3. Project Phase 1: Feature Extraction

In this phase, we have selected and implemented seven feature extraction methods for the two time series data CGMGlucose and its associated date timestamp CGMTimeStamp. Before carrying on with that, we had to do some preprocessing of the data. The CGM data were collected with a Matlab timestamp and had to be converted into Unix Timestamps. The data also had some missing values and nulls. They have to be cleansed and estimated to the standard deviation of the respective time series. Estimation is important since replacing the values with zero will result in noise or outliers which may have a profound impact on the model. Before the feature extraction was done, data were normalized, since the values of each time series namely InsulinBasal, InsulinBolus and CGMData are of different scales and hence difficult to compare with.

The seven feature extraction techniques that we have used are:

1. Velocity with Respect to Acceleration
2. Coefficient of Variation
3. Fast Fourier Transformation
4. Probability distribution fit
5. Discrete Wavelet transform
6. Windowed Entropy
7. Area Under Curve

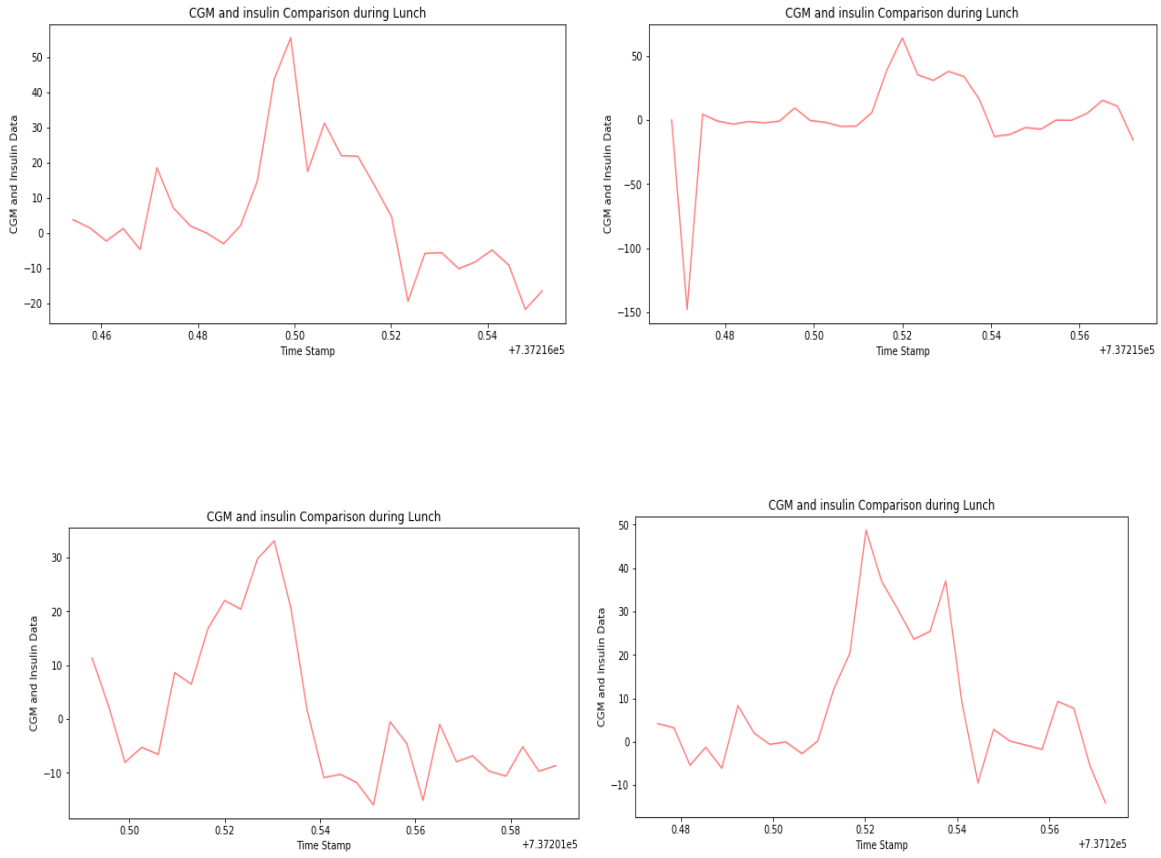
Intuition behind feature selection

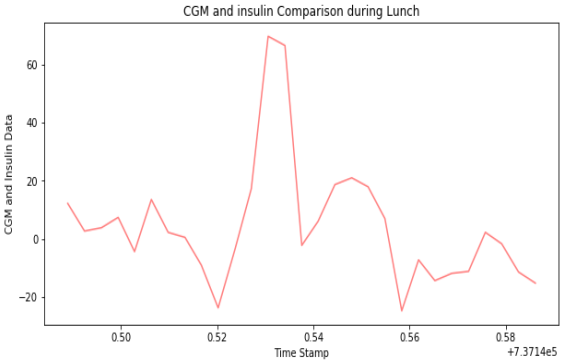
Since we are extracting features which can deduce the time interval during which a meal was intook, we need to find the spikes in Glucose levels over a period of time. The features which we selected try to approximate the descent in the CGM Glucose Data curve. Initially, the data interpretation just by using CGM_data with respect to only timestamp was vague. However, by taking into account, the Insulin Basal and Insulin Bolus series plotting and comparing them with the Glucose series, we get a specific pattern depicting how Glucose level vary. Hence, we decided to deduce various ways to co-relate these levels with each other to observe specific feature patterns. We were completely aware of the fact that plotting of values for each activity along individual feature dimension would give more suitable ground to make a reliable judgement. The contribution made by each feature extraction method and how these features are extracted is explained below.

3.1 Velocity with respect to Acceleration

Velocity can be obtained by calculating the rate of change of glucose levels with respect to time. We can simply obtain it by calculating the slope between two consecutive points in the time series. Higher rise in velocity means increased glucose levels that indicates meal intake by a patient. But CGM velocity taken as an absolute value has an inherent flaw in it. The common pattern seen in the data amongst all the patients is that they have a small meal in between during the course of the day. Hence, we end up having false positives or local minima. In order to solve this problem, we transformed velocity with respect to time. Here, at each point acceleration is observed. Once it starts getting near to 0, we can deduce that he is having lunch at a time period before the levels decelerated. This feature is useful to deduce the Hypoglycemic levels, and was validated by comparing it with the observation of insulin Bolus and Basal injections.

