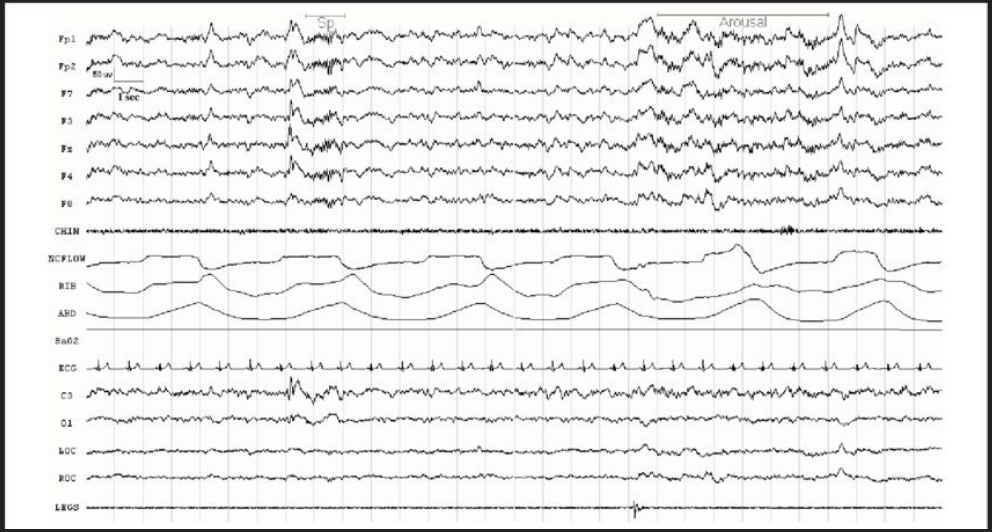
Epileptic Seizure Recognition

FSP4T10
Chua Wen Qing
Leonel Lim
Tan Wee Li

Objective: Creating the most **accurate** model for **classifying** epileptic seizure activity using **EEG readings**.

ML Problem: Classification(Binary)



Data Structuring & Cleaning

Data structuring/cleaning

	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	4
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	5
3	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

Removing unnecessary data not utilised in exploratory analysis / modelling.

Data structuring/cleaning

Separating the dataset based on the EEG readings of different classes of individuals.

Class 1: Epileptic seizure activity

Class 2: Area where tumour is located

Class 3: Healthy brain area

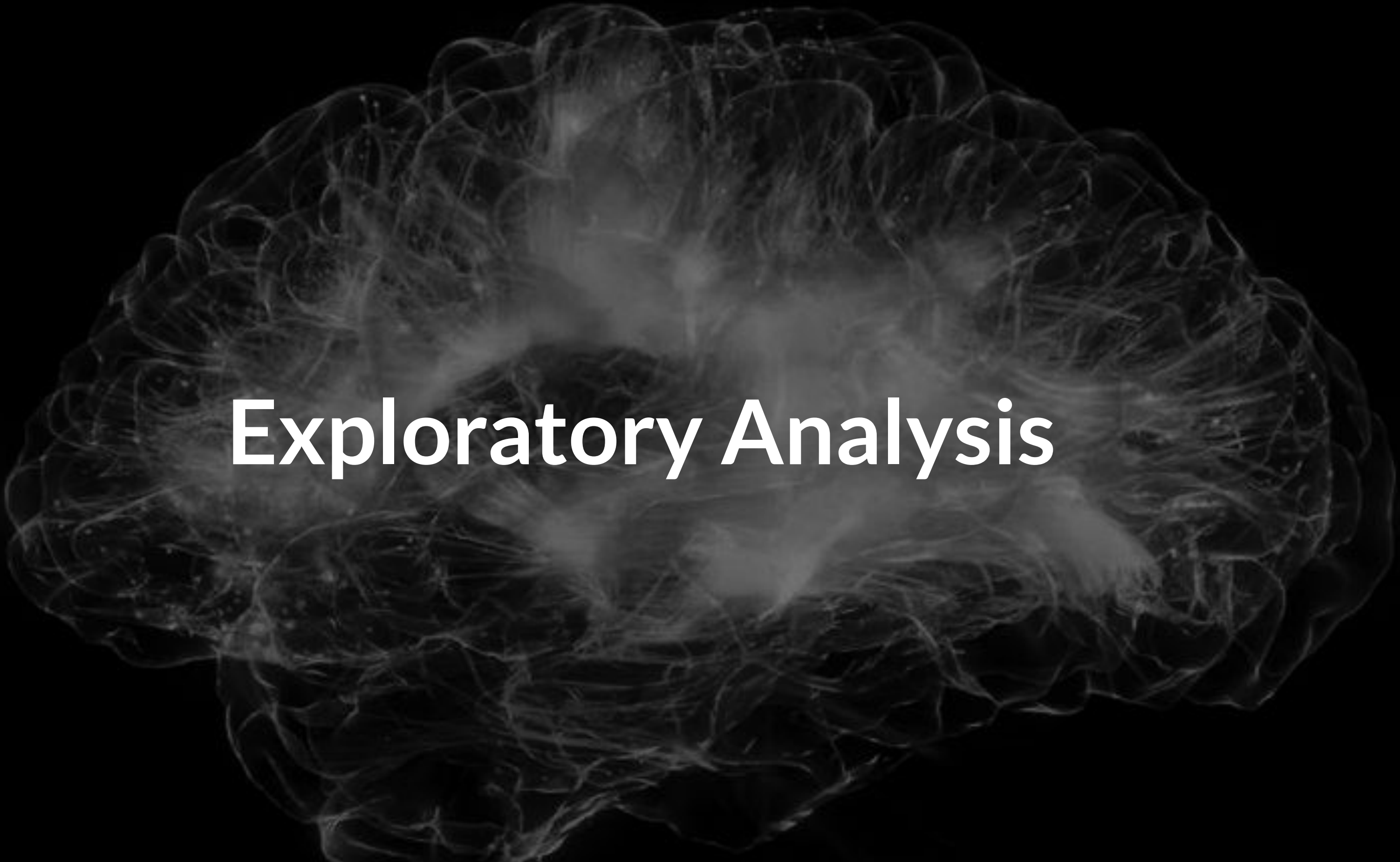
Class 4: Patient's eyes are closed

Class 5: Patient's eyes are opened.

