CZ4041 MACHINE LEARNING

GRASP AND LIFT EEG DETECTION

ASSIGNMENT REPORT

GROUP 24

Competition Ranking

Ranking: 47.8th percentile

Ranking: 181/379

Accuracy: 84.8%

GROUP MEMBERS*

BUDHKAR AKSHAY

CHEN SARAH

YEO KAISHENG NICHOLAS

*All members of this group contributed equally to the varying portions of this project.

TABLE OF CONTENTS

1. INTRODUCTION

This report shall cover the competition chosen by the group and the project analysis and results.

1.1 COMPETITION OVERVIEW

For this assignment, the details of the chosen Kaggle Competition are as follows:

Context	Description
Title	Grasp-and-Lift EEG Detection
Link	https://www.kaggle.com/c/grasp-and-lift-eeg-detection
Sponsor	WAY Consortium (Wearable interfaces for hAnd function recovery)
Duration	29 Jun 2015 to 31 Aug 2015

The objective of this competition was to explore ways to use the collected data to obtain a better and more accurate prediction of the relationship between Electroencephalography (EEG) signals and specific hand motions. Brain-Computer Interfaces (BCI) integrated into hand prosthetics would be able to apply this knowledge in giving patients the ability to evoke actions with prosthetic hands more naturally, through the observance and processing of neural activity.

1.2 DATA SETS

This following section shall detail the data that has been provided for use for this competition. There were two main data sets provided by WAY Consortium, which can be obtained from https://www.kaggle.com/c/grasp-and-lift-eeg-detection/data.

The provided data sets contained a training set and a testing set, as illustrated in Figure 1.

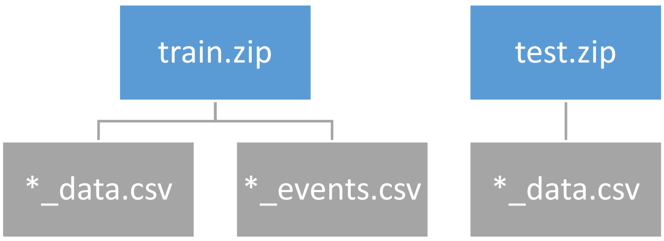

Figure 1: Breakdown of data provided

The data provided consisted of EEG readings of the subjects and the respective event of hand movement collected by WAY Consortium. The data collected can be further classified as follows:

Context	Quantity
# of Subjects	12
# of Series	10
# of Trials	~30

A detailed account of how the data was collected has been documented and can be found at: www.nature.com/articles.sdata201447.

In each trail, the subject is to perform a set directed series of events, which were described as the following series of Grasp and Lift (GAL) events:

- HandStart
- FirstDigitTouch
- BothStartLoadPhase
- LiftOff
- Replace
- BothReleased

The six mentioned events always occurred in the listed sequence and were conducted in the same environment (Figure 2).

Figure 2: Environment of Grasp and Lift Trial

In each *_data.csv file, the raw 32 channel EEG data (Figure 3) of the subject were individually collected for each channel and with a sampling rate of 500Hz (one reading per every two millisecond).

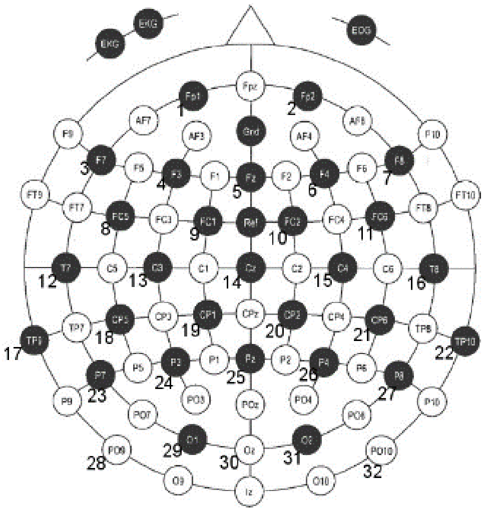

Figure 3: 32 Channels of EEG data collected

The *_events.csv files contained the ground truth frame-wise labels for all events.

1.3 PROBLEM STATEMENT

Given the provided organized data, our objective was to develop an algorithm that would process EEG data from different trials of conducted on differing individuals performing the same series of events to be able to self-train itself to accurately classify when one, if any, of the 6 GAL events as described in section 1.2.

2. CHALLENGES

The lack of prior relevant experience and knowledge in the subject, as well as the size of and variance within the dataset, posed a challenge for the team's participation in this competition.