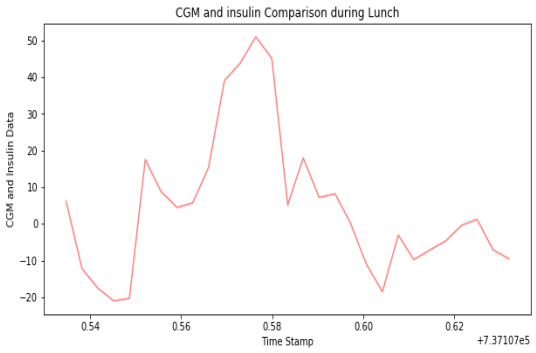
3.1.(1) Velocity graph for the timestamp acceleration:

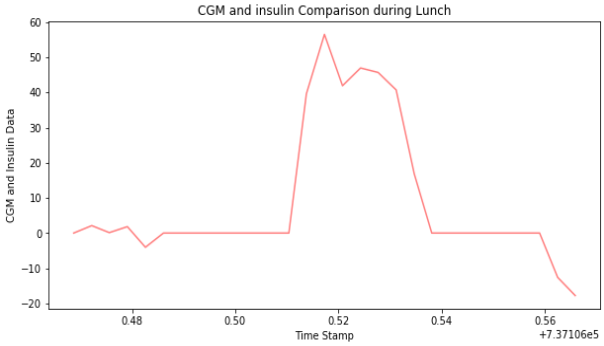



CGM and insulin Comparison during Lunch
CGM and Insulin Data
Time Stamp
+7.3714e5


CGM and insulin Comparison during Lunch
CGM and Insulin Data
Time Stamp
+7.37117e5


CGM and insulin Comparison during Lunch
CGM and Insulin Data
Time Stamp
+7.37112e5


CGM and insulin Comparison during Lunch
CGM and Insulin Data
Time Stamp
+7.37107e5


CGM and insulin Comparison during Lunch
CGM and Insulin Data
Time Stamp
+7.37106e5

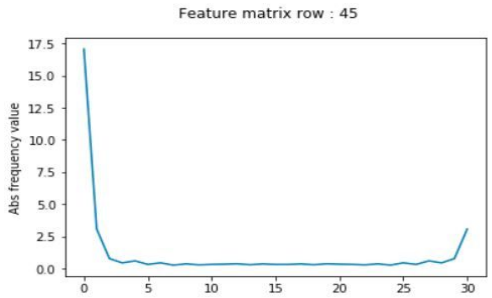
3.2 Coefficient Of Variation

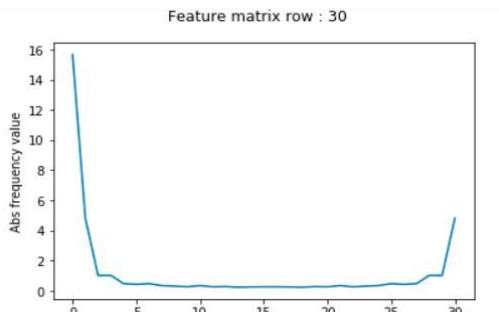
Coefficient Variance is a statistical measure that simplifies the interpretation of glycemic variability across patients with different mean unlike Standard Deviation which is proportional to the mean glucose level. That is, someone with higher glucose will have a higher Standard Deviation. Coefficient Variance helps in normalizing the Glucose variability, paving way to use a single variability measure that can be applied to people with varying mean glucose levels. This feature gives us a range of values between which we can expect a person to take his lunch. Coefficient of variation is a single parameter for GV and hence cannot be plotted.

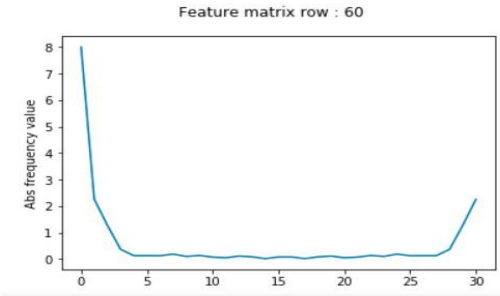
3.3 Fast Fourier Transformation

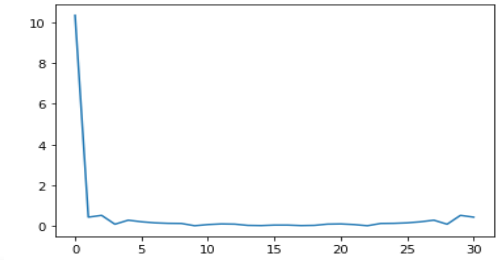
Fourier transform is a mathematical function that maps a series from time domain to frequency domain. It deconstructs the original signal into a combination of sinusoidal functions. The glycemic levels of the patients mostly follow a cyclic pattern, with blood-glucose levels steadily increasing when one is consuming a meal and reducing after the insulin is injected either by Bolus or by Basal injection. This near cyclical pattern can be made use of by adopting the fourier analysis in estimating the time period of meal consumption. Though Fourier transform is a basic engineering tool to analyze signals, there have been a lot of studies about the applicability and clinical significance of the same. The harmonic decomposition performed by Fourier transform can be used in predicting the glucose variability in absolute terms.

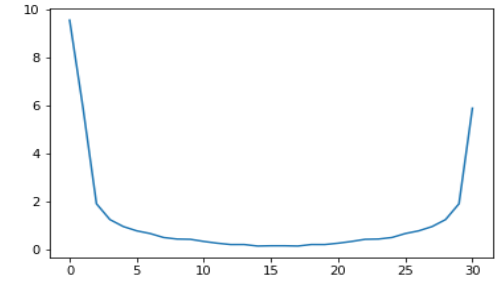
The frequency decompositions done by FFT can help in detecting the maximum significant frequency from the given data. The idea is that you could move back and forth between the period of the wave and the frequency by using the Fourier Transform. This is extremely helpful in extracting patterns which may look like a random noise. All the Glycemic level data has been sampled at a rate of 1 per 5 minutes for a total time period of two and a half hours. When FFT is applied on to a time series, it gives the amplitudes of the highest frequencies, which corresponds to the absolute values of the CGM data, from which the meal intake can be deduced.

