Classifying EEG data into successful and failed response inhibition

Graduate school of natural sciences

Master thesis

June 2024

Student name:

Lara van Cappelle

l.c.vancappelle@students.uu.nl

Thesis supervisor:

Leon Kenemans

j.l.kenemans@uu.nl

Second supervisor:

Daan van Rooij

d.vanrooij@uu.nl

ABSTRACT

Response inhibition, an essential component of cognitive control, can be studied using machine learning algorithms like Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). This projects investigates the impact of different preprocessing and regularization techniques on the performance of these classifiers. The aim of the project was to increase accuracy to significantly above chance level (50%) for the binary classification of EEG power data into successful or failed trials, derived from a Stop Signal Task (SST). Notably, frontal theta power at the FCz electrode was identified as significantly different between these two trial types. Therefore, average theta power was used for the classification algorithms. Performance evaluation showed that outlier treatment, Synthetic Minority Over-Sampling Technique (SMOTE), and L2 regularization improved classification accuracy. Average weight vector analysis showed that the focus of the classifiers was distributed across frontal electrodes as well as central and left parietal and occipital electrodes. Future research should explore individualized optimization.

Keywords: response inhibition, LDA, QDA, SST

Contents

Introduction

Response inhibition

An essential feature of human cognition and action is that they can be used to achieve certain goals. Notably, goals can change rapidly and in their entirety. Fortunately, people can readily stop their actions in response to changes in their internal state or in the environment. For example, when a red traffic light suddenly turns green upon approaching, a driver is able to stop (their intention of) braking and move their foot to the gas paddle.

The ability of stopping an action is commonly referred to as response inhibition. Response inhibition involves the cancelation or suppression of behaviors that are inappropriate, unsafe, or no longer required (Verbruggen & Logan, 2008). The ability underlies a range of behaviors crucial for adaptive functioning and has a key role in the ability to respond flexibly in ever-changing environments (Congdon et al., 2012). Moreover, failures in response inhibition have been associated with ADHD, OCD, as well as addiction (Chambers et al., 2009).

The stop-signal task

The Stop-Signal Task (SST) is commonly employed to operationalize response inhibition (Logan et al., 1984). In this task, participants respond to a Go stimulus with a button press. However, in a subset of the trials (usually 25%), the Go stimulus is followed by a Stop signal after a varying delay interval. In these trials, participants attempt to withhold their response and restrain the button press. Performance is usually measured by calculating the Stop Signal Reaction Time. Performance in the SST can be modeled as a race between a ‘go process’ and a ‘stop process’, triggered by the stimulus and the stop signal respectively. The model assumes that the response is inhibited if the stop process finishes before the go process (Verbruggen & Logan, 2008). By dynamically manipulating the Stop Signal Delay (SSD), the task difficulty can be modified in a way that every participant’s success rate is approximately 50%. To establish a reliable prepotent response, it is recommended to use a two-choice reaction time task. Furthermore, the Stop signal should be salient and is therefore often auditory (Verbruggen et al., 2019).

Neural markers of response inhibition

Combining the SST with electroencephalography (EEG) has allowed researchers to identify neural markers associated with response inhibition. Identifying these markers has several benefits. Event related potential (ERP) components can be used for the clinical evaluation of patients with inhibition-related disorders, like ADHD and OCD (Chikara & Ko, 2020). Furthermore, neural markers of response inhibition are important for the development of brain-computer interface (BCI) systems. BCI systems depend on the recognition and classification of ERPs. For example, robotic arms can be moved based on left-hand and right-hand motor imagery. However, these mental states cannot stop the movement of BCI devices. Neural markers of response inhibition could be used as a stop command in BCI devices (Chikara & Ko, 2019).

Several ERP components have been associated with the response inhibition process. Differences in latency, amplitude, and scalp topography in the N2 and P3 ERP components have been observed between successful and failed stop trials (Kok et al., 2003). In addition, N1 amplitude has been found larger for successful stop trials than failed stop trials (Bekker et al., 2005; Skippen et al., 2020).

Furthermore, extensive research has investigated which frequency band activities are associated with response inhibition. In an SST, increased beta power has been observed following the stop signal (Raud et al., 2020). Specifically, right frontal beta power has been found to increase after the stop signal and before the time of stopping and is absent on go trials. Moreover, this beta power increase was significantly larger for successful compared to failed stop trials (Wagner et al., 2018). In addition, beta-burst volume has been found predictive of successful and fast stopping (Enz et al., 2021). Furthermore, a study using the Go/No-Go paradigm for response inhibition found elevated pre-stimulus occipital alpha activity preceding false alarms, so a response to the No-Go stimulus. In addition, such errors were associated with a post-error increase in frontal theta and decrease in posterior alpha activity (Mazaheri et al., 2009). They found no significant differences for pre-stimulus theta and beta activity between false alarms and correct withholds.

Proactive response inhibition

Most EEG studies using the Go/No-Go paradigm or SST focus on the time period following stimulus presentation, the No-Go or stop stimulus. However, pre-stimulus frequency band activities may reveal relevant information regarding the response inhibition process.

It is suggested that a distinction can be made between proactive and reactive mechanisms of response inhibition (Kenemans, 2015). The proactive mechanism involves top-down control signals that potentiate inhibitory sensory-motor connections, depending on whether a motor response inhibition is anticipated (Kenemans, 2015). Similarly, a distinction is made between proactive and reactive cognitive control, an umbrella term that includes response inhibition. Reactive control processes are required for conflict resolving and overcoming interference, whilst proactive control processes are involved in preparing the system to be sensitive to future relevant features of the environment (Braver, 2012). Previous research has shown that proactive control is associated with frontoparietal theta activity (Cooper et al., 2015; Cooper et al., 2017).

The pre-stimulus time window in the SST may reflect the proactive mechanism of response inhibition.

Localization of response inhibition

Proactive response inhibition is likely to originate from the right inferior cortex (Kenemans, 2015). Indeed, an MRI study using patients with lesions in the right frontal lobe showed that response inhibition can be localized to the right inferior frontal gyrus (rIFG) (Aron et al., 2003). The extent of damage to the rIFG correlated significantly greater than other regions of interest with task performance in the SST. Furthermore, a review of several fMRI studies shows that both the inferior frontal region and the pre-supplementary motor area (pre-SMA) are consistently related to response inhibition in the Go/No-Go task and the SST (Huster et al., 2013).

EEG and Multivariate Pattern Analysis (MVPA)

As described before, previous research has used EEG to investigate response inhibition. EEG is a low cost method that records electrical activities of cortical regions in the brain through electrodes

placed on the scalp, with advantages including non-invasiveness and high temporal resolution. However, EEG signals are complex, non-linear, non-stationary, and have a low signal-to-noise ratio as well as inter-individual variability (Saeidi et al., 2021).

To overcome the low signal-to-noise ratio, traditional EEG analyses rely on averaged segments of data. Furthermore, such analyses often require a priori selection of an electrode or set of electrodes. In contrast, Multivariate Pattern Analysis (MVPA) is a machine learning approach that enables the analysis of raw or decomposed segmented data without averaging (Takacs et al., 2020). Moreover, MVPA considers the relationship between multiple variables instead of treating them as independent (Grootswagers et al., 2017), which allows training classifiers on all channels, eliminating bias associated with a priori electrode selection (Takacs et al., 2020).

Linear Discriminant Analysis (LDA)

A specific MVPA method is linear discriminant analysis (LDA). LDA is used to find linear combinations of features that most effectively separate two or more classes (Saeidi et al., 2021). This technique is usually based on the assumption that the data density follows a normal distribution, with equal covariance for all classes. LDA creates a separating hyperplane by maximizing the distance between the two classes and minimizing the distance points within each class (Saeidi et al., 2021). LDA can be used as a classifier and as a dimensionality reduction method before other classification methods. LDA stands out for its simplicity, interpretability (Carlson et al., 2003), and low computational requirements (Saeidi et al., 2021). More powerful classification method exist, such as Support Vector Machines. However, such methods are often more complicated and less interpretable. Unlike LDA, which provides a single axis of voxel weight offering a direct measure of the voxel's contribution to the classification, other methods may involve generating regions of activity space associated with a stimulus category. These regions are difficult to visualize and interpret, requiring a distinct activity map for each point within the region (Carlson et al., 2003). The main limitation of LDA is its linear nature, which may hinder effectiveness when dealing with nonlinear EEG data (Saeidi et al., 2021).

Previous findings using discriminant analyses

Nevertheless, the LDA approach has previously been used to classify EEG signals measured during an SST. Chikara and Ko (2019) compared the accuracy of several classification algorithms in classifying post-stimulus EEG data into successful and failed stop trials. They used the Phase Locking Value (PLV) method to quantify neural activities between the EEG signals of two electrodes. The PLV values were calculated using P3 waves during successful and failed stops and used as features for classification. Intra-participant classification accuracies of the algorithms were evaluated by the leave-one-out cross validation (LOOCV) method. Notably, participants had to respond with their right hand to half of the Go stimuli and with their left hand to the other half. The researchers found that performance of quadratic discriminant analysis (QDA) was best, with an average accuracy of 88.88% in the left occipital temporal cortex (electrode pair T7-O1) for right-hand response inhibition and in the right occipital temporal cortex (T8-O2) for left-hand response inhibition. Other brain regions also yielded high average accuracies. LDA performed above chance for EEG signals derived from the frontal cortex and the motor cortex (Chikara & Ko, 2019).

The same researchers (Chikara & Ko, 2020) compared classifiers using the power spectral density of post-stimulus (1-500 ms) EEG signals (1-50 Hz) as features in classifying successful and failed stop trials on a modified SST. They assessed intra-participant classification accuracies for eight channels separately (F3, F4, C3, C4, P3, P4, O1, and O2). The average classification accuracy was above chance for both the QDA and LDA classifiers at all investigated EEG channels, with average accuracies of around 69% for the QDA and 66% for the LDA classifier.

Research questions and hypotheses

The current research aims to improve an existing LDA classifier to where it can classify pre-stimulus EEG signals into successful and failed stop trials from an SST with significant accuracy. More specifically, this research will extend the LDA classifier created by Galama (2021) and modified by Caspani (2022), which has not yet been able to reach a classification accuracy above chance level (50%).

Caspani (2022) implemented two workflows. In workflow 1, all data was used for training and for testing, yielding an average accuracy of 73%. In workflow 2, LOOCV was implemented to assess the performance of the LDA. Workflow 2 resulted in an average accuracy of 50%. Despite each time leaving only one row out in workflow 2, the resulting accuracy was significantly lower than that of workflow 1. Here, a row corresponded to a specific frequency within one trial.

The first question to address is why workflow 1 yields much better accuracies than workflow 2. The hypothesis is that the weight vectors and cutoff values of Workflow 2 can differ significantly from those of Workflow 1, despite only leaving out one row at a time. This expectation arises due to the large variability in the datasets and their relatively small size, where even a single row can significantly impact the computation of the weight vector and cutoff value, leading to misclassifications and explaining the substantial difference in accuracy between the two workflows.