Data structuring/cleaning

Concatenation of each non epileptic dataset with epileptic dataset. (I.E Class 1 + Class 2)

- Able to discern which class is best able to train model in classifying epileptic seizure activity.
- Reduce bias due to large number of non-epileptic seizure data. (4 : 1)



Exploratory Analysis

Exploratory Analysis

- In the dataset, there are 11500 rows, 179 columns and no missing values
- Each row of contains 179 data points (X1, X2,. . . , X177, X178, y) for 1 second
- X1-X178 represents the EEG recording at a different point in time and y contains the category of the 178-dimensional input vector

```
In [45]: df.isnull().values.any()
Out[45]: False

In [69]: df.shape
Out[69]: (11500, 179)

In [70]: df.info()
         <class 'pandas.core.frame.DataFrame'>
         Index: 11500 entries, X21.V1.791 to X16.V1.210
         Columns: 179 entries, X1 to y
         dtypes: int64(179)
         memory usage: 16.1+ MB
```

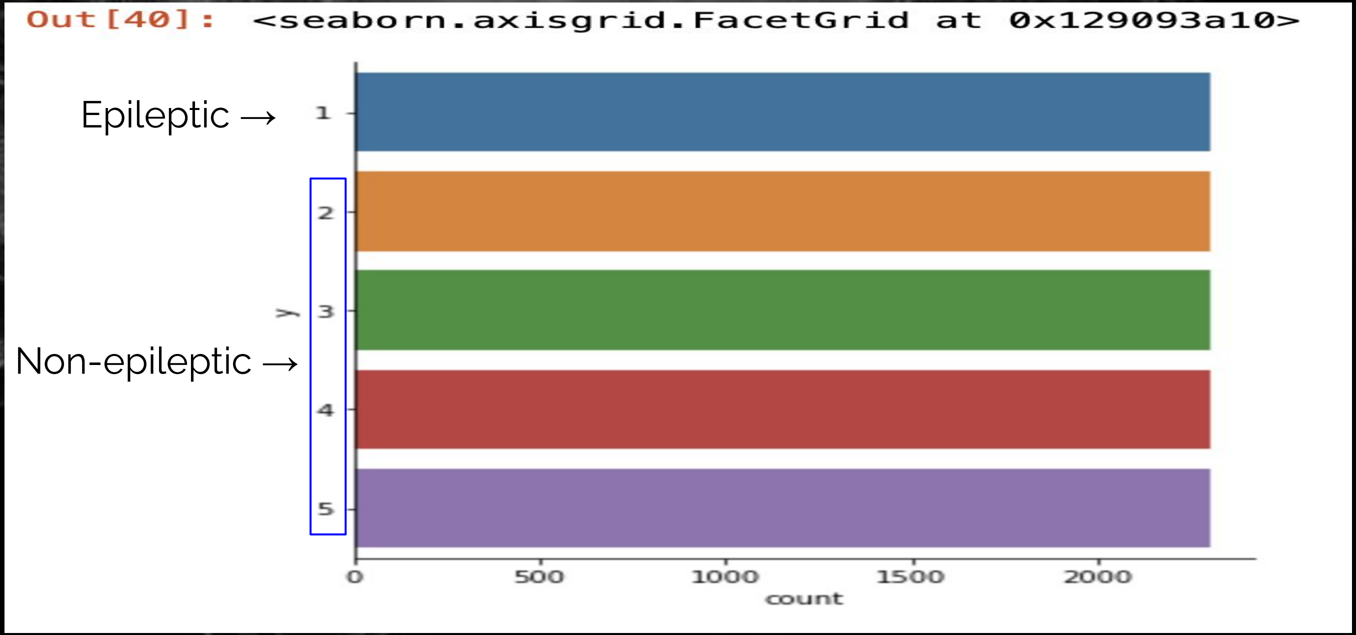
```
In [34]: df.describe()
Out[34]:
```

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
count	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.000000	11500.00000	11500.00000	11500.000000
mean	-11.581391	-10.911565	-10.187130	-9.143043	-8.009739	-7.003478	-6.502087	-6.68713	-6.55800	-6.168435
std	165.626284	166.059609	163.524317	161.269041	160.998007	161.328725	161.467837	162.11912	162.03336	160.436352
min	-1839.000000	-1838.000000	-1835.000000	-1845.000000	-1791.000000	-1757.000000	-1832.000000	-1778.00000	-1840.00000	-1867.000000
25%	-54.000000	-55.000000	-54.000000	-54.000000	-54.000000	-54.000000	-54.000000	-55.00000	-55.00000	-54.000000
50%	-8.000000	-8.000000	-7.000000	-8.000000	-8.000000	-8.000000	-8.000000	-8.00000	-7.00000	-7.000000
75%	34.000000	35.000000	36.000000	36.000000	35.000000	36.000000	35.000000	36.00000	36.00000	35.250000
max	1726.000000	1713.000000	1697.000000	1612.000000	1518.000000	1816.000000	2047.000000	2047.00000	2047.00000	2047.000000

Exploratory Analysis

- No. of instances in each category are the same
- non-epileptic seizure data : epileptic seizure data