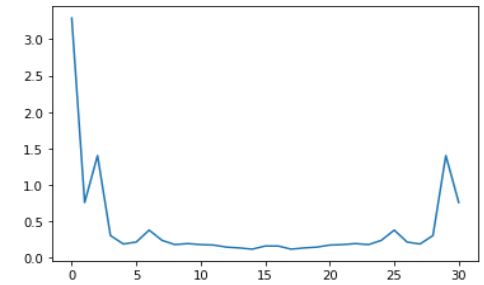

Feature matrix row : 45
Abs frequency value

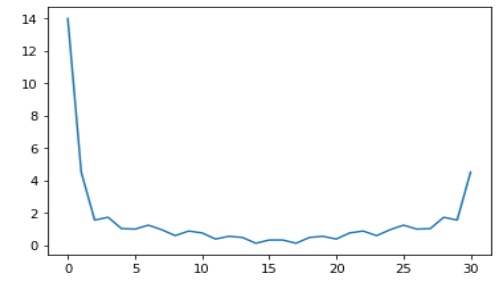

Feature matrix row : 30
Abs frequency value

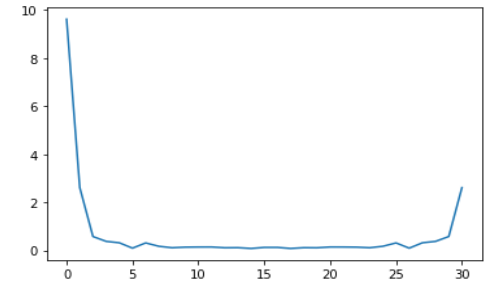

Feature matrix row : 60
Abs frequency value

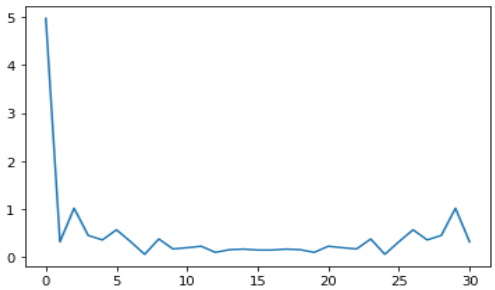


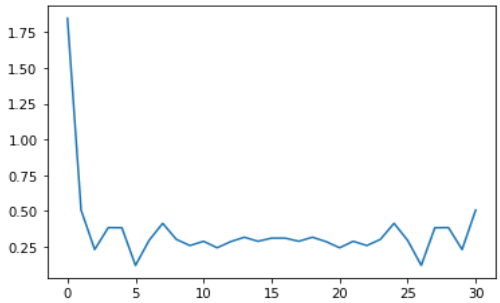










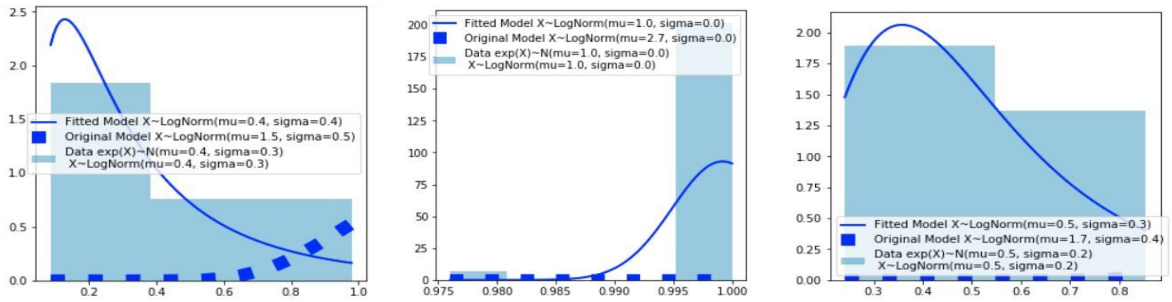


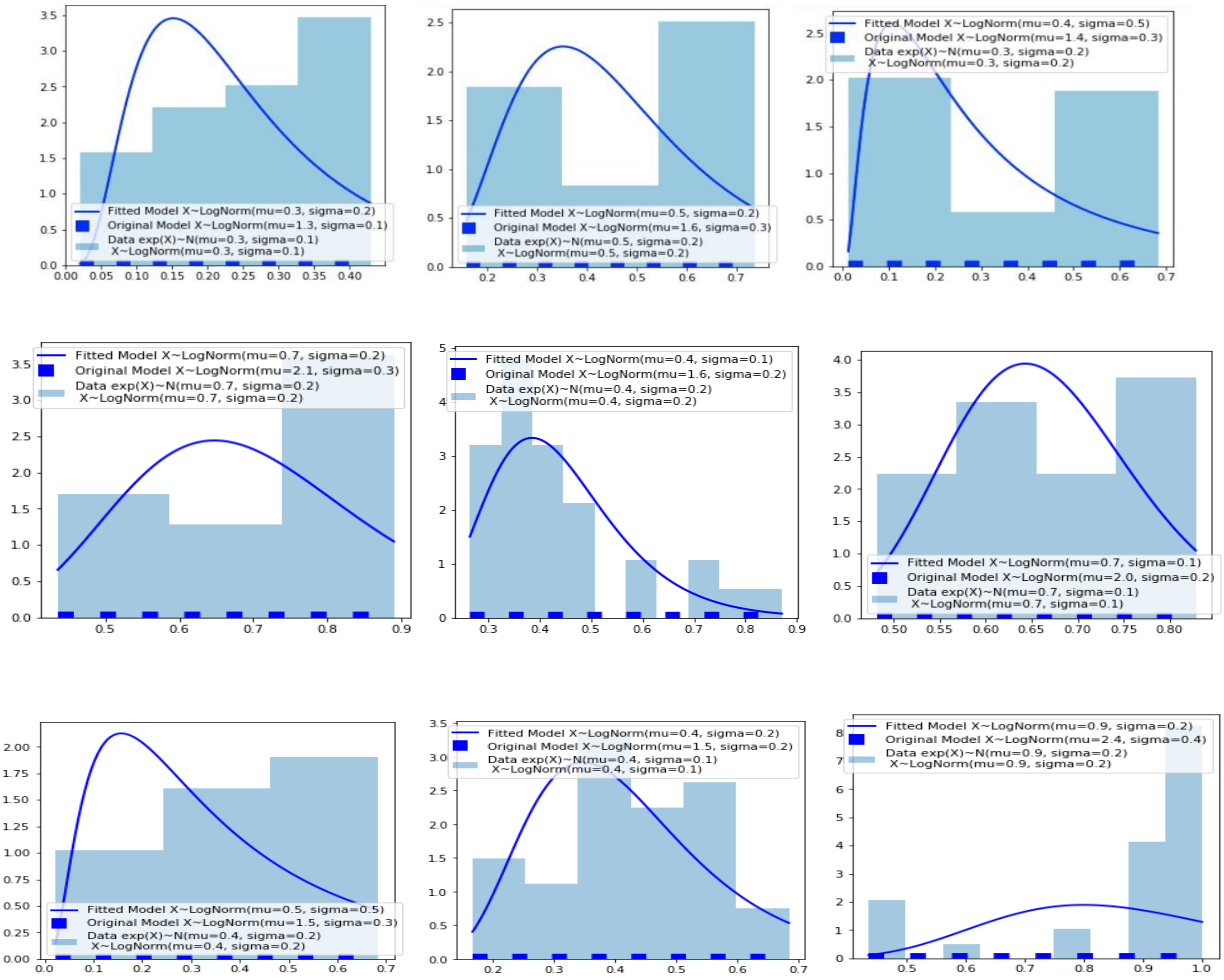
3.4 Probability Distribution Fit

Information in time series is conveyed in the form of probabilistic distribution functions. The objective is to find the best distribution that models the data. Here, we use the chi-squared test to measure the goodness of the fit. After getting the values for different distributions and comparing them, we found that the Lognorm distribution has the best value when compared to other types of distributions provided below. The chi-squared test and the KS Test provides the best value for the distribution that is neither too high nor low.

```
Distributions sorted by goodness of fit:
----------------------------------------
  Distribution   chi_square  p_value
3      lognorm    32.582732  0.11276
1        expon    64.092518  0.03960
5     pearson3    74.278849  0.00292
7      uniform    97.595647  0.00293
2        gamma   115.550519  0.08776
4         norm   119.928588  0.08402
6       triang   134.938029  0.20285
0         beta   141.610868  0.40529
9  weibull_max   561.995028  0.00000
8  weibull_min  2392.194178  0.00000
```