Based on this answer and on the assumptions of LDA, several methods will be employed in order to increase classification accuracy of the LDA classifier in combination with the LOOCV procedure. It is expected that reducing the variability of the data will improve performance. In addition, it is expected that leaving out an entire trial in the LOOCV, instead of a row that corresponds to a specific frequency within a trial, will improve performance. Furthermore, feature selection, such as selecting only the frontal electrodes, might improve performance. Finally, changing the LOOCV method to k-fold cross validation might increase performance by reducing variance.

The features that a successful classifier relies on will be investigated to give insights into the neural markers of proactive response inhibition, such as theta power. It is expected that, if a significant classification accuracy is reached, the classifier will uncover predictors of response inhibition that remain undetected when using univariate methods.

Methods

Participants

Data was acquired from Kenemans and colleagues (2023). For this dataset, 32 healthy participants were recruited from the student population at Utrecht University. All participants declared

to have normal hearing and normal or corrected-to-normal vision and signed an informed consent form which was approved by the Ethics Committee of the University Medical Centre Utrecht. Participants were paid 21 euros by means of compensation.

Materials

Stop Signal Task (SST)

The data used for the classifier were acquired from the auditory condition of the SST in Kenemans et al. (2023). Here, the primary task was a two-choice reaction time task. Participants had to discriminate between two go-stimuli by pressing either the left or right button with their left or right index finger, respectively. The go-stimuli were the letters “X” and “O” and were presented for 150 ms. On 25% of the trials, a stop signal was presented after the go stimulus. The stop signal was a 1000 Hz and 72 dB tone, presented binaurally trough in-ear headphones for 150 ms. The condition lasted for 128 trials and the trial-to-trial interval varied between 1.5 and 1.8 s. The SSD was dynamically manipulated to yield a success rate of approximately 50% for all participants.

EEG

During the task, electroencephalographic activity was recorded using the Active-Two system (Biosemi, Amsterdam, The Netherlands) with 64 Ag-AgCl electrodes. The electrodes were placed following the 10/10 system and referenced to the CMS and DRL. Signals were sampled at 2048 Hz and a low pass filter of 400 Hz was applied. EOG electrodes were placed above and below the left eye and at the outer canthi of both eyes, to record eye movements.

Procedure

Participants were asked to provide standard demographic information. After this, the EEG cap was placed and they were seated in a dark-attenuated room, approximately 90 cm from the computer screen. Participants received the instruction to respond as quickly and accurately as possible to the go-stimuli but withhold a response after a stop signal. The experiment (Kenemans et al., 2023) counterbalanced 3 conditions across participants. After each experimental block, participants received

instructions to either slow down or speed up, depending on their reaction time and success rate. Refer to Kenemans et al. (2023) for a complete description of the procedure and the SST.

Preprocessing

Preprocessing was similar to Galama (2021) and Caspani (2022). The preprocessing of the EEG data was executed using BrainVision Analyzer version 2.3. Signals were re-referenced to the right mastoid. The EEG data was down sampled to 64 HZ and a low cut-off filter of 0.5 Hz and high cutoff filter of 28.8 Hz were applied.

The first segmentation had a position of -1000 to -1500 ms relative to go-stimuli reference markers. Overlapping segments were allowed. The effects of eye blinks on the EEG data were removed with the Gratton and Coles method.

Although Galama (2021) and Caspani (2022) used no other artifact rejection procedures, the current project employed automatic artifact rejection to reduce noise in each participant's dataset. Specifically, trials were marked as bad if they contained a voltage step exceeding 50 µV/ms, if they had an amplitude above 200 µV or below -200 µV or if, within a 100 ms interval, the absolute difference between two values exceeded 100 µV or activity was below 0.5 µV. If, for a participant, more than 1/3 of failed trials or 1/3 of successful trials were marked as bad, problematic electrodes were removed until at least 2/3 of these trials were artifact free. Hereafter, all trials marked as bad were removed.

Then, the second segmentation selected all 1000 ms intervals preceding go-stimuli in stop trials. These segments were separated in failed and successful stop trials. A Fast Fourier Transformation was executed on these segments, using the half-spectrum with a resolution of 1 Hz and a periodic Hanning data window of 10 % length (with variance correction). This resulted in non-complex power values for all segments preceding the go stimuli, divided into failed and successful stop trials. An extended description of all preprocessing steps can be found in Appendix A.

Two participants were excluded from further analysis. One of these had only 6 successful trials. The data of the other participant contained artifacts in all trials and across all electrodes. An

overview of which electrodes were removed and how many trials remained for each participant can be found in Appendix B.

Initial LDA classifier

The EEG data was classified using an LDA classifier initially designed by Galama (2021), who used the frequencies 1-20 Hz, and altered by Caspani (2022), who focused on frequencies from the theta band (4-8 Hz).

In the initial approach, data was projected onto a line orthogonal to a projection vector w. To derive w, the mean vectors M_f and M_s for failed and successful stop trails were computed by summing all values of the trials X_i per class and dividing by the number of trials n ($M_f = \left(\frac{1}{n_f}\right)\sum_{i,X_i \in C_f} X_i$, $M_s = \left(\frac{1}{n_s}\right)\sum_{i,X_i \in C_s} X_i$). Covariance matrices S_f and S_s for each class were calculated, and the within-class covariance matrix S_w was found by adding these covariance matrices together ($S_w = S_f + S_s$). The weight vector w was calculated by $w = S_w^{-1}(M_f - M_s)$. The discriminant criterion c was computed by projecting w onto the average of the mean vectors ($c = w \cdot (\frac{M_f + M_s}{2})$). A data point d was classified by projecting w on d by taking the dot product of the two vectors and comparing it to the discriminant criterion c:

$$w \cdot d < c \Rightarrow classify\ as\ success$$

$$w \cdot d > c \ \Rightarrow \ classify\ as\ fail$$

Caspani (2022) introduced several modifications. Firstly, the frequency range was changed to 4-8 Hz (theta band), and outliers were removed from the data. Caspani (2022) then implemented six different LDA workflows:

1. The LDA classifier was trained on all data within each participant.
2. Leave-one-out cross-validation (LOOCV) was applied.
3. Cells containing outliers were filled with the mean of the non-outlier values of the same column, followed by LOOCV.
4. The dataset was pivoted to bring frequency range information into the column names, followed by LOOCV.

5. A combination of workflow 3 and 4 was used.
6. Only data from the FCz electrode was considered.

Additionally, all these workflows were applied to data from only the frontal electrodes.

Further modifications by Caspani (2022) of workflow 2 included applying a Principal Component Analysis (PCA) to the preprocessed data, replacing the LOOCV procedure by holdout cross-validation, setting all off-diagonal elements of S_w to zero to only consider variances, averaging the preprocessed data, and modifying the weight vector computation to $w = M_f - M_s$. However, none of these modifications resulted in classification accuracies above chance.

Workflow 1 vs. workflow 2

To investigate why there is a substantial difference in accuracy between workflow 1 and 2 by Caspani (2022), a single participant was selected for further analysis from the dataset. Participant 1 was selected from the dataset without artifact rejection. The preprocessing steps were consistent with those in Caspani (2022), including the frequency range of 4-8 Hz, the inclusion of all electrodes, and the removal of rows containing values greater than 5. It is important to note that Caspani (2022) separated each trial into four rows, corresponding to each theta frequency. This means that accuracy was assessed over each row, which corresponds to a specific trial and frequency combination. Therefore, this section uses the term row instead of trial, which would not be the appropriate term. Classes were balanced by randomly removing rows from the majority class. To ensure that the same rows were removed for workflow 1 and 2, a fixed random seed was set. Removing the same rows for both workflows promotes reliable comparison.

To asses the variability in the dataset of participant 1, a boxplot for each electrode was created.

The accuracy for each workflow was computed. Additionally, one-sample t-tests were performed to test the significance of the averaged accuracy against chance level (0.5). For these tests, the accuracy of each single row was used, meaning that the rows were the source of variance.

For the comparison between the 2 workflows, it should be noted that workflow 1 involved training the LDA classifier on the entire dataset of participant 1, resulting in a single weight vector

and cutoff value. Workflow 2 implemented LOOCV, generating 216 different weight vectors and cutoff values, each calculated after leaving one of the 216 rows out.

For workflow 1, the single weight vector and cutoff value (discriminant criterion c) were recorded. For workflow 2, descriptive statistics of the 216 different weight vectors and cutoff values were computed. Specifically, the analysis included the largest weights in workflow 1, representing electrodes P1, P2, and P03.

The cutoff values from workflow 2 were compared to the cutoff value of workflow 1 using one-sample t-tests, both when considering all rows and when only considering the misclassified rows. These misclassified rows were rows that were correctly classified in workflow 1 but misclassified in workflow 2 and might thus show a significant difference in cutoff value or weight vector. Cosine similarity was used to compare the 216 weight vectors from workflow 2 to the weight vector of workflow 1, with the mean cosine similarity computed both for all rows and for the misclassified rows.

Sixty out of 216 rows were correctly classified in workflow 1 but misclassified in workflow 2. A detailed analysis of nine of these rows was conducted by displaying their corresponding cutoff values, and weights for P1, P2, and PO3. These rows included three with extremely low cutoff values, three with cutoff values close to the one from workflow 1, and three with extremely high cutoff values.

For the three rows with cutoff values very close to that of workflow 1, the dot products of these rows with their weight vectors were computed. Additionally, the dot products with slightly altered weights and the weight vector from workflow 1 were calculated to understand how specific weights influence classification.

LDA classifier development

To improve the accuracy of the existing LDA classifier, several methods were employed.

Based on the observation that some weights were irrationally large and that these weights corresponded to the electrodes with the lowest power values and least amount of variance, the computation of the weights in the LDA was altered. The quadratic nature of EEG power values

indicates that the weight vector needs to be normalized by the square root of the covariance matrices. This normalization approach adjusts for the influence of low-variance electrodes, which can disproportionately affect the weights.

Based on the difference between workflow 1 and 2 by Caspani (2022), modifications to the preprocessing of the data were considered, including outlier detection and the before mentioned artifact rejection. In addition, regularization techniques for the LDA were considered. LDA requires that a dataset has less features than samples. Therefore, an oversampling technique as well as outlier replacement were considered to keep as many samples as possible, as well as feature selection to reduce the number of features. Furthermore, LDA assumes a normal distribution, therefore a normalization technique was considered. Finally, because LDA assumes that the data is linearly separable whilst the collected data might not be linearly separable, the use of a Quadratic Discriminant Analysis (QDA) was considered.

A link to the code implementing all that is described in this methods section can be found in Appendix C.

Computation of LDA weights

Similar to the initial LDA, the mean vectors M_f and M_s for failed and successful stop trails were computed by summing all values of the trials X_i per class and dividing by the number of trials n ($M_f = \left(\frac{1}{n_f}\right)\sum_{i,X_i \in C_f} X_i$, $M_s = \left(\frac{1}{n_s}\right)\sum_{i,X_i \in C_s} X_i$). Covariance matrices S_f and S_s for each class were calculated. However, the square root of these matrices were used and the within-class covariance matrix S_w was found by taking the average of these covariance matrices ($S_w = \frac{\sqrt{S_f}+\sqrt{S_s}}{2}$), instead of adding them together. The weight vector w was calculated by $w = S_w^{-1}(M_f - M_s)$. The discriminant criterion c was computed by projecting w onto the average of the mean vectors ($c = w \cdot (\frac{M_f+M_s}{2})$).

