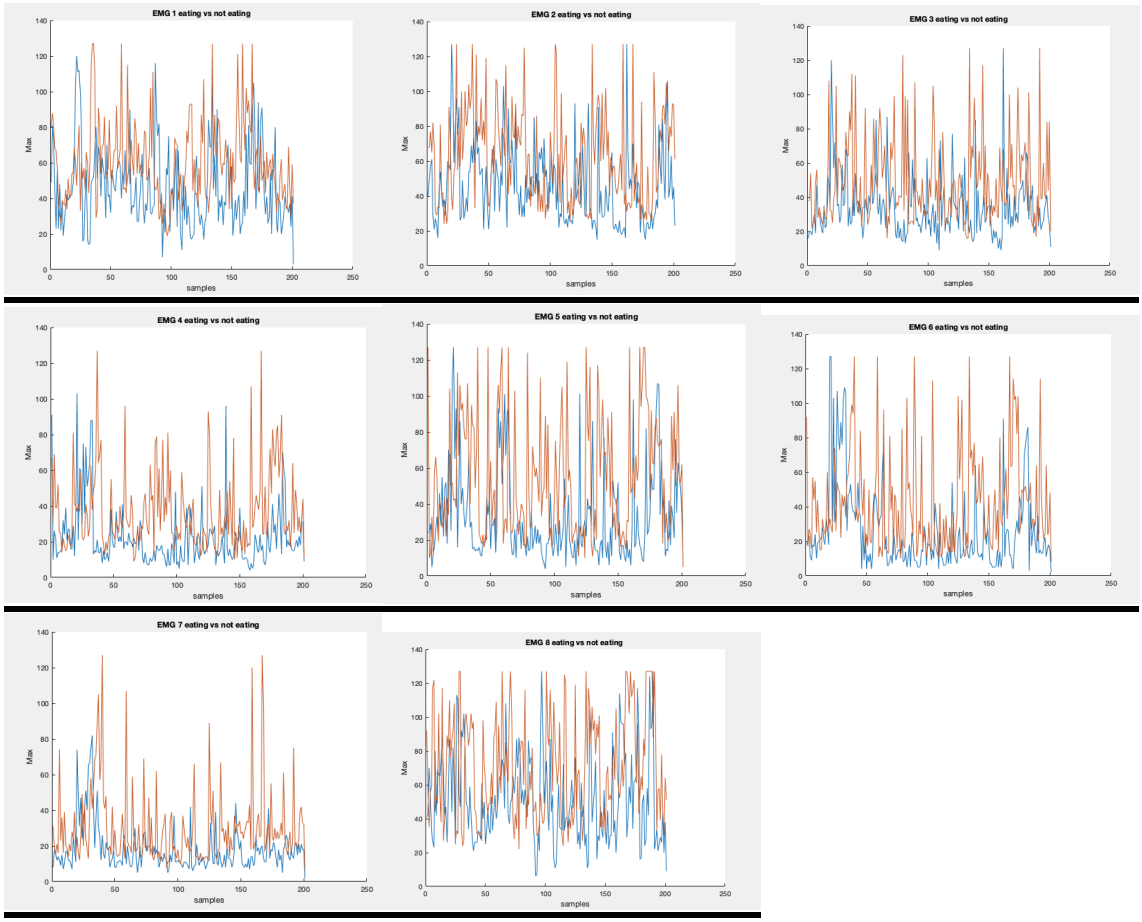


Fig.8: Comparison of Max plots between the two activities Eat Food vs Non Eating : EMG

5. First Fourier Transformation

INTUITION :

It is a technique which transforms the data points in our sample space over a period of time and divides it into frequency components.
Choosing FFT as a feature helped us get a better spread of the data across the time domain with its frequency components intact hence ensuring that the data preserves all the properties even after transformation.