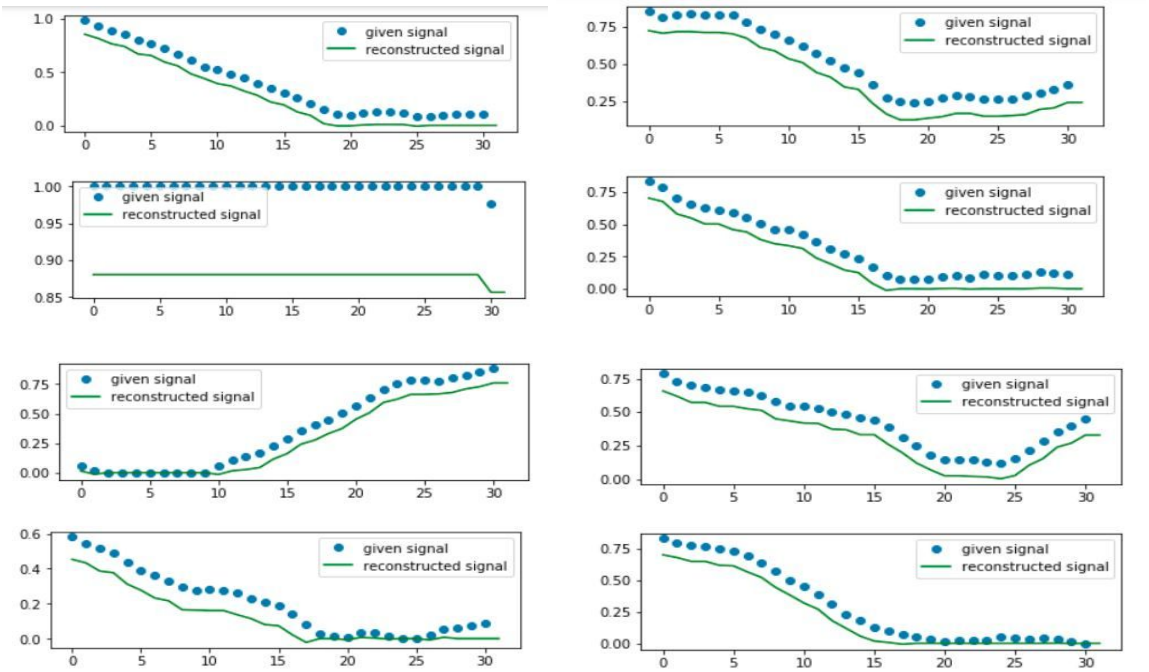
The lognorm distribution fitted data can be used to derive two parameters of the distribution namely the Mean and Variance of the lognorm distribution.





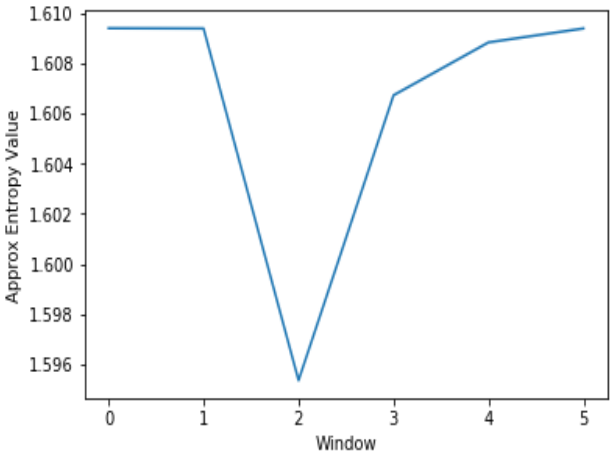
3.5 Discrete Wavelet Transform

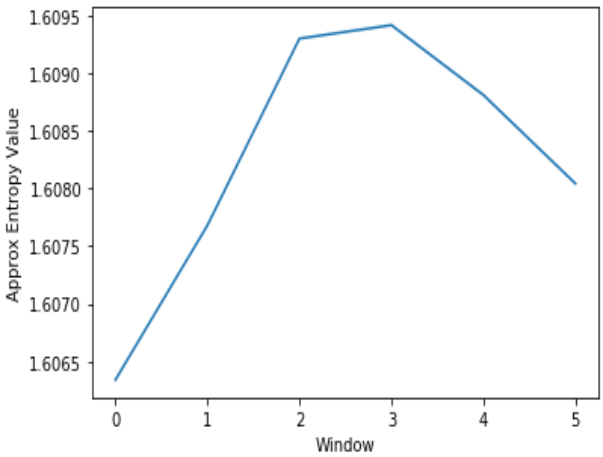
Wavelet transforms analyses signal whose frequency varies over time. The time domain signal is decomposed by contractions and expansions of the wavelet function by applying small windows at higher frequencies and large windows at low frequencies. When we consider the CGM data, we find that peaks occur at different times in 2.5 hour interval. To analyse this, we use the Discrete Wavelet Transform. The wavelet transform contains information on both the time frequency and location of a signal, and hence wavelet transform can give a better result for meal detection.

