CSE 572 Data Mining Project - Phase 1 Report

Dr. Ayan Banerjee
Ira A. Fulton Schools of Engineering
Arizona State University

Submitted by:
Ganesh Ashok Yallankar (1217797087)

Introduction:

The aim of the project is to predict the meal intake time for a given patient. This phase of the project deals with finding different types of features for the data available. We have data readings from simulations (and patient readings) which gives the continuous glucose monitorings, basal and bolus insulin readings with respect to the time. The data is a time series based data.

Input:

Five cell arrays:

- **a)** The first cell array has tissue glucose levels every 5 mins for 2.5 hrs during a lunch meal. The data starts from 30 mins before meal intake as continues up to 2hrs after the start of meal consumption. There are several such time series for one subject.
- **b)** The second cell array has timestamps of each time series in the first cell array.
- **c)** The third cell array has insulin basal infusion input time series at different times during the 2.5 hr time interval.
- **d)** The fourth cell array has time stamps for each basal or bolus insulin delivery time-series.
- **e)** The fifth cell array has an insulin bolus infusion input time series at different times during the 2.5 hr time interval.

Steps:

1. Preprocessing the data.
2. Extracting 4 different types of time series features from only the CGM data cell array and CGM timestamp cell array.
3. Show and justify the features from the previous step.
4. Create a feature matrix where each row is a collection of features from each time series.
5. Perform Principal Component Analysis on the above feature matrix and derive the new feature matrix and plot them.
6. Justify the top five features from the above PCA feature matrix.

1. Data Preprocessing

On the first glace over the data, I could notice the missing values and some additional values in the data that is:

- Some data readings have more values.
- Some data readings are missing values.

To minimize the missing data and standardize the data, we made the following changes in data preprocessing:

- Since there were more than 30 readings, Only the first 30 readings are selected and others are discarded.
- For missing data, If data is missing for more than 5 data points, they are discarded, otherwise, the missing data points are extrapolated.

2. Feature Extraction

I chose the following features for extraction:

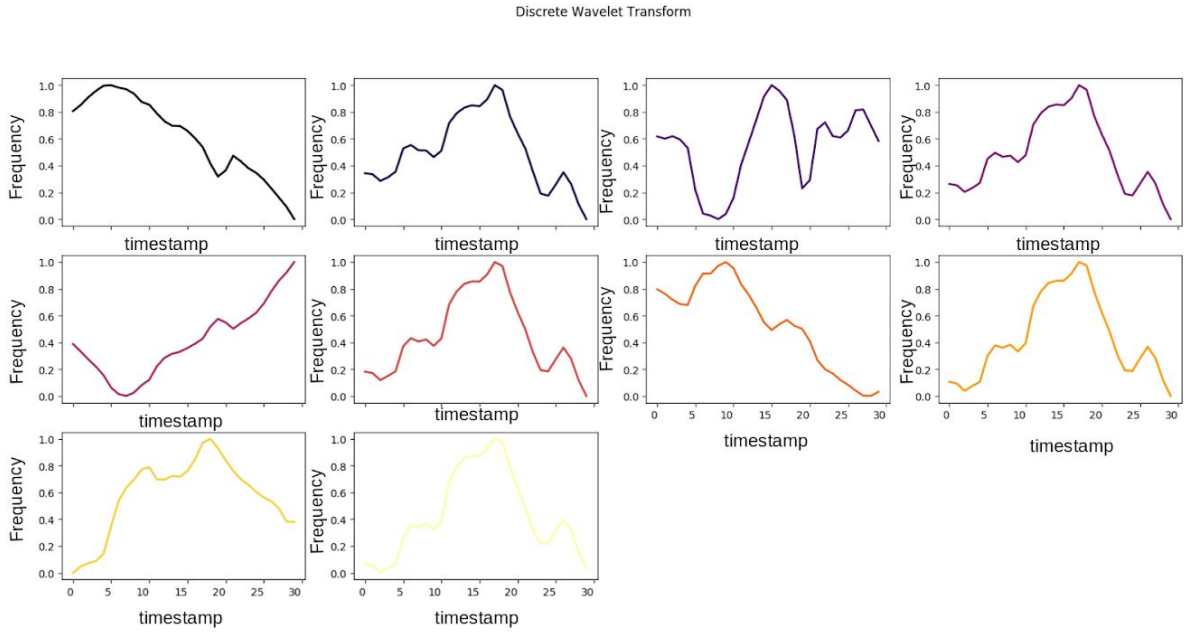
1. Discrete Wavelet Transform.
2. Discrete Fourier Transform.
3. Power spectral density.
4. Time series features.