Fig.9: Comparison of FFT plots between the two activities Eat Food vs Non Eating : IMU

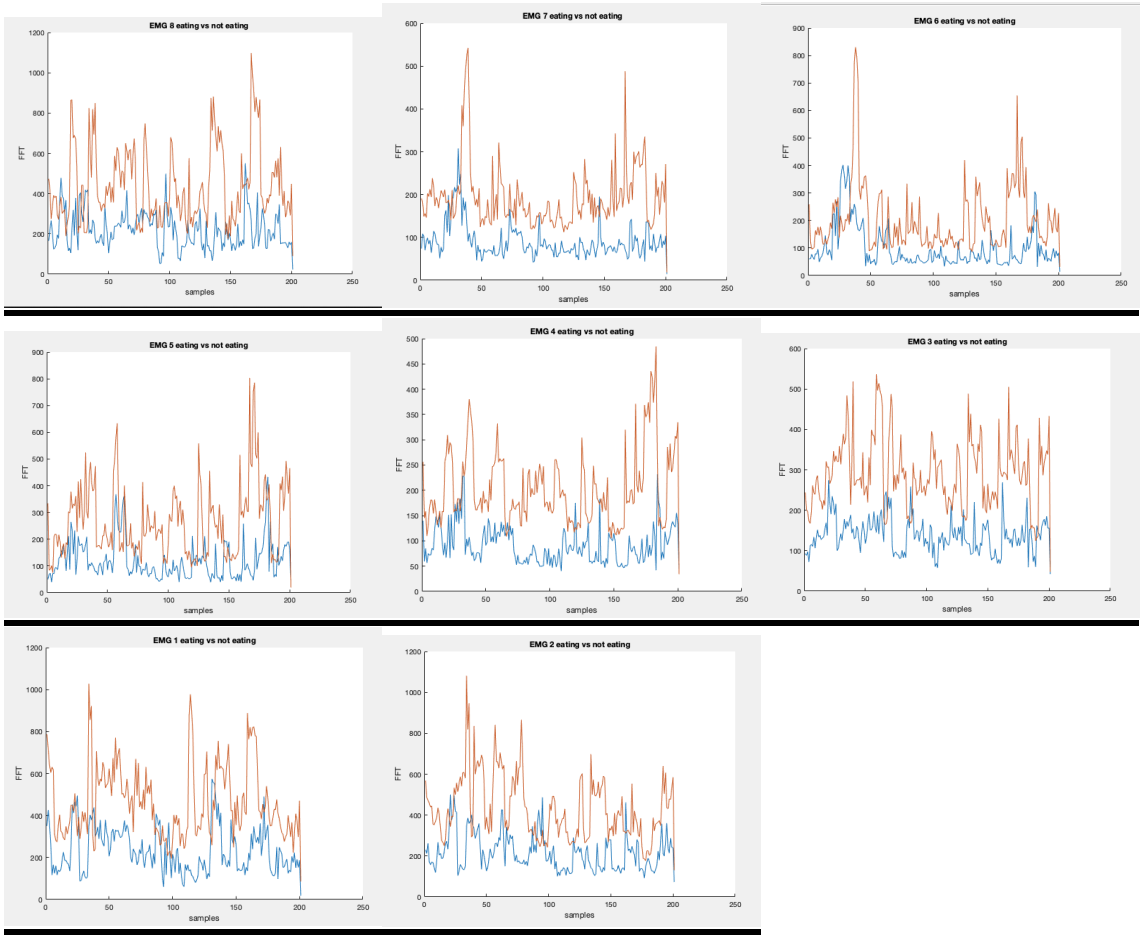


Fig.10: Comparison of FFT plots between the two activities Eat Food vs Non Eating : EMG

PHASE 3

The aim of this task is to change the original feature space into a reduced feature space by keeping only those features which showed maximum distance between the two classes.

To achieve this, we used the Principle Component Analysis (PCA) technique.

Subtask 1: Arranging the feature matrix

The total number of features extracted are 18(emg 10 , imu 8 features.). These features will become the columns of the feature matrix. Each row will correspond to each of the activities (eating n non eating) We get 2 feature matrices, one for each type of activity (EatFood and Non eating). On applying PCA, we obtain the principal components which give us the directions along which our variance is preserved while reducing the number of dimensions.

