PROJECT 1 REPORT

CSE 572: Data Mining *(Spring'20)*

Atin Singhal

ASU ID# 1217358454

Tasks

1. Extract 4 different types of time series features from only the CGM data cell array and CGM timestamp cell array.

Before using the Feature Extraction Techniques, the data is being preprocessed as follows:

i. Since majority of the files have data in the first 30 columns, only initial 30 columns are chosen for all the patients.
ii. If a row contains less than 95% of data, it is dropped. For example, if a meal instance has 30 columns. 95% of 30 means 28 data points. So, if any row has more than 2 NaN values, it is dropped.
iii. Data interpolation is done using pandas df.fillna() function on the remaining data. The interpolation methods used are 'pad' and 'bfill'. 'bfill' is used to fill the gap with the next valid observation. 'pad' is used to propagate the last valid observation in the gap.

After preprocessing, the following features are extracted:

a. Fourier Transform (FFT)
b. Root Mean Square (RMS)
c. Power Spectral Density (PSD)
d. Discrete Wavelet Transform (DWT)

2. For each time series explain why you chose such feature.

This section includes the intuition for selecting the features.

2.1 Fourier Transform (FFT)

Fourier Transform is used for filtering data. The intuition behind selecting Fourier transform was that it will help by removing the noise and help in identifying meaningful patterns from the data. For this data, top 4 values of the coefficient are extracted and passed in the Feature Matrix.

Function used: scipy.fftpack.fft(series)

2.2 Root Mean Square (RMS)

The RMS function provides the Root Mean Square value of the array passed into it. With the help of RMS, one can tell the magnitude of a set of values & can get a sense for typical size of values being recorded by the sensor.

Function used: Wrote the code manually in Python.

2.3 Power Spectral Density (PSD)

Power Spectral Density describes the distribution of power into frequency components composing that signal. It can show which variations are strong and which are weak. So, the intuition was that using PSD would help in identifying variations in the CGM series data.

Function used: scipy.signal.welch(series)

2.4 Discrete Wavelet Transform (DWT)

Discrete Wavelet Transform decomposes raw sensor data into wavelets. It is used for reducing noise from the data. The intuition behind choosing this feature was that it will help in identifying any patterns in the data by removing the noise. Top 7 values are chosen for the feature matrix.

Function used: pywt.dwt()

3. Show values of each of the features and argue that your intuition in step b is validated or disproved.

Values of Features:

3.1 Fourier Transform (FFT)

The values obtained for Fourier Transform were as follows:

Table 1: Top 4 Values for FFT for Patient 1

FFT1	FFT2	FFT3	FFT4
5105	1296.97412	1296.97412	322.133334
9207	787.565762	787.565762	181.919398
5858	1200.81596	1200.81596	225.159411
4587	1090.84366	1090.84366	275.813133
4305	344.707806	344.707806	79.0347041
4831	1095.14255	1095.14255	213.16176
4250	672.400058	672.400058	160.299966
4611	1126.41191	1126.41191	242.156063
4377	627.159961	627.159961	199.299606
4604	464.550406	464.550406	110.315508
5260	1051.92754	1051.92754	232.662365
4459	1340.74311	1340.74311	292.511895
4691	741.971605	741.971605	135.979846
4553	850.16052	850.16052	114.324581
3058	206.945465	206.945465	146.958491
4461	898.533665	898.533665	239.625908
5289	768.84098	768.84098	139.405204
4472	1258.7084	1258.7084	324.153668
3746	781.372855	781.372855	241.095015
4552	778.695464	778.695464	81.4063765
4652	619.967008	619.967008	104.303376
2740	95.5825707	95.5825707	53.5931775
4723	578.215218	578.215218	87.6257076
3331	522.995056	522.995056	195.813742
4799	765.102439	765.102439	158.481606
3110	376.901393	376.901393	73.4306233
2943	224.582789	224.582789	143.905167
4199	230.302005	230.302005	172.471433
4286	544.830606	544.830606	178.925095
4255	653.509453	653.509453	98.5875232
2996	291.009391	291.009391	98.7142744
3619	179.318928	179.318928	111.474824
3124	582.429686	582.429686	194.381838

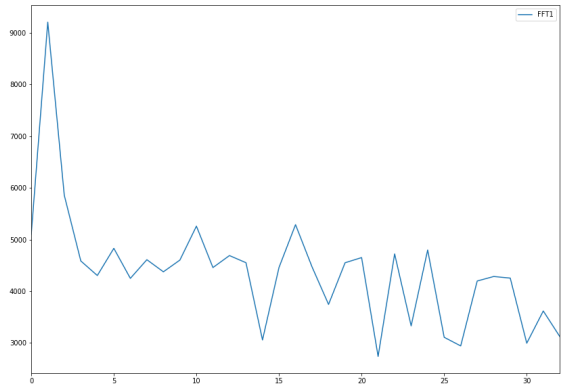

Figure 1: FFT1 Plot for Patient 1

The intuition seems to be validated. After applying FFT, the data seems filtered. There is a high peak in the graph and then the value drops significantly. I think this variation can be studied to see if meal was taken or not.

