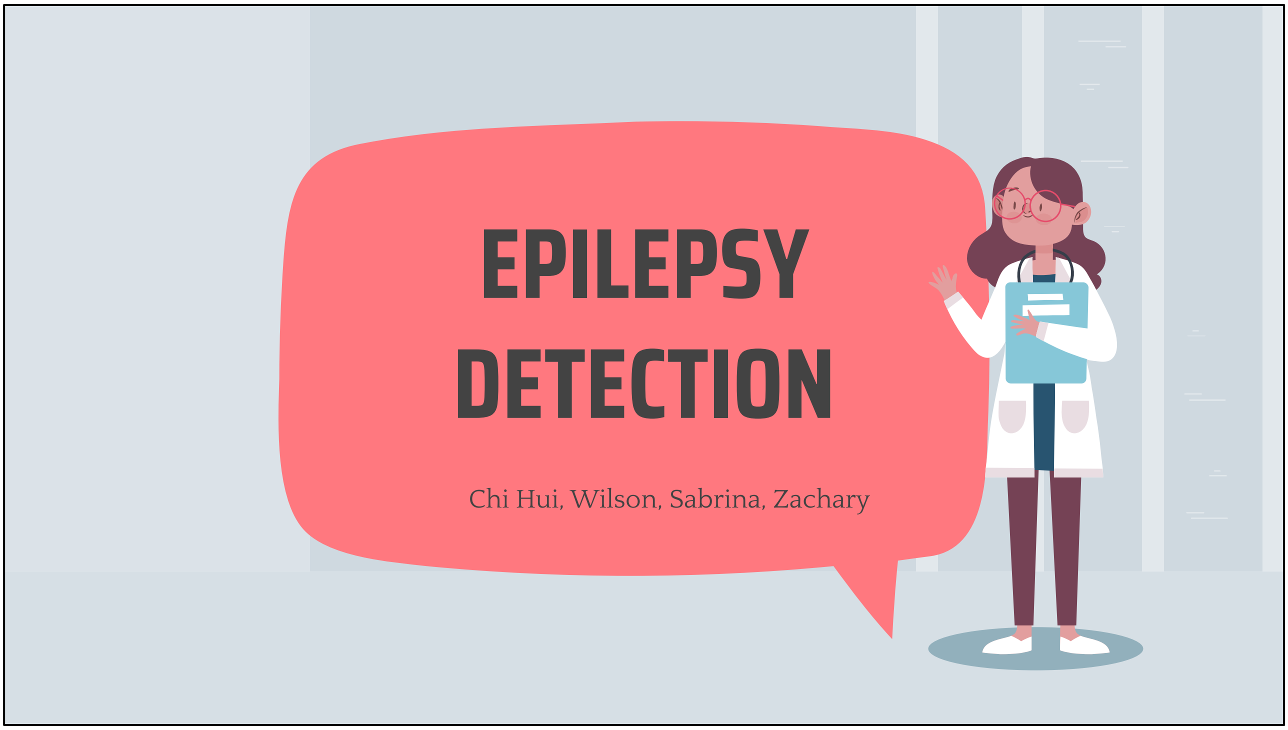


EPILEPSY DETECTION

Chi Hui, Wilson, Sabrina, Zachary

Good afternoon Prof and fellow classmates. I am Chi Hui and these are my teammates Wilson, Zachary, and Sabrina. We have chosen the Problem on Epilepsy Detection.

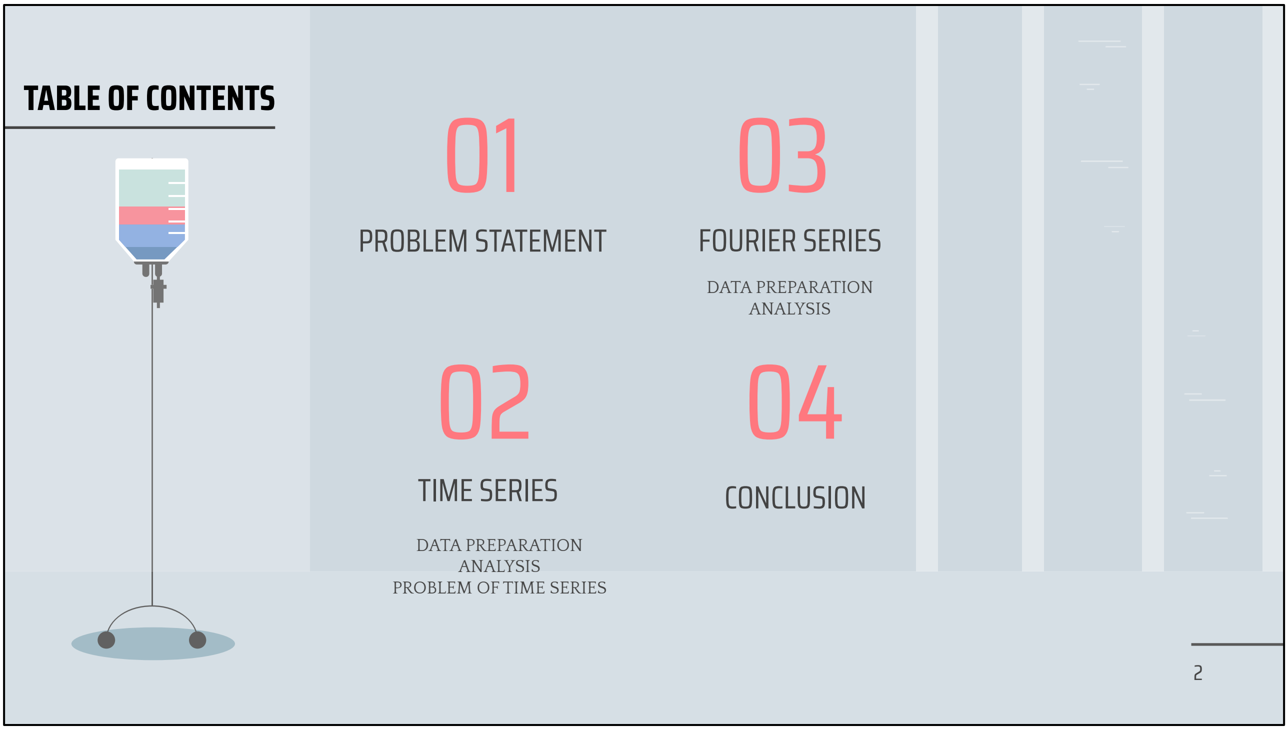


TABLE OF CONTENTS

01

PROBLEM STATEMENT

02

TIME SERIES

DATA PREPARATION
ANALYSIS
PROBLEM OF TIME SERIES

03

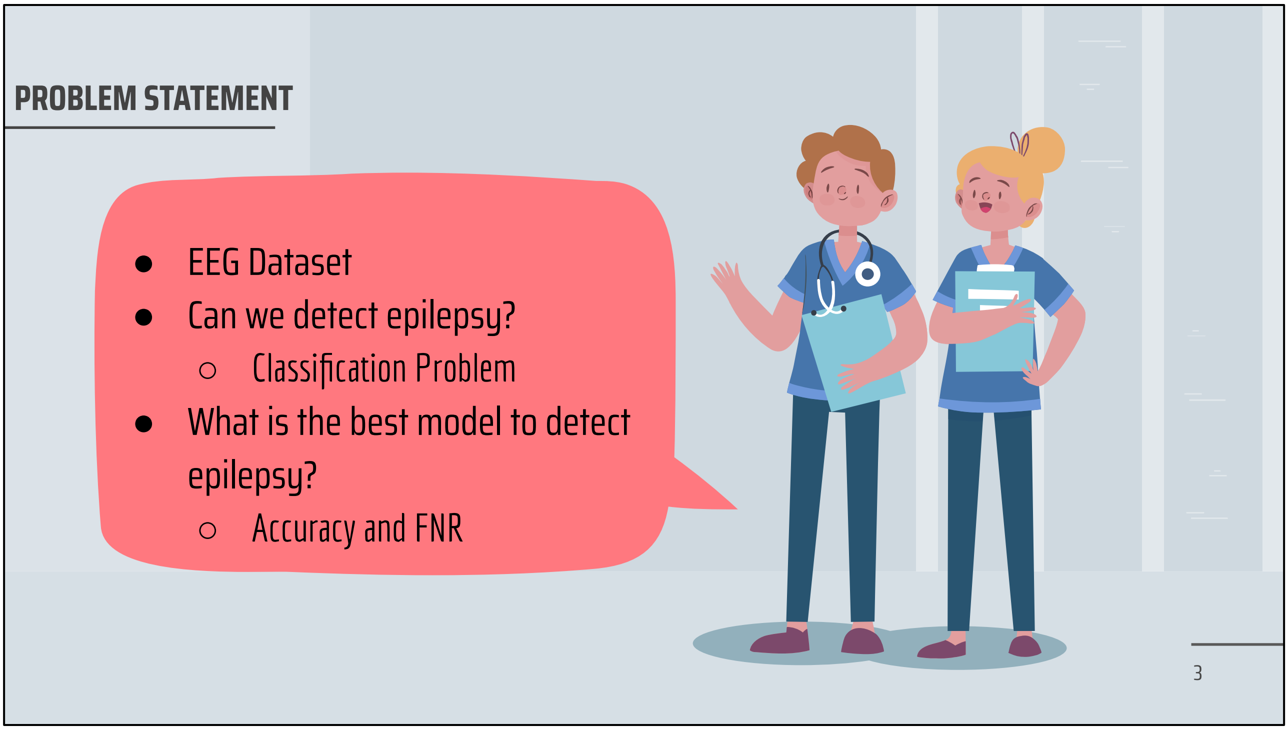
FOURIER SERIES

DATA PREPARATION
ANALYSIS

04

CONCLUSION

We will first talk about our Problem statement followed by how we plan to tackle. On how we approach this problem, we did it in two different ways. Firstly, we treated the data as a time series and secondly, as a fourier series. We would end off by concluding with our finding.

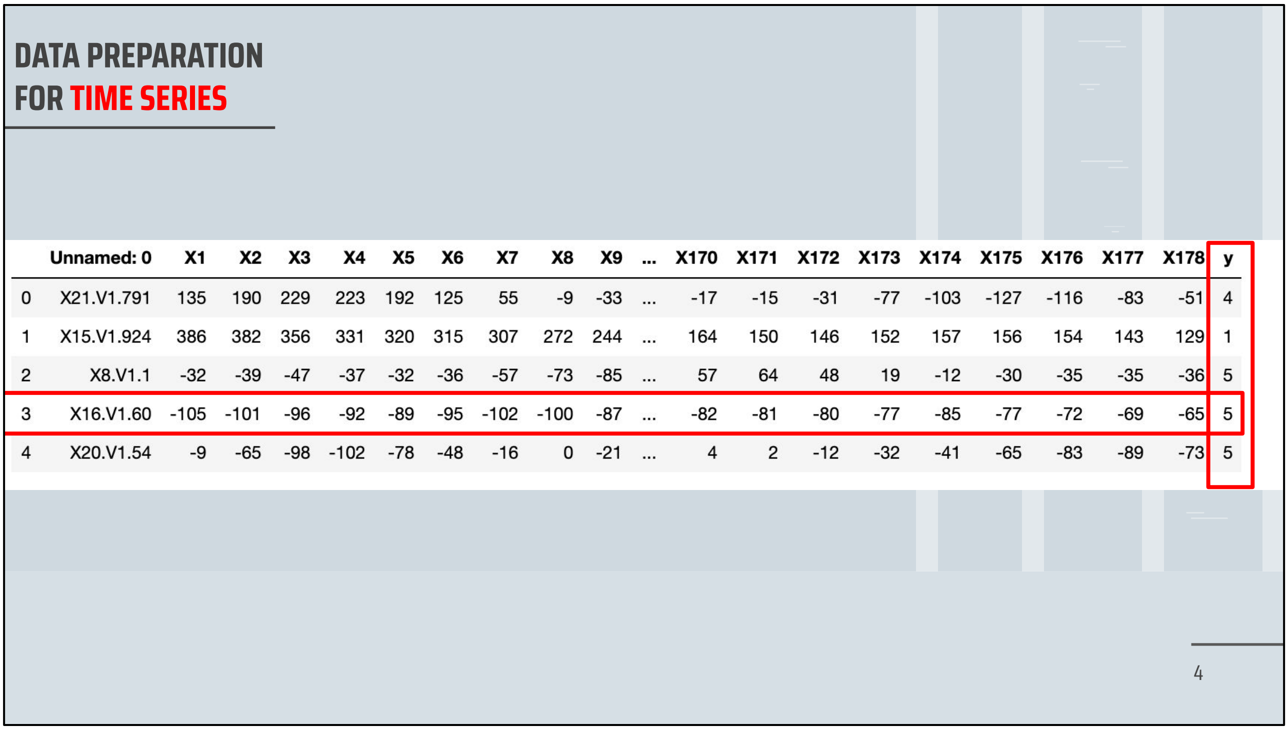


PROBLEM STATEMENT

- EEG Dataset
- Can we detect epilepsy?
  - Classification Problem
- What is the best model to detect epilepsy?
  - Accuracy and FNR

Our group had chosen the EEG Dataset.