Subtask 2: Execution of PCA

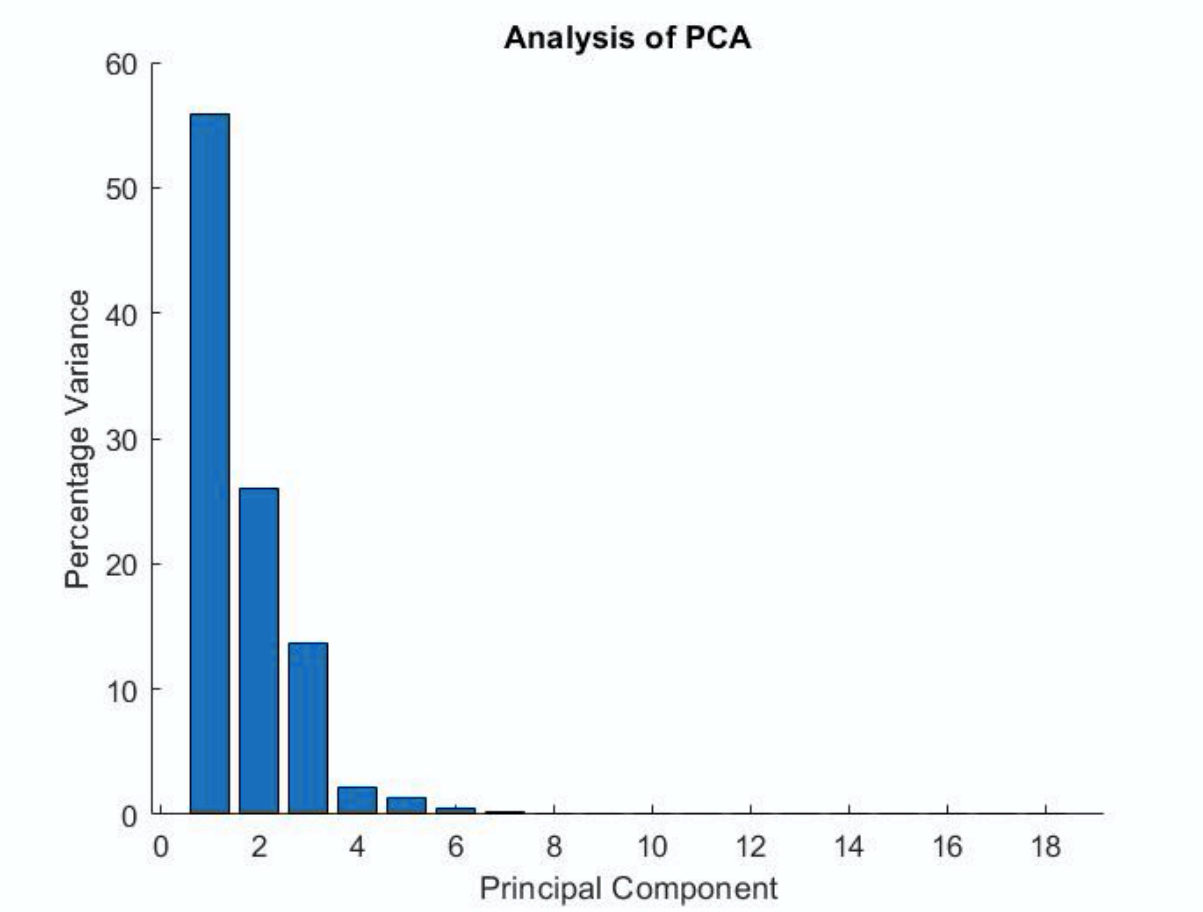
We wrote the code in MATLAB, which is named, *pcaA.m* .
The pca function of MATLAB defined as [coeff,score,latent] = pca(X), returns:

- **Coeff:** The coefficient matrix (18-by-18 matrix) as **'coeff'**. For our n-by-d feature matrix, i.e. 402 x 18 feature matrix X, the corresponding coefficient 18-by-18 matrix is calculated and stored in the variable 'coeff'. Each column of coeff contains loading/weights for one principal component, and the columns are in descending order of component variance, latent.
- **Score:** The principal component scores as **'score'**. Principal component scores are the representations of X in the principal component space, where rows of score correspond to observations, and columns corresponds to components.
- **Latent** :The principal component variances as **'latent'**, which stores the eigenvalues of the covariance matrix of X.

After applying PCA,we obtained :