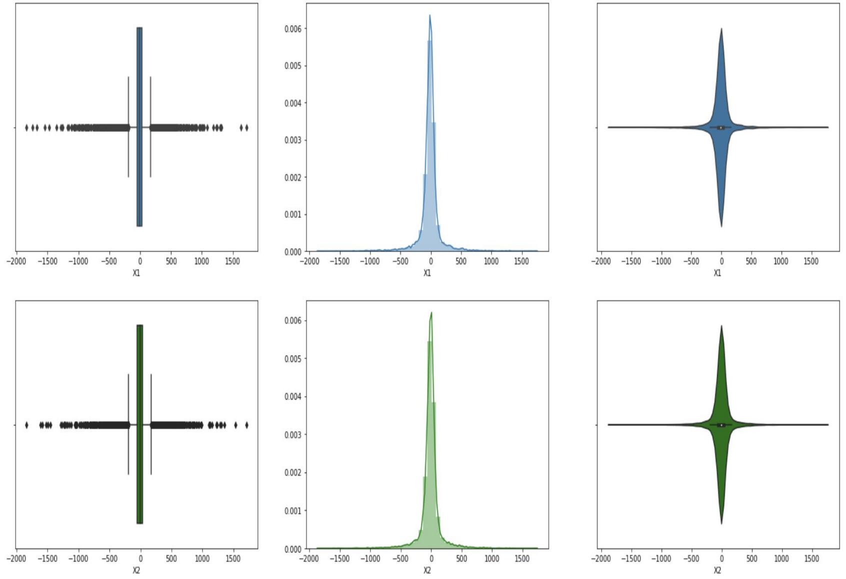
= 4 : 1



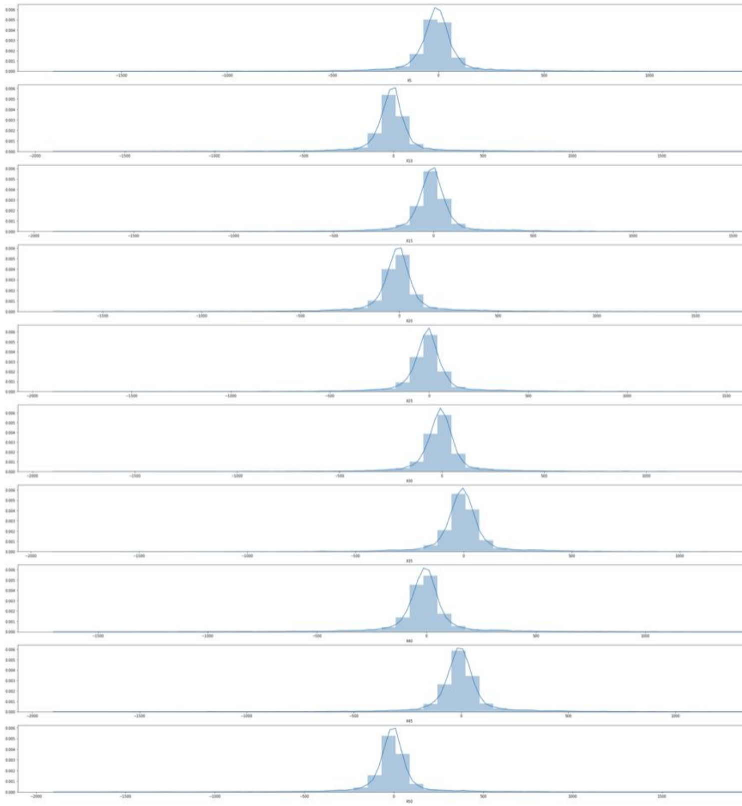
Exploratory Analysis

- EEG values are normally distributed across any variable from X1-X178

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x128e34ed0>

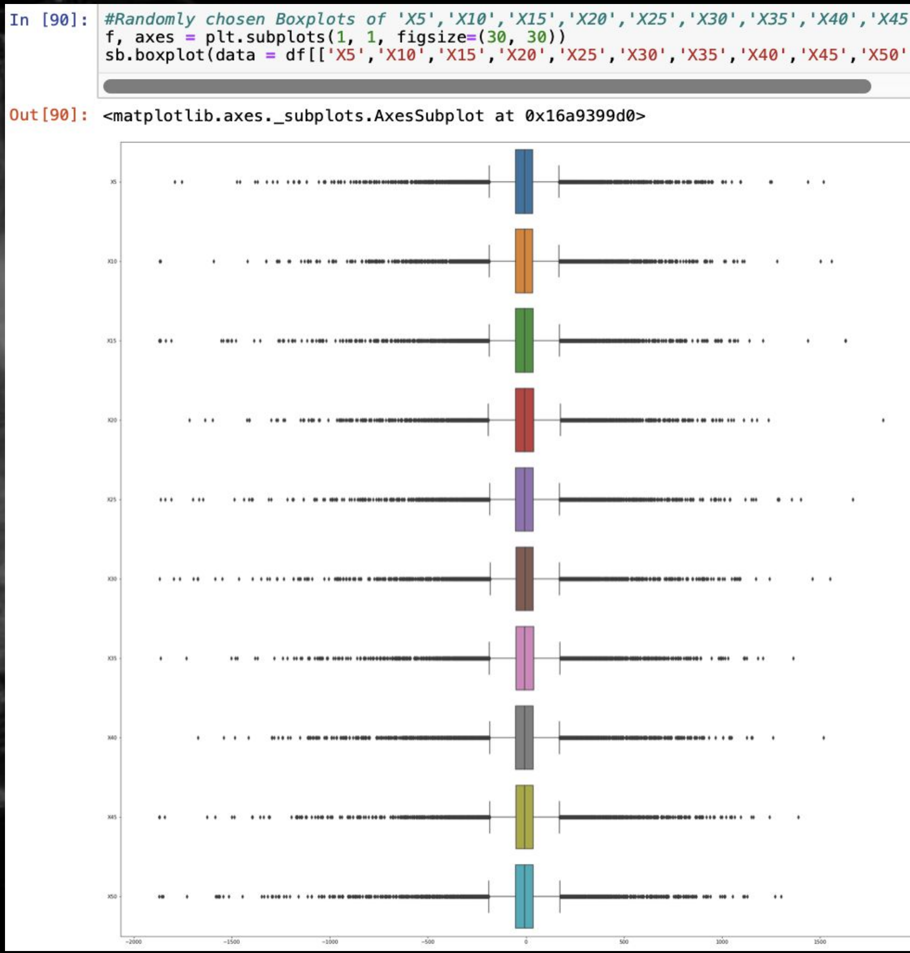



```
In [96]: #Plot randomly chosen distplots of 'X5','X10','X15','X20','X25','X30','X35','X4
         f, axes = plt.subplots(10,1, figsize=(40, 40))
         x=0
         for i in ['X5','X10','X15','X20','X25','X30','X35','X40','X45','X50']:
             sb.distplot(df[i], ax = axes[x])
             x+=1
```