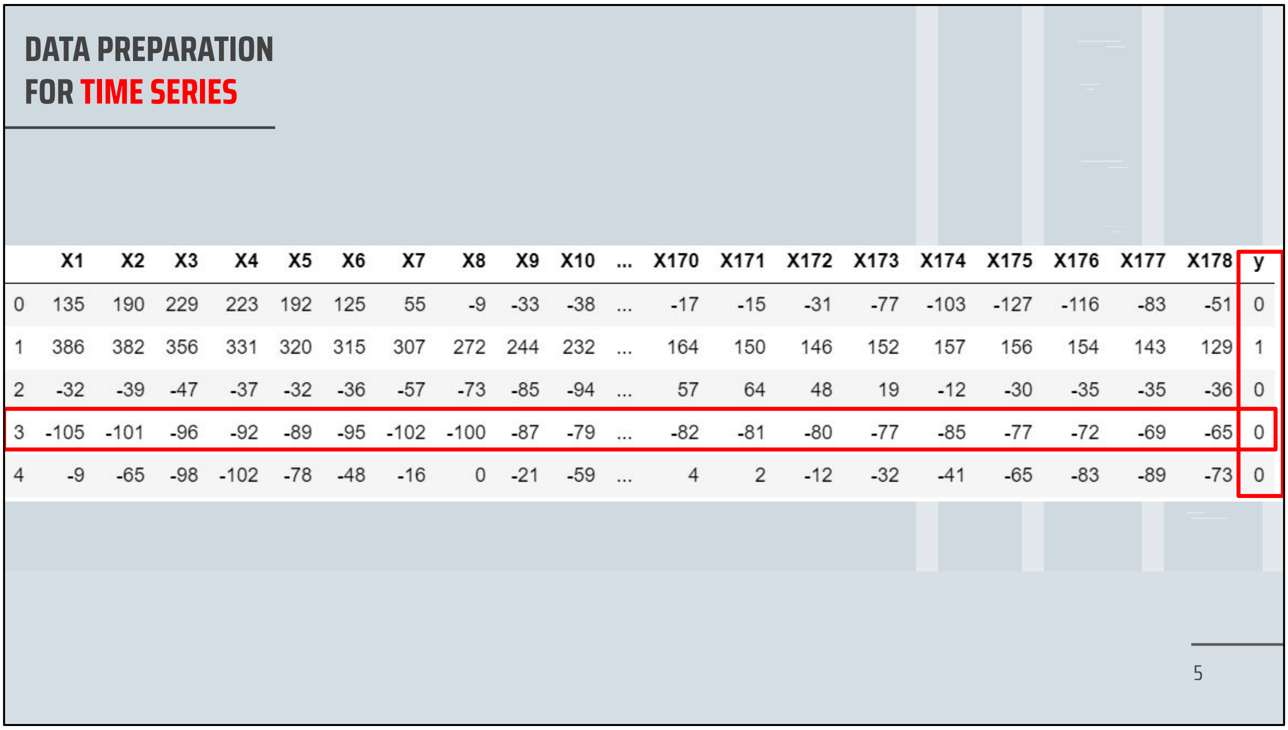
Epilepsy affects more than 65 million people worldwide. Maybe you would the one tmr! Thus this is an important issue and we are determined to find out how we can best detect epilepsy. Taking our problem and looking at it from a DS perspective, this is a classification problem, where we are classifying epilepsy cases and non-epilepsy cases. We extrapolated this problem and decided to determine the best model for detecting epilepsy using machine learning. Our metrics was based on classification accuracy and FNR. We set the FNR as an important variable because we want to minimise the number of undetected epilepsy. Now we would be going through the dataset in detail.



DATA PREPARATION FOR TIME SERIES

	Unnamed: 0	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	X21.V1.791	135	190	229	223	192	125	55	-9	-33	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
1	X15.V1.924	386	382	356	331	320	315	307	272	244	...	164	150	146	152	157	156	154	143	129	1
2	X8.V1.1	-32	-39	-47	-37	-32	-36	-57	-73	-85	...	57	64	48	19	-12	-30	-35	-35	-36	5
3	X16.V1.60	-105	-101	-96	-92	-89	-95	-102	-100	-87	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	X20.V1.54	-9	-65	-98	-102	-78	-48	-16	0	-21	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

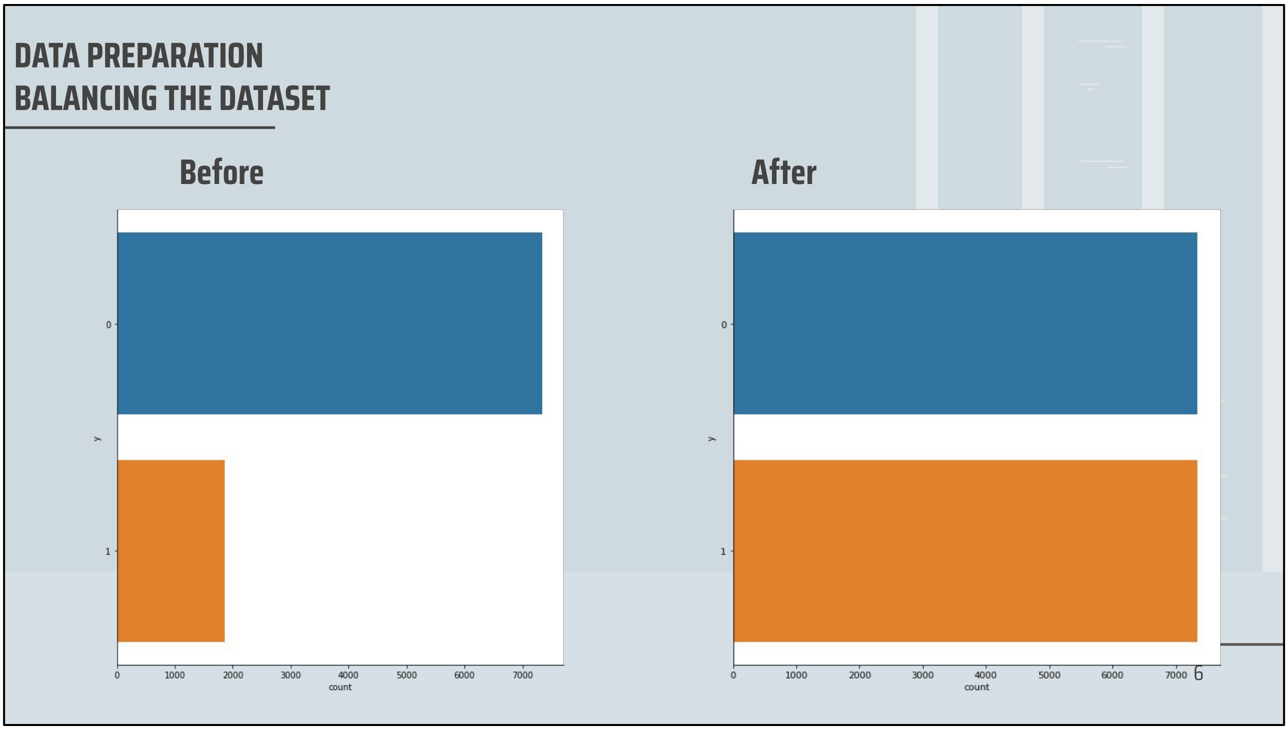
The dataset includes 4097 electroencephalogram (EEG) readings per patient over 23.5 seconds, with 500 patients in total. The 4097 data points were then shuffled and divided equally into 23 chunks per patient, each chunk is translated into one row in the dataset. Each and every row contains 178 readings. The last column is the labelled from 1 - 5. 1 is for epilepsy and 2,3,4,5 are all non-epilepsy data points. Our dataset is not clean to solve our Machine Learning problem. It has unnecessary features such as the first column and the Y labels that is not necessary to process the data.



DATA PREPARATION FOR TIME SERIES

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	135	190	229	223	192	125	55	-9	-33	-38	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	0
1	386	382	356	331	320	315	307	272	244	232	...	164	150	146	152	157	156	154	143	129	1
2	-32	-39	-47	-37	-32	-36	-57	-73	-85	-94	...	57	64	48	19	-12	-30	-35	-35	-36	0
3	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	0
4	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	...	4	2	-12	-32	-41	-65	-83	-89	-73	0

In the preparation of the dataset we drop the first column. We also converted the 'y' label from (1,2,3,4,5) to 1 and 0 since we are only interested in detecting whether there would be epilepsy or not. 1 represents epilepsy (case 1), 0 represents no epilepsy (case 2,3,4,5). This help us in achieving a better fit in our models.