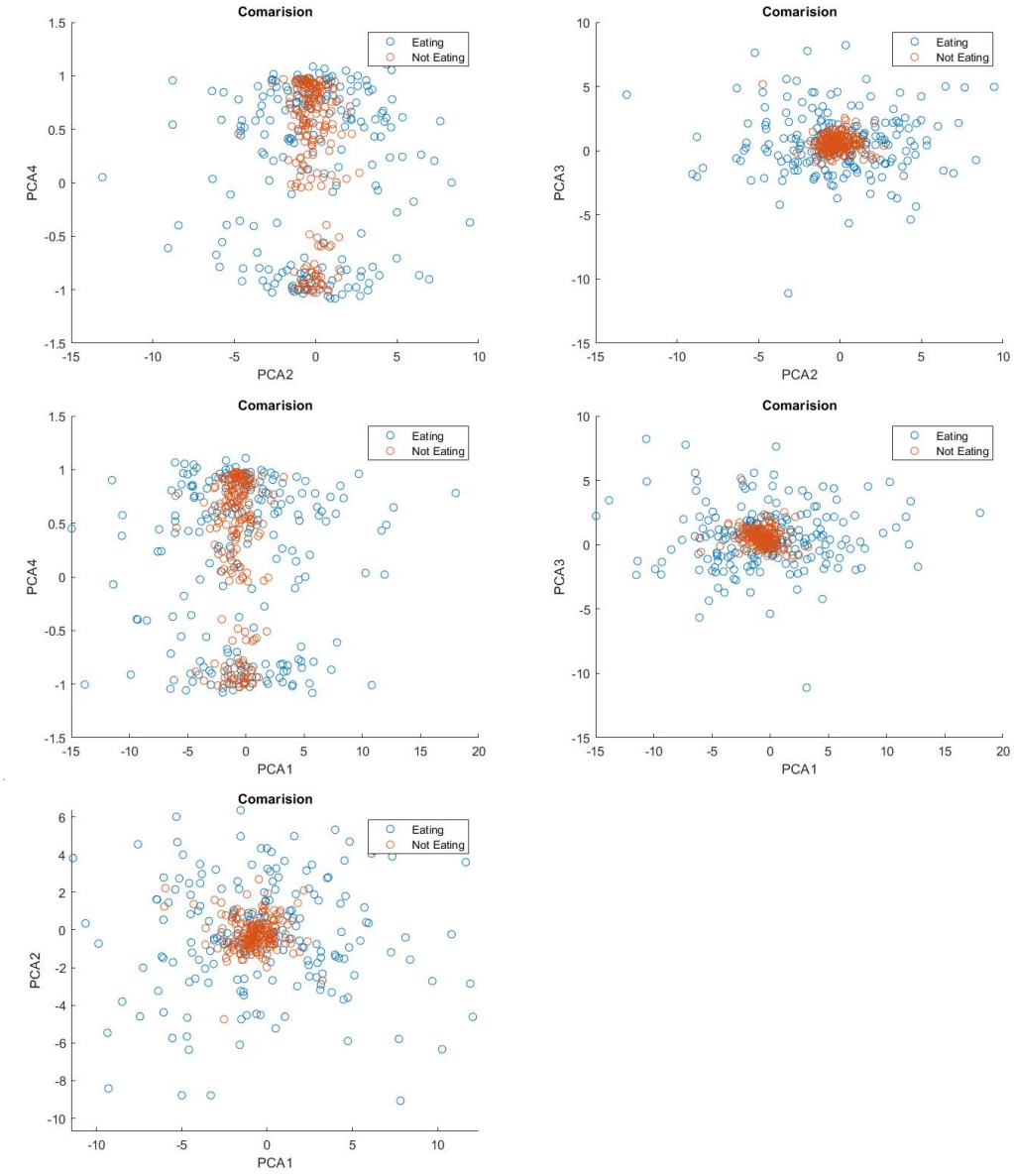
- 55.92% variance for Principal Component 1,
- 26.05% variance for Principal Component 2,
- 13.61% variance for Principal Component 3,
- 2.16% variance for Principal Component 4.

Thus, we selected the first four principal components and retained **97.744%** variance.



Subtask 3: Making sense of the PCA's Eigenvectors

The values of an eigen vector define the contribution of a feature towards the direction of the new principal component defined by the eigen vector.



We can clearly see that eating and non eating classes being separated in 2 co-centric clusters(as specified by different colours in fig above).

Subtask 4: Results of PCA

After performing PCA on the given data, we were able to reduce the featre space from 18 to 4, while retaining 97.744% of the original information and were still able to classify the data effectively.

After applying PCA we can see that the eating and non eating activities forms two different clusters with same centre. Thus linear classifiers like logistic regression won't be able to perform well. But SVM with different kernel would give impressive classification results.

Subtask 5: Argue whether doing PCA was helpful or not

YES.

Given the nature of the data, collected from different sensors it is highly susceptible to noise. PCA can help reduce the effect of noise on classifier. We choose 4 principle components from 18 so the effect of noise and outliers can be mitigated.