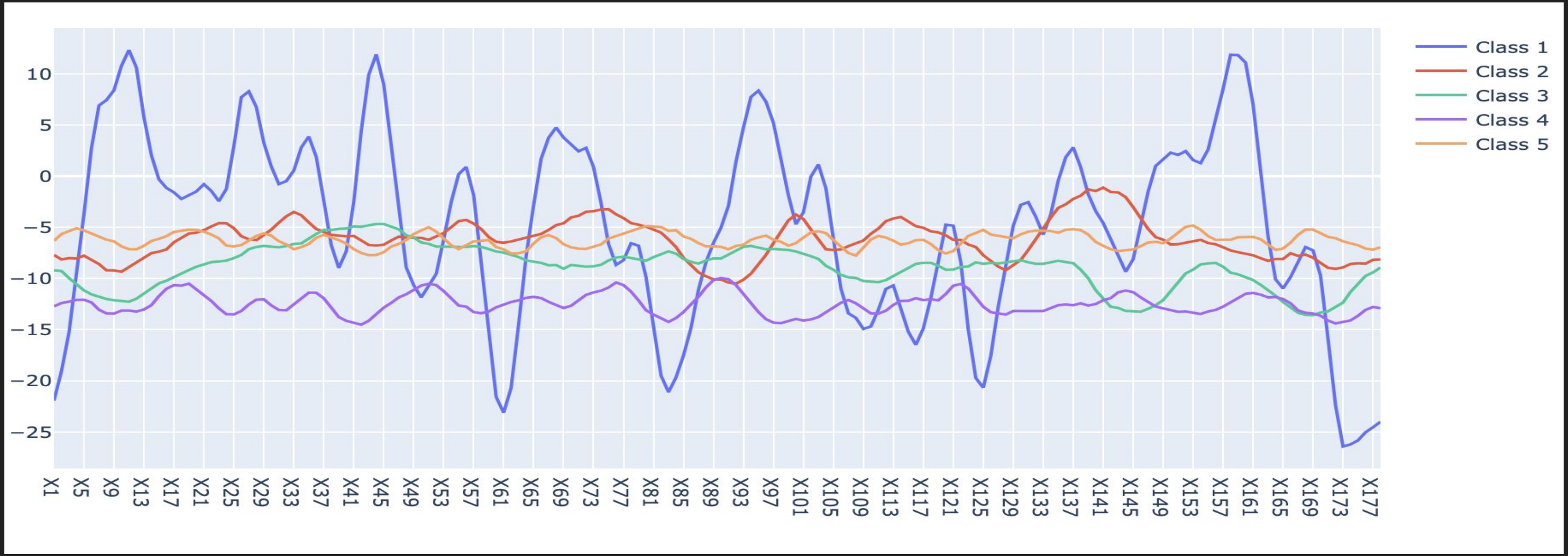
Exploratory Analysis

- A box plot of any attribute (X1-X178) might look like that there are many outliers.
- However, these are maximum/minimum values of recorded voltages and represent relevant information to determine the degree of seizure.
- Besides, these data points are well distributed within the curve.
- Therefore, these data points should not be treated as real outliers.