2.1 BACKGROUND KNOWLEDGE

The interpretation and understanding proved to be a difficult task in itself due to the lack of the team members' knowledge in the field of EEG, psychology, and biology. The task required knowledge regarding how EEG experiments are conducted, how relevant information is extracted, and what range of frequencies of data may be considered relevant to motor actions, specifically, hand motions. Prior to forming a methodology for this task, the team conducted research on EEG and EEG-processing through the readings of multiple scholarly articles and research papers. Such research also enabled the team to outline additional challenges and points of emphasis in forming a methodology to complete this task.

In the interpretation of EEG data, it was learned that there are certain limitations of EEG, such as that it is only able to measure blurred cortical activities due to the diffusion that occurs through the skull and the skin of the subject. EEG signals are thus generally considered to be contaminated by noise from various sources such as ocular and myoelectric signals. Which areas of the brain and what frequency ranges are deemed relevant also vary by different academic and professional sources, however after comparison between multiple sources, the team decided on a range and area of focus for the completion for this task.

2.2 DATASET COMPLEXITY AND SIZE

With relation to the dataset, the size and complexity of the data itself also proved to be an obstacle with regards to pre-processing and interpretation of the data. The size of the data and events file for a single series of one individual was approximately 20 MB. Recordings of electrical currents were taken every 2ms at each of the 32 channels (Figure 3). For this reason, running any scripts to process the data would take a large quantity of time; training models using the entirety of the dataset had to be completed overnight, and just testing the model on Kaggle often took a long time as well.

The team had to decide how to narrow down the data to highlight and train the model on only relevant data values, which required that the data be pre-processed and filtered. Because the range of values was so varied and different across the different locations of the skull, and each individual's brain emitted differing frequencies and voltages of electrical signals, training could not be completed using the entirety of the data to apply for all subjects, but rather, training for individual subjects and individual actions had to done separately, as it was decided that the variance in each individual's behaviour for different action events would disrupt the accuracy of the model and not be applicable to all subjects.

This also meant that for the data to be understood and interpreted, any jumps in the frequencies would need to be smoothed so that the data would not include as many disruptions, and the model would be able to focus on patterns and not just individual frequency values.

2.3 TECHNICAL LIMITATIONS

The team chose to pursue the models described in the Section 3.3, Identifying Classification Models, in part due to the technical obstacles faced by the time. The team would have liked to experiment with more deep learning and with Convolutional Neural Network (CNN) models, but due to the computational expense and the inability to successfully test these models through the team members' computer systems, the team was not able to experiment with these models. Instead, we focused on the models that we had found performed well to adjust them for better accuracy and results.

3. METHODOLOGY

Given the dataset consisting of the data and corresponding events for the first eight series, the data to be submitted was to predict the events for Series Nine and Ten.

Our methodology used for this completion has been outlined as follows:

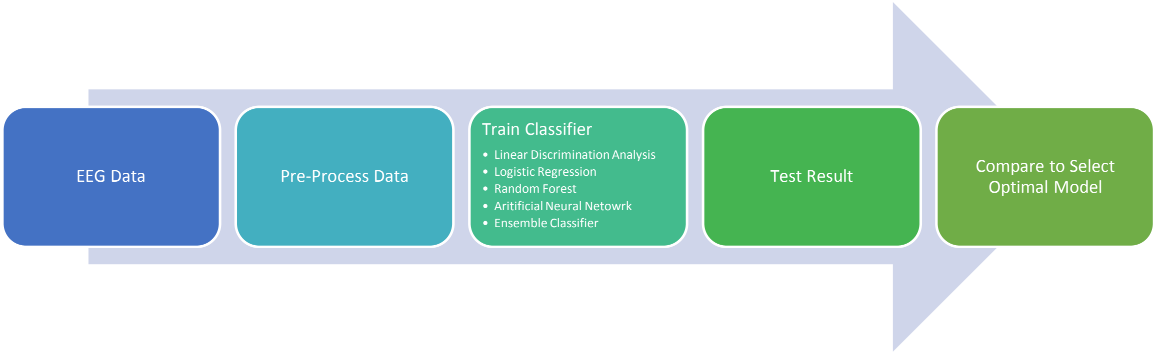

Figure 4: Strategy Process

3.1 SOFTWARE TOOLS

Python was the programming language used for this Machine Learning Assignment.

Anaconda [1], an Open Data Science Platform powered by Python was used as it contained the key packages required for data mining and manipulation such as scikit learn for machine learning and pandas for data manipulation.

3.2 PRE-PROCESSING EEG DATA

EEG is the recording of electrical activity diffused from the cortex; it measures voltage fluctuations resulting from ionic current flows from the cortex.

According to The Journal of Clinical Neurophysiology [2], EEG signals can be classified into different rhythms, as according to frequency, as seen in Table 1.