Outlier treatment

Although the dataset was subjected to artifact rejection, outlier treatment was used on top of it because this might further improve classification accuracy. Even after artifact rejection, there can still be outliers in the data. Handling these outliers can further reduce variability in the data and make the data distribution more normal, which aligns better with the assumptions of LDA.

Caspani (2022) addressed outliers by either removing rows with values above 5 or replacing such cells with the column mean. This threshold is arbitrary and the removal of rows reduced the datasets considerably in size.

To statistically detect outliers, the current project uses the Interquartile Range (IQR) method. For each electrode-frequency column, the first quartile (Q1), median, and third quartile (Q3) were identified, and the IQR was calculated (Q3 - Q1). Upper (Q3 + 1.5 * IQR) and lower (Q1 - 1.5 * IQR) bounds were set, with outliers above or below these bounds. Instead of removal, outliers were winsorized by replacing values above the upper bound with the upper bound and values below the lower bound with the lower bound.

Normalization

Given that LDA assumes normally distributed data and the current data does not follow a normal distribution, normalization seemed necessary. Since the datasets contain power values (amplitudes squared), a log transformation was applied to each column in the dataset.

Oversampling

Most participants' datasets have an unequal number of successful and failed trials. To prevent classifier bias towards the majority class, Galema (2021) balanced the data by randomly removing trials from the majority class. To retain more data, the current project used the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) to oversample the minority class, generating synthetic samples close to the original samples in feature space. When this oversampling step was used, it was placed after making the train and test split in the data and applied to only the train set, to ensure that no data leakage was possible.

Regularization techniques

The LDA classifier by Caspani (2022) showed overfitting, evidenced by a significant drop in accuracy from 73% (using all data) to 50% (using LOOCV). To address this, a variant of LDA was implemented where only the variances of the data were used, setting all off-diagonal values in the covariance matrices to zero.

Another regularization method was also considered to prevent overfitting. L2 regularization, also known as ridge regression, adds a penalty term to the computation of the weight vector. This penalty term is the sum of the squares of the covariance matrix combined with a regularization parameter λ, which controls the strength of the penalty.

Feature selection

By means of feature selection, the approach that was implemented across all methods was averaging the theta frequency band (4-8 Hz), where the average of frequencies 4, 5, 6, and 7 Hz was used. In addition, using only the 26 frontal electrodes (Fp1, AF7, AF3, F1, F3, F5, F7, FT7, FC5, FC3, FC1, Fpz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FCz, FC2, FC4, FC6, FT8) for classification was considered.

Cross validation

In addition to the LOOCV method, k-fold cross validation was considered to evaluate performance. The dataset was randomly divided into five folds. Each fold served as a test set once and the remaining four folds were used for training.

QDA classifier

QDA was implemented to capture non-linear relationships and potentially improve classification performance.

For each trial in the participant's dataset, the LOOCV approach was used. One trial was left out as the test sample, while the remaining trials formed the training set.

For the training set, the mean vector μ_c and covariance matrix Σ_c were computed for each class c (failed and successful stops) across all features. The square root of the covariance matrices were used.

Next, for each class c, the quadratic discriminant function Q_c was computed using the formula: $Q_c(x) = -\frac{1}{2}\ln\left|\sqrt{\Sigma_c}\right| - \frac{1}{2}(x-\mu_c)^T\sqrt{\Sigma_c}^{-1}(x-\mu_c)$, where:

- x represents the feature vector of the left-out trial,
- μ_c is the mean vector of class c,
- Σ_c is the covariance matrix of class c,
- $|\Sigma_c|$ denotes the determinant of Σ_c,
- Σ_c^{-1} represents the inverse of Σ_c.

For the left-out trial x, the quadratic discriminant function Q_c was calculated for each class c. The class $\hat{y}$ that maximized the quadratic discriminant function Q_c was assigned to the left-out trial. This can be represented as: $\hat{y} = \arg\max_c Q_c(x)$

This process was repeated for each trial in the dataset. The number of correct classifications was then counted to evaluate the overall accuracy of the QDA classifier.

Repeated measures ANOVA

To evaluate the difference in frontal theta power between failed and successful trials, a repeated measures ANOVA was performed across all participants. This was a replication of the approach used by Caspani (2022) but with data incorporating artifact rejection and a different outlier treatment. The analysis focused on the FCz electrode, representing the frontal area, and used frequencies of 4, 5, 6, and 7 Hz. Outliers were detected using the Interquartile Range (IQR) method and replaced using winsorization. For each participant, trials were averaged across class and frequency, resulting in averaged FCz values. These values were then put into the repeated measures ANOVA.

Performance assessment

Accuracy evaluation

To compare the effectiveness of different preprocessing and classification techniques, each algorithm was run ten times. This was necessary due to the balancing of classes procedure, which is random. Balancing was performed either by randomly removing trials from the majority class or by randomly duplicating trials from the minority class using SMOTE. This randomness could lead to variability in the results; thus, running the algorithms multiple times ensured a more reliable comparison of their performance.

For each run, each trial was assigned an accuracy score of 0 (incorrect) or 1 (correct). These scores were averaged to determine the accuracy for a single run. The accuracies from the ten runs were then averaged to obtain a final accuracy score per participant. To assess whether these average accuracies per participant were significantly different from chance (50%), a one-sample t-test was performed on the accuracy scores of all trials from the ten runs.

For instance, if a participant had 150 trials, a total of 1500 accuracy scores were used in the t-test, compared against a chance level of 50%. Because some trials were included in every run, there were duplicates in the t-test, which could artificially inflate the degrees of freedom and affect the significance level. To account for this, an alpha level of 0.0001 was set to ensure the test remained stringent despite the presence of duplicate data points.

In addition, across participant accuracies were evaluated. Per classifier, the average accuracies for each participant was compared to a chance level accuracy value for each participant using a paired t-test. Thus, 30 accuracy values were compared to 30 values of 0.5 to test whether the average accuracy of a classifier was significantly different from chance level performance.

Analysis of weight vectors

For the methods that resulted in the most participants with significantly above chance accuracy, weight vectors were visualized. Per method, for participants that had an accuracy significantly higher than chance, the average weight vector over the ten runs was examined. These

weight vectors were visualized by plotting the weight components onto scalp electrodes. For participants with removed electrodes, the associated weights were set to zero for visualization.

In addition, averaged weight vectors across participants with significant classification accuracy were visualized. For computing this average weight vector, the weights with value zero due to missing electrodes were ignored.

For LDA, the weight vector was directly computed from the training data and represents the importance of each feature (electrode) in distinguishing between failed and successful stop trials. The average weight vector for each participant was obtained by averaging the weight vectors across the ten runs. These averaged weight vectors were visualized to identify the important electrodes.

Unlike LDA, QDA models quadratic decision boundaries and does not directly use linear weights. However, it is possible to derive weights that relate to the linear terms of the QDA decision function by focusing on the linear part of the discriminant function: $x^T\Sigma_c^{-1}\mu_c$. Here, $\Sigma_c^{-1}\mu_c$ can be seen as an equivalent to the linear weights in LDA, where Σ_c represents the covariance matrix and μ_c denotes the mean vector for class c. This way, per class weight vectors were computed for each participant. These two weight vectors were then combined by averaging them to produce a single weight vector per participant. The resulting weight vectors were averaged across the ten runs. These average weight vectors for QDA were visualized similarly to LDA to identify the important electrodes for successful classifications.

Results

Comparison of workflow 1 and workflow 2 for participant 1

Variability in the data of participant 1 is displayed in Figure 1, which shows a box plot for each of the 64 scalp electrodes.

Table 1 displays the accuracies and their statistical significance for workflow 1 and workflow 2 for participant 1.

The cutoff value for participant 1 workflow 1 was 0.1187. The descriptive statistics of the cutoff values of workflow 2 for participant 1 are summarized in Table 2.

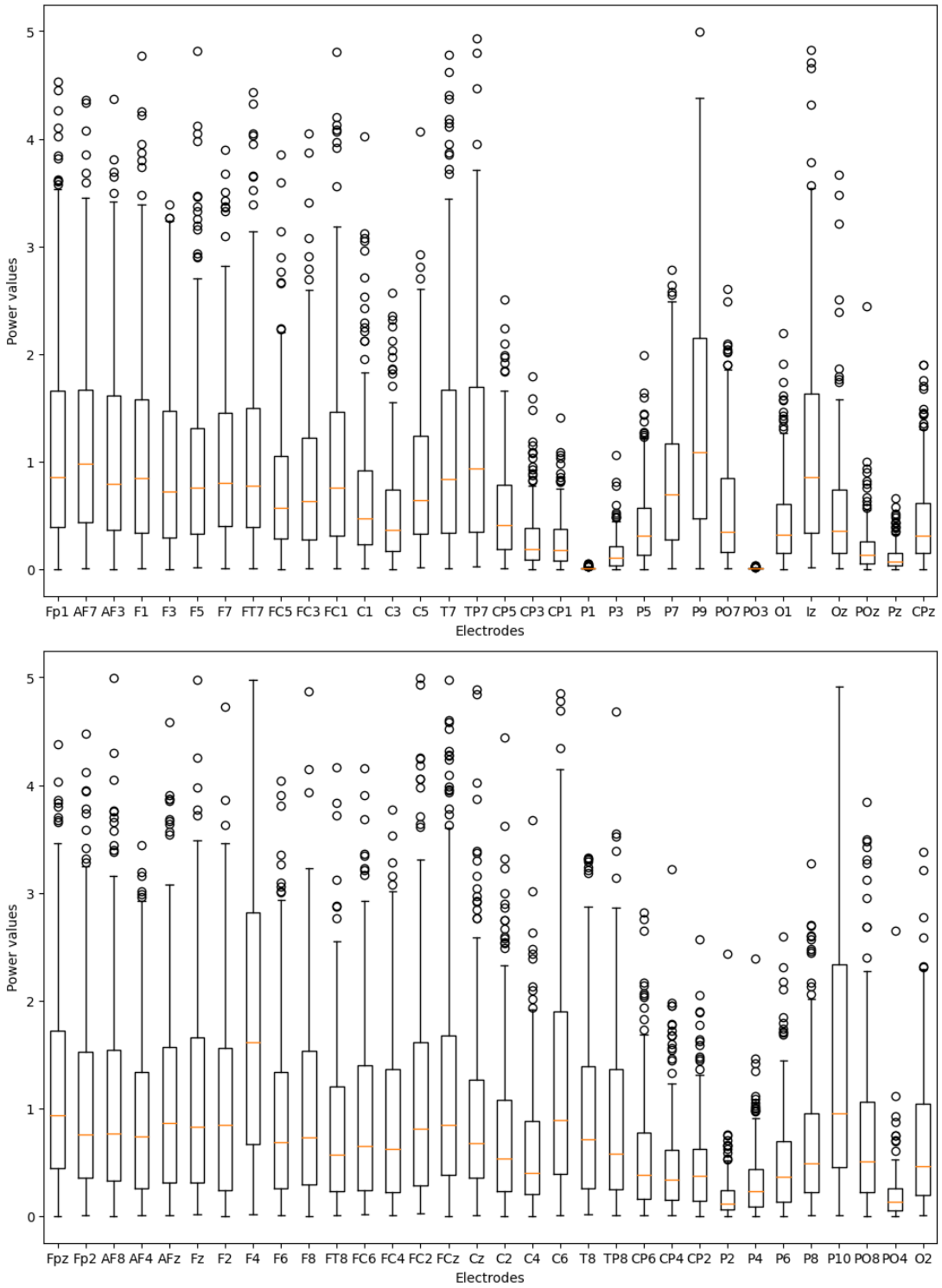
Table 1

Accuracies for participant 1 in workflow 1 and 2

Workflow	Accuracy	P-value
1	0.75	< 0.001
2	0.48	0.50

Figure 1

Box Plots of EEG Power Values across the 64 Scalp Electrodes



Note. Boxes represent Q1 to Q3 of the data, with a line at the median. The whiskers extend from the box to the farthest datapoint lying within 1.5 * IQR. The circles represent flier data points.