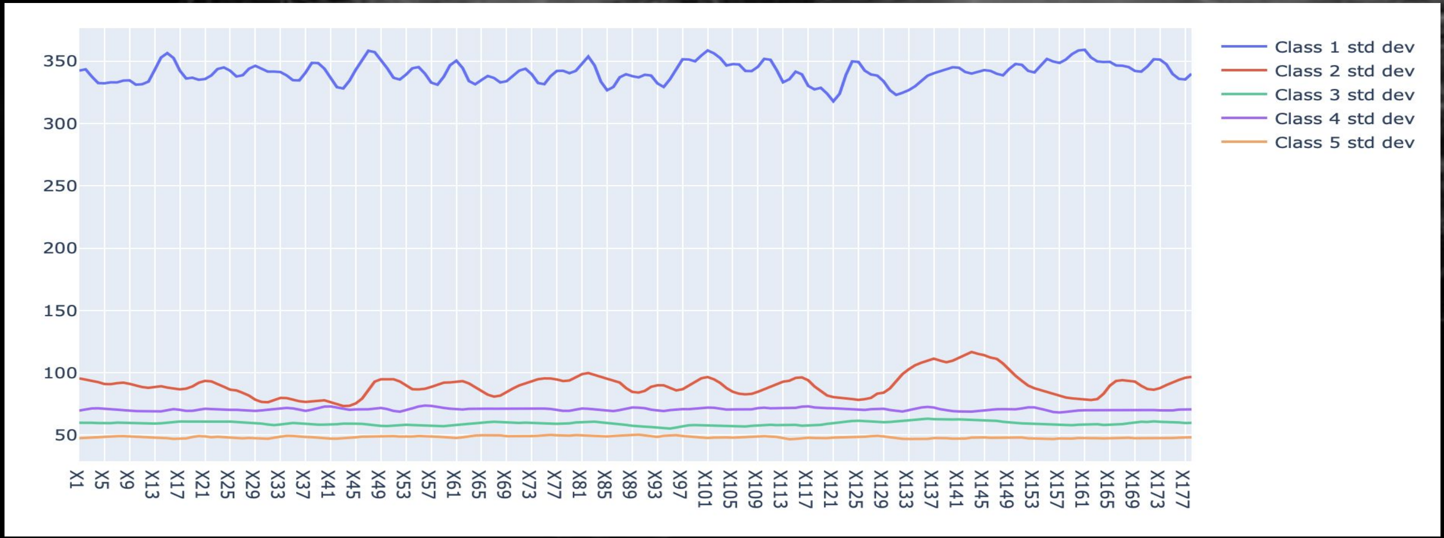
Exploratory Analysis

- Plot line graph using mean values for every timestamp, for the different classes
- Class 1 has much larger fluctuations as compared to the other classes



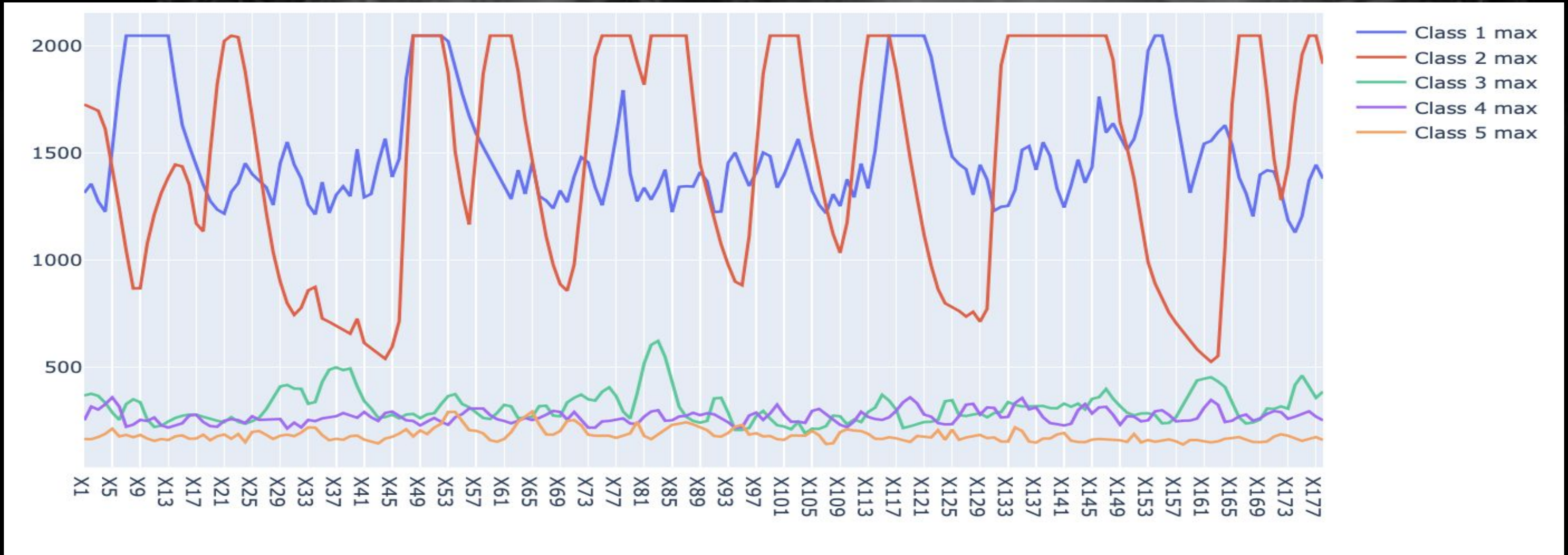
Exploratory Analysis

- Plot line graph using standard deviation for every timestamp, for the different classes
- Huge disparity between standard deviation of the EEG values for class 1 and the other classes, clearly distinguishes the seizure class from the rest
- Class 5 std dev < class 3 std dev < class 4 std dev < class 2 std dev
- Class 2 std dev shows slightly larger fluctuations as compared to class 3, 4 and 5



