MINI PROJECT:
Epilepsy Data Set

DONE BY:

Zechary Hoe Wen
Lee Yun Fai
Goh Yi Min

TABLE OF CONTENTS

01: Introduction To Given Dataset

Overview

- 5 different folders
- Each with 100 files
- 1 file ≡ 1 person

Sampling

- 1 file → Recording of brain activity for 23.6 seconds
- Each data point = Value of EEG at different point in time

In Total

- 500 Individuals
- 4097 data points each
- For 23.5 seconds

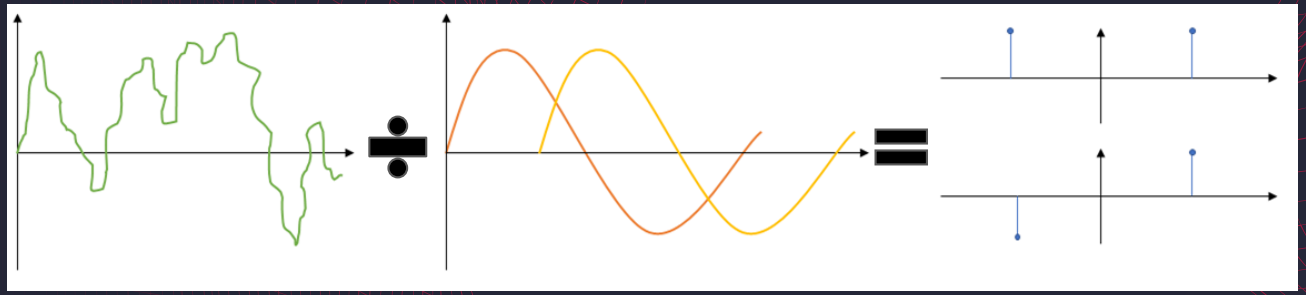
5 Categories:

1	Recorded EEG during patient's seizure
2	Recorded EEG from patient's brain tumor area
3	Recorded EEG from patient's healthy brain area
4	Recorded EEG when patient's eyes closed
5	Recorded EEG when patient's eyes open

Our Objective:

To classify the discrete categorical bins for each of the y labels 1,2,3,4,5.

02: Fast Fourier Transform

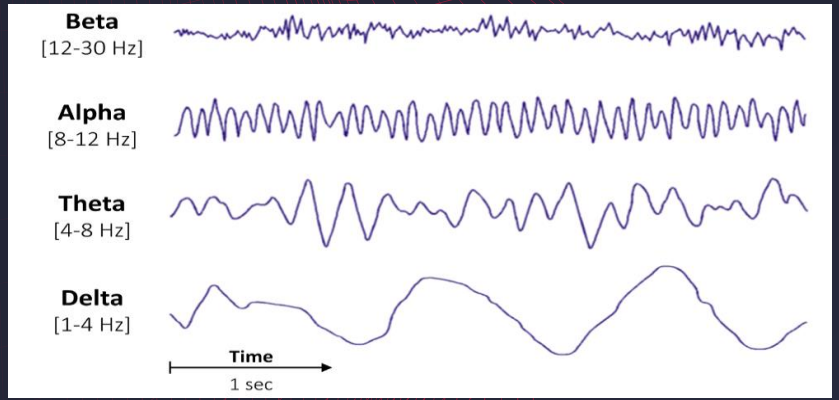


Key Idea:
Any continuous signal can be broken down into sinusoids and fully described with just:

- **AMPLITUDE**
- **FREQUENCY**
- **PHASE.**

What we did:

1. **Use FFT to categorise EEG Waves into 4 bands:**

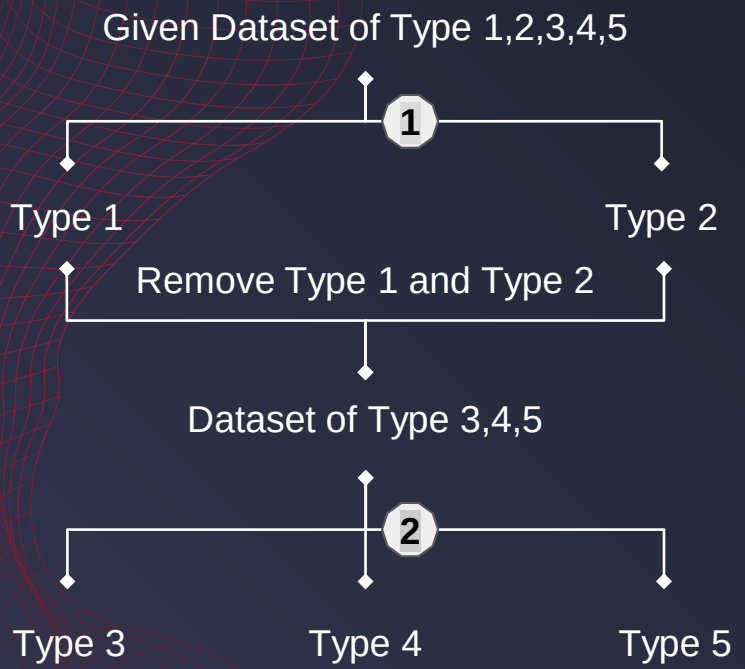


2. **Apply various filters on dataset for pre & post processing:**
   - Zero Padding using Gaussian function
   - Smoothing using Butterworth Low Pass Filter

2. **Produce the variables for Machine Learning use:**
   - Max, Min, Mean, Standard Deviation, Variance, Zero Crossings, …
   - Export as new .csv file

03: Machine Learning: Summarised Approach

Our double-layered classification approach:



Use new .csv file to continue Data processing for features extraction, recognition & classification:

→ Countplot, Scatterplot, Heatmap, Boxplot, KDE, Pairwise Correlation Scatterplot, Parallel Coordinates, 3D Plot

Classify Type 1 and 2 from given dataset:

→ Random Forest Classifier, Support Vector Machine, Logistic Regression, K Nearest Neighbours, Naive Bayes, Principal Component Analysis

Multiclass classification of Type 3,4,5:

→ Decision Tree, Extreme Gradient Boosting: XGBoost, Adaboost, Model Stacking: Vecstack

+

→Same as Type 1 & 2 classification: Random Forest Classifier, Support Vector Machine, Logistic Regression, K Nearest Neighbours, Naive Bayes, Principal Component Analysis

03: Machine Learning: Categorising Type I and 2

Random Forest	Support Vector Machine	Naives Bayes
- Made up of many decision tree. 1. Random sampling of training data points when building trees. 1. Random subsets of features considered when splitting nodes.	1. Plotting of data as a point in the n-dimensional space. 1. Perform classification by finding the hyperplane that differentiates the classes.	- Assumption of independence among predictors. 1. By finding the probability of each feature that will lead the result to being true. 1. Use the Naive Bayesian equation to calculate the probability.

03: Machine Learning: Categorising Type I and 2

K- Nearest Neighbours (KNN)	Logistic Regression	Principal Component Analysis
Assumes similar things are in close proximity. 1. Initialize K neighbours 1. Calculate and relate each point to nearest centroid 1. Recompute the centroid of cluster for each cluster of data point 1. Iterate step 2 and 3.	Used when dependent variable is categorical. It predicts the probability of occurrence by utilizing a logit function. 1. Predict probability of binary response based on one or more independent variables 1. Predict an outcome which has two values such as 0 or 1, pass or fail, yes or no etc.	PCA combines output variables in a specific way, and drop the least important variable. 1. Reduce the number of variables of a data set, while preserving as much information as possible.

03: Machine Learning: Categorising Type I and 2

Decision Tree	Adaboost	XGBoost	Vecstack
Partitions made in the data space methodically and proceed with binary decision. 1. Splitting the training set into subsets and find the gini coefficient. 1. The again split the subset again and find the gini coefficient. 1. Continue for the rest of the variables.	Ensembling reduces noise, bias and variance when training datasets. 1. Uses sequential ensemble (Boosting) 2. More models are used when previous models fail 3. More weight is given to better performing models 4. Tries to reduce bias and better for short decision trees	1. Uses Parallel Ensemble (Bagging) 2. Models are independent 3. Weighted Average is equal in XGBoost 4. Prone to overfitting but more versatile	1. Using more than one first level models to make predictions 1. Use the predictions as features to fit one or more second level model.

04: Outcome
Summary of classification:

	TYPE 1 CLASSIFICATION	TYPE 2 CLASSIFICATION	TYPE 3,4,5 CLASSIFICATION
Random Forest	**97.35**	79.13	84.13
Support Vector Machine	97.17	**79.57**	**84.93**
Naives Bayes	95.91	66.79	79.57
KNN	94.61	75.60	84.20
Logistics Regression	80.78	74.24	78.91
Principal Component Analysis	71.00	-	-
Decision Tree	-	-	78.67
Adaboost	-	-	78.09
XGBoost	-	-	83.25
Vecstack	-	-	83.01

Joint Statement:

With regards to choosing the problem, all of us chose this Epilepsy Data Set together. We were able to find information on how to categorise Type 1 and Type 2 online, on whether or not the person was experiencing seizure or brain tumor. However, to categorise Type 3,4,5 was still unchartered territory and to allow us the opportunity to explore it made this dataset very interesting. Thus, together, we chose the objective to classify the discrete categorical bins for each of the y labels 1,2,3,4,5 for Type 1,2,3,4,5 respectively.