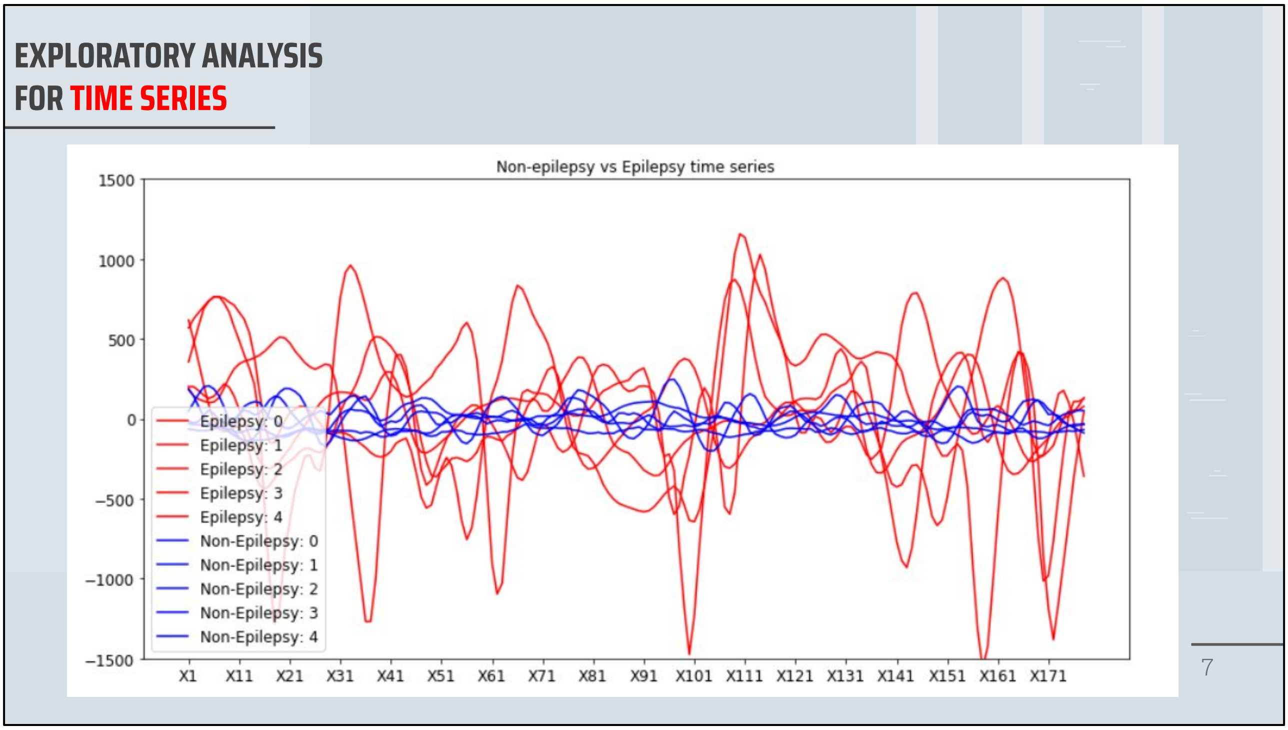


DATA PREPARATION
BALANCING THE DATASET

Before

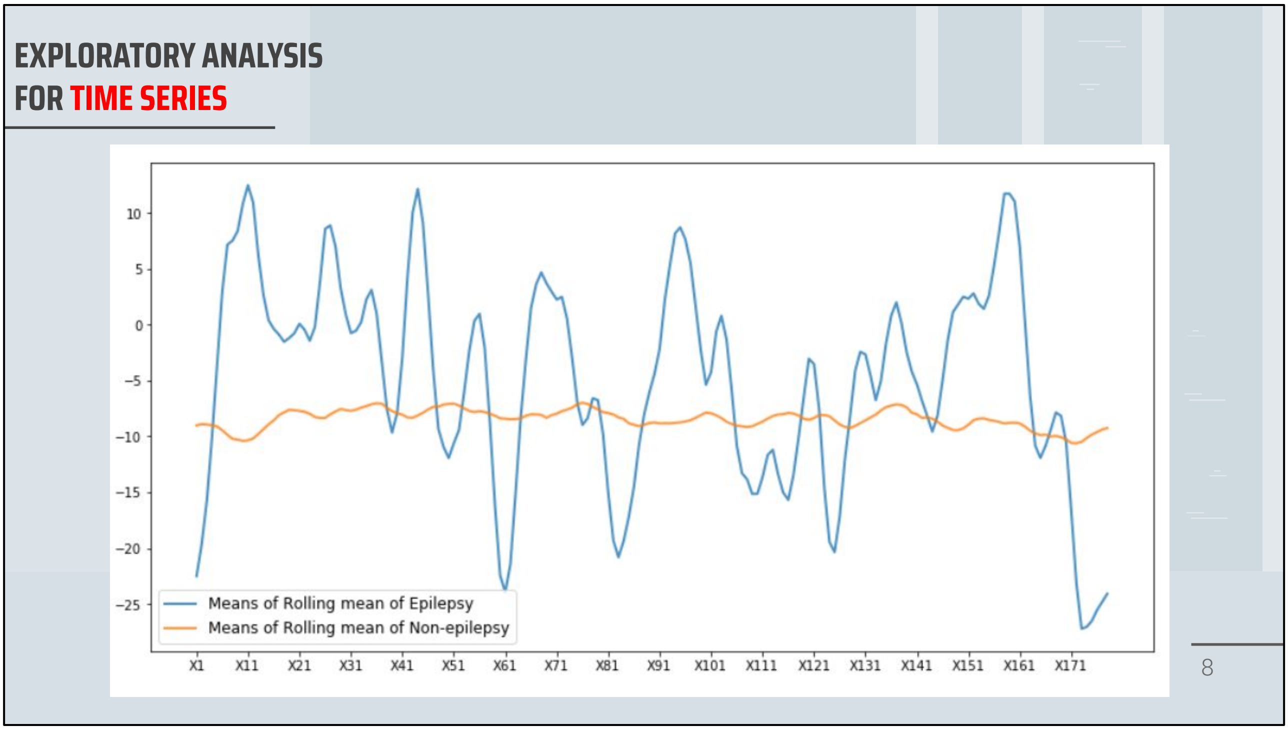
After

During data preparation, we also realised that the dataset is unbalanced and the ratio of non-epilepsy(0) VS epilepsy(1) is 1:4.
In order to better train our model, we need to ensure the dataset is balanced to reduce biases. Our group decided to use the Synthetic Minority Over-sampling Technique (SMOTE) to balance our dataset. This is an oversampling method by creating additional points on the minority data which is 1(epilepsy) and adding it to the original dataset. After applying this method, we have a balance dataset.



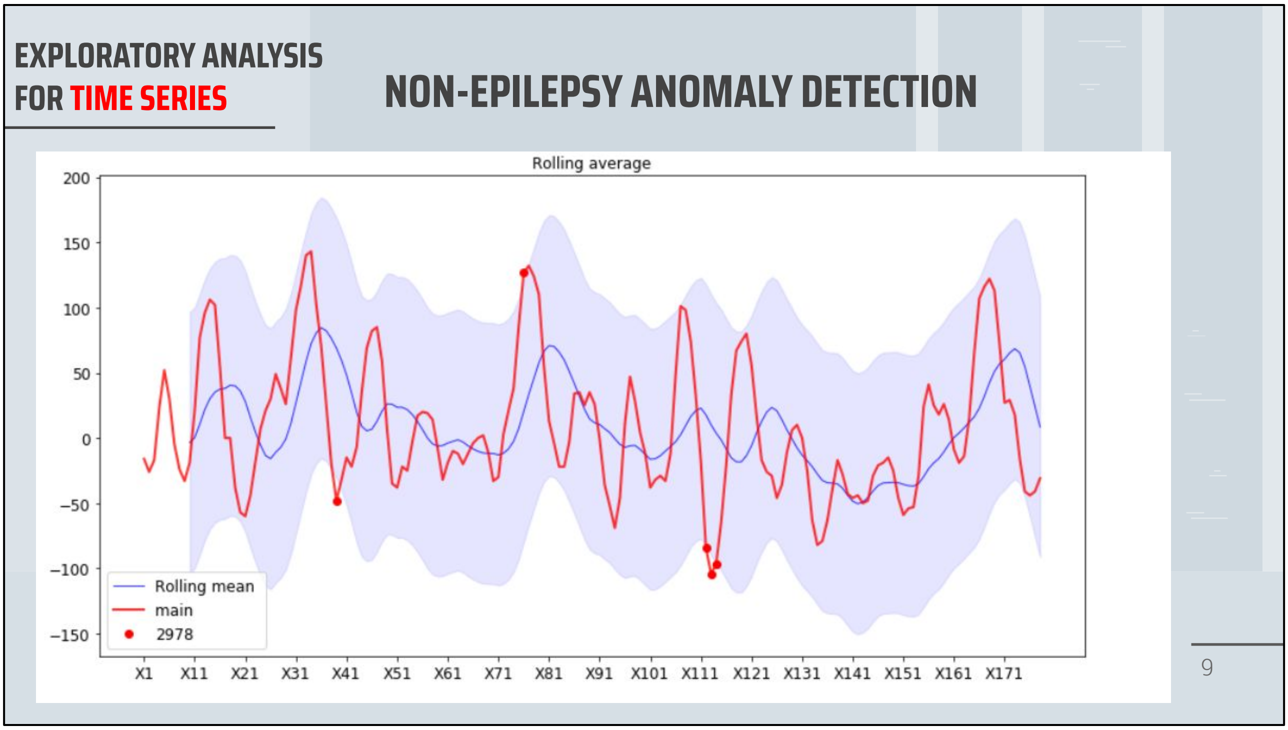
EXPLORATORY ANALYSIS FOR TIME SERIES

Now for the exploratory analysis. We choose five random points from the epilepsy and non epilepsy dataset and plotted them together. From here we can visualise a big differences between epilepsy and non epilepsy. We can clearly see that epilepsy data is much extreme.



EXPLORATORY ANALYSIS FOR TIME SERIES

We decided to calculate the rolling means of our epilepsy and non-epilepsy dataset as this is a time series data.Zachary would now explain how we use this rolling means.



EXPLORATORY ANALYSIS
FOR TIME SERIES

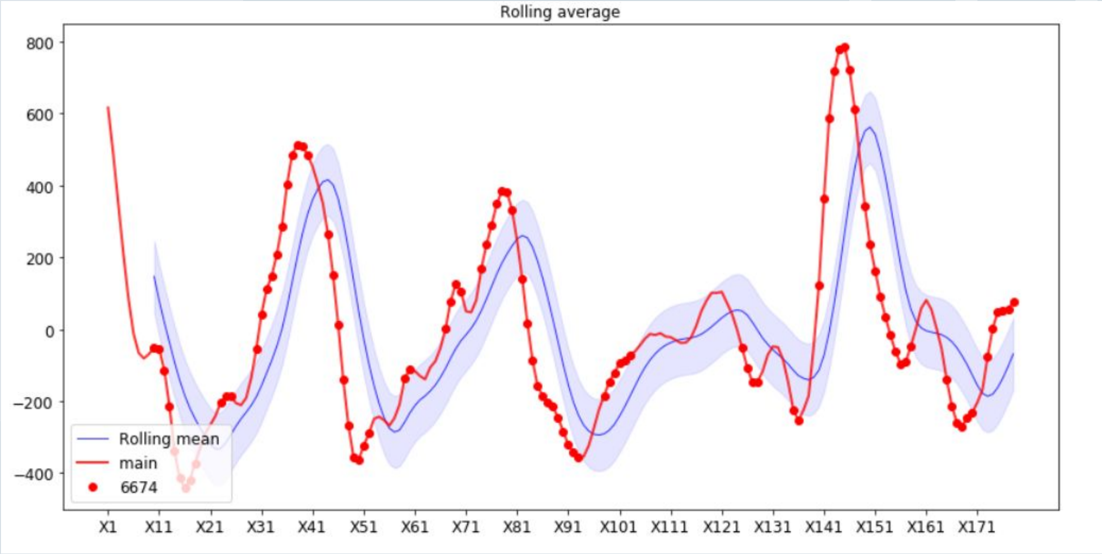
NON-EPILEPSY ANOMALY DETECTION

So how do we use this rolling means? We used the rolling mean to get upper and lower bound which is represented by the blue area. We did that by adding and subtracting a constant.

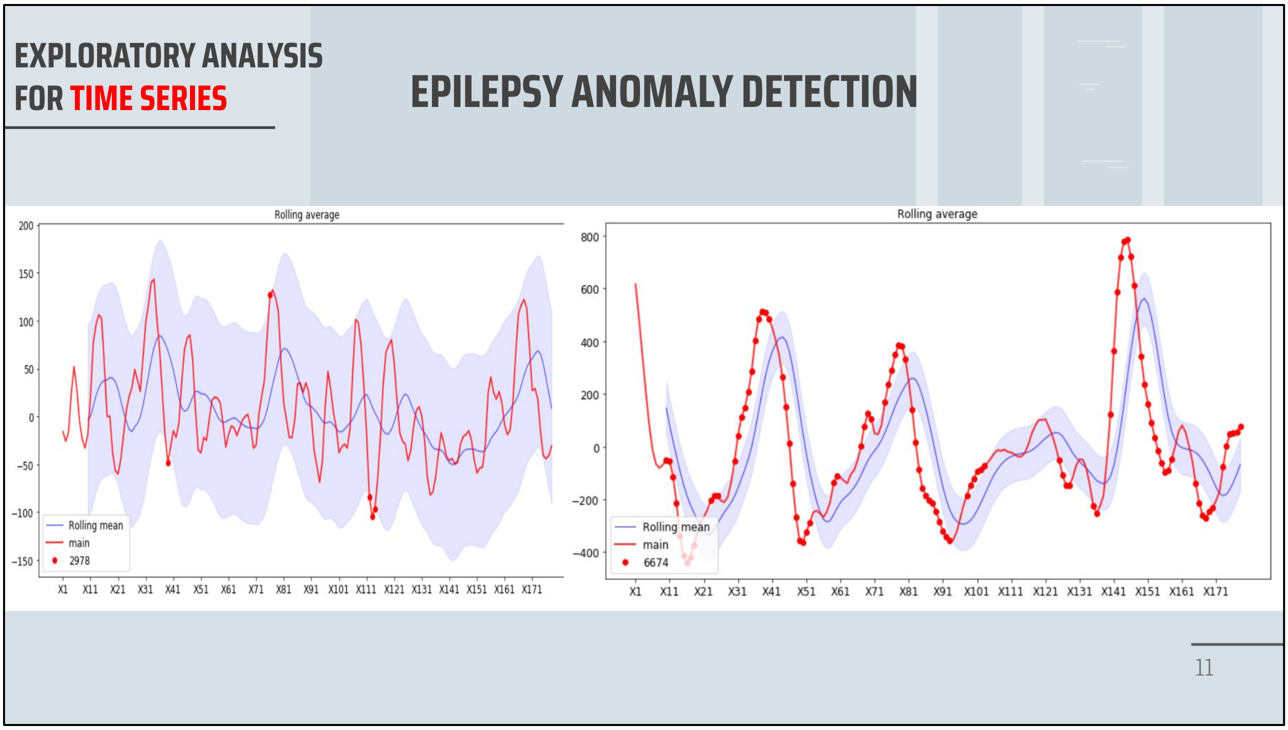
In this non-epilepsy case, we plotted the actual data in red against the upper and lower bounds. We can then classify anomalies as any points that lies outside the boundary. On the graph, anomalies are plotted as red dots.