Rhythm	Frequency Range (Hz)*	Description
Alpha	8-13	Particularly evident during the absence of visual stimuli
Beta	14-30	Frontal region of brain, observed during concentration of subject
Gamma	30-100	Motor activities
Delta	0.5-4	Observed at stage 3 and 4 of sleep
Theta	4-8	Observed during light sleep and during hypnosis
Mu	7-11	Used in Motor Imagery (MI) BCI paradigm

*The frequency range defined usually differs slightly by author.

In this experiment, the main objective was to collect the EEG data associated with GAL hand movements. The hand Motor Imagery of a healthy subject results in a desynchronization (ERD) of Mu rhythms in contralateral EEG. Hand Motion Imagery also results in synchronization (ERS) of Beta rhythm in ipsilateral EEG.

In this scenario, the type of rhythm to consider can thus be constrained to the Mu and Beta rhythms, which lie within the range of 7-30 Hz.

LOW PASS FILTER

A healthy human brain generates waves at frequencies ranging from 0.5Hz to 100Hz. To filter out any values over 30 Hz, a bandpass filter was considered to filter out unwanted frequencies, as they are not of relevance in detecting hand motions.

Designing a bandpass filter with a lower bound very close to the value of 0 Hz value is difficult, especially with the in-built function in Python, which did not accept values very close to zero. The data provided included multiple instances of such values, so we decided to use a low pass filter instead, filtering out any value above the upper bound of 30 Hz.

The frequency of values in 0 – 7 Hz range was minimal, so this change did not allow for such values to significantly impact our pre-processing. Most of the values that would fall within that range are only observed in sleep or hypnosis, both of which are not present within an experiment on hand motions. We used a Butterworth filter to implement the third order low pass filter for uniform sensitivity in our desired subset of data, after rejecting the higher-valued frequencies.

We wanted to use the low pass filter as a way to smooth our data while still exposing the features for the classification models. To achieve this, the values obtained after passing the data through a low pass filter were concatenated to the pre-existing values.

All of our calculations involved normalized values (so that they laid between 0 and 1). To further annunciate the effect of smoothing, we decided to raise the low pass filtered values to the 8^{th} power, and once again concatenate it to the pre-existing matrix, already filled with the initial values and the low-pass filter values. This also helped in increasing the variance of the signal during hand motions and while at rest, distinguishing features for model classification. This new matrix was now used to both train and test the models.

3.3 IDENTIFYING CLASSIFICATION METHODS

Five classification methods were identified to be used:

- Linear Discriminant Analysis (LDA)
- Logistic Regression (LR)
- Random Forest Classification (RF)
- Stochastic Gradient Decent (SGD)
- Ensemble Classification

3.3.1 LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis classifier was selected as one of the models to use because it performed both the functions of dimensionality and feature reduction as well as classification. The LDA classifier is also commonly used for biological datasets for analysis and machine learning.

As previously mentioned in Section 2, one of the challenges that the team faced in this task was the large size of the data; because LDA maximizes class separation in its modelling, it was one of our preferred models of consideration. The large size of the dataset along with the structure of the LDA model was expected to yield high accuracy in large scale classification.

The LDA classifier was imported from scikit learn:

```
from sklearn.lda import LDA
```

3.3.2 LOGISTIC REGRESSION

The group decided to use a Logistic Regression model as one of our experimental models. For classification purposes, a Logistic Regression was more appealing than a linear model because we expected the data to fit a logistic model more than a linear model. In addition, while training the model, a logistic regression model would better be able to be more flexible to fit to the data. The logistic regression model would be less prone to disruptions from outliers captured by the EEG technology while recording. The Logistic Regression classifier would also be able to produce more precise results with such a large sample size, and so was considered as one of our optimal models.

The LR classifier was imported from scikit learn:

```
from sklearn.linear_model import LogisticRegression
```

3.3.3 RANDOM FOREST CLASSIFICATION

The RF classifier was identified as:

- It is a type of ensemble classifier
- RF has methods for balancing errors in datasets where classes are imbalanced
- RF is resistant to overtraining and outliers
- RF is very accurate for large datasets

The RF algorithm uses a boostrap aggregating method also known as bagging. This algorithm combines random decision trees with bagging and have each model in the ensemble vote with equal weight to achieve very high classification accuracy.

The RF classifier was imported from scikit learn:

```
from sklearn.ensemble import RandomForestClassifier
```

3.3.4 STOCHASTIC GRADIENT DESCENT

The SGD classifier was identified as:

- SGD is efficient with large scale datasets
- SGD provides the flexibility of tuning its parameters for better results

The SGD classifier was imported from scikit learn:

```
from sklearn.linear_model import SGDClassifier
```

The SGD Classifier implements a plain stochastic gradient descent learning routine, allowing different loss functions and penalties as variables to fine tune classifier results.

