**Thesis Title:**

*Classification of the optimal push and relax visual imagery EEG features by standard machine learning and deep learning models for the use in Brain Computer Interface systems*

**Research Questions:**

What are the data cleaning and artifact detection techniques which achieve maximum data volume-quality balance?

What are the feature extraction techniques best representing push and relax visual imagery electroencephalogram data in the feature space of machine learning models in the context of BCI systems?

What are the advantages of employing DL architectures in developing visual imagery based BCI systems compared to traditional ML (TML) methodologies?

**Hypothesis 1:**

The cleaning and artifact detection techniques which achieve maximum data volume-quality balance are Independent Component Analysis (ICA) and peak-to-peak amplitude.

**Hypothesis 2:**

Feature extraction techniques for VI based EEGs that optimise ML models’ performance are Power Spectral Density, Common Spatial Patterns (CSP) and Principal Component Analysis (PCA).

**Hypothesis 3:**

In the context of moderate datasets, TML models achieve higher accuracy in classifying visual imagery EEG signals than DL models when fed pre-processed and feature-extracted data.

**Hypothesis 4:**

Raw, noisy visual imagery EEG data can be fed into a deep learning model and achieve a satisfactory classification accuracy.

**Research aim:**

The aim of this work is to explore and outline efficient ways in which EEG data can be pre-processed and classified for BCI systems. Main points of exploration are establishing a systematic procedure for feature-engineering with the potential for its automation, determining features representing VI most strongly as well as compare TML models’ performance to that of DL models.

**Rationale:**

The main rationale for posing the above research questions is ameliorating the quality of life of LIS patients by contributing to the improved functioning of BCI systems.

**Commentary on the code:**

* First, loading all 30 datasets
* Confirming reference electrodes, selecting the 14 electrodes of interest, setting 10-20 international montage
* 0.16 Hz high-pass filtering it to remove OC drifts
  + Is it built-into Emotiv Epoc X? If so, is re-filtering it OK?
* Notch filtering it at 50 Hz to remove powerline noise
  + Is it built-into Emotiv Epoc X? If so, is re-filtering it OK?
* Annotating the raw data so that we have 5 instances of push VI, 5 instances of relax VI per dataset
* Creating events from annotations and then epochs from events
  + We have 150 relax/push VI instances
  + Each instance originally lasted 10 seconds, but was cropped to 0.5 s – 9.5 s to remove edge artifacts
* Applying different cleaning & artifact detection mixes
  + Mix 1: ICA 🡪 PTP with 200 µV
  + Mix 2: PTP with 200 µV 🡪 ICA
    - Rationale: preventing ICA components to be dominated by large artifacts by deleting them first with PTP
  + Mix 4: ICA 🡪 Autoreject
  + Mix 5: Autoreject 🡪 ICA
    - Rationale: preventing ICA components to be dominated by large artifacts by deleting them first with PTP
* Feature extraction:
  + Frequency over time information
  + PyEEG features
  + CSP explanation
  + PCA explanation
  + RGB data for images of spectrograms
* Data exploration to inform feature engineering
  + Visualizing averaged epochs per class to look for timing of differences
  + Plotting powers of alpha, beta, gamma across space and across time
  + Calculating absolute values of power differences per channel per frequency band per timepoint between the two classes
    - Dividing the time dimension into 3 parts, calculating average of the power difference of a given frequency band per channel per 3 parts
    - Plotting it to inform feature selection in terms of frequencies, timing and space
  + Running sliding estimator on raw epochs to inform feature selection in terms of timing
  + Running sliding estimator on time-frequency data (TFD) to inform feature selection in terms of timing
* Feature engineering
  + Cropping raw epochs time-wise based on sliding estimator run on raw epochs
  + Cropping TFD time-wise based on sliding estimator
    - Extracting timepoints that achieved above 60% accuracy in sliding estimator
  + Cropping TFD time-wise based on the calculated differences per frequency band per segments (I, II, III) of the epoch
  + Cropping TFD space-wise based on the calculated differences per frequency band per channel per segments (I, II, III) of the epoch
* Classification
  + The below 96 input modalities will be classified per cleaning mix by LDA, SVM, 2 DL models
  + Giving in total 384 different input modalities per classifier
  + The below tables will be condensed and feature modalities categorized more broadly