3.2 Root Mean Square (RMS)

The values obtained for RMS were as follows:

Table 2: RMS Values for Patient 1

RMS
183.060919
309.343768
204.095566
162.799365
144.482871
170.356978
145.683447
164.133787
149.662175
155.14982
182.875732
163.409404
160.709365
157.178349
102.806939
155.712235
180.399464
161.797404
131.432365
156.267079
157.985442
91.5332362
159.868592
114.208435
164.383393
105.255562
99.741165
141.27314
145.775169
145.386267
101.07423
121.081653
108.873627

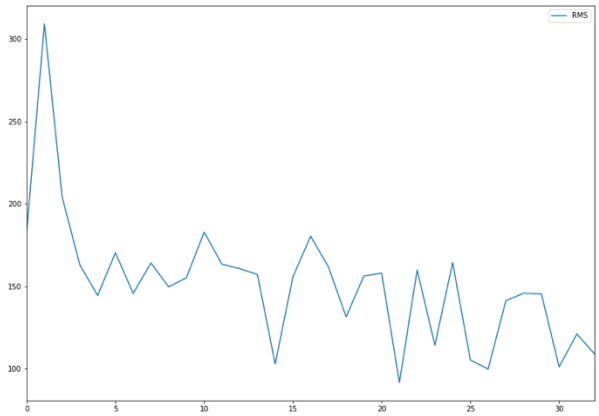

Figure 2: RMS Plot for Patient 1

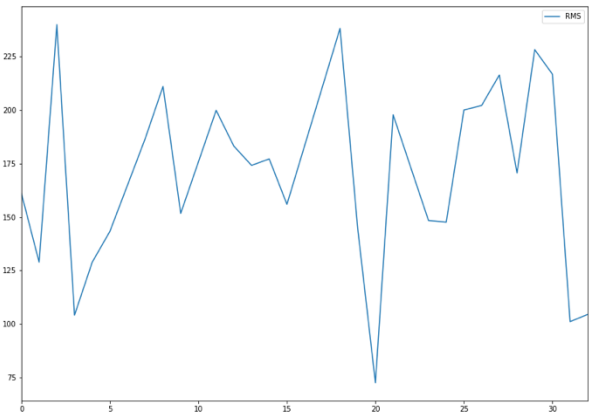

Figure 3: RMS Plot for Patient 2

The intuition seems to be validated. There is a high peak in the graph and then the value is dropped significantly. This could be used to identify if meal was taken or not.

3.3 Power Spectral Density (PSD)

The values obtained for PSD were as follows:

Table 3: Values for PSD for Patient 1

PSD
897.833131
11761.8774
0.02731798
928.551499
1250.83049
164.935709
0.00172279
1594.34567
0.06298605
4122.02942
108.821724
2877.64394
2.63909523
314.503791
0.34420339
1482.10456
87.8665075
1387.14486
171.650914
2701.5723
2409.68091
4.82638912
97.7691015
1204.01656
9.47949454
770.357554
167.185669
49.2802325
48.8341373
6881.03556
1201.68936
560.011524
1736.68761

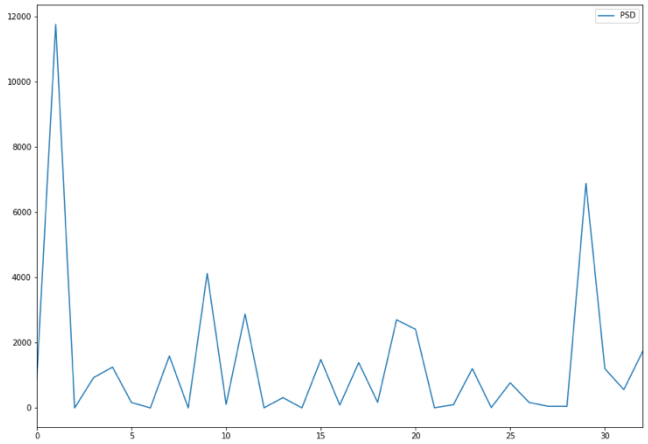

Figure 4: Power Spectral Density for Patient 1

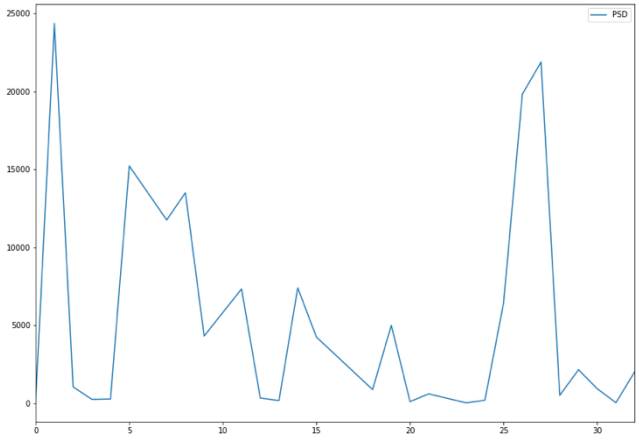

Figure 5: Power Spectral Density for Patient 2

There is variation shown in the graph. The strong and weak variation should be able to differentiate between the patient eating meal or not. There is a high peak in the graph and then the value is dropped significantly. The intuition seems to be validated. I think this could be used to identify if meal was taken or not.

3.4 Discrete Wavelet Transform (DWT)

The intuition seems to be validated. The variation in the graph (highs & lows) can denote the meal intake or no meal intake. There is a high peak in the graph and then the value drops significantly. So, I think that the data can be further be analyzed using this feature to make predictions.