Table 2

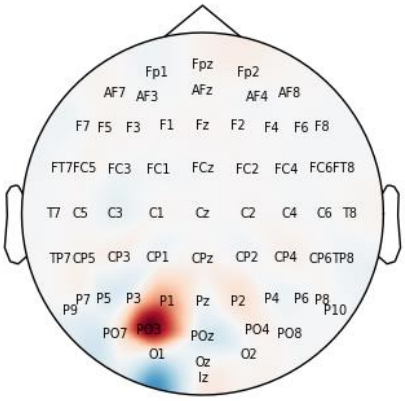
Descriptive Statistics of the Cutoff Values of Workflow 2 for participant 1

Min	25%	Mean	75%	Max	SD
-0.230	0.091	0.119	0.149	0.316	0.060

The weight vector of workflow 1 is plotted in Figure 2, with specific weights for electrodes P1, P2, and P03 being 10.05, 3.02, and 22.22, respectively. Descriptive statistics for these weights in workflow 2 are provided in Table 3.

Figure 2

Representation of the Weight Vector for Participant 1 Workflow 1



Note. Red indicates positive weight values, blue indicates negative weight values.

Table 3

Descriptive Statistics of Weights P1, P2, and PO3 of Workflow 2 for Participant 1

Weight	Min	25%	Mean	75%	Max	SD
P1	-5.97	9.05	10.08	11.27	25.02	2.83
P2	0.63	2.86	3.03	3.21	5.65	0.44
PO3	12.89	20.67	22.23	23.41	37.98	2.88

A one-sample t-test comparing the cutoff value of workflow 1 with the cutoff values from workflow 2 indicated no significant difference ($p = 0.89$). This lack of significant difference was present even when considering only the misclassified trials in workflow 2 ($p = 0.90$).

The mean cosine similarity between the weight vectors of workflow 2 and the weight vector of workflow 1 is 0.99, suggesting a very high degree of similarity. This high similarity remained even when analyzing only the weight vectors of misclassified trials (cosine similarity = 0.99).

Table 4 presents the cutoff values and weights for electrodes P1, P2, and PO3 derived from workflow 2 for nine specific rows. These rows included three with very low cutoff values, three with cutoff value close that of workflow 1, and three with very high cutoff values. All these rows were correctly classified in workflow 1 but misclassified in workflow 2.

Table 4

Cutoff Value and P1, P2, and P03 Weights for 9 Misclassified Trials in Workflow 2

Trial	Cutoff value	P1 weight	P2 weight	PO3 weight
35	0.002	10.67	3.56	19.33
68	0.014	3.51	3.25	27.63
109	0.024	7.32	3.69	23.60
159	0.119	12.50	3.10	18.75
164	0.123	8.78	2.29	21.16
116	0.123	7.17	2.12	23.58
133	0.204	11.42	2.25	24.05
152	0.266	11.91	3.47	21.12
137	0.306	13.13	5.60	22.34

Note. These rows were selected from the 60 trials that were correctly classified in workflow 1 but misclassified in workflow 2.

Trials 159, 164, and 116, which have cutoff values close to that of workflow 1, were misclassified as failed trials in workflow 2. The dot product of these data rows and their respective weight vectors resulted in values higher than the cutoff values (0.547, 0.560, 0.587, respectively). Changing the weights for P1, P2, and PO3 to those from workflow 1 did not alter the classification (0.534, 0.627, 0.582). However, using the entire weight vector from workflow 1, so changing the weights for all electrodes, resulted in correct classifications, with dot products of -0.168, -0.319, and -0.377, respectively. These values were lower than the cutoff value, thus correctly classifying the rows as successful trials.

Repeated measures ANOVA

A repeated measures ANOVA was conducted to assess the effects of class and frequency on frontal theta power taken from the FCz electrode. Results of the ANOVA are summarized in Table 5. There was a significant main effect of class, $F(1, 29) = 17.47$, $p < 0.001$, $\eta^2 = 0.004$. Participants exhibited significantly different FCz values between class conditions. However, no significant main effect of frequency condition was observed, $F(3, 87) = 1.32$, $p = 0.27$, $\eta^2 = 0.005$, indicating that there were no significant differences in FCz values across frequency conditions. In addition, there was no significant interaction between class and frequency condition, $F(3, 87) = 0.48$, $p = 0.695$, $\eta^2 = 0.0002$. Sphericity assumption was violated for the frequency condition effect ($\varepsilon = 0.481$), but not for the class effect ($\varepsilon = 1.000$).

Table 5

Repeated Measures ANOVA Results Assessing the Effects of Class and Frequency on Theta Power from the FCz Electrode

	Sum of Squares	Ddof1	Ddof2	Mean Square	F-statistic	p-value	ng2	eps
class	2.032	1	29	2.032	17.474	0.00025	0.0035	1.000
frequency	2.856	3	87	0.952	1.324	0.27189	0.0049	0.481
Class * frequency	0.110	3	87	0.037	0.483	0.69496	0.0002	0.764

Performance assessment

Accuracy evaluation

The average accuracies per participant for various preprocessing techniques applied to a variances-only LDA classification algorithm using averaged theta band data and all 64 electrodes as features are shown in Table 6. Each algorithm was run ten times to account for variability due to the random balancing of classes.

Table 7 shows the average accuracies per participant for different regularization techniques applied to the LDA as well as the SMOTE method, with outlier treatment, again using averaged theta band data and all 64 electrodes as features.

Table 8 displays the average accuracies per participant QDA combined with different regularization techniques as well as the SMOTE method, with outlier treatment and using averaged theta and all electrodes.

Furthermore, methods that reached relatively high accuracies, LDA with L2 regularization and SMOTE and QDA with L2 regularization and SMOTE, were applied to only the frontal electrodes. In addition, these methods were also tested using k-fold cross validation instead of LOOCV, on all electrodes. These results are shown in Table 9.

Finally, across participant accuracies were evaluated using paired t-tests. The results of these tests are displayed in Table 10.

Table 6

Average Per Participant Accuracies for LDA Variances Only, Average Theta, 64 Electrodes

Participant	No outlier treatment	Outlier treatment	Outlier treatment and log-transformation	Outlier treatment and SMOTE
Pp01	0,59	0,60	0,58	0,59
Pp02	0,46	0,51	0,50	0,52
Pp03	0,50	0,50	0,54	0,51
Pp04	0,56	0,56	0,55	0,56
Pp05	0,53	0,54	0,56	0,54
Pp06	0,50	0,51	0,52	0,51
Pp07	0,61	0,61	0,62	0,62
Pp09	0,57	0,56	0,60	0,58
Pp10	0,54	0,55	0,58	0,56
Pp101	0,57	0,57	0,57	0,58
Pp102	0,49	0,45	0,51	0,45
Pp103	0,51	0,51	0,50	0,54
Pp106	0,59	0,54	0,54	0,52
Pp107	0,53	0,53	0,51	0,53
Pp108	0,55	0,54	0,55	0,54
Pp109	0,54	0,53	0,52	0,54
Pp11	0,62	0,62	0,60	0,69
Pp110	0,49	0,50	0,44	0,50
Pp111	0,49	0,54	0,52	0,53
Pp112	0,47	0,45	0,46	0,47
Pp12	0,51	0,50	0,49	0,50
Pp13	0,55	0,58	0,58	0,59
Pp14	0,41	0,47	0,43	0,47
Pp15	0,56	0,57	0,57	0,56
Pp16	0,56	0,55	0,59	0,55
Pp17	0,54	0,56	0,55	0,57
Pp19	0,49	0,46	0,43	0,45
Pp20	0,55	0,57	0,59	0,57
Pp21	0,48	0,53	0,53	0,53
Pp22	0,52	0,53	0,54	0,55
Average	**0.53**	**0.54**	**0.54**	**0.54**

Note. Colored cells represent accuracies that significantly differ from chance level (50%). Green cells contain values significantly higher and red cells contain values significantly lower than 50%.

Table 7

، Average Per Participant Accuracies for LDA, Average Theta, Outlier Treatment, 64 electrodes

Participant	No regularization	L2 regularization (λ = 10)	L2 regularization (λ = 10) and SMOTE
Pp01	0,55	0,59	0,59
Pp02	0,50	0,56	0,60
Pp03	0,50	0,52	0,54
Pp04	0,55	0,55	0,57
Pp05	0,51	0,54	0,54
Pp06	0,48	0,50	0,51
Pp07	0,61	0,61	0,61
Pp09	0,54	0,52	0,54
Pp10	0,52	0,54	0,55
Pp101	0,54	0,57	0,57
Pp102	0,43	0,46	0,45
Pp103	0,53	0,52	0,52
Pp106	0,55	0,57	0,55
Pp107	0,53	0,54	0,54
Pp108	0,55	0,55	0,55
Pp109	0,55	0,52	0,53
Pp11	0,52	0,58	0,69
Pp110	0,50	0,50	0,49
Pp111	0,52	0,53	0,53
Pp112	0,44	0,44	0,47
Pp12	0,50	0,49	0,49
Pp13	0,58	0,57	0,57
Pp14	0,43	0,51	0,47
Pp15	0,55	0,56	0,55
Pp16	0,60	0,56	0,55
Pp17	0,57	0,55	0,56
Pp19	0,48	0,47	0,46
Pp20	0,60	0,59	0,59
Pp21	0,51	0,54	0,54
Pp22	0,52	0,54	0,56
Average	**0.53**	**0.54**	**0.54**

Note. Colored cells represent accuracies that significantly differ from chance level (50%). Green cells contain values significantly higher and red cells contain values significantly lower than 50%.