EXPLORATORY ANALYSIS FOR TIME SERIES

EPILEPSY ANOMALY DETECTION



This is an example of an epilepsy dataset.

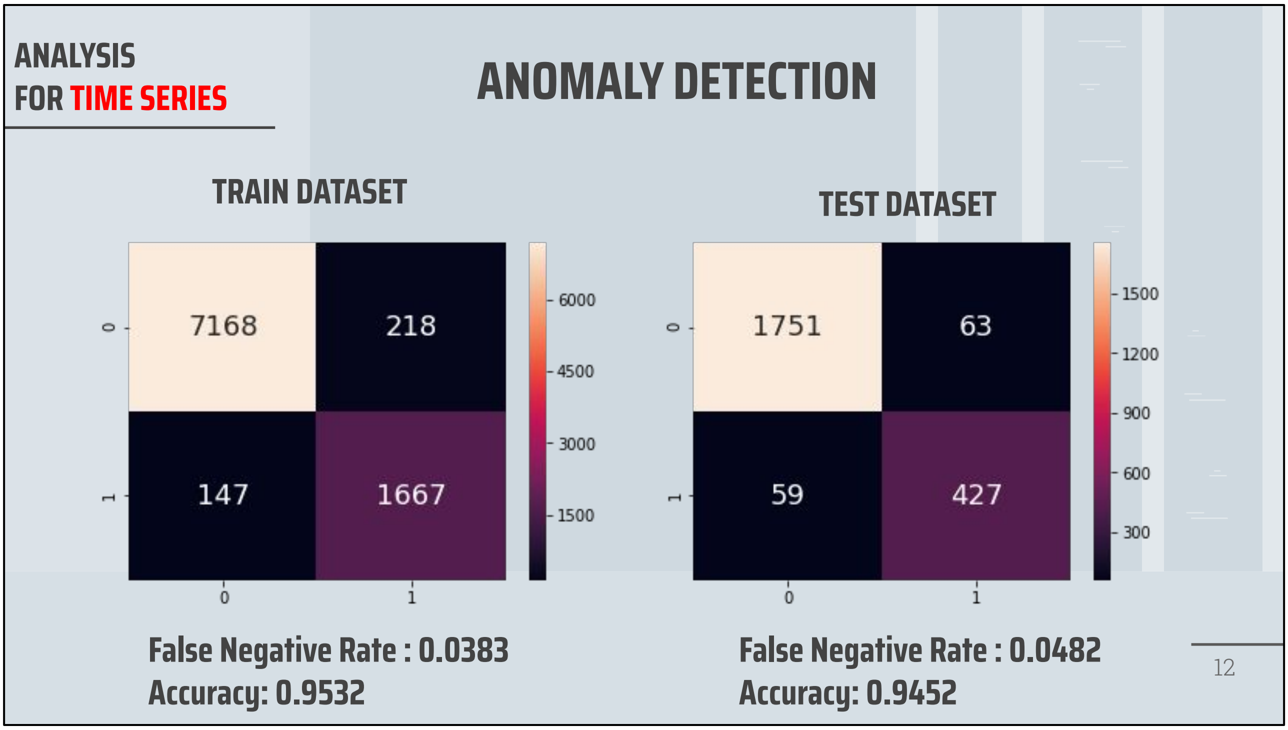


EXPLORATORY ANALYSIS
FOR TIME SERIES

EPILEPSY ANOMALY DETECTION

Compared to the non-epilepsy graph on the left, there are more anomalies in the epilepsy graph as there’s relatively more red dots.

Based on this comparison we hypothesized that the epilepsy dataset would have more abnormalities that non-epilepsy, which we could use to classify epilepsy and non-epilepsy.



ANALYSIS
FOR TIME SERIES

ANOMALY DETECTION

TRAIN DATASET

	0	1
0	7168	218
1	147	1667

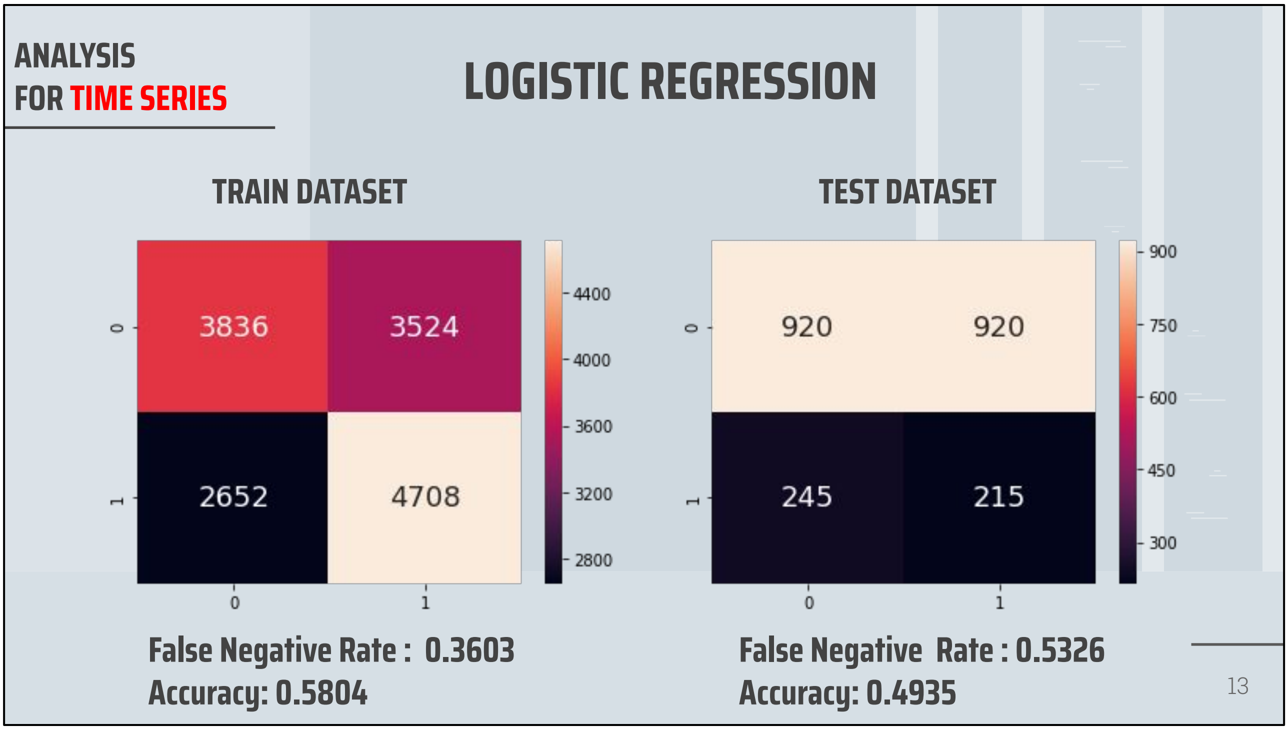
False Negative Rate : 0.0383
Accuracy: 0.9532

TEST DATASET

	0	1
0	1751	63
1	59	427

False Negative Rate : 0.0482
Accuracy: 0.9452

Four different models: Anomaly Detection, Logistic Regression, KNN, GBC



ANALYSIS
FOR TIME SERIES

LOGISTIC REGRESSION

TRAIN DATASET

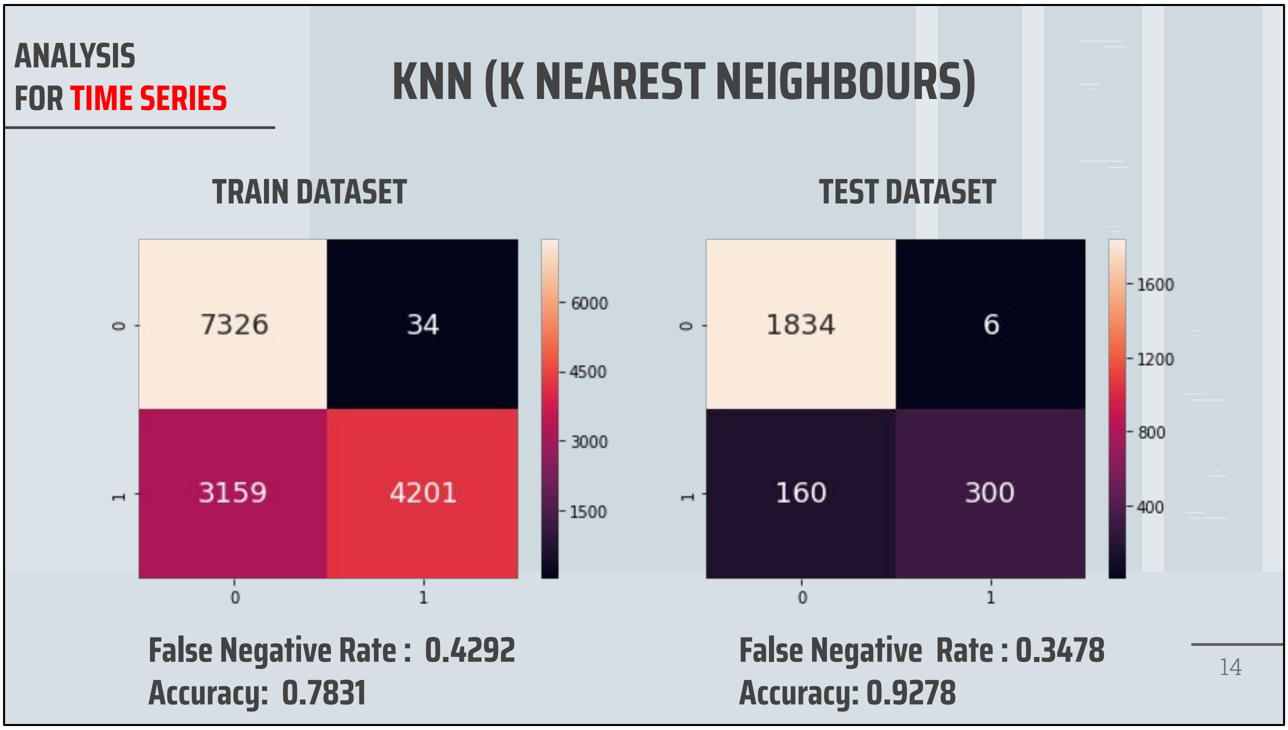
False Negative Rate : 0.3603
Accuracy: 0.5804

TEST DATASET

False Negative Rate : 0.5326
Accuracy: 0.4935

Moving on we also tried different machine learning models to detect epilepsy.