The values obtained for DWT were as follows:

Table 4: Value for DWT

DWT1	DWT2	DWT3	DWT4	DWT5	DWT6	DWT7
363.452886	366.281313	357.796031	344.361002	328.097546	306.884343	283.549819
376.887914	406.586399	433.456457	453.255447	478.004184	486.489465	490.732106
331.63308	357.088924	376.180808	385.373196	378.302128	358.503138	328.804653
324.562013	309.71277	304.055916	300.520382	294.863528	287.085353	253.144228
193.747258	197.989899	211.424928	220.617316	222.738636	226.981277	231.931024
302.641702	303.348809	299.106168	310.419877	316.783838	304.055916	289.206674
234.052345	241.830519	247.487373	260.922402	263.750829	251.022907	229.809704
323.854906	324.562013	320.319372	313.955411	296.984848	273.650324	241.830519
263.750829	260.922402	257.386868	250.315801	248.901587	247.487373	237.587878
188.797511	192.333044	199.404112	210.010714	234.052345	260.922402	270.821897
309.71277	311.834091	317.490945	324.562013	321.733585	320.319372	308.298557
323.854906	335.168614	342.239682	333.047294	311.126984	276.478751	229.102597
267.286363	274.357431	275.771645	277.892965	283.549819	276.478751	255.265548
225.567063	255.265548	277.892965	287.79246	289.91378	280.014285	263.043723
171.826948	168.291414	168.998521	162.63456	148.492424	137.178716	138.592929
295.570635	303.348809	302.641702	288.499567	267.99347	245.366053	224.15285
306.884343	307.59145	305.470129	307.59145	310.419877	301.934596	283.549819
110.308658	98.9949494	89.8025612	97.5807358	121.622366	156.270599	188.090404
255.972655	253.144228	231.931024	219.910209	222.738636	228.39549	220.617316
197.989899	220.617316	248.901587	279.307179	289.91378	280.014285	270.821897
185.261977	212.132034	248.19448	276.478751	284.256926	259.508189	259.508189
130.814755	131.521861	132.228968	128.693434	127.279221	127.986327	134.350288
218.495995	244.658946	275.064538	287.085353	279.307179	262.336616	242.537626
110.308658	154.149278	190.211724	205.060967	202.939646	191.625938	195.161472
255.972655	279.307179	294.863528	295.570635	282.135606	277.892965	265.165043
138.592929	146.371104	153.442172	164.048773	178.898016	188.090404	184.55487
100.409163	117.379726	143.542677	159.099026	162.63456	157.684812	146.371104
141.421356	170.412734	188.090404	185.261977	188.797511	202.232539	202.939646
263.043723	246.07316	238.294985	238.294985	243.244733	239.709199	231.931024
232.638131	217.081782	199.404112	188.797511	172.534055	152.027958	137.178716
110.308658	111.015765	120.91526	130.814755	135.057395	136.471609	137.178716
150.613744	164.75588	179.605122	180.312229	181.726443	187.383297	184.55487
229.102597	217.788889	208.5965	190.918831	173.948268	166.170094	148.492424

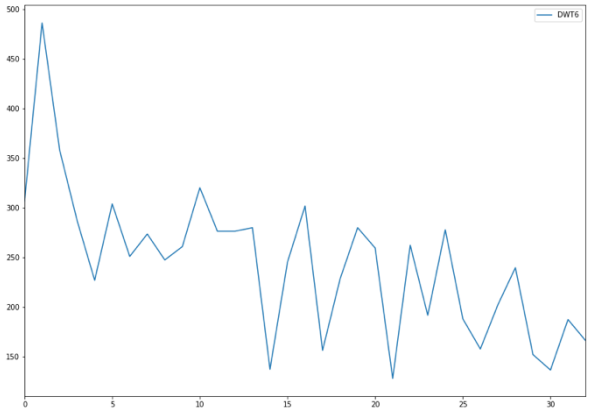

Figure 6: DWT Plot for Patient

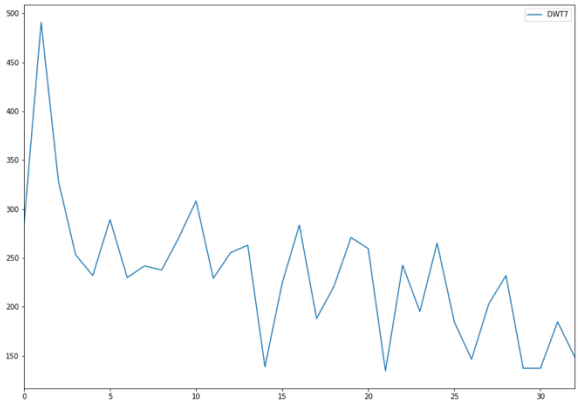

Figure 7: DWT Plot for Patient

4. Create a feature matrix where each row is a collection of features from each time series.

There are 33 time series and the feature length after concatenation of the 4 types of features is 13 (4 for FFT, 1 for RMS, 1 for PSD and 7 for DWT). Hence, the feature matrix size is 33x13. (This data is for Patient 1)

Below is a screenshot of the Feature Matrix that is passed to PCA.