Table 8

Average Per Participant Accuracies QDA, Average Theta, Outlier Treatment, 64 electrodes

Participant	No regularization	Variances only	L2 regularization ($\lambda = 10$)	L2 regularization ($\lambda = 10$) and SMOTE
Pp01	0,55	0,60	0,59	0,57
Pp02	0,44	0,49	0,58	0,60
Pp03	0,50	0,50	0,54	0,53
Pp04	0,44	0,58	0,51	0,58
Pp05	0,54	0,54	0,53	0,55
Pp06	0,45	0,52	0,48	0,52
Pp07	0,51	0,61	0,59	0,59
Pp09	0,50	0,55	0,55	0,54
Pp10	0,46	0,54	0,52	0,52
Pp101	0,53	0,57	0,56	0,58
Pp102	0,54	0,45	0,43	0,42
Pp103	0,52	0,52	0,49	0,52
Pp106	0,53	0,54	0,55	0,56
Pp107	0,54	0,55	0,51	0,51
Pp108	0,55	0,54	0,55	0,54
Pp109	0,53	0,52	0,52	0,63
Pp11	0,99	0,59	0,60	0,81
Pp110	0,45	0,49	0,48	0,59
Pp111	0,55	0,53	0,53	0,51
Pp112	0,46	0,46	0,46	0,56
Pp12	0,55	0,54	0,48	0,50
Pp13	0,48	0,57	0,57	0,59
Pp14	0,44	0,42	0,45	0,46
Pp15	0,44	0,58	0,56	0,57
Pp16	0,61	0,56	0,59	0,60
Pp17	0,60	0,55	0,54	0,55
Pp19	0,40	0,44	0,48	0,53
Pp20	0,60	0,57	0,58	0,59
Pp21	0,58	0,51	0,53	0,53
Pp22	0,56	0,55	0,56	0,57
Average	**0.53**	**0.53**	**0.53**	**0.56**

Note. Colored cells represent accuracies that significantly differ from chance level (50%). Green cells contain values significantly higher and red cells contain values significantly lower than 50%.

Table 9

Average Per Participant Accuracies for LDA and QDA with L2 Regularization ($\lambda = 10$) and SMOTE

Participant	LDA frontal electrodes	QDA frontal electrodes	LDA k-fold cv (k = 5)	QDA k-fold cv (k = 5)
Pp01	0,60	0,60	0,58	0,55
Pp02	0,59	0,59	0,57	0,61
Pp03	0,57	0,57	0,50	0,48
Pp04	0,56	0,56	0,57	0,52
Pp05	0,56	0,56	0,55	0,51
Pp06	0,50	0,52	0,48	0,50
Pp07	0,59	0,60	0,61	0,55
Pp09	0,54	0,56	0,50	0,48
Pp10	0,54	0,48	0,54	0,51
Pp101	0,59	0,59	0,57	0,59
Pp102	0,49	0,42	0,45	0,44
Pp103	0,53	0,50	0,50	0,48
Pp106	0,56	0,56	0,56	0,56
Pp107	0,53	0,52	0,52	0,55
Pp108	0,55	0,55	0,54	0,56
Pp109	0,49	0,62	0,51	0,62
Pp11	0,66	0,76	0,63	0,79
Pp110	0,50	0,57	0,45	0,58
Pp111	0,50	0,50	0,52	0,52
Pp112	0,45	0,49	0,47	0,54
Pp12	0,48	0,47	0,52	0,46
Pp13	0,57	0,58	0,58	0,57
Pp14	0,53	0,51	0,45	0,46
Pp15	0,56	0,57	0,53	0,55
Pp16	0,55	0,59	0,55	0,62
Pp17	0,55	0,55	0,55	0,55
Pp19	0,49	0,50	0,50	0,51
Pp20	0,57	0,58	0,58	0,58
Pp21	0,54	0,54	0,57	0,57
Pp22	0,56	0,57	0,58	0,59
Average	**0.54**	**0.55**	**0.53**	**0.55**

Note. Colored cells represent accuracies that significantly differ from chance level (50%). Green cells contain values significantly higher and red cells contain values significantly lower than 50%.

Table 10

Paired t-test Results for Across Participant Accuracy Evaluation

Classifier	t	df	p-value	95% CI lower	95% CI upper	mean difference
LDA var. only, no outlier treatment	3,4138	29	0,0019	0,0118	0,0469	0,0293
LDA var. only	4,3559	29	0,0002	0,0187	0,0519	0,0353
LDA var. only + log-transformation	3,8267	29	0,0006	0,0163	0,0537	0,0350
LDA_var. only + SMOTE	4,4641	29	0,0001	0,0219	0,0590	0,0405
LDA	3,0373	29	0,0050	0,0083	0,0428	0,0256
LDA + L2 regularization	5,1939	29	<0,0001	0,0221	0,0508	0,0364
LDA + L2 regularization + SMOTE	4,7360	29	0,0001	0,0245	0,0618	0,0431
LDA + L2 regularization + SMOTE + frontal electrodes	5,6638	29	<0,0001	0,0284	0,0606	0,0445
LDA + L2 regularization + SMOTE + k-fold cv	4,0277	29	0,0004	0,0172	0,0526	0,0349
QDA	1,4505	29	0,1577	-0,0113	0,0662	0,0275
QDA var. only	3,7637	29	0,0008	0,0149	0,0505	0,0327
QDA + L2 regularization	3,6195	29	0,0011	0,0130	0,0469	0,0299
QDA + L2 regularization + SMOTE	4,9581	29	<0,0001	0,0341	0,0819	0,0580
QDA + L2 regularization + SMOTE + frontal electrodes	4,6866	29	0,0001	0,0293	0,0747	0,0520
QDA + L2 regularization + SMOTE + k-fold cv	3,8147	29	0,0007	0,0212	0,0701	0,0456

Analysis of weight vectors

Weight vectors were visualized by plotting the weight components onto scalp electrodes.

Figure 3 displays the weight vectors for participants that reached an average accuracy significantly above chance level when using the LDA variances only with outlier treatment.

The weight vectors for participants that reached a significant average accuracy when using the LDA with L2 regularization are shown in Figure 4.

The linear weights derived from the QDA were also plotted onto scalp electrodes. Figure 5 shows the weight vectors for participants that reached an average accuracy significantly higher than 50% using the QDA variances only with outlier treatment.

In addition, weight vectors for participants with significant accuracy when using the QDA with L2 regularization and outlier treatment are shown in Figure 6.

Finally, the averaged weight vectors across participants with significant accuracy for these methods are displayed in Figure 7. Here, average weight vectors are displayed for the LDA with outlier treatment using variances only (A) or with L2 regularization (B) as well as for the QDA with outlier treatment and variances only (C) or with L2 regularization (D).

Weight vectors for algorithms that included the SMOTE method were visually very similar to the same methods without SMOTE, therefore those weight vectors are not shown.

Figure 3

Weight Vectors for Participants with Classification Accuracy Significantly Above 50% for the LDA Variances Only and Outlier Treatment

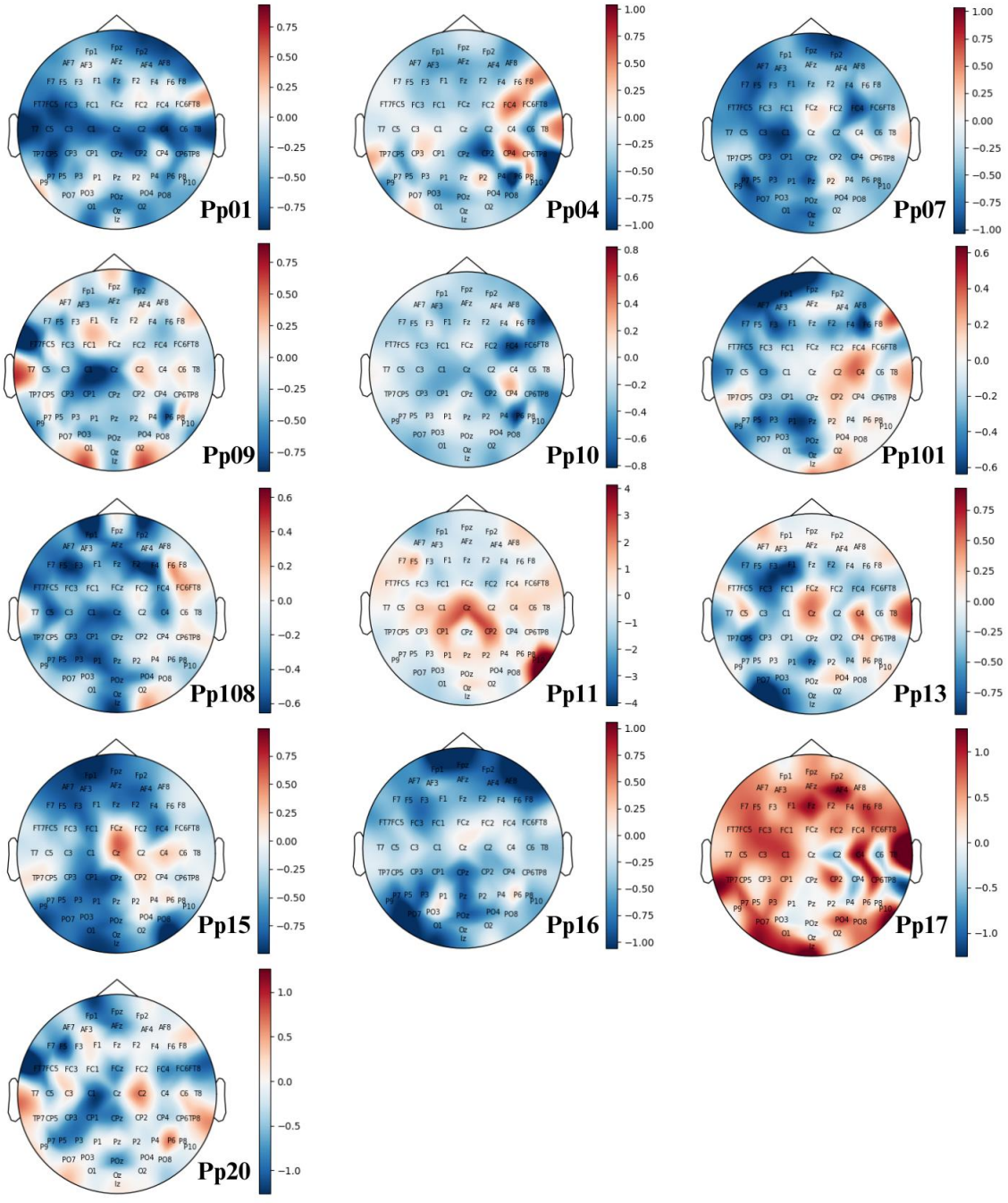


Figure 4

Weight Vectors for Participants with Classification Accuracy Significantly Above 50% for the LDA with L2 Regularization and Outlier Treatment