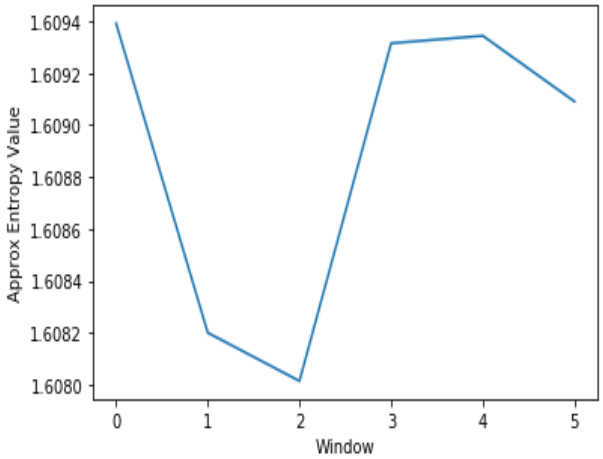


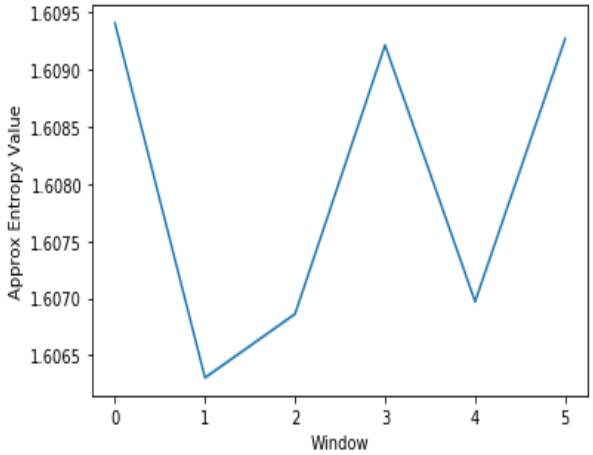
3.6 Windowed Entropy

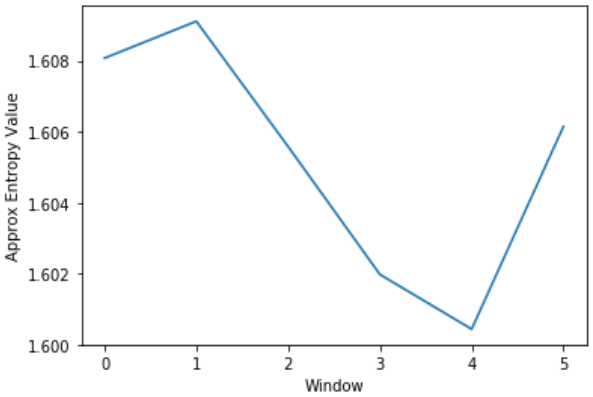
In statistics, an **approximate entropy** (**ApEn**) is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data. Moments such as mean and variance only tells us about the distribution of data in general. Hence we need a measure like Entropy to measure the randomness which is quite helpful in forecasting based on time series data. Applying entropy as a whole over a series will give a valued measure of the randomness. We adopted the same and split the entire series into different windows and calculated the entropy for every window. The window which shows the maximum randomness corresponds to the period of meal intake. With a window size of 5, we deduced that the lowest entropy value correlated with the window in which the person took in a meal. The results can be seen below.

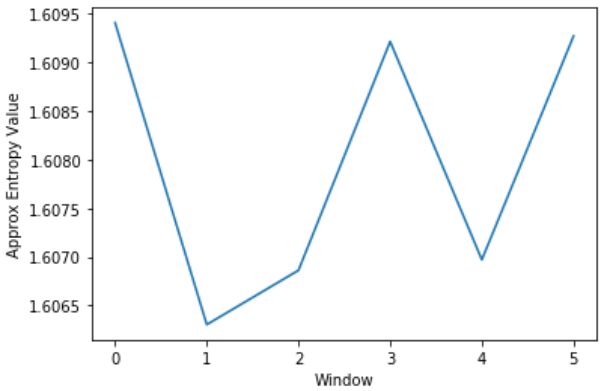

Approx Entropy Value
1.610
1.608
1.606
1.604
1.602
1.600
1.598
1.596
0
1
2
3
4
5
Window

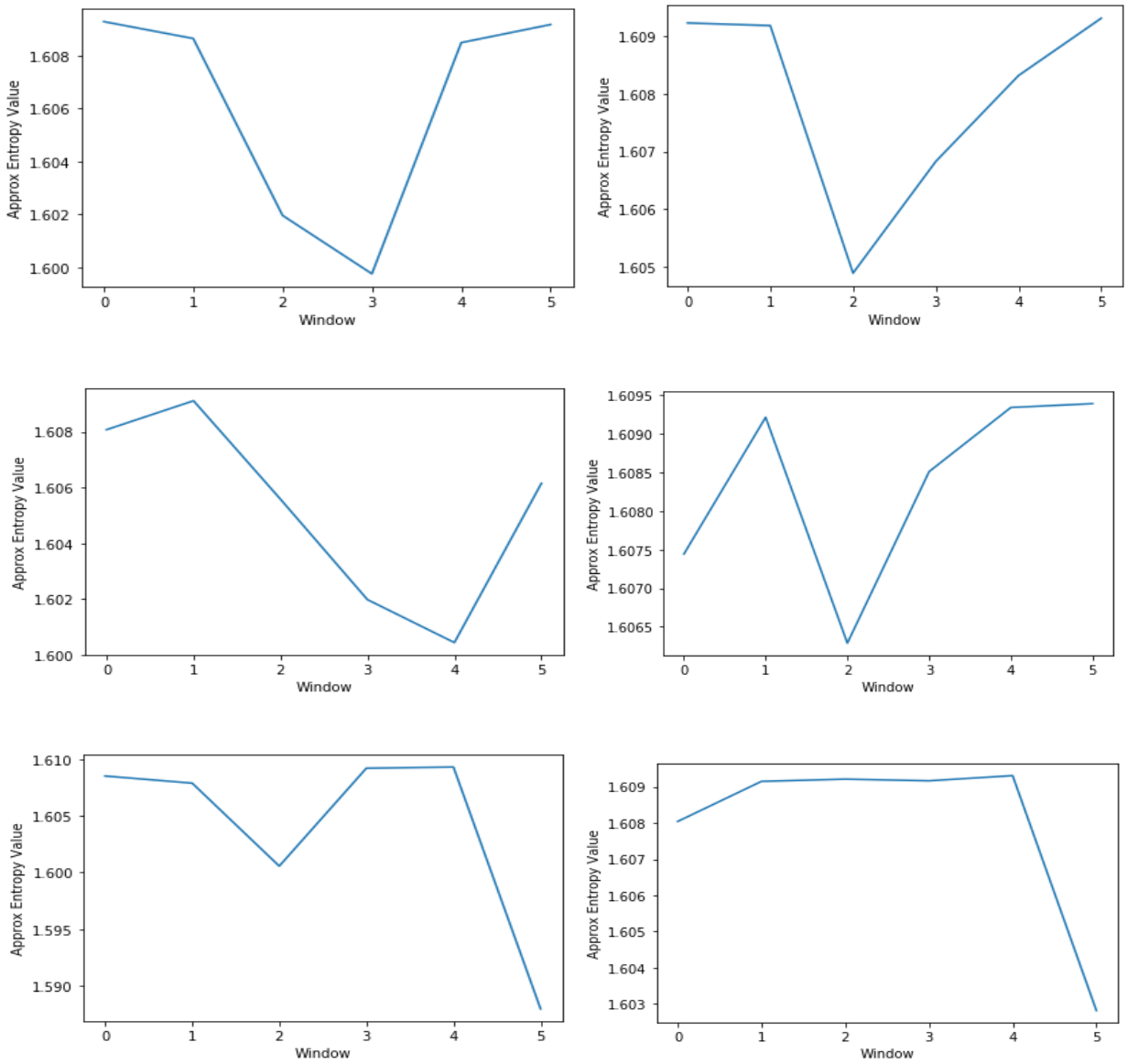

Approx Entropy Value
1.6095
1.6090
1.6085
1.6080
1.6075
1.6070
1.6065
0
1
2
3
4
5
Window