Table 5: Feature Matrix

FFT1	FFT2	FFT3	FFT4	RMS	PSD	DWT1	DWT2	DWT3	DWT4	DWT5	DWT6	DWT7
5105	1296.97412	1296.97412	322.133334	183.060919	897.833131	363.452886	366.281313	357.796031	344.361002	328.097546	306.884343	283.549819
9207	787.565762	787.565762	181.919398	309.343768	11761.8774	376.887914	406.586399	433.456457	453.255447	478.004184	486.489465	490.732106
5858	1200.81596	1200.81596	225.159411	204.095566	0.02731798	331.63308	357.088924	376.180808	385.373196	378.302128	358.503138	328.804653
4587	1090.84366	1090.84366	275.813133	162.799365	928.551499	324.562013	309.71277	304.055916	300.520382	294.863528	287.085353	253.144228
4305	344.707806	344.707806	79.0347041	144.482871	1250.83049	193.747258	197.989899	211.424928	220.617316	222.738636	226.981277	231.931024
4831	1095.14255	1095.14255	213.16176	170.356978	164.935709	302.641702	303.348809	299.106168	310.419877	316.783838	304.055916	289.206674
4250	672.400058	672.400058	160.299966	145.683447	0.00172279	234.052345	241.830519	247.487373	260.922402	263.750829	251.022907	229.809704
4611	1126.41191	1126.41191	242.156063	164.133787	1594.34567	323.854906	324.562013	320.319372	313.955411	296.984848	273.650324	241.830519
4377	627.159961	627.159961	199.299606	149.662175	0.06298605	263.750829	260.922402	257.386868	250.315801	248.901587	247.487373	237.587878
4604	464.550406	464.550406	110.315508	155.14982	4122.02942	188.797511	192.333044	199.404112	210.010714	234.052345	260.922402	270.821897
5260	1051.92754	1051.92754	232.662365	182.875732	108.821724	309.71277	311.834091	317.490945	324.562013	321.733585	320.319372	308.298557
4459	1340.74311	1340.74311	292.511895	163.409404	2877.64394	323.854906	335.168614	342.239682	333.047294	311.126984	276.478751	229.102597
4691	741.971605	741.971605	135.979846	160.709365	2.63909523	267.286363	274.357431	275.771645	277.892965	283.549819	276.478751	255.265548
4553	850.16052	850.16052	114.324581	157.178349	314.503791	225.567063	255.265548	277.892965	287.79246	289.91378	280.014285	263.043723
3058	206.945465	206.945465	146.958491	102.806939	0.34420339	171.826948	168.291414	168.998521	162.63456	148.492424	137.178716	138.592929
4461	898.533665	898.533665	239.625908	155.712235	1482.10456	295.570635	303.348809	302.641702	288.499567	267.99347	245.366053	224.15285
5289	768.84098	768.84098	139.405204	180.399464	87.8665075	306.884343	307.59145	305.470129	307.59145	310.419877	301.934596	283.549819
4472	1258.7084	1258.7084	324.153668	161.797404	1387.14486	110.308658	98.9949494	89.8025612	97.5807358	121.622366	156.270599	188.090404
3746	781.372855	781.372855	241.095015	131.432365	171.650914	255.972655	253.144228	231.931024	219.910209	222.738636	228.39549	220.617316
4552	778.695464	778.695464	81.4063765	156.267079	2701.5723	197.989899	220.617316	248.901587	279.307179	289.91378	280.014285	270.821897
4652	619.967008	619.967008	104.303376	157.985442	2409.68091	185.261977	212.132034	248.19448	276.478751	284.256926	259.508189	259.508189
2740	95.5825707	95.5825707	53.5931775	91.5332362	4.82638912	130.814755	131.521861	132.228968	128.693434	127.279221	127.986327	134.350288
4723	578.215218	578.215218	87.6257076	159.868592	97.7691015	218.495995	244.658946	275.064538	287.085353	279.307179	262.336616	242.537626
3331	522.995056	522.995056	195.813742	114.208435	1204.01656	110.308658	154.149278	190.211724	205.060967	202.939646	191.625938	195.161472
4799	765.102439	765.102439	158.481606	164.383393	9.47949454	255.972655	279.307179	294.863528	295.570635	282.135606	277.892965	265.165043
3110	376.901393	376.901393	73.4306233	105.255562	770.357554	138.592929	146.371104	153.442172	164.048773	178.898016	188.090404	184.55487
2943	224.582789	224.582789	143.905167	99.741165	167.185669	100.409163	117.379726	143.542677	159.099026	162.63456	157.684812	146.371104
4199	230.302005	230.302005	172.471433	141.27314	49.2802325	141.421356	170.412734	188.090404	185.261977	188.797511	202.232539	202.939646
4286	544.830606	544.830606	178.925095	145.775169	48.8341373	263.043723	246.07316	238.294985	238.294985	243.244733	239.709199	231.931024
4255	653.509453	653.509453	98.5875232	145.386267	6881.03556	232.638131	217.081782	199.404112	188.797511	172.534055	152.027958	137.178716
2996	291.009391	291.009391	98.7142744	101.07423	1201.68936	110.308658	111.015765	120.91526	130.814755	135.057395	136.471609	137.178716
3619	179.318928	179.318928	111.474824	121.081653	560.011524	150.613744	164.75588	179.605122	180.312229	181.726443	187.383297	184.55487
3124	582.429686	582.429686	194.381838	108.873627	1736.68761	229.102597	217.788889	208.5965	190.918831	173.948268	166.170094	148.492424

5. Provide this feature matrix to PCA and derive the new feature matrix. Chose the top 5 features and plot them for each time series.