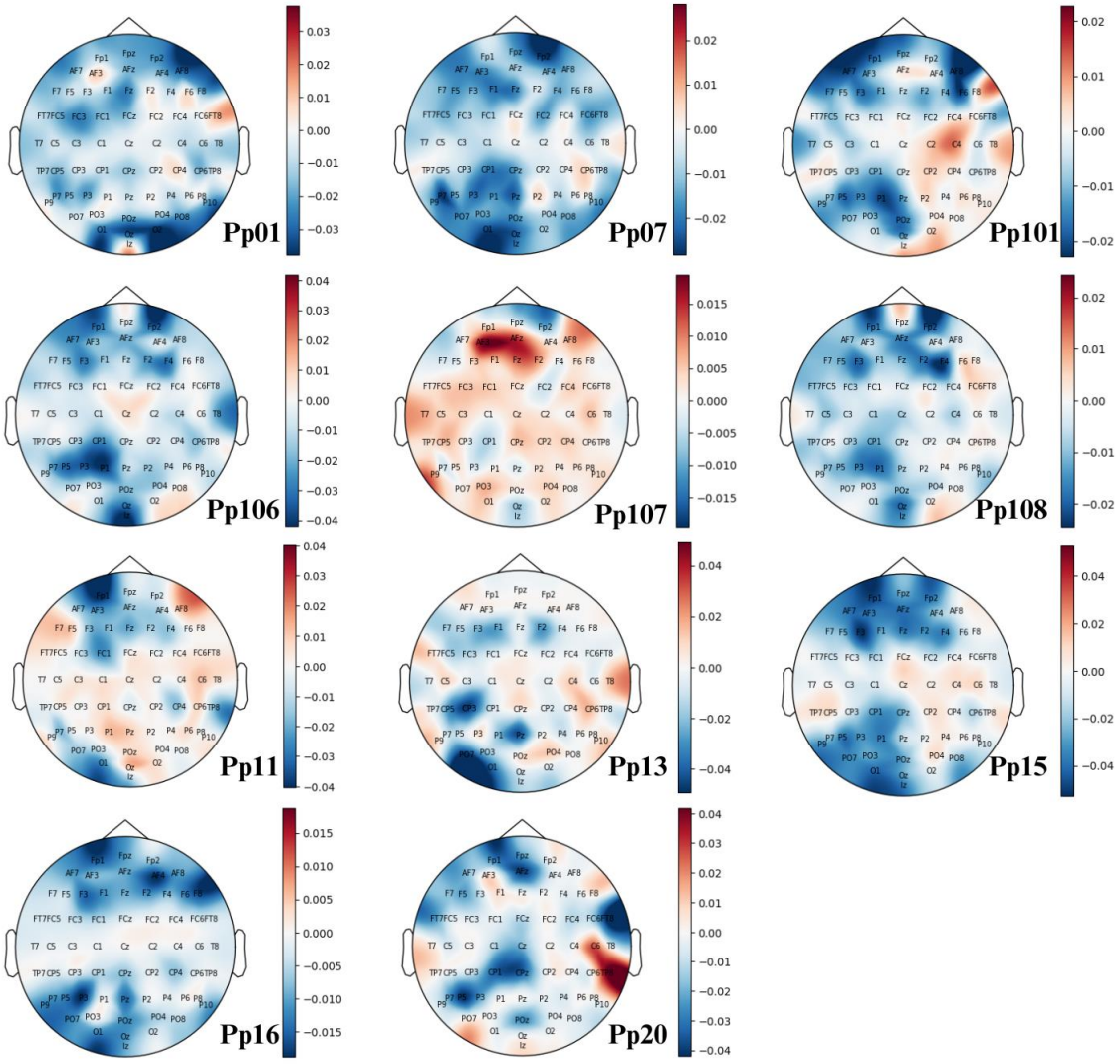


Figure 5

Weight Vectors for Participants with Classification Accuracy Significantly Above 50% for the QDA Variances Only and Outlier Treatment

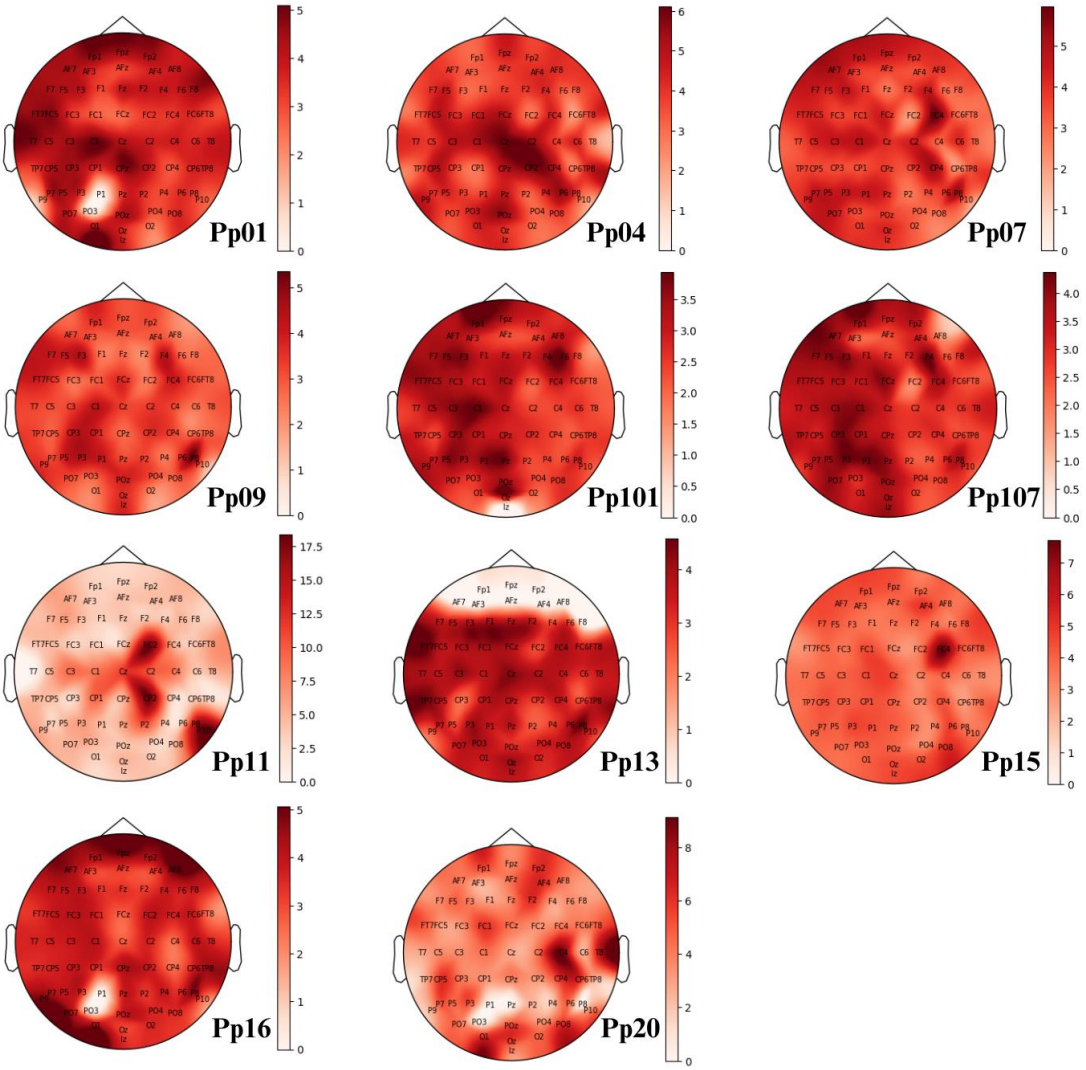


Figure 6

Weight Vectors for Participants with Classification Accuracy Significantly Above 50% for the QDA with L2 Regularization and Outlier Treatment

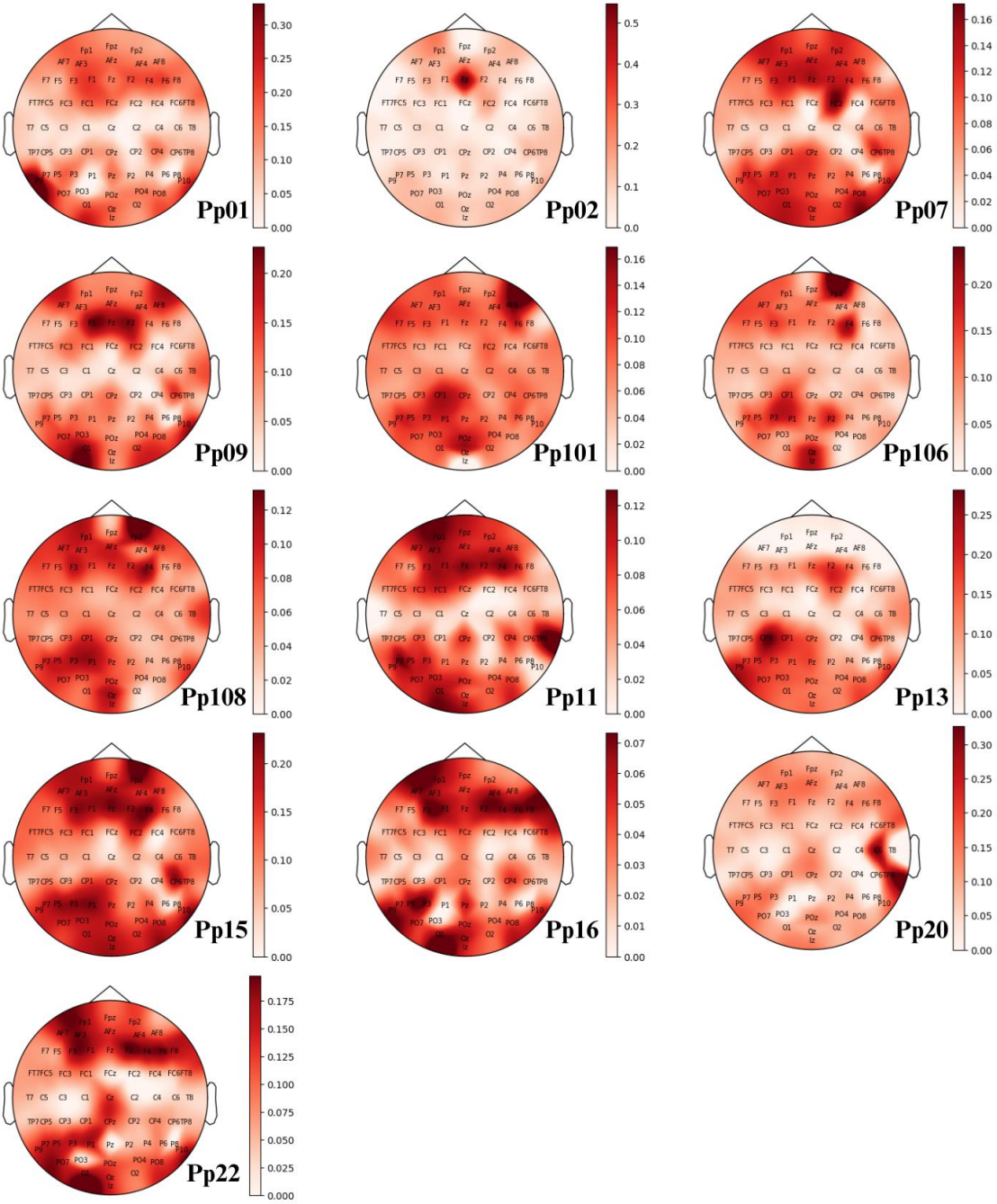
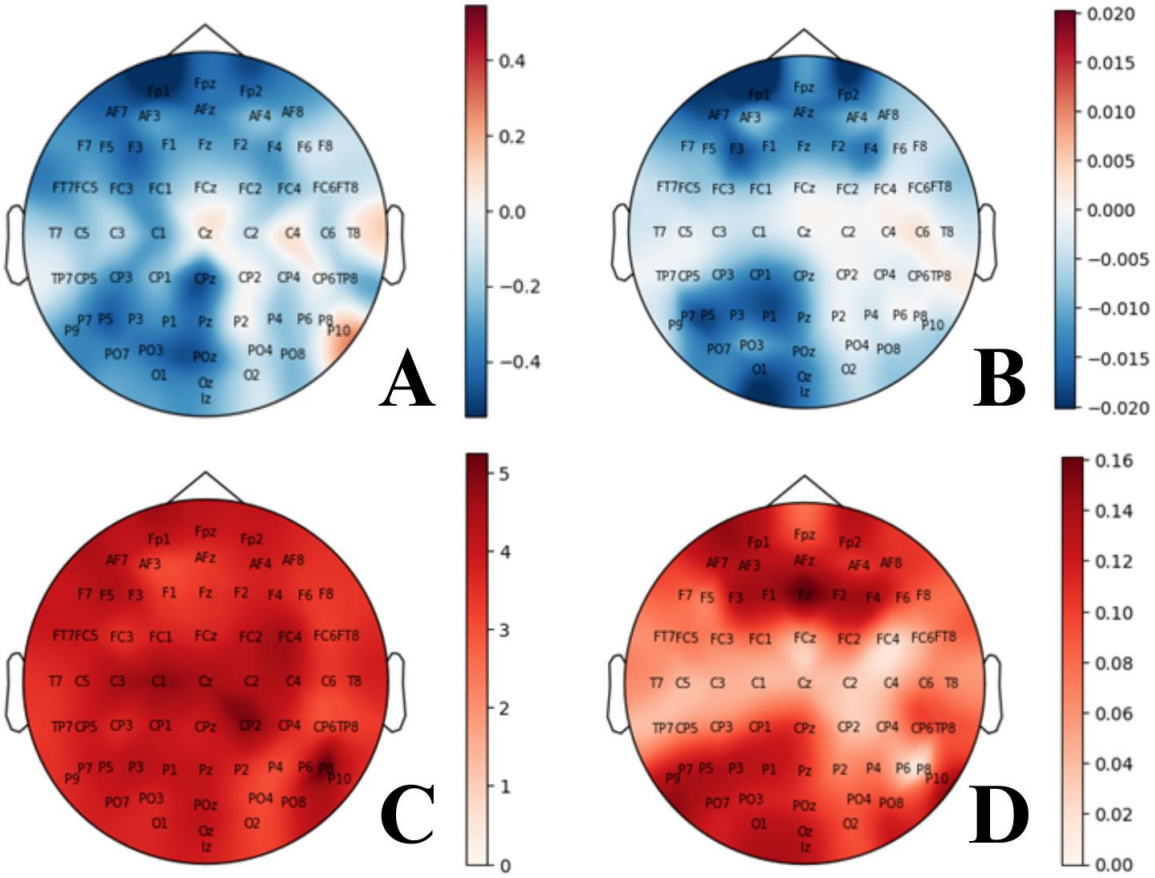