Although we only learned linear regression, we chose to use logistic regression as linear regression is useful when the response variable is continuous in nature, whereas logistic regression is suited for classifying a binary response variable, in this case, epilepsy or not.
Furthermore, logistic regression performs better when the data separation is noticeable. In our exploratory analysis, we identified that the variance of epilepsy was much greater than non-epilepsy. In train dataset, the fnr is 0.8203 and accuracy 0.8203



ANALYSIS
FOR TIME SERIES

KNN (K NEAREST NEIGHBOURS)

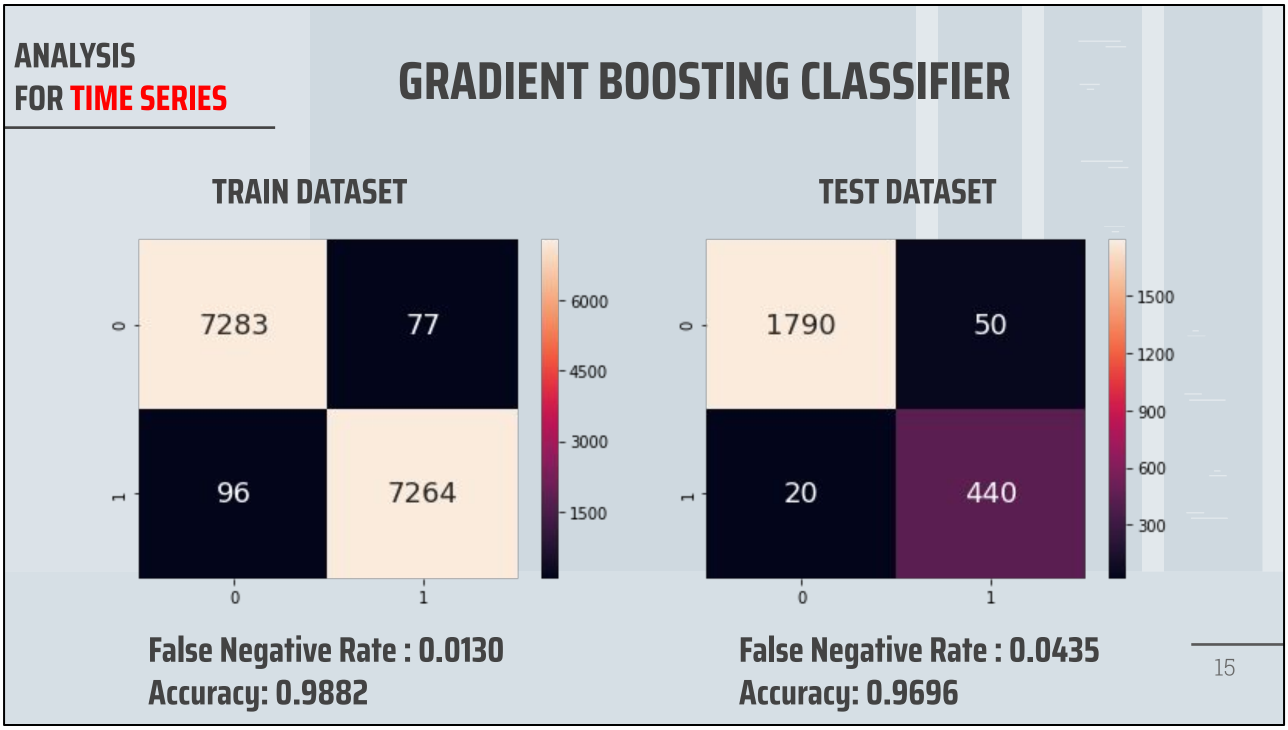
TRAIN DATASET

False Negative Rate : 0.4292
Accuracy: 0.7831

TEST DATASET

False Negative Rate : 0.3478
Accuracy: 0.9278

The next model we used is knn.
KNN is one of the most common methods used in classification, so we decided to use it as a point of comparison and test if it would give better results.
K nearest neighbours makes its selection based off of the proximity to other data points regardless of what feature the numerical values represent and predict from these data points.



ANALYSIS
FOR TIME SERIES

GRADIENT BOOSTING CLASSIFIER

TRAIN DATASET

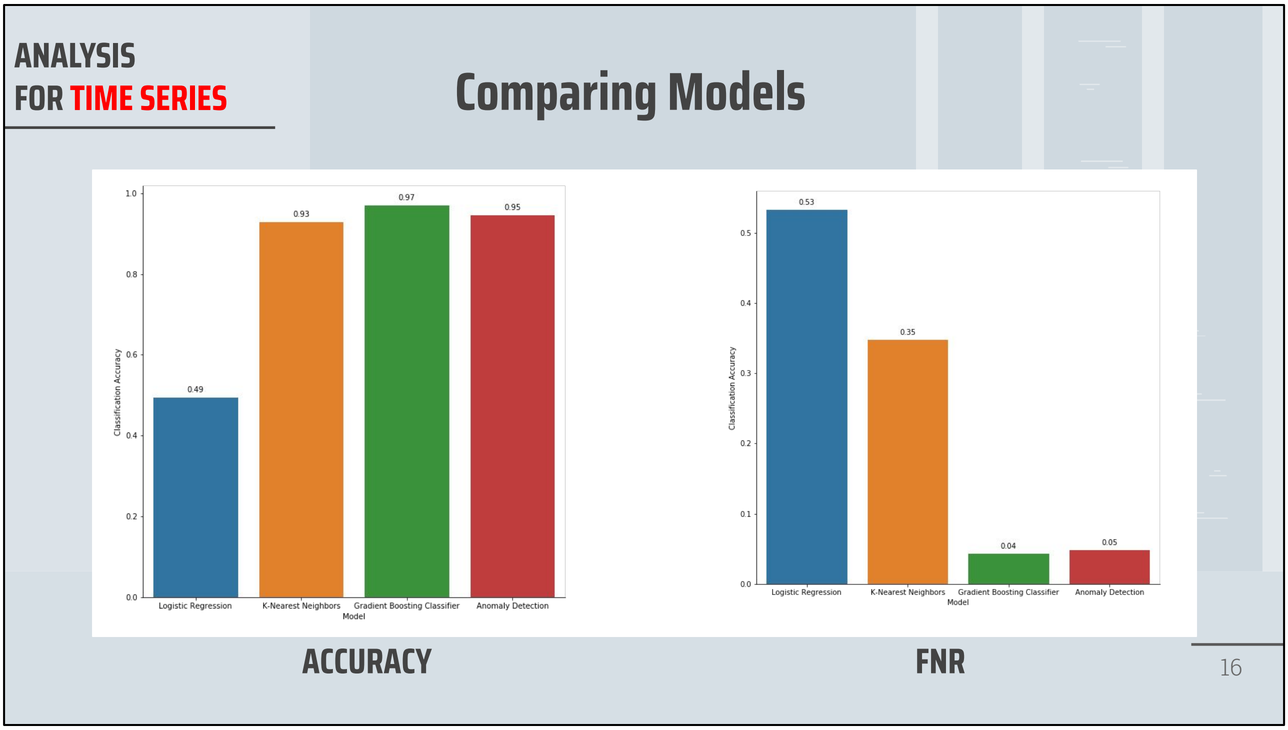
False Negative Rate : 0.0130
Accuracy: 0.9882

TEST DATASET

False Negative Rate : 0.0435
Accuracy: 0.9696

The next model is used is gbc.

We tried different techniques previously, such as linear classifiers with Logistic Regression, naive techniques with KNN, and we wanted to try with decision trees. Gradient Boosting is a decision tree that classifies based on the errors, the algo will optimise the tree based on the previous errors. Additionally, it is quite robust and well suited to non-linear problems such as our problem which deals with Time Series data.



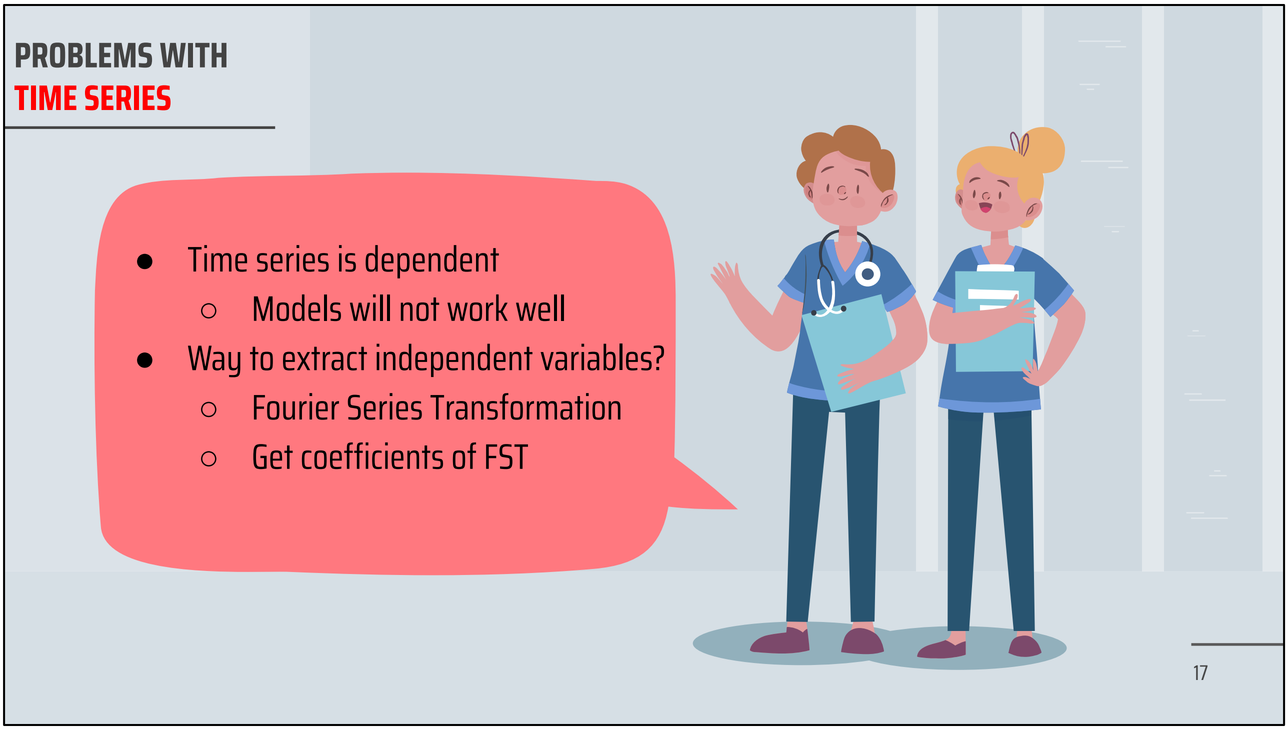
ANALYSIS
FOR TIME SERIES

Comparing Models

ACCURACY

FNR

Based on our analysis of time series data, logistic regression and KNN are the worse models as the accuracy is low and the fnr is high. We think that GB is the best has it has a high accuracy and lowest fnr. As we can see from the graphs the gbc in green has a acc of around 2 times of knn and has the lowest fnr rate



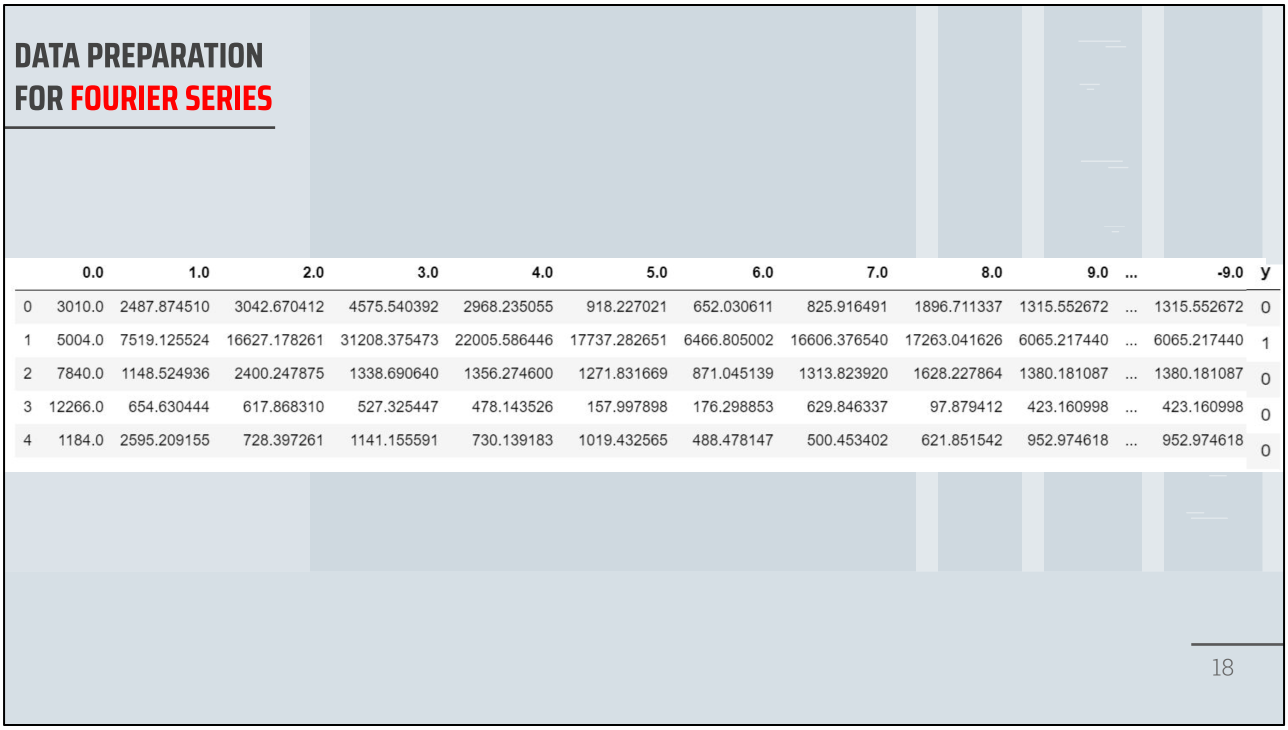
PROBLEMS WITH TIME SERIES

- Time series is dependent
  - Models will not work well
- Way to extract independent variables?
  - Fourier Series Transformation
  - Get coefficients of FST

As we can see some of the models earlier low acc and high fnr

Data are dependent on the past data points while the models take in independent variable. Therefore, the models would not work as well.