After providing the feature matrix to PCA the following new feature matrix is obtained:

Table 6: Feature Matrix obtained after applying PCA

FFT1	FFT2	FFT3	FFT4	RMS	PSD	DWT1	DWT2	DWT3	DWT4	DWT5	DWT6	DWT7
0.29213168	0.24573016	0.24573016	0.15816319	0.30105938	0.12823999	0.29605682	0.30905018	0.31237395	0.31323239	0.31432219	0.31116111	0.29550172
0.24618392	-0.4208747	-0.4208747	-0.5301728	0.18392746	0.40094758	-0.1222686	-0.068157	0.00361581	0.05884155	0.11381084	0.15506215	0.22222388
0.19878586	0.20815326	0.20815326	0.31779828	0.21792634	0.70089528	-0.1660027	-0.205701	-0.2400631	-0.2362388	-0.1885259	-0.1207261	-0.0257117
-0.2369235	-0.0976807	-0.0976807	-0.067036	-0.2304683	0.48112977	0.46390664	0.35651425	0.23693942	0.0938251	-0.0668315	-0.2343233	-0.4111688
-0.1342952	0.45187876	0.45187876	-0.7072666	-0.1031118	0.10053905	-0.1345419	-0.0966723	-0.0306243	0.08584866	0.09451534	-0.013358	-0.0879957

The shape of the new feature matrix obtained is 5x13. Below are the scatter plots obtained from the Top 5 features extracted using PCA. (This data is for Patient 1)

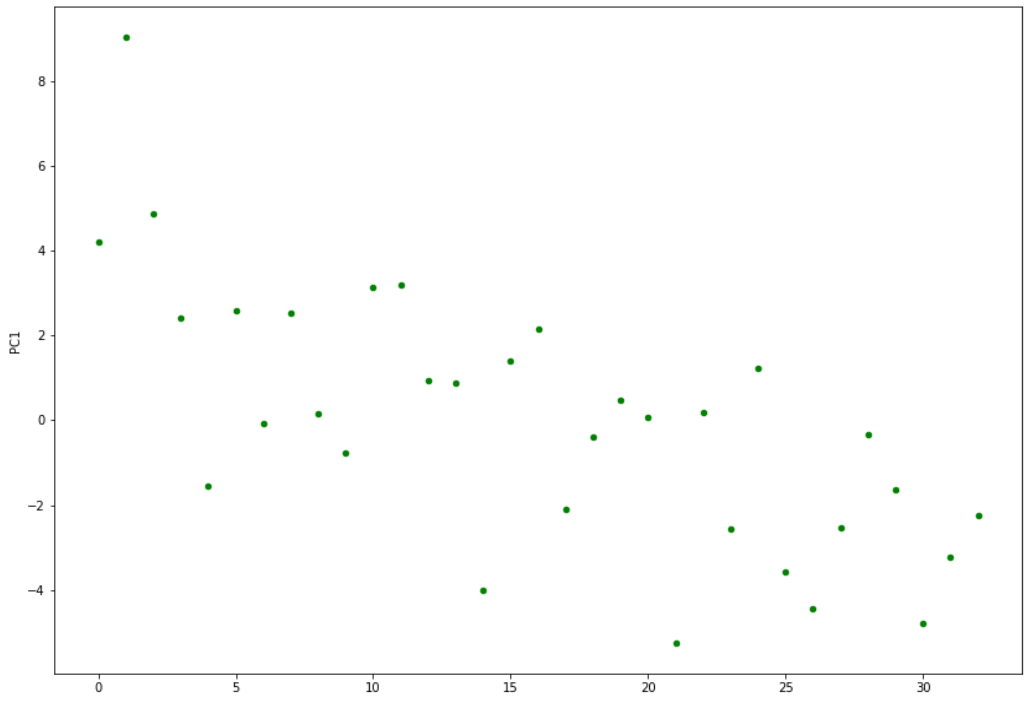

Figure 8: Scatter Plot for PCA1 (Patient 1)