3. Features choice and Intuition

Discrete Wavelet Transform

Why did I choose the feature:

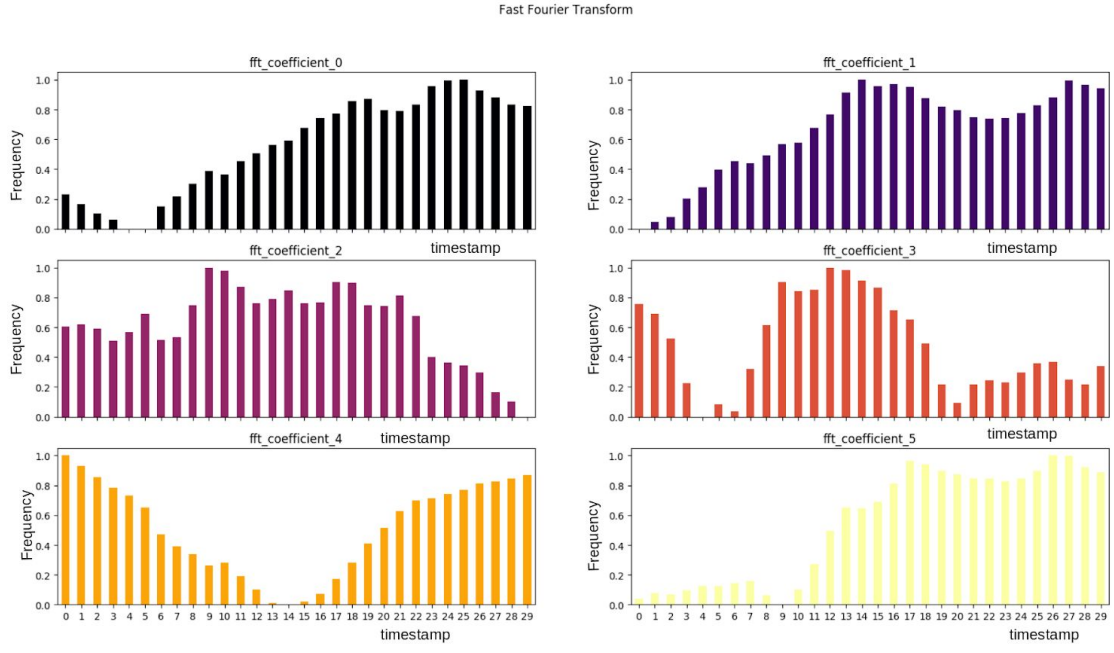
In the discrete wavelet transform, the wavelets are discretely sampled. A key advantage is a temporal resolution: it captures both frequency and location information (location in time). As our input data is in the form of time series this feature suits our requirement. Here, the data is normalized between 0 and 1. And we get 10 features from DWT, wherein 5 are discrete components and 5 approximate components.



Discrete Fourier Transform

Why did I choose the feature:

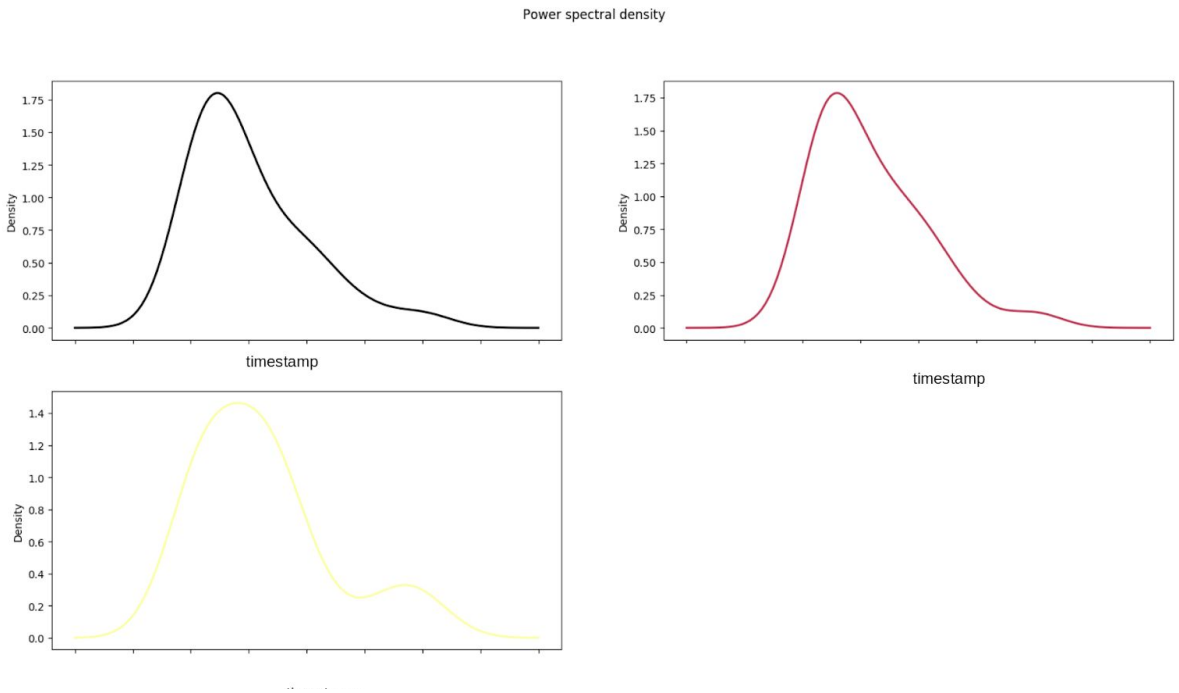
Discrete Fourier Transform converts equally-spaced samples of data into a same-length sequence of equally-spaced samples in terms of frequency. Since the input data is equally spaced for every 5 mins, this feature fits well. I used the Fast Fourier Transform algorithm to get the coefficients. Data is normalized between 0 and 1.



Power Spectral Density

Why did I choose the feature:

Power spectral density describes how the power of a signal or time series is distributed over frequency, compared to Fourier Transform, this feature can be used over continuous time-series data. I used PSD to get three different components from the data. Data is again normalized between 0 and 1.



Time-series Features

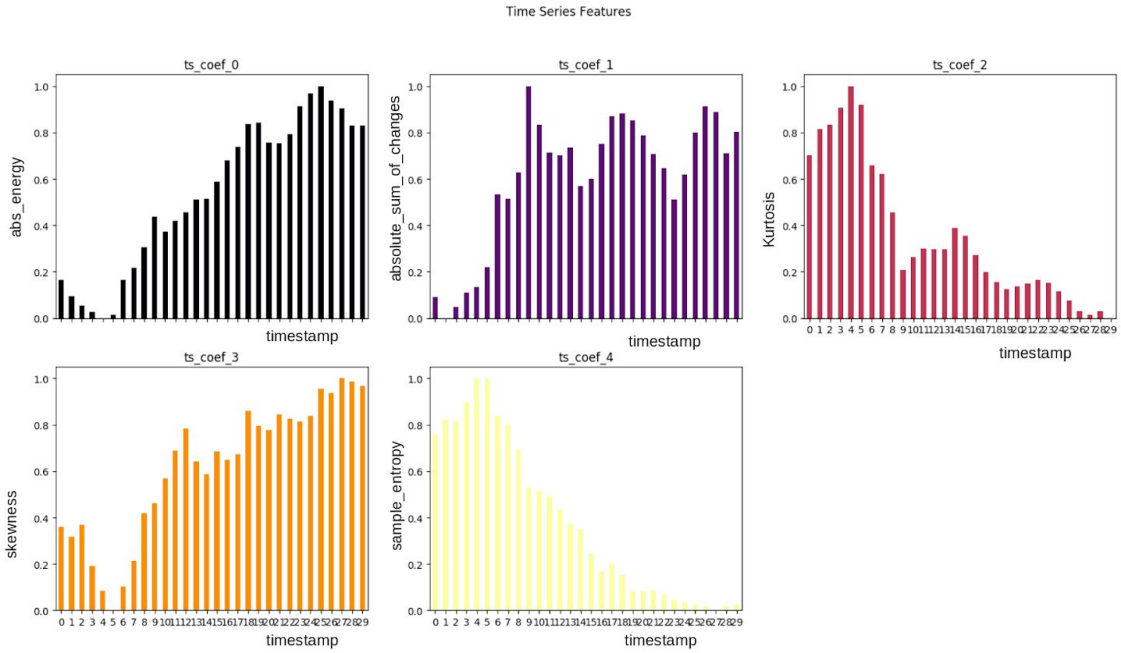
Why did I choose the feature:

Here, I have considered various characteristics of time series data namely -

- **Absolute Sum of Changes:** Shows rate of change in time-series over a period of time. Here, High values depict glucose spikes.
- **Absolute Energy:** This characteristic shows the increase in glucose value. The area under the curve gives the change.