Using EpilepsyFFT.ipynb, with exploratory data analysis / visualisation to understand the data, Zechary plotted the countplot, looked at the data.csv file and promptly realised that with just the given dataset, it was not possible to categorise the types accordingly without preparing the dataset to suit our objective. Yun Fai and Yi Min agreed.

To prepare the dataset to suit our specific problem definition, we decided to make our own variables using the values of EEG given. All three of us worked on the code for Fast Fourier Transform and each produced 6-7 variables (Max, Min, ZeroXings, RelativeAlpha…) to be used for machine learning afterwards. Altogether, we agreed on 20 variables and after making sure the code is right, we exported them into My_ML_DF.csv file.

We then proceeded for another bout of exploratory data analysis / visualisation on our new My_ML_DF.csv file. This time much more useful than before. Yi Min plotted sample Time Domain graphs for all the signals with different axes (EEG values in frequency domain & Magnitude v Frequency in Hz). Yun Fai computed the power spectral density using Welch's method and colour coded the EEG bands (alpha, beta, theta, deta) accordingly to visualise them. Zechary showed convolution of time signal with Gaussian functions. He plotted time and frequency domain charts to show the labelled peaks and valleys.

Further preparing the dataset to suit our specific problem definition, Zechary did zero padding, Yun Fai did smoothing using the Butterworth Low Pass Filter and Yi Min used the Blackman Windowing Technique. Yi Min plotted the peak frequencies after the Butterworth Low Pass Filter, Zechary found for the row data after convolution, the number of peaks detected and the first 3 max values of the peaks. Yun Fai found the number of zero crossings and printed them beside the plotted graph.

Going on to EpilepsyML.ipynb, for further data visualisation, Zechary plotted the scatter plots for all Type 1-5, Heatmap for correlation and visualised the uni-variate distributions of number of peaks and zero crossings via boxplots, histograms and violin plots. Yi Min plotted distributions of all statistical variables, boxplots of wave bands and KDE plots. Yun Fai plotted correlation plots for all parameters and the Pairwise Correlation Scatter Plots. Yi Min and Yun Fai worked together to produce the Parallel Coordinates Graphs for different variables we deemed to be possible to give a pattern of sort for a hint towards Machine Learning. Last of our data visualisation attempts before Machine Learning, Zechary plotted a 3D plot with 3 axes to explore if the additional axis would be able to show the data better for our dataset. We agreed that it was a good additional way to visualise and kept it within our Jupyter Notebook file.

Joint Statement:

Moving on to the use of data science / machine learning to solve the problem, our group understood that without trying, it is impossible to know which machine learning technique is best suited for our dataset. Therefore, we had chosen 10 different Machine Learning techniques and applied them via our double-layered classification approach. Starting with techniques that can be applied on both binary and multi-class datasets, Zechary applied the Random Forest Classifier, Support Vector Machine (SVM), Naive Bayes and Principal Component Analysis (PCA). Following, Yun Fai applied the K-Nearest Neighbour and Yi Min applied the Logistic Regression. For the multi-class classification of Type 3,4,5 only, Yun Fai and Yi Min used 2 different approaches respectively in the decision tree to find out the best set of variables to use before trying to improve the classification accuracy of the decision trees. Ultimately, we decided to use Approach 1 with all the variables in the decision tree. Yun Fai improved on the classification accuracy via boosting algorithms in XGBoost and Adaboost while Yi Min improved via Model Stacking i.e Vecstack. With all that done, we reached the end of the project with the summarised results displayed in Slide 8.

At the end of this project, we think it is fair to say that we have learnt a lot more beyond the course. To understand the EEG, all three of our group members read up a lot on medical literature to know the EEG differences between healthy individuals, brain tumor patients, epilepsy patients and so on. To prepare the data set, we researched and went about in our own direction to expand the number of variables, not by a few but by 20 of them for use in Machine Learning. To deepen our understanding of our own data, we went above and beyond for data visualisation, using self-learnt methods such as Pairwise Correlation Scatter Plots, Parallel Coordinate Graphs and 3D Plots. Especially for the Machine Learning section, we invested a tremendous amount of time into learning new Machine Learning techniques and applied the new knowledge into this problem.

Overall, all three of us had put in equally great amounts of effort. We worked with a clear direction towards attaining our objective and possessed open, honest communication amongst us. Each of our strengths made up for the other's weaknesses and honestly, it was a fun and interesting project that we enjoyed doing.