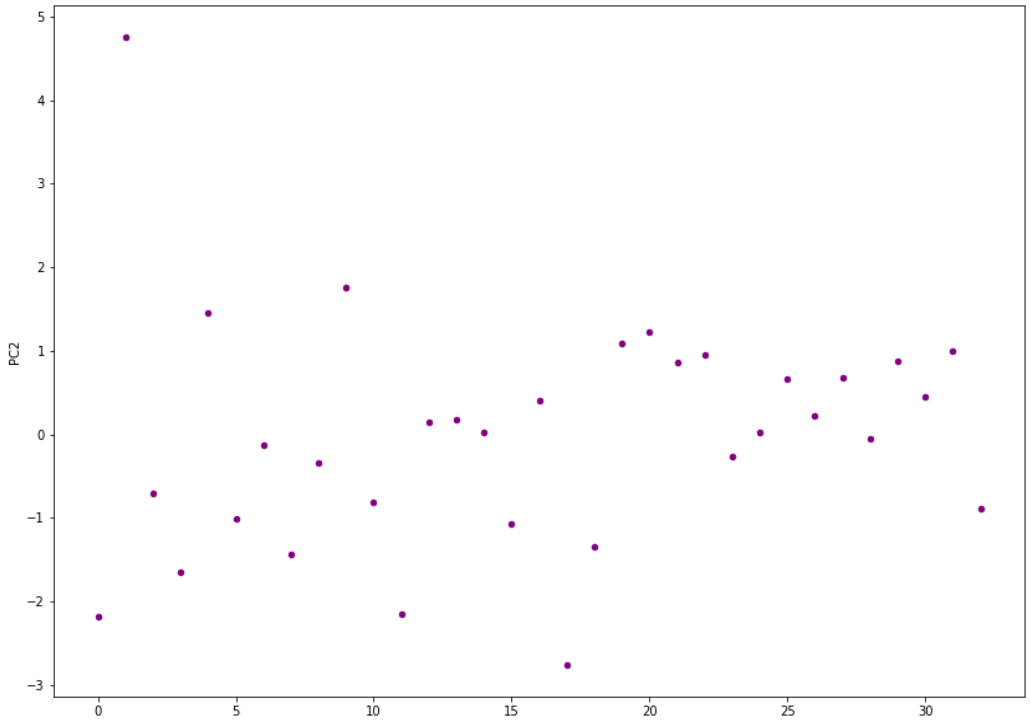

Figure 9: Scatter Plot for PCA2 (Patient 1)

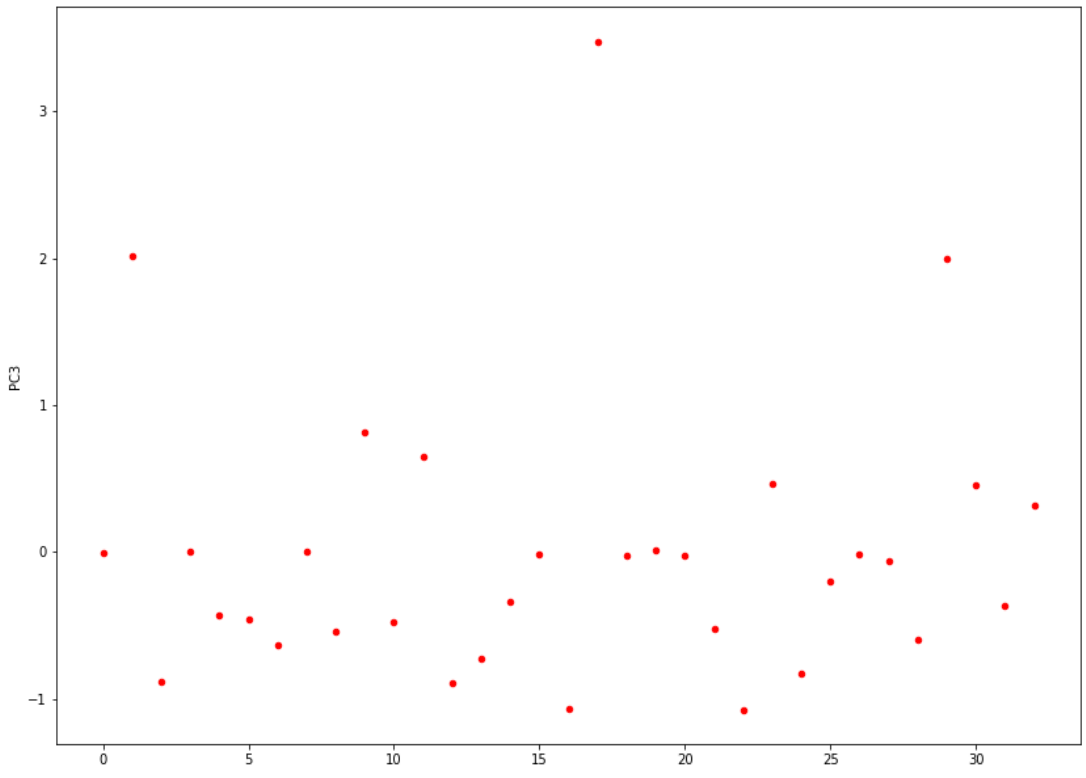

Figure 10: Scatter Plot for PCA3 (Patient 1)