- **Kurtosis:** Is a measure of the "tailedness" of the probability distribution of a real-valued random variable. We can capture sharp spikes in glucose with this.
- **Skewness:** It is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. We can get a measure of spikes in the time range.
- **Sample Entropy:** It is used for assessing the complexity of time-series signals. We can find the complexity of the data.

Data is normalized between 0 and 1 here.



4. Feature Matrix

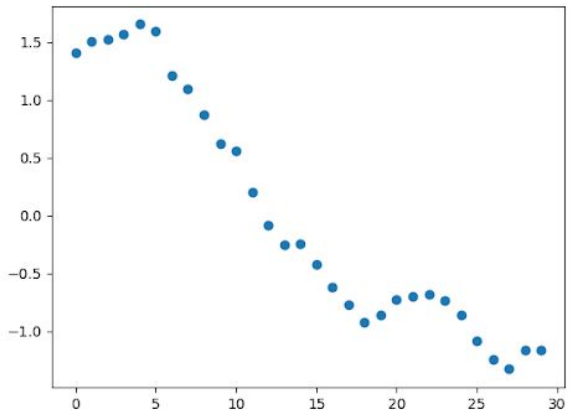
The feature matrix is generated from the above-mentioned features and stored in the file - “all_features.csv”. Here we can find time series data for 24 features and 30 time-series data. Resulting in a feature matrix of 30*24. The data in the matrix is normalized between 0 and 1.

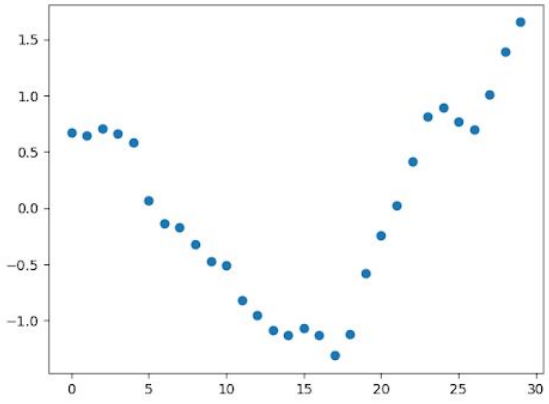
5. PCA Matrix and top 5 features

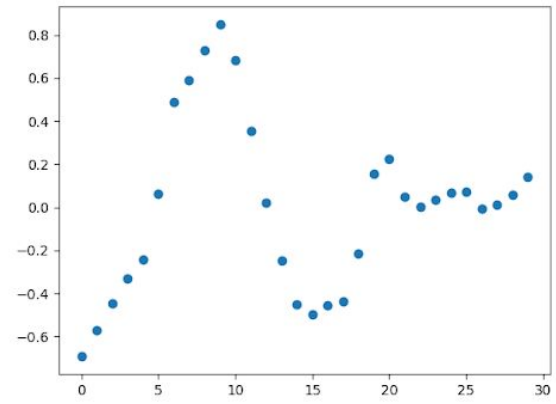
Principal components analysis is performed on the feature matrix (30*24) from the previous step using *sklearn.decomposition.PCA module* in python. The resultant matrix is stored in the "pca_all_features.csv" file. We choose 5 components from PCA. The top 5 features are:

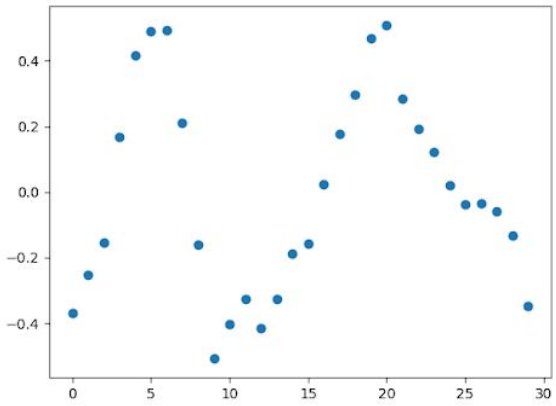
1. fft_coefficient_5.
2. fft_coefficient_0.
3. ts_coef_0.
4. ts_coef_3.
5. fft_coefficient_1.