Figure 7

Average Weight Vectors Across Participants with Classification Accuracy Significantly Above 50%



Note. The figure shows the average weight vector for the LDA with outlier treatment using variances only (A) or L2 regularization (B) as well as for the QDA with outlier treatment using variances only (C) or L2 regularization (D).

Discussion

In this section, there will first be an explanation for the observed difference in performance between workflow 1 and workflow 2 by Caspani (2022). Second, the ANOVA results will be interpreted. Next, the accuracies of the different classification algorithms as well as their visualized weight vectors will be discussed. Finally, the implications of these results will be considered as well as limitations to the current project, and future directions.

Comparison of workflow 1 and workflow 2

To investigate accuracy difference between Caspani's 2022 workflow 1 and workflow 2, data from a single participant (participant 01) without artifact rejection was analyzed. For this participant, workflow 1 achieved an accuracy of 75%, which is significantly higher than chance level performance, while workflow 2 reached only 48%, which is not significantly different from chance.

The boxplot analysis in Figure 1 reveals significant variability across electrodes, with values ranging from close to zero to the maximum threshold of 5. This variability could lead to considerable differences in weight vectors and cutoff values when a single trial is left out. Despite this, no significant difference was observed between the average of weight vectors and cutoff values of workflow 2 and the single weight vector and cutoff value of workflow 1, even when focusing on misclassified trials. However, the considerable variability in the cutoff values of workflow 2, ranging from -0.230 to 0.316, may cause misclassifications.

The most influential weights for classification were those representing the P1, P2 and P03 electrodes. Notably, altering these weights for trials with cutoff values close to that of workflow 2 did not significantly impact the dot product results to get correct classification. Only using the entire weight vector of workflow 1 lead to correct classification. This highlights influence of the weight vector as a whole, as opposed to individual weights, on classification outcomes.

A considerable observation is that the largest weights corresponded to the electrodes with the lowest power values and least variance, as indicated by Figure 1, which was unexpected.

To summarize, the accuracy difference between workflow 1 and workflow 2 can attributed to a high variability in workflow 2's cutoff values as well as some variability its weights. Notably, even minor alterations across all weights can significantly impact classification outcomes. This variability was caused by a large variability in the dataset as well as a model that is prone to overfitting, which means that leaving one row out can result in a very different weight vector and cutoff value.

To attenuate the influence of data variability on the weight vector and cutoff value, several techniques can be used. Artifact removal, normalization, and outlier treatment could reduce the variability in the dataset. In addition, regularization techniques could decrease the sensitivity of the

classifier making it less prone to overfitting. Furthermore, altering the computation of the weights by using the square root of the covariance matrices could adjust for the disproportionate influence of electrodes with low variance.

Interpretation of ANOVA results

The repeated measures ANOVA showed a significant main effect of class on frontal theta power at the FCz electrode, indicating that participants exhibited significantly different FCz values between successful and failed stop trials. This suggests that frontal theta power is a discriminating feature between successful and failed stop trials, supporting its use for the improvement of classification accuracy of the LDA algorithm. The observed effect size, although small ($\eta^2 = 0.004$), emphasizes the consistency of this difference across all participants.

In contrast, the lack of a significant main effect of frequency condition implies that there were no meaningful differences in FCz values across the four frequencies (4,5,6 and 7 Hz).

In addition, the absence of a significant interaction between class and frequency condition indicates that the theta band, as a whole, may be a more reliable indicator of trial type than specific frequencies within the theta range.

Interpretation of accuracy evaluations

The initial exploration of the LDA classifier using only the variances, displayed in Table 6, provided insight into the impact of various preprocessing techniques on classification performance. Overall, outlier treatment improved classification, with 13 out of 30 participants achieving significantly above-chance accuracy compared to 10 without outlier treatment. Log-transformation did not improve the algorithm. The SMOTE method did further improve the algorithm, as indicated by an increase in participants with significant above-chance accuracies to 15.

The exploration of combining LDA with L2 regularization, shown in Table 7, revealed improved performance compared to a regular LDA. The regular LDA performed worse than the LDA using variances only, with only eight participants yielding an accuracy significantly above chance. Combining the regular LDA with L2 regularization resulted in performance closer to that of LDA

with variances only. Adding the SMOTE method lead to performance very similar to that of the LDA with variances only and SMOTE.

QDA classifier performance, displayed in Table 8, indicates that its performance can be better than that of LDA. The highest number of participants (18) with accuracies significantly above chance was reached with the QDA combined with L2 regularization and the SMOTE method.

Finally, as displayed in Table 9, selecting only the frontal electrodes for LDA and QDA did not lead to further improvements. Replacing the LOOCV by k-fold cross validation decreased the performance of both classifiers slightly. However, computational time decreased significantly. Running these algorithms with five-fold cross validation instead of LOOCV resulted in approximately a fivefold reduction of computational time.

Across participant accuracy evaluation, shown in Table 10, indicates that the combination of QDA with L2 regularization and SMOTE performed best. This method yielded a mean difference of 5.8% against chance level performance ($p < 0.0001$). All other methods, except for the regular QDA ($p = 1.577$)., achieved across-participant accuracies significantly above chance level.

These observations indicate that it is important to ensure large datasets and select the appropriate preprocessing and regularization techniques to improve classifier performance. The IQR method and winsorization, by means of outlier treatment, improved performance and handled outliers without reducing the datasets in size. Normalizing the data by adding a log-transformation decreased performance, indicating that this approach may introduce complexities without substantial benefits. The SMOTE method for class balancing did further improve performance for both LDA and QDA, highlighting the importance of large datasets, as SMOTE results in larger datasets than balancing by removal of trials from the majority class. However, SMOTE involves data augmentation which introduces increased computational complexity and the risk of overfitting. Therefore, a better approach to improve classifier performance would be to ensure that each participant reaches a large number of both failed and successful trials in the SST. For both the LDA and QDA, performance was best when combining SMOTE with L2 regularization, suggesting that L2 regularization is a good method to prevent overfitting. Selecting only the frontal electrodes did not lead to improvements. This finding suggests that LDA and QDA in combination with L2 regularization were able to adjust the

weights in a way that is better than manual feature selection. Weighing the considerable decrease in computational time against the slight decrease in performance when using k-fold cross validation, it seems that k-fold cross validation is a viable option when computational time is a consideration.

Interpretation of weight vectors

It was expected that frontal theta was predictive of successful response inhibition in the SST, with higher power values in the theta range for successful trials. Trials were classified as successful when their dot product with the weight vector was below the cutoff value. This means that negative weight values are expected at the frontal electrodes for the LDA.

Individual weight vectors for the LDA, displayed in Figure 3 and 4, did not show this pattern. While some participants had large negative weights across the frontal electrodes, these weights also appeared in more posterior regions. Furthermore, some participants did not have large negative weights across the frontal electrodes at all. Finally, each participant's weight vector differed considerably, which complicates their interpretation.

Notably, individual weight vectors derived here differ from those in by Caspani (2022) (Figure 2), with weights now being more distributed. This suggests that the modification to the computation of the weights, which included taking the square root of the covariance matrices, was effective in adjusting disproportionate the influence of low-variance electrodes.

QDA does not directly use linear weights. However, such weights were derived using the linear terms of the QDA decision function in order to visualize which electrodes were important for successful classification. Per class weigh vectors were computed and those were averaged per participants. The resulting weights are different to interpret than the LDA weights. Here, large weights indicate that the corresponding electrodes were important for classification. Thus, large weights are expected at the frontal electrodes.

As displayed in Figure 5 and 6, for most participants such a pattern was not clearly visible. Again, some participants had large weights across the frontal electrodes but also in more posterior regions and some participants did not show large weights across the frontal electrodes. Notably, for the QDA variances only, for most participants the weights were widely distributed. The QDA with L2

regularization shows more of a focus on specific electrodes. Interestingly, for participant 02, a single large weight can be found at the Fz electrode. This finding is most consistent with expectations, although it is present in only one of the participants.

The average weight vectors across participants, displayed in Figure 7, did show a focus on the frontal electrodes. In addition, large weights were distributed across the central and left parietal and occipital electrodes. These findings are consistent across three of the four visualized average weight vectors. Only the QDA with variances only (Figure 7C) does not show such a pattern or any clear pattern at all.

Thus, when classifying pre-stimulus theta power data into successful and failed response inhibition there is an overall focus on frontal and parietal electrodes. This is consistent with previous research associating proactive control with frontoparietal theta activity (Cooper et al., 2015; Cooper et al., 2017). Furthermore, this finding also aligns with previous research associating the inferior frontal region and pre-SMA to response inhibition (Huster et al., 2013).