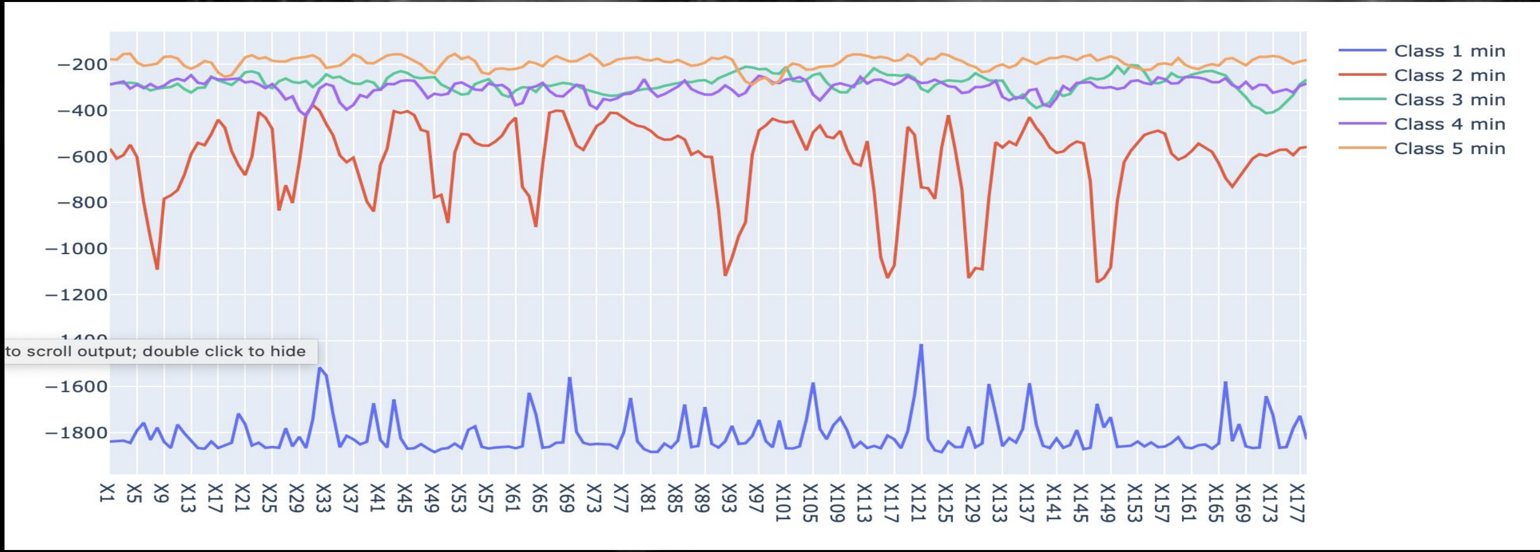
Exploratory Analysis

- Plot line graph using maximum value for every timestamp, for the different classes
- Max EEG values for every timestamp for class 1 and 2 are much higher than that of class 3, 4 and 5
- Interestingly, although both class 1 and 2 had similar peak maximum EEG values of just above 2000 microvolts, class 2 showed greater fluctuations than class 1

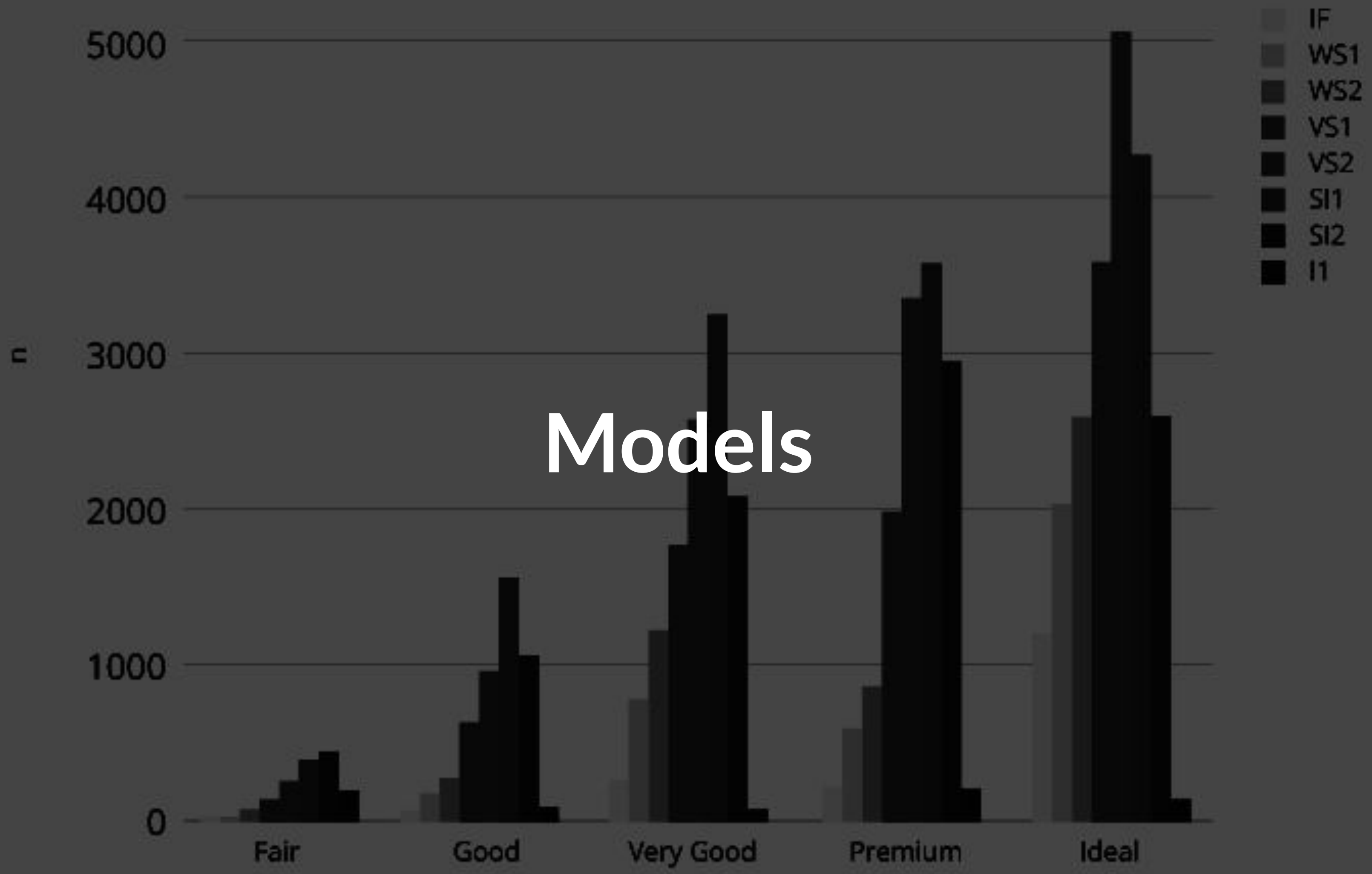


Exploratory Analysis

- Plot line graph using minimum value for every timestamp, for the different classes
- It is clear that class one has the lowest minimum EEG value
- Although the minimum value for class 2 is not as low as class 1, it showed greater fluctuations than all the other classes