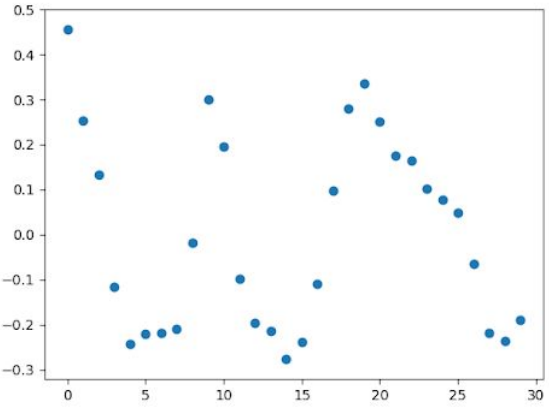
The following are the plotted graph for the top 5 features from PCA.







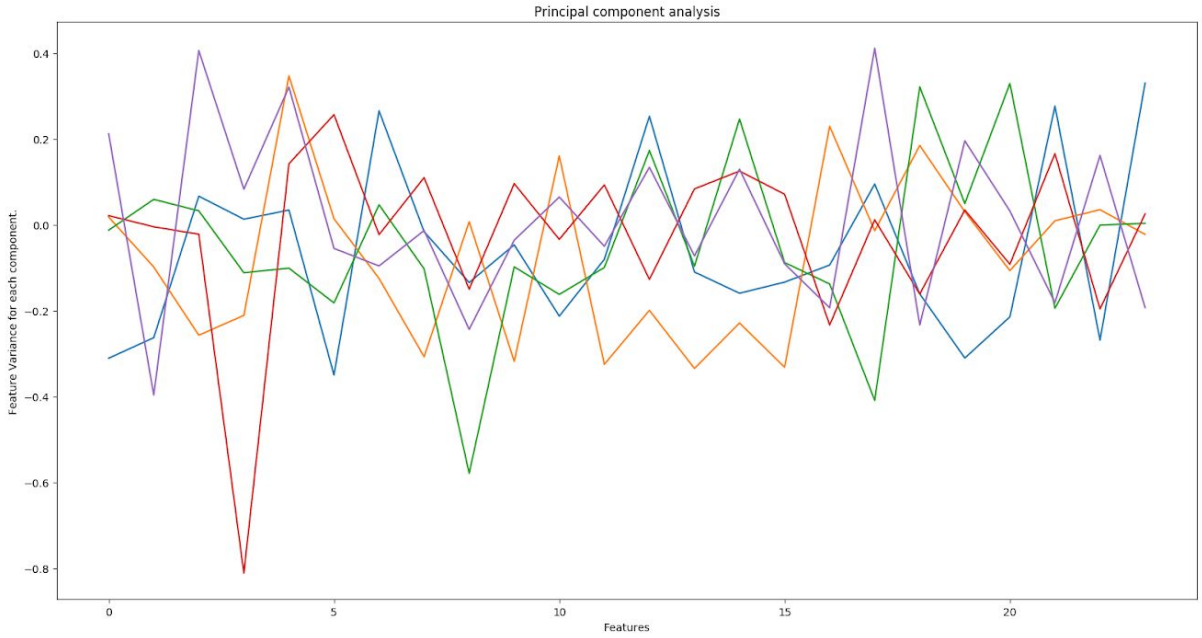




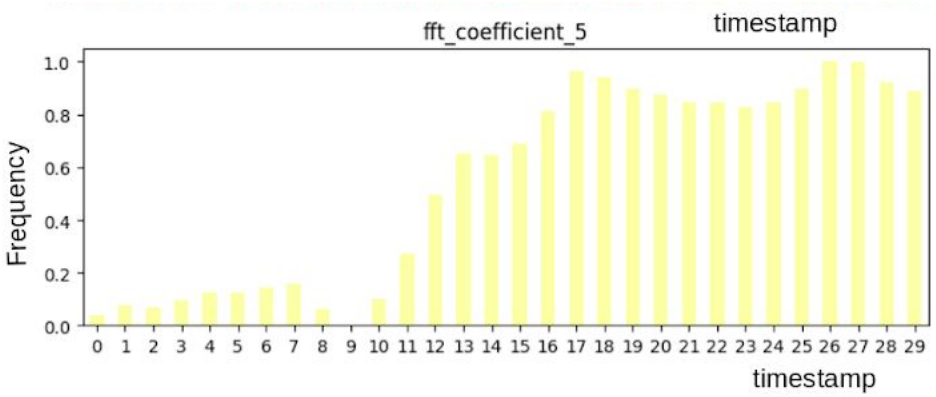
PCA produces a transformed feature set with five orthogonal directions of maximum variance from the original feature matrix.