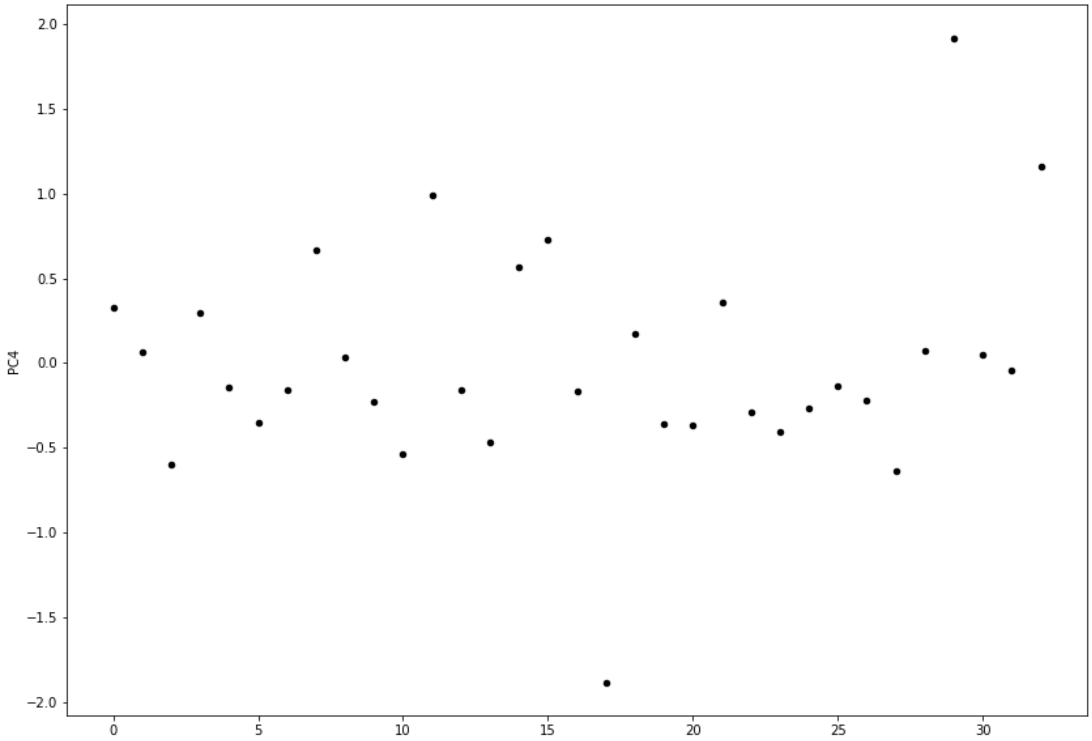

Figure 11: Scatter Plot for PCA4 (Patient 1)

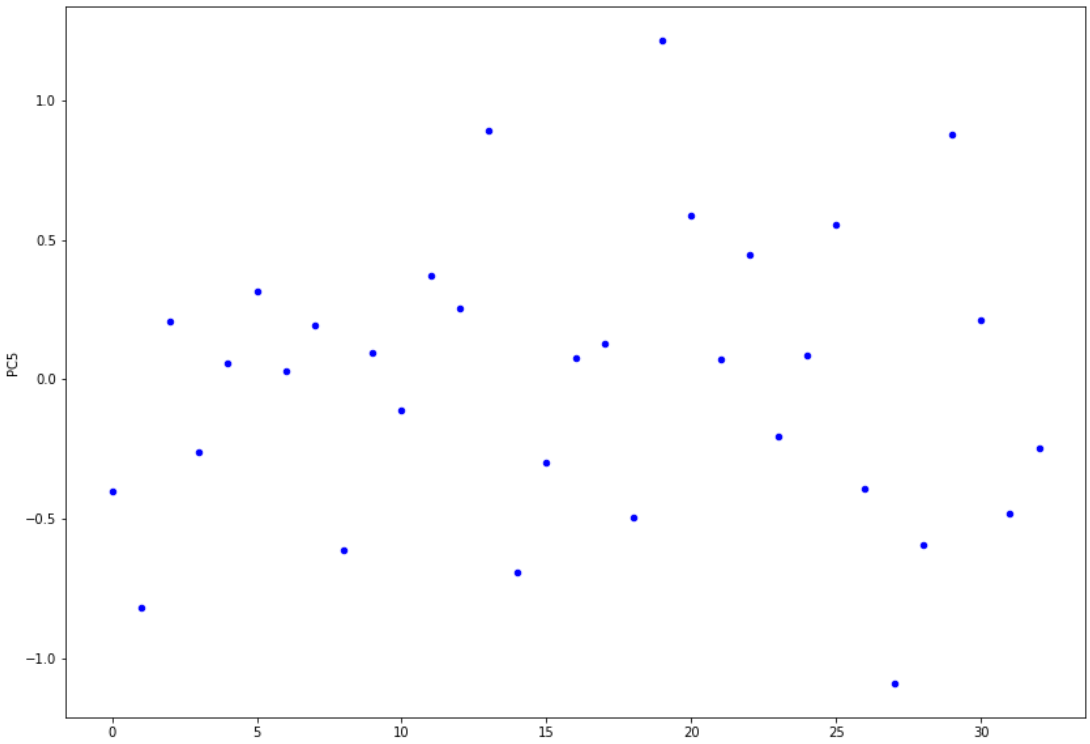

Figure 12: Scatter Plot for PCA5 (Patient 1)