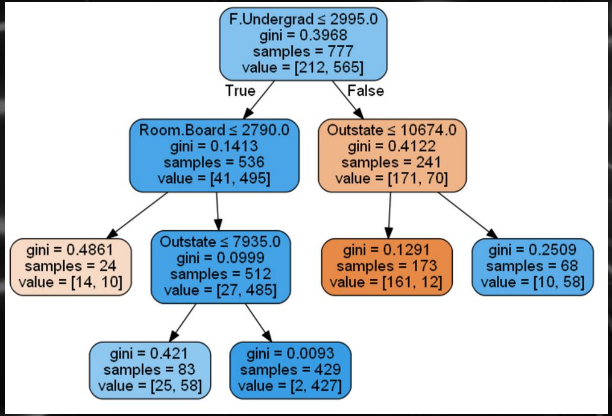


Models

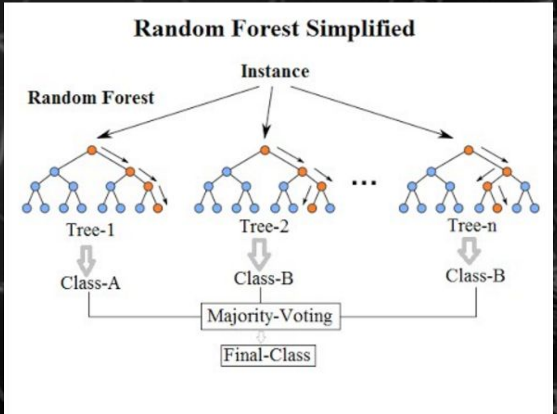

5000
4000
3000
2000
1000
0
n
Fair
Good
Very Good
Premium
Ideal
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1

Models:

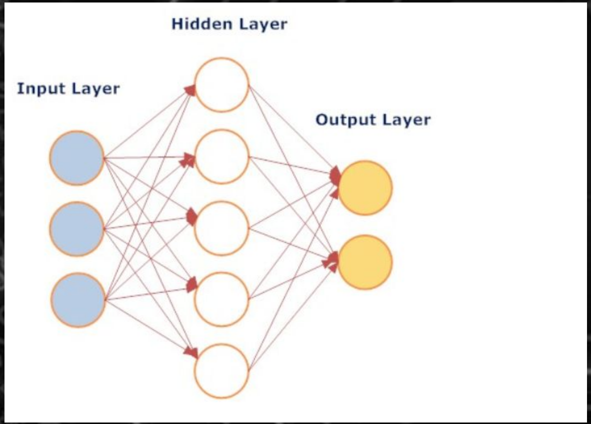
Decision Tree



Random Forest



Neural Network



Hypothesis:

Neural Network is the best model in detecting epileptic seizure activity

Decision Tree

Purpose: For classification between epileptic seizure activity or normal brain activity based on EEG readings.

How: Analyse the EEG readings of the train dataset, and create a model which classifies EEG readings under 2 groups (epileptic vs nonepileptic)

Why: Decision tree is more intuitive and simpler to understand.

Conclusion: Accuracy of 92-95%

Random Forest

Purpose: For classification between epileptic seizure activity or normal brain activity based on EEG readings.

How: Made up of many decision trees that:

- Uses random sampling to create individual decision trees
- Considers random subset of features when splitting nodes
- Final classification based on averaging each individual tree

Why: Decision tree is able to classify epileptic seizure to a greater accuracy.

Conclusion: Accuracy of 97-99%

Neural Network(Class 2/3 vs Class 1)

Model 1: Normal Neural Network Model

Purpose: to compare the effectiveness of Neural Network Models vs Decision tree/Random Forest Models

How: feedforward neural network model which maps sets of input data onto a set of outputs by going through many **layers of nodes in a directed graph**

Conclusion: Test accuracy: 0.93-0.95%
Not significantly better than Decision Tree models! --------->LSTM Neural Network Model

GaussianDropout()

```
model = keras.Sequential([
    keras.layers.Flatten(input_shape=(178,1)),
    keras.layers.Dense(1000, activation=tf.nn.relu),
    keras.layers.GaussianDropout(0.1),
    keras.layers.Dense(1000, activation=tf.nn.relu),
    keras.layers.GaussianDropout(0.5),
    keras.layers.Dense(1000, activation=tf.nn.relu),
    keras.layers.GaussianDropout(0.1),
    #keras.layers.Dense(1000, activation=tf.nn.relu),
    #keras.layers.Dropout(0.1),
    keras.layers.Dense(1, activation=tf.nn.sigmoid),
])

model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

model.fit(trainX, trainy, epochs=200, batch_size=320

test_loss, test_acc = model.evaluate(testX, testy)
```

Neural Network(Class 2/3 vs Class 1)

Model 2: Long short-term memory(LSTM) Neural Network model

Purpose: A type of recurrent neural network that are able to learn and remember over long sequences of input data, more suited for time series EEG data as compared to a normal neural network model

How: They can store information about previous values and exploit the time dependencies between our datapoints.

Conclusion: Test accuracy: 0.98-0.99~%
Much more accurate than all the models used before!

```
def create_model():
    model = Sequential()
    model.add(LSTM(100, return_sequences=False, input_shape=(178, 1)))
    model.add(Dropout(0.5))
    #model.add(LSTM(100)) dramatically worse results
    model.add(Dense(1, activation='sigmoid'))

    model.compile(loss='binary_crossentropy',
                  optimizer='rmsprop',
                  metrics=['accuracy'])

    model.fit(trainX3, trainy3, batch_size=16, epochs=50)
    #score = model.evaluate(testX, testy, batch_size=16)
    test_loss, test_acc = model.evaluate(testX3, testy3)
    print('Test accuracy:', test_acc)

    return model
```

LSTM neural network model is the best at classification of EEG data in epileptic & non-epileptic activity

Cross Validation of Models

K-Fold Cross Validation

-How : Used for classification models and if target variable is binary

- A resampling procedure
- Compares predicted values vs actual values
- Scores the model based on the accuracy of classification

Cross Validation Score

Decision Tree

```
Classification Accuracy (Group 3 Test Set)      : 0.941304347826087
Cross validation score :  0.8653870556745733
```