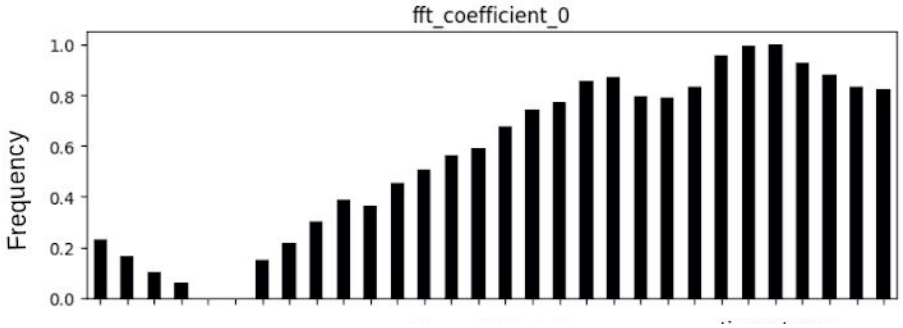
6. Justification for the top 5 features from PCA

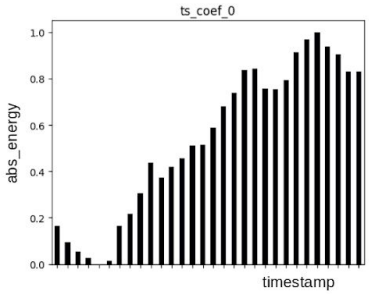
The below graph depicts the variance of 24 features. Its a plot of all five components from PCA.

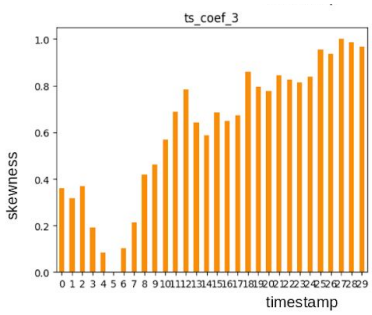


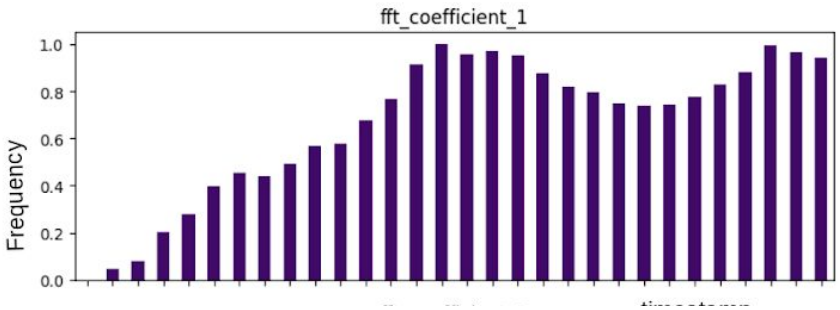
The features in the above graph are: "fft_coefficient_5", "fft_coefficient_0", "ts_coef_0", "ts_coef_3", "fft_coefficient_1".
Plots of all five features are:


fft_coefficient_5
timestamp
Frequency
timestamp


fft_coefficient_0
Frequency


ts_coef_0
abs_energy
timestamp


ts_coef_3
skewness
timestamp


fft_coefficient_1
Frequency

The above features are chosen as they are mostly scattered in component 1 and contributing to the variance of the component.
Discrete Fourier tranform features provides frequent peaks in given data hence "fft_coefficient_5", "fft_coefficient_0", "fft_coefficient_1" are chosen.
Absolute Energy shows spikes in glucose levels that is the feature - "ts_coef_0". Even Skewness provides the spikes that happened during the beginning and end of time series - "ts_coef_3". Hence these features are chosen.