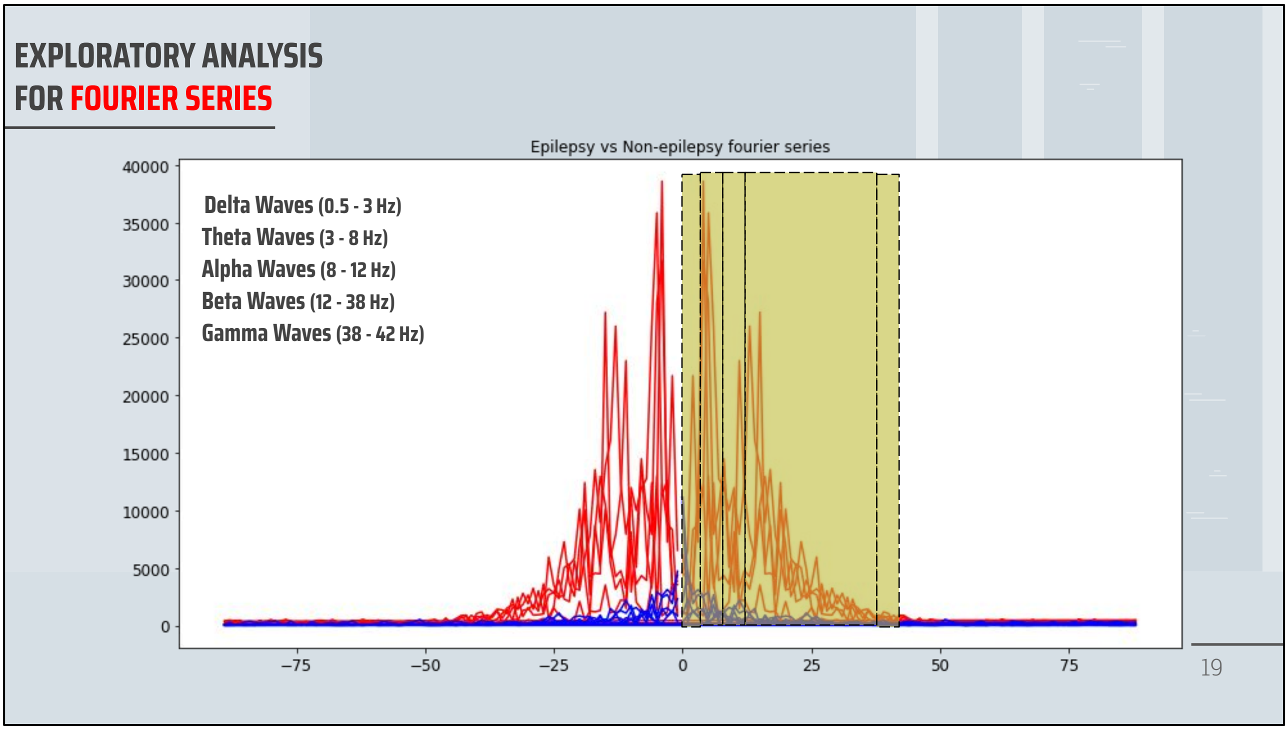
In order to better fit the model, we converted the time series data into the freq domain using ftt. The freq coeff will used as the independent variables



DATA PREPARATION FOR FOURIER SERIES

	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	...	-9.0	y
0	3010.0	2487.874510	3042.670412	4575.540392	2968.235055	918.227021	652.030611	825.916491	1896.711337	1315.552672	...	1315.552672	0
1	5004.0	7519.125524	16627.178261	31208.375473	22005.586446	17737.282651	6466.805002	16606.376540	17263.041626	6065.217440	...	6065.217440	1
2	7840.0	1148.524936	2400.247875	1338.690640	1356.274600	1271.831669	871.045139	1313.823920	1628.227864	1380.181087	...	1380.181087	0
3	12266.0	654.630444	617.868310	527.325447	478.143526	157.997898	176.298853	629.846337	97.879412	423.160998	...	423.160998	0
4	1184.0	2595.209155	728.397261	1141.155591	730.139183	1019.432565	488.478147	500.453402	621.851542	952.974618	...	952.974618	0

Created another data set by converting the time series to freq series
Using numpy.ftt.ftt, we convert the data set to freq coeff
Changed the domain to its freq

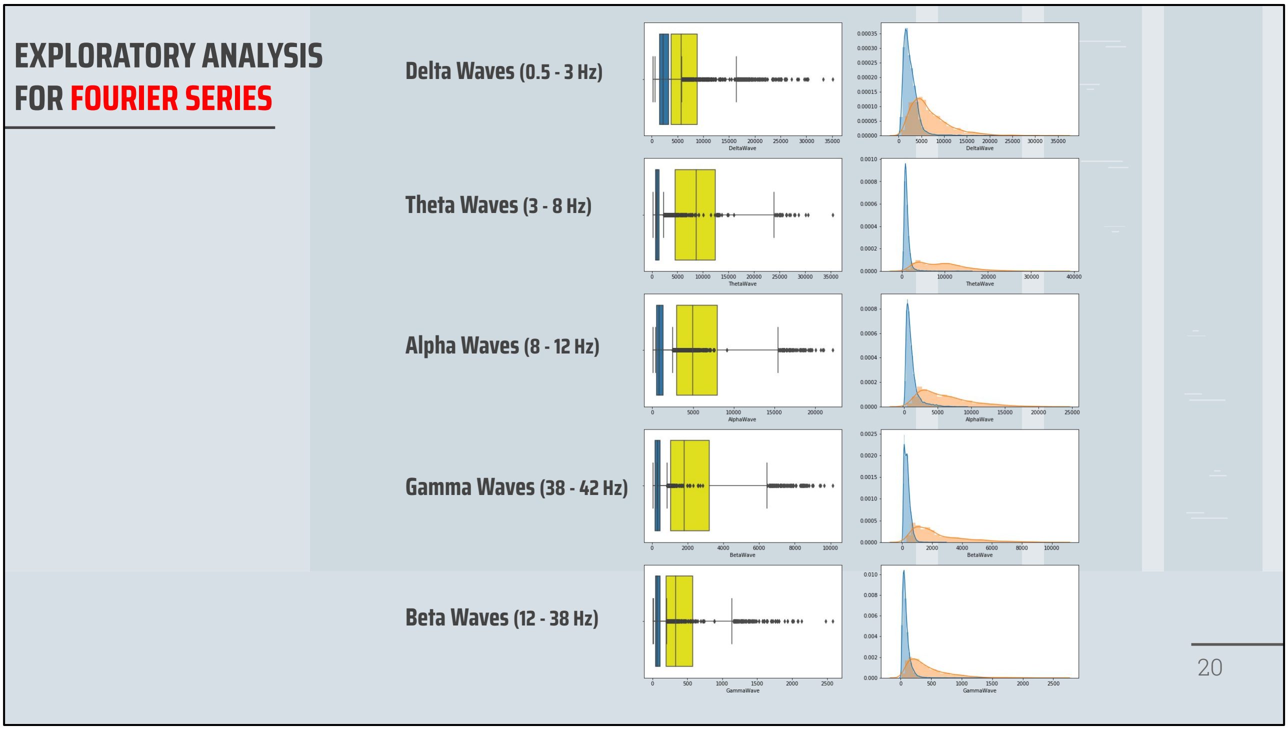


EXPLORATORY ANALYSIS FOR FOURIER SERIES

We plotted 5 random epi and non-epi data set

From the graph we can see that the epi data set has a higher magnitude. From the data set we extracted the 5 brain waves bandwith..

- Delta Waves (0.5 to 3 Hz)
- Theta Waves (3 to 8 Hz)
- Alpha Waves (8 to 12 Hz)
- Beta Waves (12 to 38 Hz)
- Gamma Waves (38 to 42 hz)



EXPLORATORY ANALYSIS FOR FOURIER SERIES

Delta Waves (0.5 - 3 Hz)

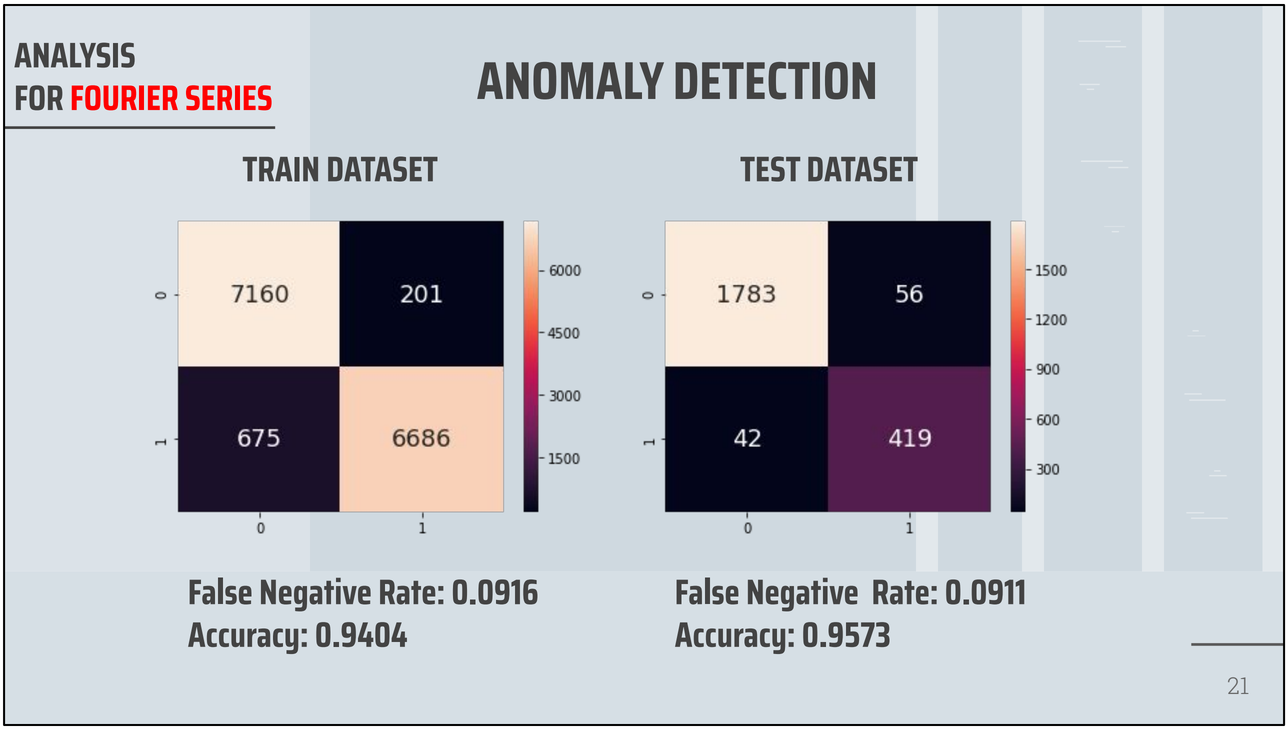
Theta Waves (3 - 8 Hz)

Alpha Waves (8 - 12 Hz)

Gamma Waves (38 - 42 Hz)

Beta Waves (12 - 38 Hz)

Yellow: Epilepsy
Blue: Non- epilepsy
Most of the wave except the delta wave, the 1 quartile of the epilepsy data is greater than the upper outlier of the non epilepsy data