|  |  |  |  |
| --- | --- | --- | --- |
| CLEANING MIX 1  \*raw\_uncropped means epochs were not cropped before feature extraction  \*raw\_cropped means epochs were cropped before feature extraction | | | |
| Input modality | Description | Mean Accuracy | Standard Deviation |
| raw\_uncropped\_alpha | Alpha frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_beta | Beta frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_gamma | Gamma frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_alpha\_beta | Alpha & beta frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_alpha\_beta\_gamma | Alpha, beta and gamma frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_alpha\_gamma | Alpha & gamma frequencies vectors from all timepoints |  |  |
| raw\_uncropped\_alpha\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_beta\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_alpha\_beta\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_alpha\_beta\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_alpha\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_uncropped\_alpha\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_beta\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_beta\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_beta\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_beta\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_beta\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_beta\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_TFD\_space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_beta\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_alpha\_beta\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_time\_space\_mix1 | Best timepoints with best performing channels |  |  |
| raw\_uncropped\_time\_space\_mix2 | Best timepoints with best performing channels |  |  |
| raw\_uncropped\_time\_space\_mix3 | Best timepoints with best performing channels |  |  |
| raw\_uncropped\_time\_space\_mix4 | Best timepoints with best performing channels |  |  |
| raw\_uncropped\_alpha\_beta\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_alpha\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_uncropped\_pyeeg\_all |  |  |  |
| raw\_uncropped\_pyeeg\_all |  |  |  |
| raw\_uncropped\_fractal |  |  |  |
| raw\_uncropped\_pyeeg\_all |  |  |  |
| raw\_uncropped\_pyeeg\_1 |  |  |  |
| raw\_uncropped\_pyeeg\_2 |  |  |  |
| raw\_uncropped\_pyeeg\_3 |  |  |  |
| raw\_uncropped\_pyeeg\_4 |  |  |  |
| raw\_uncropped\_pyeeg\_5 |  |  |  |
| raw\_uncropped\_pyeeg\_6 |  |  |  |
| raw\_uncropped\_pyeeg\_7 |  |  |  |
| raw\_uncropped\_pyeeg\_7 |  |  |  |
| raw\_uncropped\_PCA |  |  |  |
| Raw\_uncropped\_CSP |  |  |  |
| raw\_cropped\_alpha | Alpha frequencies vectors from all timepoints |  |  |
| raw\_cropped\_beta | Beta frequencies vectors from all timepoints |  |  |
| raw\_cropped\_gamma | Gamma frequencies vectors from all timepoints |  |  |
| raw\_cropped\_alpha\_beta | Alpha & beta frequencies vectors from all timepoints |  |  |
| raw\_cropped\_alpha\_beta\_gamma | Alpha, beta and gamma frequencies vectors from all timepoints |  |  |
| raw\_cropped\_alpha\_gamma | Alpha & gamma frequencies vectors from all timepoints |  |  |
| raw\_cropped\_alpha\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_beta\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_alpha\_beta\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_alpha\_beta\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_alpha\_gamma\_TFD\_time1 | Timepoints where SE was over X % |  |  |
| raw\_cropped\_alpha\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_cropped\_beta\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_cropped\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_beta\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_beta\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_gamma\_TFD\_time2 | Timepoints where absolute differences were the biggest |  |  |
| raw\_uncropped\_alpha\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_beta\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_beta\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_beta\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_gamma\_TFD\_space1 | Channels where absolute differences were the biggest |  |  |
| raw\_cropped\_alpha\_TFD\_space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_beta\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_alpha\_beta\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_time\_space\_mix1 | Best timepoints with best performing channels |  |  |
| raw\_cropped\_time\_space\_mix2 | Best timepoints with best performing channels |  |  |
| raw\_cropped\_time\_space\_mix3 | Best timepoints with best performing channels |  |  |
| raw\_cropped\_time\_space\_mix4 | Best timepoints with best performing channels |  |  |
| raw\_cropped\_alpha\_beta\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_alpha\_gamma\_TFD\_ space2 | Except channels where differences were lowest |  |  |
| raw\_cropped\_pyeeg\_all |  |  |  |
| raw\_cropped\_pyeeg\_all |  |  |  |
| raw\_cropped\_fractal |  |  |  |
| raw\_cropped\_pyeeg\_all |  |  |  |
| raw\_cropped\_pyeeg\_1 |  |  |  |
| raw\_cropped\_pyeeg\_2 |  |  |  |
| raw\_cropped\_pyeeg\_3 |  |  |  |
| raw\_cropped\_pyeeg\_4 |  |  |  |
| raw\_cropped\_pyeeg\_5 |  |  |  |
| raw\_cropped\_pyeeg\_6 |  |  |  |
| raw\_cropped\_pyeeg\_7 |  |  |  |
| raw\_cropped\_pyeeg\_7 |  |  |  |
| raw\_cropped\_PCA |  |  |  |
| raw\_cropped\_CSP |  |  |  |

* + The above table will be more broadly categorized into the below classes, score averages of individual input modalities will be calculated per category:
    - Alpha frequencies
    - Beta frequencies
    - Gamma frequencies
    - Frequency mixes
    - pyEEG
    - PCA
    - CSP
  + Descriptive statistics:
    - Cross validation scores of all the runs within a category will be taken
    - Averages and SDs will be calculated
  + Raw data will be run through convolutional and recurrent networks
  + Spectrograms images will be run through convolutional networks
* Discussion
  + Commentary on the results
  + Describing advantages and disadvantages of various approaches (keeping in mind the need for automation and time-efficiency)
    - Can automated feature engineering improve accuracy by suggesting appropriate frequency bands, channels, timepoints?
    - Frequency powers, CSP, PCA, pyEEG – are those good feature extracting techniques?
    - Which classifiers are best for VI?
    - Are DL models promising? Why? Why not?
* Conclusion