Approx Entropy Value
1.6094
1.6092
1.6090
1.6088
1.6086
1.6084
1.6082
1.6080
0
1
2
3
4
5
Window


Approx Entropy Value
1.6095
1.6090
1.6085
1.6080
1.6075
1.6070
1.6065
0
1
2
3
4
5
Window


Approx Entropy Value
1.608
1.606
1.604
1.602
1.600
0
1
2
3
4
5
Window


Approx Entropy Value
1.6095
1.6090
1.6085
1.6080
1.6075
1.6070
1.6065
0
1
2
3
4
5
Window



3.7 Area Under Curve

AUC curve is a performance measurement for classification problem at various thresholds settings. AUC can be used as single figure for measuring performance as well as for classifying the data has to whether meal has been taken or not. A data with peak has an AUC that is probably different from the data without glucose peak i.e, when the graph is stationary. The area can also be used to derive true positive rate and false positive rate which can be used to evaluate the model.

4. Feature matrix & Dimensionality Reduction

4.1. Subtask 1: Arranging the feature matrix

Principal Component Analysis (PCA) takes in matrix of dimensions o*f where o corresponds to the number of data samples and f corresponds to the number of features or the combined length of the feature vector. We constructed a feature matrix of dimensions 186*20. Total number of samples were 186 and the combined length of all seven features amounted to 20. This has a higher dimension. Hence it is necessary to reduce the dimensions for better matching and prediction.

```
Feature_Matrix_New
```

```
matrix([[2.36000000e+02, 2.58000000e+02, 5.57963196e+00, 2.00829428e+00,
         1.67902510e+00, 1.07592323e+00, 8.00772636e-01, 8.56779939e-01,
         6.62282582e-01, 6.10420422e-01, 4.01720202e-02, 3.25827320e+01,
         1.28081440e+00, 6.59293776e-01, 5.46604766e-01, 4.89188097e-01,
         3.71373599e-01, 2.07534387e-01, 1.50659893e-01, 1.59535165e+00],
        [3.24000000e+02, 3.49000000e+02, 2.40000000e-02, 2.40000000e-02,
         2.40000000e-02, 2.40000000e-02, 2.40000000e-02, 2.40000000e-02,
         2.40000000e-02, 2.40000000e-02, 1.04166335e-01, 1.29813173e+03,
         1.28081440e+00, 2.57186953e-01, 3.00800980e-01, 3.69514290e-01,
         3.29081347e-01, 2.38456320e-01, 2.34308484e-01, 1.60634042e+00],
        [2.48000000e+02, 2.69000000e+02, 4.82159456e+00, 1.01955467e+00,
         1.02877207e+00, 4.76629029e-01, 4.31620055e-01, 4.76248971e-01,
         3.51420010e-01, 3.13516246e-01, 5.22938375e-02, 3.21541558e+02,
         1.28081440e+00, 5.09968372e-02, 5.45482374e-02, 1.27878847e-01,
         1.32198224e-01, 5.32365010e-02, 3.34836441e-02, 1.59852747e+00],
        [2.11000000e+02, 2.22000000e+02, 4.67680611e+00, 1.24572574e+00,
         1.34641306e+00, 8.38667936e-01, 7.60000233e-01, 7.63447603e-01,
         5.89978143e-01, 5.20780040e-01, 3.30099846e-02, 5.06584007e+01,
         1.28081440e+00, 6.77471233e-01, 5.43911026e-01, 4.51205349e-01,
```

4.2. Subtask 2 : Execution of PCA

Principal Component Analysis is a method that is used to prevent the curse of dimensionality by extracting only features that show high variance in the data set. Highly correlated variables are abstracted from the feature vector to form a significant set of variables known as principal components which define the variance of the data set. PCA reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).