ANALYSIS
FOR FOURIER SERIES

ANOMALY DETECTION

TRAIN DATASET

	0	1
0	7160	201
1	675	6686

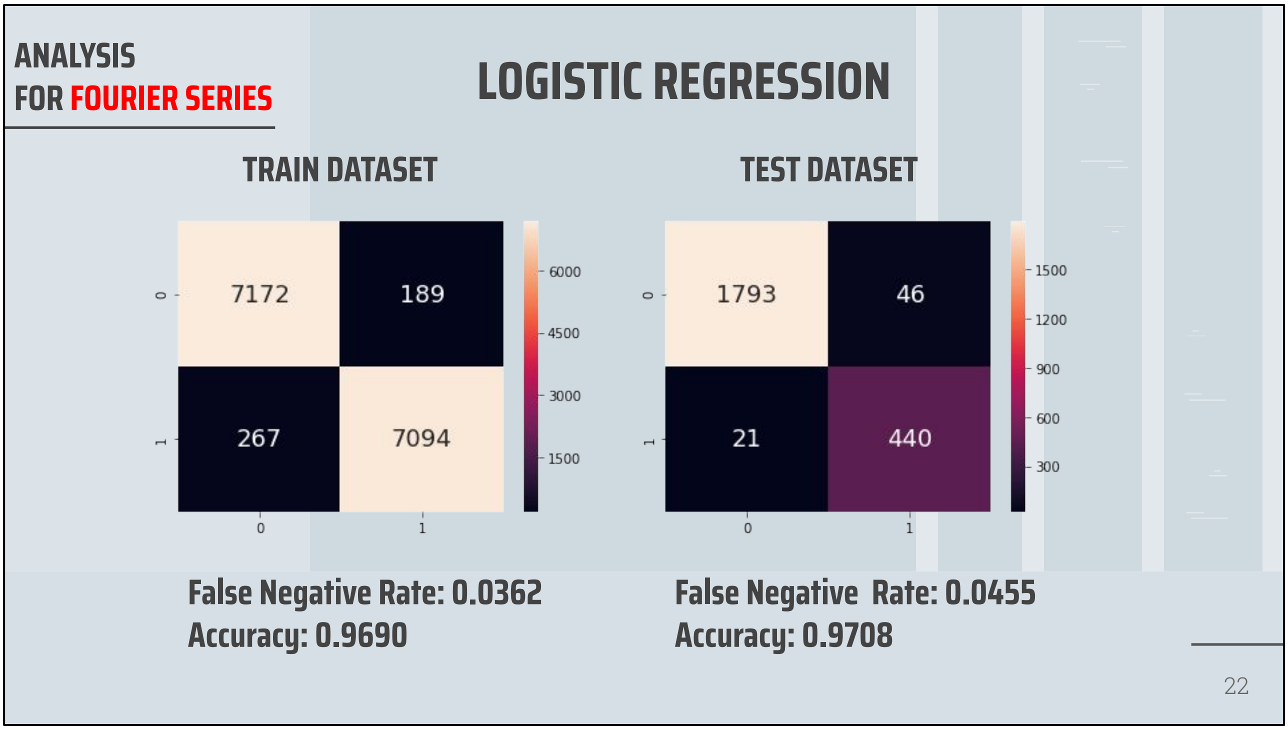
False Negative Rate: 0.0916
Accuracy: 0.9404

TEST DATASET

	0	1
0	1783	56
1	42	419

False Negative Rate: 0.0911
Accuracy: 0.9573

Just like the time series data, we also used anomaly detection. Our method was different as we are no longer using time series data so
We detected anomalies by considering if the data is above the outliers of the non-epilepsy dataset. If it is, it is considered an epilepsy case.
We noticed the test dataset had an high accuracy and low fnr rate.



ANALYSIS
FOR FOURIER SERIES

LOGISTIC REGRESSION

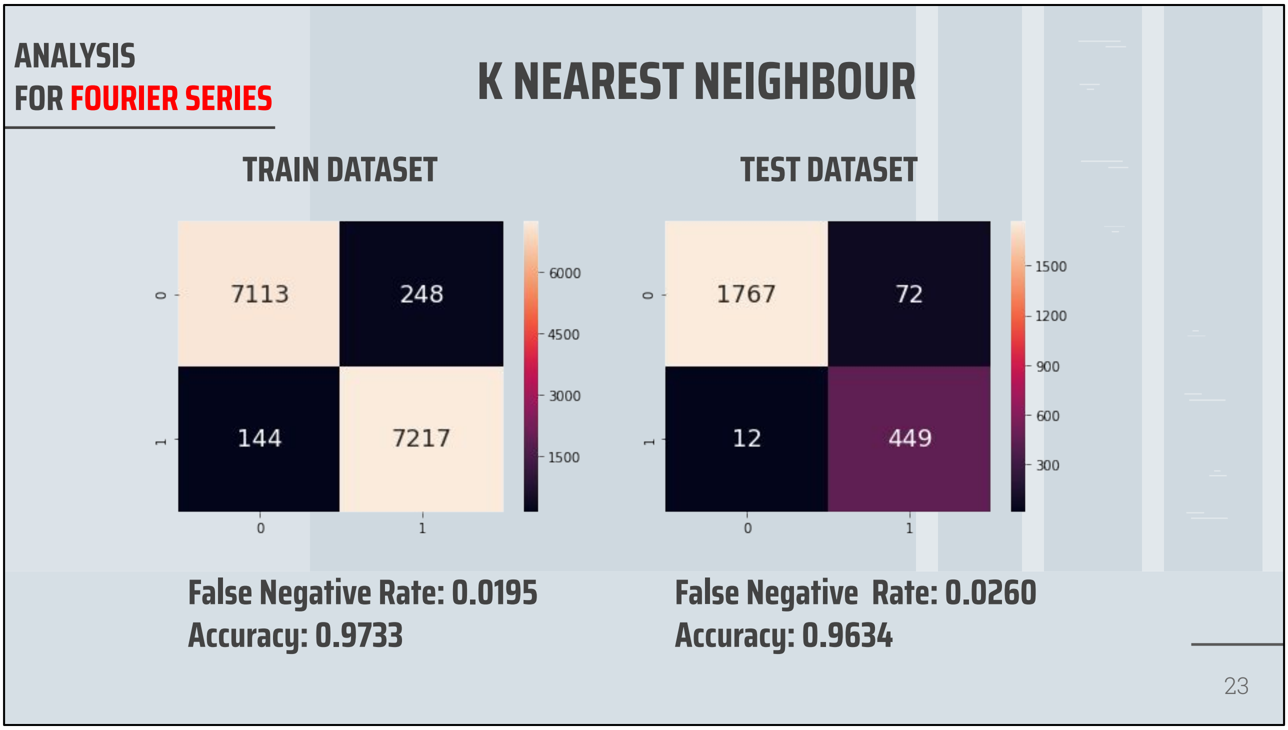
TRAIN DATASET

False Negative Rate: 0.0362
Accuracy: 0.9690

TEST DATASET

False Negative Rate: 0.0455
Accuracy: 0.9708

Next, we also used the logistic regression on the fourier series dataset.
We obtained high accuracy and low fnr from the test dataset. If you remember the accuracy and fnr from the test dataset in the time series, it was 49% and 0.53 respectively. Hence, using the logistic regression on fourier series data did definitely improve the accuracy and lower the fnr.



ANALYSIS
FOR FOURIER SERIES

K NEAREST NEIGHBOUR

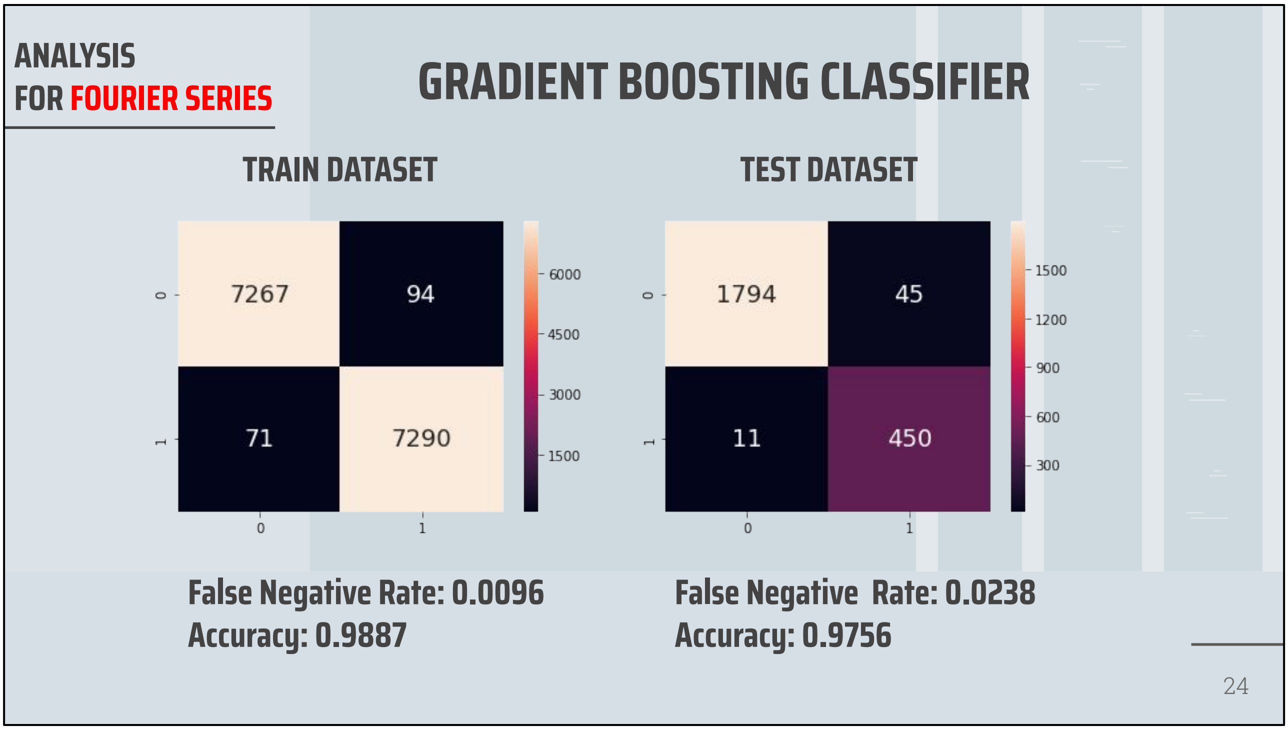
TRAIN DATASET

TEST DATASET

False Negative Rate: 0.0195
Accuracy: 0.9733

False Negative Rate: 0.0260
Accuracy: 0.9634

We also used the knn model on the fourier series data. We obtained a relatively high accuracy and relatively low FNR.



ANALYSIS
FOR FOURIER SERIES

GRADIENT BOOSTING CLASSIFIER

TRAIN DATASET

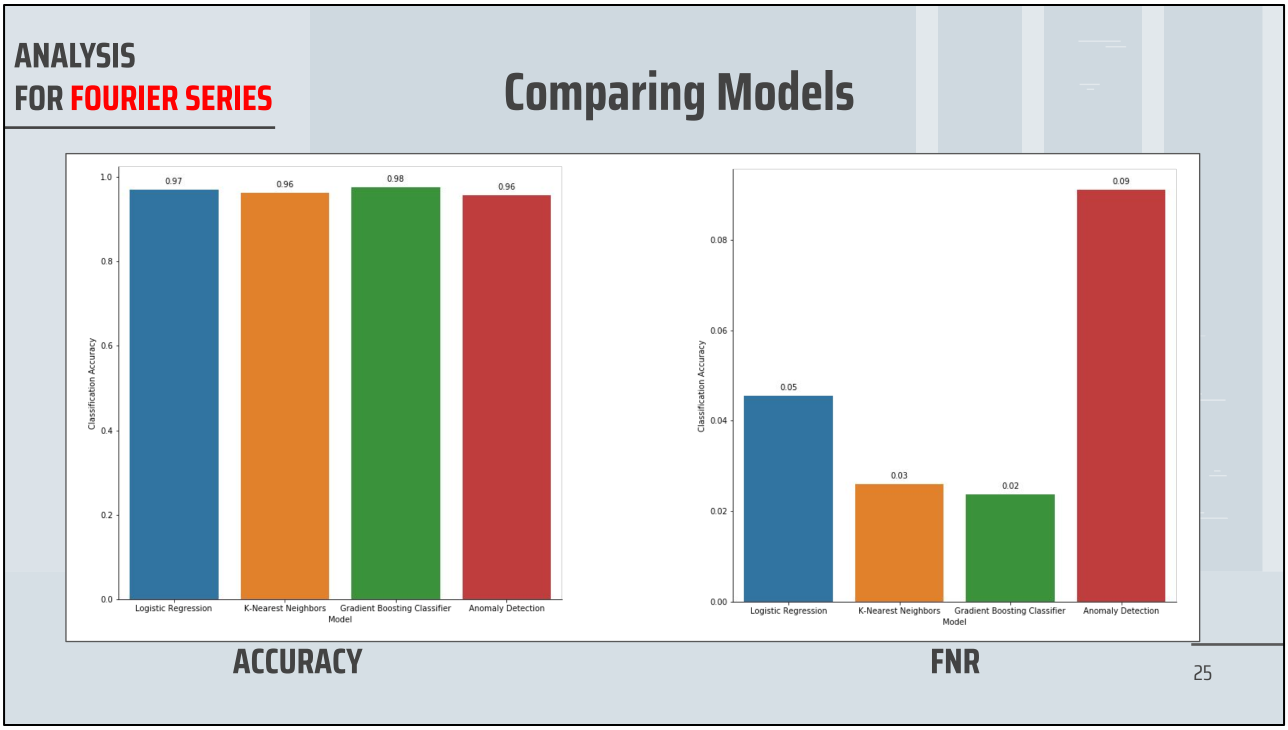
TEST DATASET

False Negative Rate: 0.0096
Accuracy: 0.9887

False Negative Rate: 0.0238
Accuracy: 0.9756

The last model used is the gradient boosting classifier on the fourier series data. As mentioned before, gradient boosting uses decision trees to analyse the data.