Random Forest:

```
ROC AUC Score for Group 3 Test Set:  0.9987887046710576
Cross validation score :  0.9500667868830561
```

Neural Network:

```
Test accuracy: 0.9847826361656189
0.9934782683849335
```
(Cross validation score)

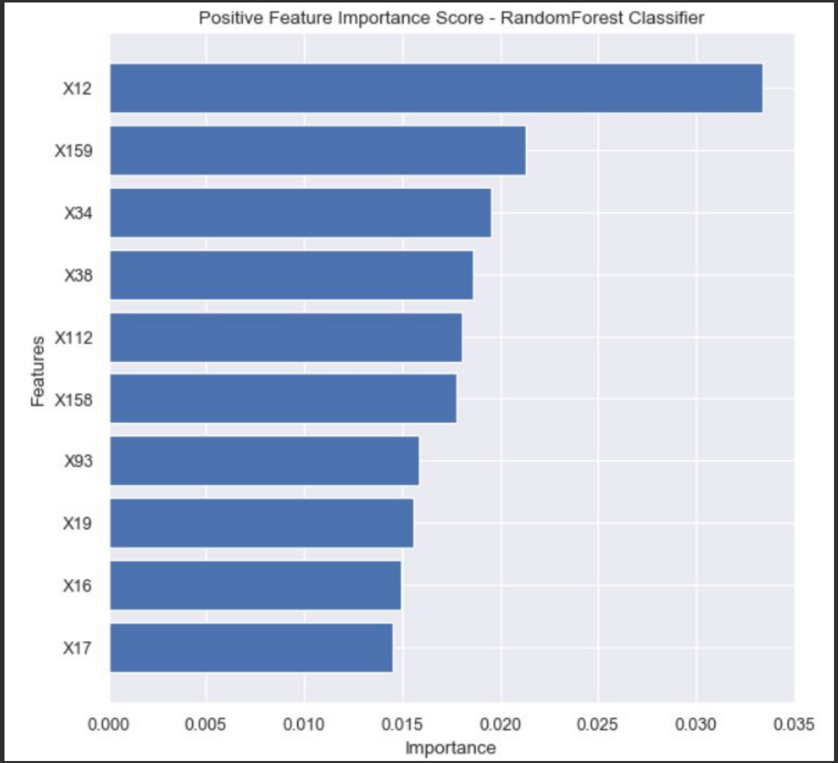
Conclusion

Project Outcome

Feature Importance

Q: Are there any features (X1-X178) that are more important than another?

A: Theoretically, **No!**

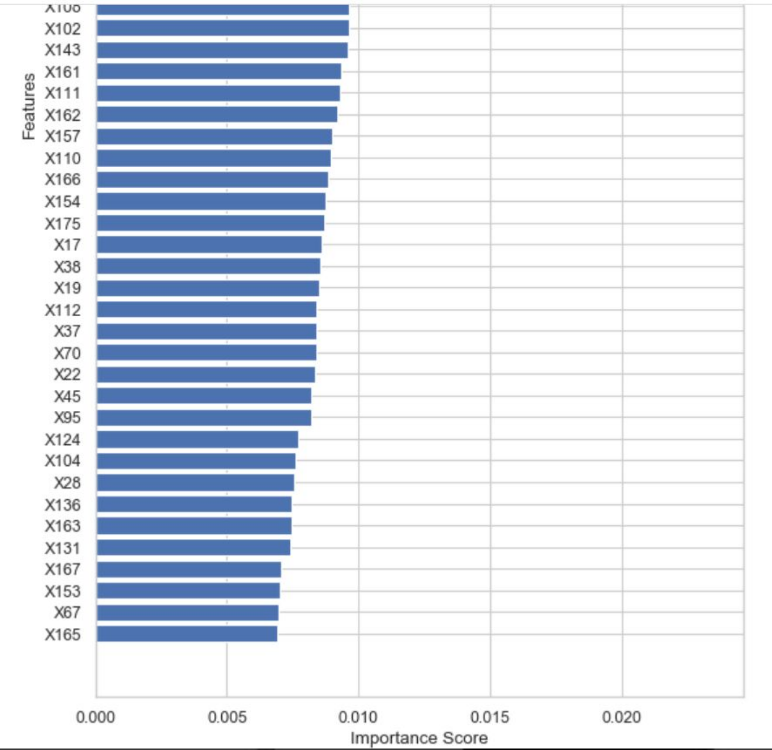


Exploratory Analysis

Feature Extraction

Q: Are there any features(X1-X178) that are more important/removable?

A: Theoretically, No!



Project Outcome

(Recall)

Objective: Creating the most **accurate model** for classifying epileptic seizure activity using **EEG readings**.

ML Problem : Classification

- Epileptic seizure readings have higher fluctuations compared to normal brain activity readings.
- Best Model in classifying epileptic readings → **LSTM Neural Network**
- Best Variable building an accurate model → **Class 3**

Contributions

Wen Qing:

- Formulation of problem & objective
- Neural Network Models
- Cross Validation of models

Leonel:

- Data structuring
- Exploratory Analysis of data
- Data Visualisation

Wee Li:

- Data cleaning/structuring
- Decision Tree model
- Random Forest model

Thank You!

References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4045570/
2. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Fast fourier transform	(i) Good tool for stationary signal processing (ii) It is more appropriate for narrowband signal, such as sine wave (iii) It has an enhanced speed over virtually all other available methods in real-time applications	(i) Weakness in analyzing nonstationary signals such as EEG (ii) It does not have good spectral estimation and cannot be employed for analysis of short EEG signals (iii) FFT cannot reveal the localized spikes and complexes that are typical among epileptic seizures in EEG signals (iv) FFT suffers from large noise sensitivity, and it does not have shorter duration data record	Frequency domain	Narrowband, stationary signals