In our implementation, `loss="hinge"` was used, which is equivalent to a linear Support Vector Machine. This loss function is a lazy implementation which only updates the model parameters if an example violates the margin constraint, making training very efficient and may result in sparser models.

The concrete penalty for the SGD classifier was set at `penalty="l2"`.

3.3.5 ENSEMBLE CLASSIFICATION

Besides the use of RF as an ensemble classifier as mentioned in 2.3.3, another ensemble classifier was tested using equal voting between the Linear Discriminant Analysis and Logistic Regression models. The two models were chosen as for ensemble classification as:

- Averaging predictions often reduces overfit
- Ideal to have a smooth separation between classes
- A single model's prediction can be a little rough around the edges
- LDA and LR models performed decently individually

3.4 MODEL FITTING

To increase the time efficiency of identifying an optimal model, different sizes of sub-samples of the training data were used to train the models. All models were initially trained with a sub-sample size of 100. After comparing the results, the team decided to again train the better-performing models with a sample size of 1,000, and then once again with the full training dataset. All trials were tested with the provided two series of data (Series 9 & 10 as described in section 1.2) that WAY Consortium had designated as testing data.

4. RESULTS

For this competition, submission on Kaggle was evaluated using Mean-Column wise Area under Curve (AUC) on the test dataset. Rankings were made through comparisons with submissions made while the competition was still active out of the total of 379 applicable submissions.

4.1 SUBSAMPLE (100) OUTCOMES

The five models that were tested with a subsample of 100 can be found below, in order of best-performing to worst-performing. Of these models, the Random Forest Classification and Artificial Neural Network performed the least well and trained for the longest duration. Linear Discriminant Analysis and Logistic Regression both performed relatively similar to one another and comparatively, better than Random Forest Classification and the Artificial Neural Network model that we had implemented.

Model	Accuracy (%)	Ranking (Place)	Ranking (Nth Percentile)
Linear Discriminant Analysis	83.7	226	59.6
Logistic Regression	82.8	233	61.5
Random Forest Classification	78.4	261	68.9
Artificial Neural Network	54.2	334	88.1

4.2 SUBSAMPLE (1000) OUTCOMES

The team continued to test the two better performing models by training the models with a subsample of the data of 1000 instead. Contrary to our expectations, it appeared that the Logistic Regression model trained with a subsample of 100 performed better than that trained with a subsample of 1000. The Linear Discriminant Analysis tests did not follow the same pattern. Instead, the LDA model performed only slightly better when trained with 1000 subsamples.

Model	Accuracy (%)	Ranking (Place)	Ranking (Nth Percentile)
Linear Discriminant Analysis	83.9	224	59.1
Logistic Regression	77.4	262	69.1

4.3 FULL TRAINING DATASET OUTCOMES

In addition to training the Linear Discriminant Analysis and Logistic Regression models with the full training set, the team tested the ensemble model with both LDA and LR. LDA still performed the best out of these three models, followed by the Ensemble Classification, and finally, the Logistic Regression model. Although all three models performed similarly to one another in accuracy, in terms of rankings, the Ensemble Classification and Logistic Regression models performed more similarly. Comparatively, it seems that our trained LDA model performed the best out of all of our experiments.

Model	Accuracy (%)	Ranking (Place)	Ranking (Nth Percentile)
Linear Discriminant Analysis	84.8	181	47.8
Ensemble (LDA + LR)	84.7	225	59.4
Logistic Regression	84.0	226	59.6

Linear Discriminant Analysis was also the fastest to test. Ensemble classification was slower than both LDA and LR, due to the fact that it performed training and testing with both models and made comparisons between results from both.

5. CONCLUSION

In retrospect, learning and working with EEG data has proven to be a challenge, yet also a very interesting experience for the group. While exploring other related works, we have seen how CNN could be utilized to produce a better result, yet requires substantial amounts of processing power to train the model, which, as mentioned in Section 2.3, was a limitation for our group.

This competition allowed for us to see how knowledge learned in class can be applied to real-world situations and incorporated into human life for the betterment of society. In this scenario, we were able to work on an interdisciplinary task and learn about a different and more specialized field of study.

REFERENCES

[1] “Continuum | Home,” Continuum Analytics, [Online]. Available: https://www.continuum.io/.

[2] C. B. J. E. F. M. A. S. W. S. Noachtar*, “Guidelines of the International Federation of Clinical Physiology (EEG Suppl. 52),” *Recommendations for the Practice of Clinical Neurophysiology,* pp. 21-40.