After using this model, we noticed that the accuracy for the test dataset is highest and has the lowest FNR rate.



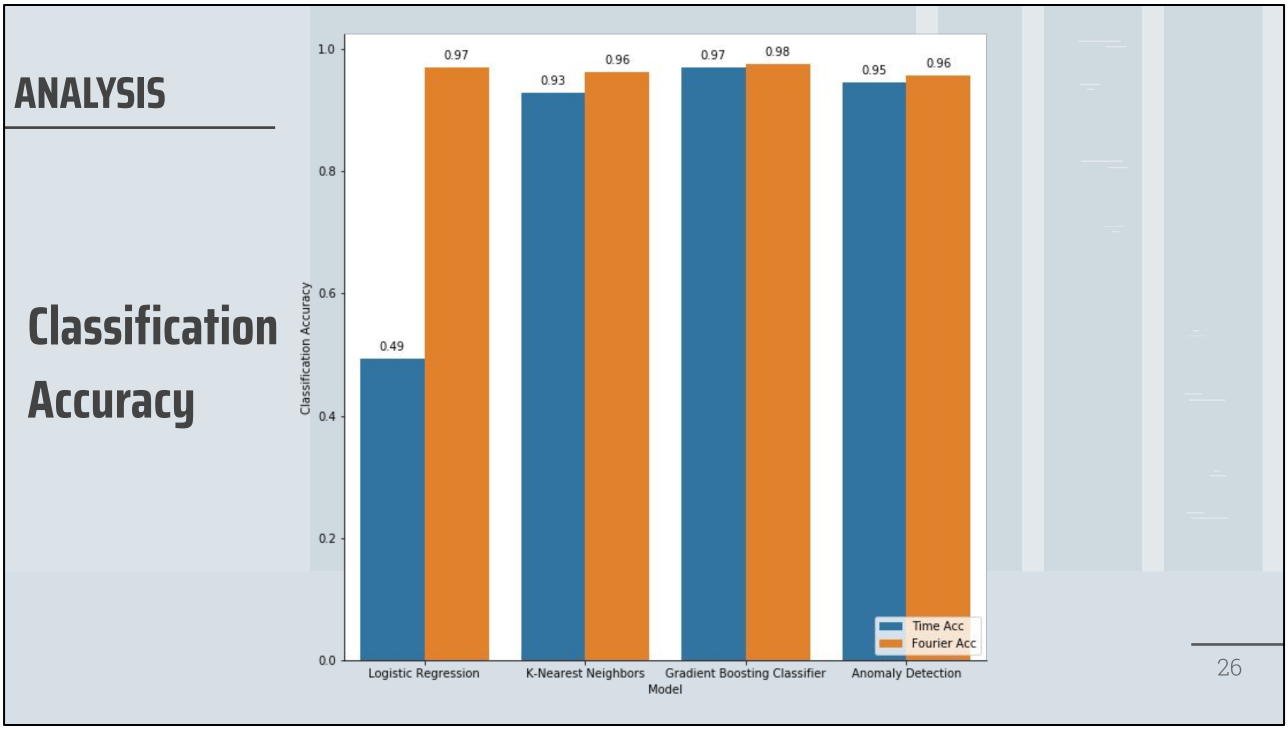
ANALYSIS
FOR FOURIER SERIES

Comparing Models

ACCURACY

FNR

This is the graphical representation of the accuracy and fnr of the various models, the accuracy of all models used are above 95% and most of the models have a low FNR, lower than 0.1. However, it is evident that GB has the highest accuracy of 97.6%. For the FNR, it is quite obvious the GB has the lowest fnr of 0.0238 compared to the rest and hence GB is the best model to be used.

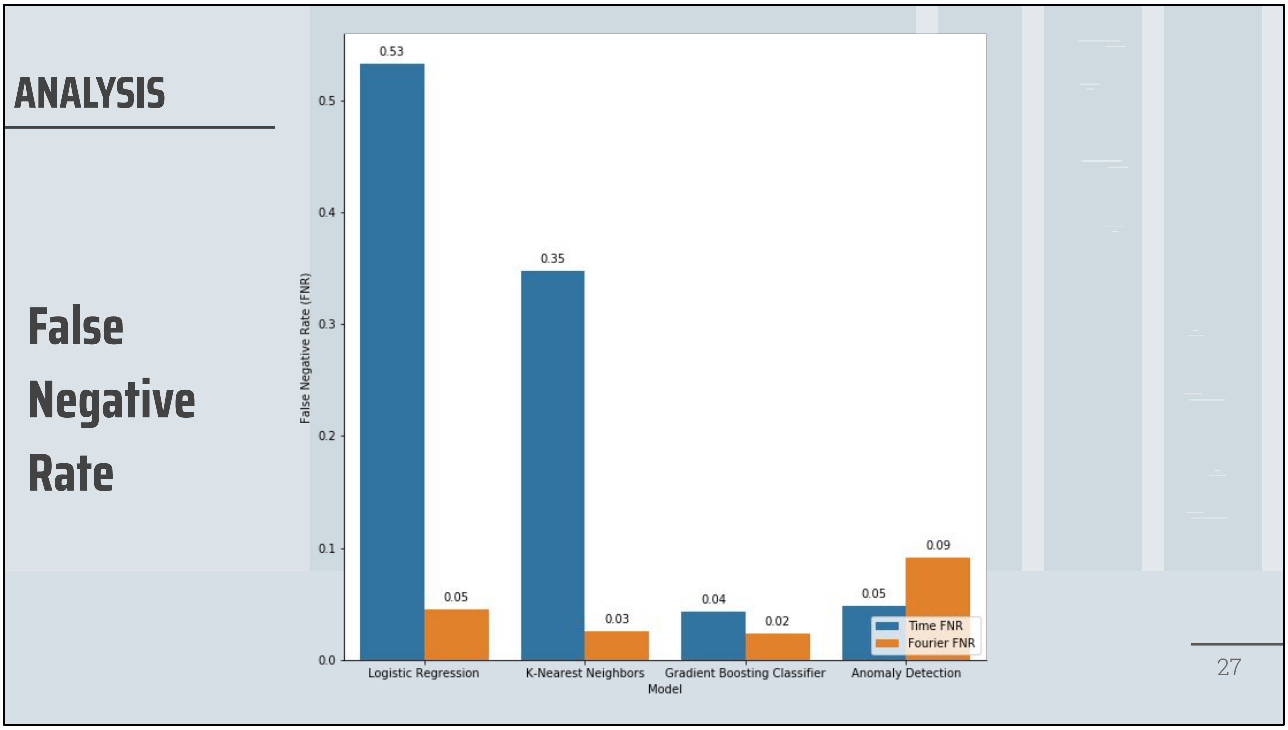


ANALYSIS

Classification Accuracy

By converting to Fourier Series we noticed that the accuracy and FNR improved a lot.

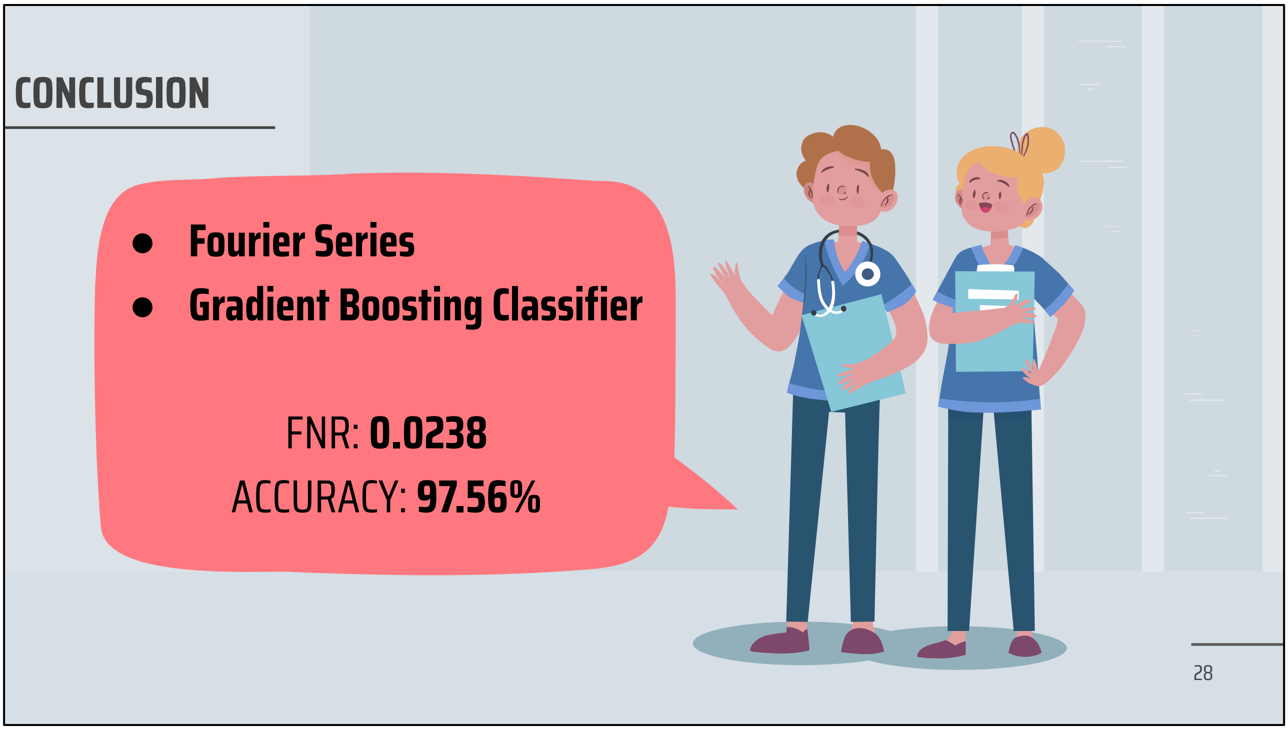
Comparing the models used between time series and fourier series, the best method to accurately predict if there is epilepsy would be the Fourier Series with Gradient Boosting Classifier as it has the highest accuracy as seen from the graph.



ANALYSIS

False Negative Rate

This is graph for the comparison of the FNR rate. Blue represents the time series FNR and orange represents fourier series FNR. As observed, the model with lowest fnr would also be the fourier series dataset analysed with GBC followed closely by KNN



CONCLUSION

- **Fourier Series**
- **Gradient Boosting Classifier**

FNR: **0.0238**

ACCURACY: **97.56%**

In conclusion, to best predict epilepsy, data has to be converted to the fourier series instead of using time series and data has to be analysed using the gradient boosting classifier. This is as it brings the highest accuracy rate and lowest false negative rate, means the lowest probability of predicting there is no epilepsy when there is in fact epilepsy.

INDIVIDUAL CONTRIBUTIONS

- Time Series - Zach, Chi Hui
- Fourier Series - Wilson, Sabrina

GITHUB LINK

https://github.com/zvarellalee/CZ1015-FS4T05-Epilepsy-Detection-Mini-Project