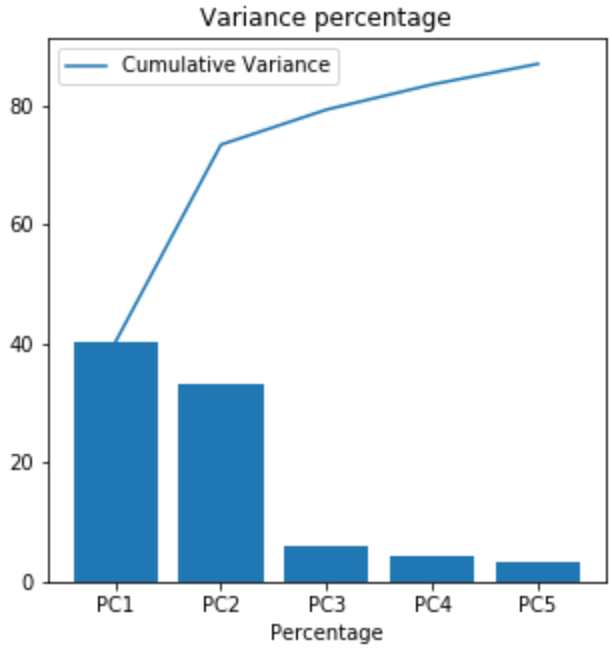
PCA gets a Data matrix of dimensions o*f (o - number of data objects, f- number of features or attributes corresponding to each object) as input. It then computes the covariance matrix of dimensions f*f, which tells how every feature correlates with each other. If two or more features are correlated, then they are redundant. PCA tries to decompose this matrix using Singular valued decomposition or Eigen Decomposition. It computes the EigenValues and EigenVectors for the covariance matrix, which forms the new ortho-normal basis and can be called as principal components or latent space. Eigenvalues are arranged in descending order of value and the top n values and its corresponding vectors are chosen. They represent the reduced principal dimensions. By multiplying the original data matrix with it, we project the data on to the latent space, which can conveniently represent the data with less dimensions and retaining maximum variance among them.

```
reduced_matrix
```

```
)]: array([[ 1.45586248, -0.27582086,  0.37396133, -0.40363757, -0.02683407],
           [-0.76026193,  1.41039206, -0.58899203, -0.54069655, -0.05518246],
           [ 0.64554246,  0.80814631,  0.28605413, -0.19759383, -0.02740415],
           [ 1.05259116, -0.13394608,  0.42950008, -0.17605828,  0.05435404],
           [-0.86481167,  0.24093562,  0.25081374,  0.0044762 , -0.36011693],
           [ 0.95890552,  0.14493624,  0.26758034, -0.11743914,  0.03047041],
           [ 0.24339145,  0.67785388,  0.61701737,  0.26107987,  0.10286472],
           [ 1.16716483,  0.27585453, -0.51204512,  0.07999155,  0.02726483],
           [-0.06887999, -0.26678713,  0.06377684, -0.17974892, -0.11237527],
           [-0.53656857,  0.65071252,  0.15995141,  0.35819865, -0.42737757],
           [-0.03787891,  0.18644935, -0.37591243, -0.17707942,  0.49310743],
           [ 1.53750991,  0.31814328, -0.44624779,  0.00605393, -0.11690922],
           [-0.1366337 ,  0.17236179, -0.12685249, -0.094945  , -0.0366153 ],
           [-0.49355042,  0.25760201,  0.4173887 , -0.1599658 ,  0.18361361],
           [-0.39663676,  0.0162418 , -0.09866127,  0.38735235, -0.30651714],
           [ 0.6145641 ,  0.75376716,  0.01086451,  0.11121983, -0.38183506],
           [-0.46857292, -0.15159354, -0.01376415, -0.45584708,  0.08461782],
           [ 1.01337061, -1.01072199, -0.32942458,  0.19431559, -0.03413698],
           [ 0.60804424,  0.38843985,  0.05844938,  0.38322445,  0.36034242],
           [-0.51743613,  0.19936353, -0.52800452,  0.09256916,  0.26590662],
           [-0.74993097,  0.15733   , -0.1722546 , -0.03640044,  0.11011627],
           [-1.04045825, -0.43321214,  0.29946675,  0.19640369, -0.19970647],
           [-0.7802847 , -0.27627616,  0.2478252 , -0.26452393, -0.07209778],
           [-0.39200422,  0.31531267,  0.71460787,  0.49194075,  0.50556918],
           [-0.33401134,  0.13331964, -0.01953751, -0.20554373,  0.00326224],
           [-1.12431507, -0.1378116 , -0.34044036,  0.34724385,  0.10167105],
           [ 0.78479412, -0.61025396, -0.44888055,  0.56534645, -0.06751249],
           [ 0.75327574, -1.03745512,  0.17831713, -0.16544522, -0.33691987],
           [-0.0530292 ,  0.26170045, -0.24880153,  0.11923267, -0.23833402],
           [ 0.36910461, -0.59636407, -0.08484361, -0.09242949,  0.51576225],
           [-0.87358943, -0.6807066 ,  0.02907605,  0.0819987 ,  0.06269127],
           [-1.43909937, -0.89303036, -0.16958962, -0.17847542,  0.00791462],
           [-0.13616767, -0.86488308,  0.09960131, -0.23481786, -0.10965399]])
```

```
.]: PCA_Variance = pca.explained_variance_ratio_
```

4.3. Subtask 3: Results of PCA

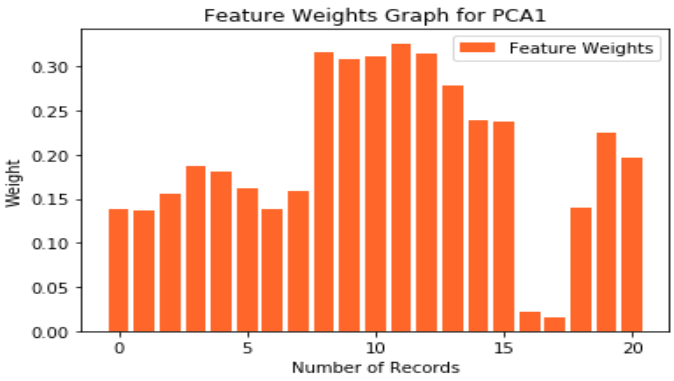


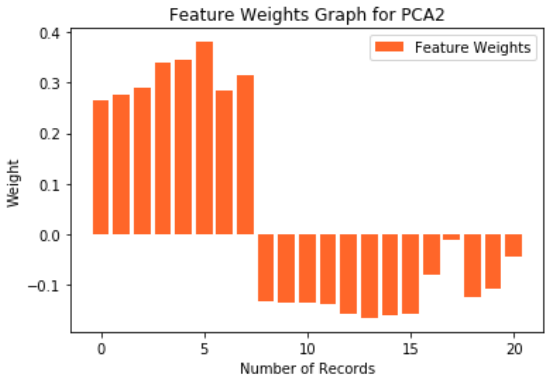
We took the top five eigenvalues from the PCA and calculated their ratio and plotted them on a graph like the one seen above. It shows that the top five latent features obtained contributes to approximately 93% of the total variance of the data. Hence it is rather sufficient to represent the data as a combination of five features rather than twenty.

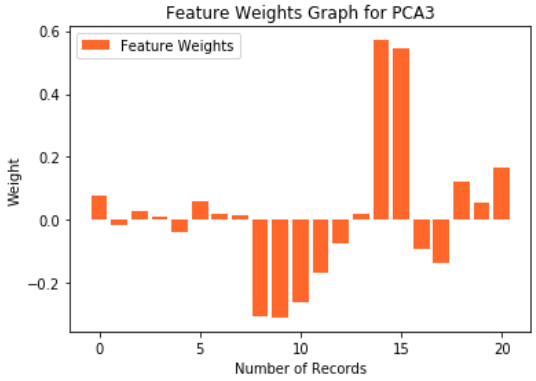
4.4 Explanation for Choosing the top 5 features.

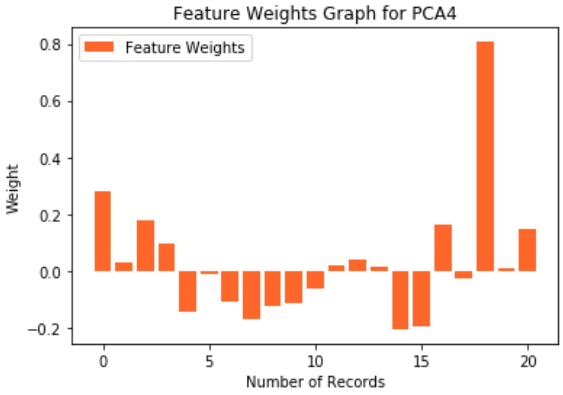
The weightage of each feature after PCA can be visualized from the below histogram. As we can see, the histogram shows 5 different peak bars in each plot that corresponds to the maximum weighted feature that can be used as a feature.