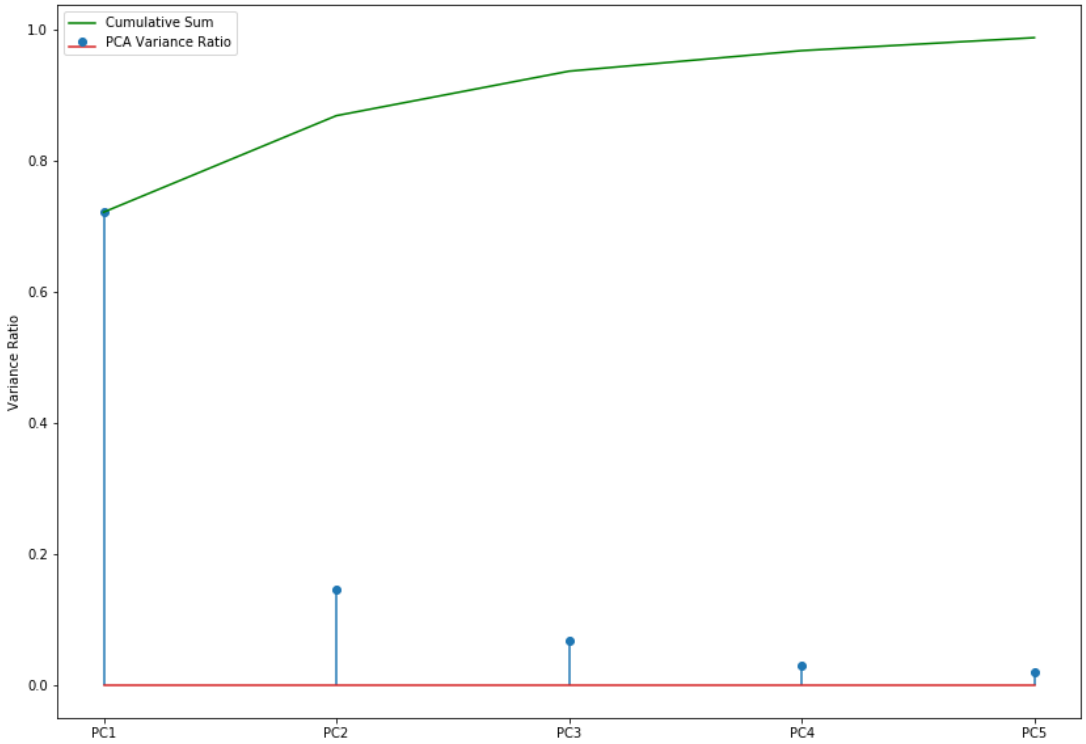

Figure 13: PCA v/s Variance Ratio (Patient 1)

From the obtained graphs for the top 5 components, we can see the data points on the PCA Plots are scattered. This means the features selected for this project help in dimension reduction. In this case, we can see that 98.77% of the original data can be represented using the Top 5 features instead of using the original data.

6. For each feature in the top 5 argue why it is chosen as a top five feature in PCA?

We can use a Heat Plot map to visualize the top features and the PCA components. Using this we can see which feature is significant.

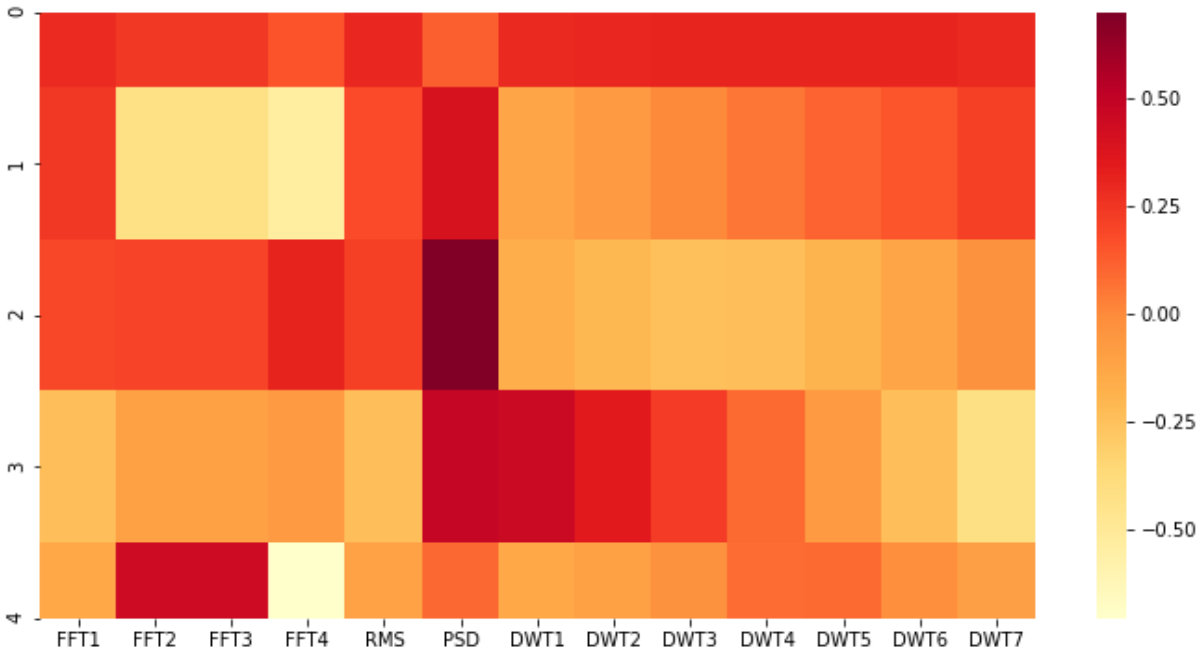

Figure 14: Heat Plot for New Feature Matrix (obtained after PCA)- Patient 1

Table 7: Table Mapping Most Significant Features to PCA Components (Patient 1)

PRINCIPAL COMPONENT#	SIGNIFICANT FEATURE
1	FFT, RMS, DWT
2	PSD
3	PSD
4	PSD
5	FFT

From the heat plot & table above, we can see that FFT, RMS & DWT are shown as significant features for PC1; PSD is shown as significant feature for PC2, PC3 & PC4; and FFT is shown as a significant feature for PC5.