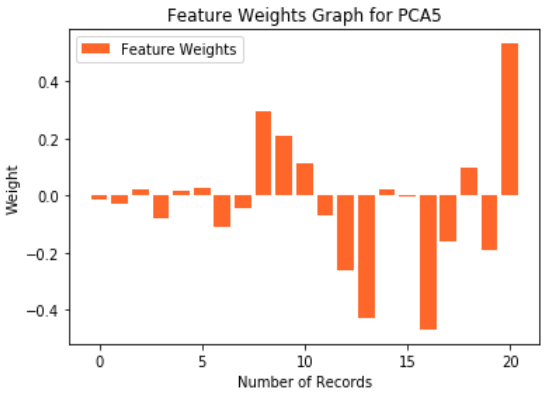
In graph 1 below, the fourth feature of DWT has maximum weightage and can be used as an important feature for meal detection. Likewise, in graph 2, 6th feature of FFT tops. In graph 3, CGM velocity with respect to acceleration is the maximum. In graph 4, AUC can be used. Finally, in graph 5, Log normal distribution standard deviation is the maximum.



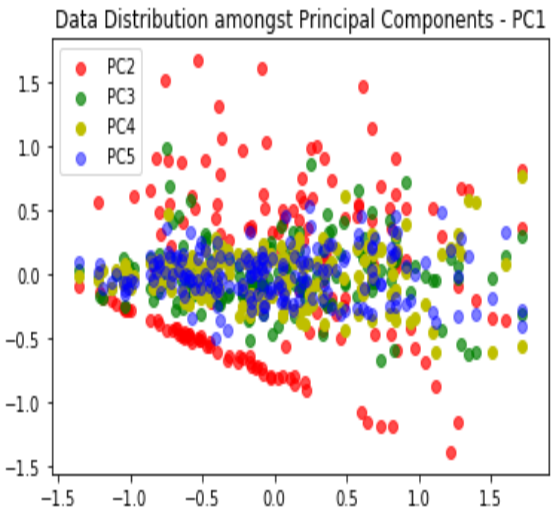


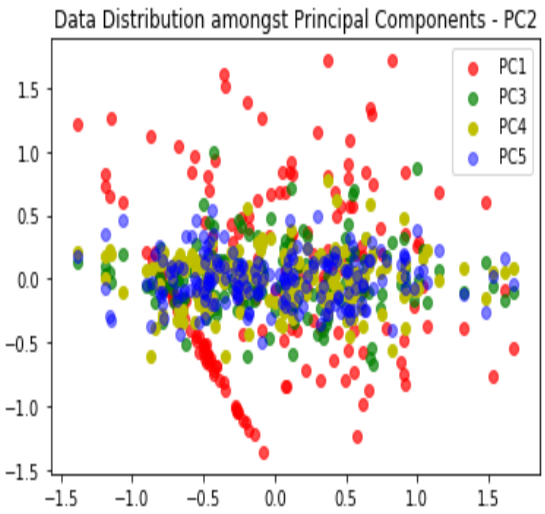


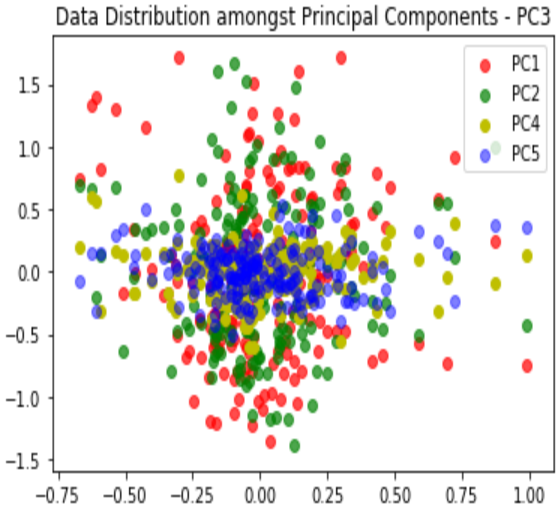


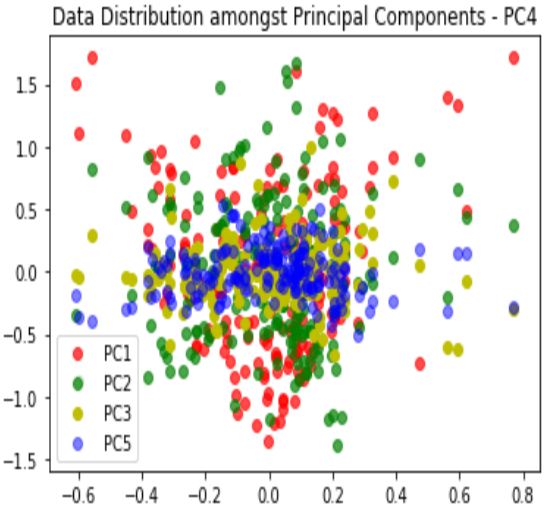


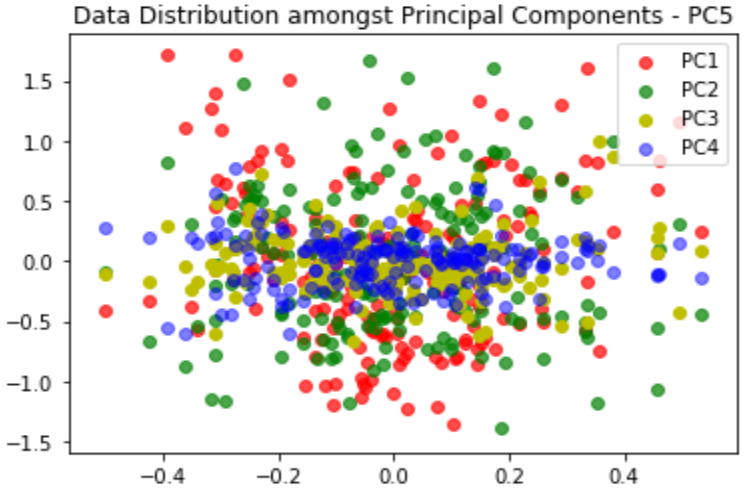
The principal components contribute to the bulk of variance in the entire dataset. We tried to visualise it by plotting the spread of data points along a principal component on x-axis against all other principal components for all patients. As seen in the below graphs, the principal component 1 indicated in red color has the data spread out the maximum indicating high variance and higher discrimination power. Though the variance decreases as we move down to other principal components, it can be shown that the distribution of data among the five principal components is varied enough to have higher discrimination power and hence can result in better analysis and prediction without errors.


Data Distribution amongst Principal Components - PC1
PC2
PC3
PC4
PC5


Data Distribution amongst Principal Components - PC2
PC1
PC3
PC4
PC5


Data Distribution amongst Principal Components - PC3
PC1
PC2
PC4
PC5


Data Distribution amongst Principal Components - PC4
PC1
PC2
PC3
PC5


Data Distribution amongst Principal Components - PC5
PC1
PC2
PC3
PC4