Limitations

The current project has a few limitations. First, while classification accuracies significantly higher than chance level were reached, they remained relatively low, with a maximum average accuracy of 56%. Although statistically significant, these low accuracies are likely not practically significant. Consequently, the visualized weight vectors might not be very meaningful. The weight vectors for participants with classification accuracies significantly above chance were visualized. However, even those participants had many misclassifications. Therefore, these weight vectors might not accurately capture the electrodes of importance in distinguishing between successful and failed response inhibition. Another consideration is that each classification algorithm reached significantly below chance level accuracies for at least one of the 30 participants, indicating that the classifiers may lack robustness. Finally, the observed variability in classifier outcomes across repeated runs suggests a degree of instability in the classifiers.

Conclusion

The results of the current project show the importance of selecting appropriate preprocessing and regularization techniques to improve classifier performance. Results were consistently improved when outlier treatment was employed. In addition, SMOTE further improved accuracy, but its application needs to be weighed against the possibility of overfitting and its computational load. Ensuring large datasets during experimentation is probably a better option. The effectiveness of L2 regularization was shown by improved accuracies. Feature selection of the frontal electrodes did not improve classification accuracy. While k-fold cross validation decreases classification accuracies slightly, its computational efficiency makes it a practical consideration. The best performance was obtained when QDA was combined with SMOTE and L2 regularization.

The observed variability in classifier performance across participants suggests that individualized optimization may be beneficial. However, individualized optimization will likely cost considerable computational resources and may lack generalizability. A one-size-fits-all approach may be less accurate for some individuals but offers more consistency and generalizability. In addition, such an approach is easier to implement and prevents ongoing optimization for every participant.

Future research should focus on further improving preprocessing and regularization methods as well as explore (extends of) individualized optimization, in order to investigate the neural markers of proactive response inhibition. Uncovering the predictors of response inhibition, that may remain undetected by univariate methods, offers insights into the neural mechanisms underlying this cognitive process.

References

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience, 6*(2). https://doi.org/10.1038/nn1003

Bekker, E. M., Kenemans, J. L., Hoeksma, M. R., Talsma, D., & Verbaten, M. N. (2005). The pure electrophysiology of stopping. *International Journal of Psychophysiology, 55*(2). https://doi.org/10.1016/j.ijpsycho.2004.07.005

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. In *Trends in Cognitive Sciences* (Vol. 16, Issue 2). https://doi.org/10.1016/j.tics.2011.12.010

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of Activity in the Categorical Representations of Objects. *Journal of Cognitive Neuroscience*, *15*(5). https://doi.org/10.1162/089892903322307429

Caspani, M. (2022). *Binary classification of EEG data to overt behavior of motor inhibition*. Utrecht University.

Chambers, C. D., Garavan, H., & Bellgrove, M. A. (2009). *Insights into the neural basis of response inhibition from cognitive and clinical neuroscience*. https://doi.org/10.1016/j.neubiorev.2008.08.016

Chikara, R. K., & Ko, L. W. (2019). Neural activities classification of human inhibitory control using hierarchical model. *Sensors (Switzerland)*, *19*(17). https://doi.org/10.3390/s19173791

Chikara, R. K., & Ko, L. W. (2020). Prediction of human inhibition brain function with inter-subject and intra-subject variability. *Brain Sciences*, *10*(10). https://doi.org/10.3390/brainsci10100726

Congdon, E., Mumford, J. A., Cohen, J. R., Galvan, A., Canli, T., & Poldrack, R. A. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, *3*(FEB). https://doi.org/10.3389/fpsyg.2012.00037

Cooper, P. S., Wong, A. S. W., Fulham, W. R., Thienel, R., Mansfield, E., Michie, P. T., & Karayanidis, F. (2015). Theta frontoparietal connectivity associated with proactive and reactive cognitive control processes. *NeuroImage*, *108*. https://doi.org/10.1016/j.neuroimage.2014.12.028

Cooper, P. S., Wong, A. S. W., McKewen, M., Michie, P. T., & Karayanidis, F. (2017). Frontoparietal theta oscillations during proactive control are associated with goal-updating and reduced behavioral variability. *Biological Psychology*, *129*. https://doi.org/10.1016/j.biopsycho.2017.09.008

Enz, N., Ruddy, K. L., Rueda-Delgado, L. M., & Whelan, R. (2021). Volume of β-bursts, but not their rate, predicts successful response inhibition. *Journal of Neuroscience*, *41*(23). https://doi.org/10.1523/JNEUROSCI.2231-20.2021

Galama, T. (2021). *Bottom-up prediction of cognitive control features in EEG signals. Linear classification of pre-stimulus stop-signal task response inhibition*.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4). https://doi.org/10.1162/jocn_a_01068

Huster, R. J., Enriquez-Geppert, S., Lavallee, C. F., Falkenstein, M., & Herrmann, C. S. (2013). Electroencephalography of response inhibition tasks: Functional networks and cognitive contributions. In *International Journal of Psychophysiology* (Vol. 87, Issue 3). https://doi.org/10.1016/j.ijpsycho.2012.08.001

Kenemans, J. L. (2015). Specific proactive and generic reactive inhibition. In *Neuroscience and Biobehavioral Reviews* (Vol. 56). https://doi.org/10.1016/j.neubiorev.2015.06.011

Kenemans, J. L., Schutte, I., Van Bijnen, S., & Logemann, H. N. A. (2023). How salience enhances inhibitory control: An analysis of electro-cortical mechanisms. *Biological Psychology*, *177*. https://doi.org/10.1016/j.biopsycho.2023.108505

Kok, A., Ramautar, J. R., De Ruiter, M. B., Band, G. P. H., & Ridderinkhof, K. R. (2003). *ERP components associated with successful and unsuccessful stopping in a stop-signal task*. https://doi.org/10.1046/j.1469-8986.2003.00127.x

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice reaction time responses: A model and a method. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2). https://doi.org/10.1037/0096-1523.10.2.276

Mazaheri, A., Nieuwenhuis, I. L. C., Van Dijk, H., & Jensen, O. (2009). Prestimulus alpha and mu activity predicts failure to inhibit motor responses. *Human Brain Mapping*, *30*(6). https://doi.org/10.1002/hbm.20763

Raud, L., Huster, R. J., Ivry, R. B., Labruna, L., Messel, M. S., & Greenhouse, I. (2020). A single mechanism for global and selective response inhibition under the influence of motor preparation. *Journal of Neuroscience*, *40*(41). https://doi.org/10.1523/JNEUROSCI.0607-20.2020

Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., & Al-Juaid, A. (2021). Neural decoding of eeg signals with machine learning: A systematic review. In *Brain Sciences* (Vol. 11, Issue 11). https://doi.org/10.3390/brainsci11111525

Skippen, P., Fulham, W. R., Michie, P. T., Matzke, D., Heathcote, A., & Karayanidis, F. (2020). Reconsidering electrophysiological markers of response inhibition in light of trigger failures in the stop-signal task. *Psychophysiology*, *57*(10). https://doi.org/10.1111/psyp.13619

Takacs, A., Mückschel, M., Roessner, V., & Beste, C. (2020). Decoding Stimulus-Response Representations and Their Stability Using EEG-Based Multivariate Pattern Analysis. *Cerebral Cortex Communications*, *1*, 1–12. https://doi.org/10.1093/texcom/tgaa016

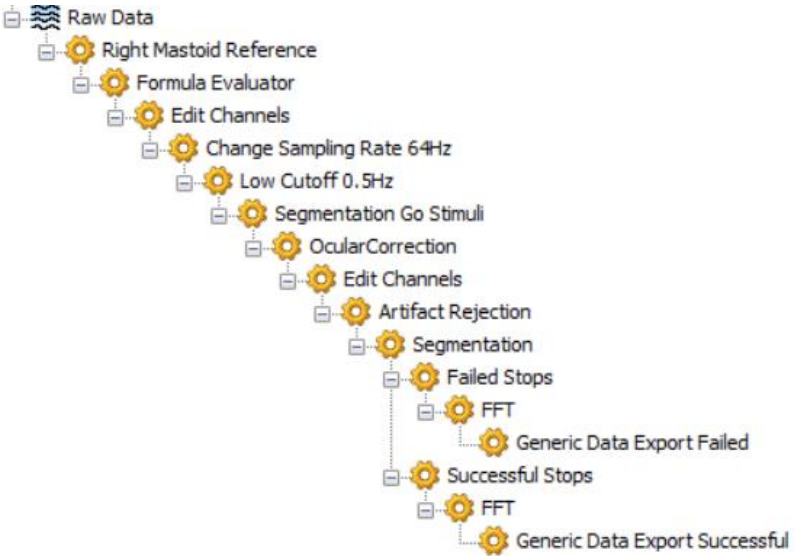
Verbruggen, F., Aron, A. R., Band, G. P. H., Beste, C., Bissett, P. G., Brockett, A. T., Brown, J. W., Chamberlain, S. R., Chambers, C. D., Colonius, H., Colzato, L. S., Corneil, B. D., Coxon, J. P., Dupuis, A., Eagle, D. M., Garavan, H., Greenhouse, I., Heathcote, A., Huster, R. J., … Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *ELife*, *8*. https://doi.org/10.7554/eLife.46323

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. In *Trends in Cognitive Sciences* (Vol. 12, Issue 11). https://doi.org/10.1016/j.tics.2008.07.005

Wagner, J., Wessel, J. R., Ghahremani, A., & Aron, A. R. (2018). Establishing a right frontal beta signature for stopping action in scalp EEG: Implications for testing inhibitory control in other task contexts. *Journal of Cognitive Neuroscience*, *30*(1). https://doi.org/10.1162/jocn_a_01183

Appendix A

The preprocessing steps performed using BrainVision Analyzer 2.3.




```
*** New Reference ***
Selected channels to include into the new reference:
EXG5
The implicit reference is included in the calculation of the new reference.

Channels to which the new reference applies to:
AF3 AF4 AF7 AF8 AFz C1 C2 C3 C4 C5
C6 CP1 CP2 CP3 CP4 CP5 CP6 CPz Cz F1
F2 F3 F4 F5 F6 F7 F8 FC1 FC2 FC3
FC4 FC5 FC6 FCz Fp1 Fp2 Fpz FT7 FT8 Fz
Iz O1 O2 Oz P1 P10 P2 P3 P4 P5
P6 P7 P8 P9 PO3 PO4 PO7 PO8 POz Pz
T7 T8 TP7 TP8
Remaining (non rereferenced) channels are kept.

Name of the new reference channel: right mastoid

*** Formula Evaluator ***

The following formulas were calculated:
VEOG = EXG4 - EXG3    Unit: µV
HEOG = EXG1 - EXG2    Unit: µV

The remaining channels were kept.
The new channels are on top.

*** Edit Channels ***

The following channels have been disabled:
EXG1   EXG2   EXG3   EXG4
EXG5   EXG6   EXG7   EXG8
Status

*** Change Sampling Rate / Resolution ***
Conversion is based on spline interpolation.
New Sampling Rate [Hz]: 64
New Sampling Interval [µs]: 15625
Data was filtered before downsampling with 28.8Hz, 24dB/oct.

*** IIR Filters ***
Zero phase shift Butterworth filters.
```

```
Data stored in continuous filter cache.

Global filter settings:
Low cutoff: 0.5 Hz, time constant 0.31830988618379069, order 2
High cutoff: 28.8 Hz, order 2
Notch filter: ---

*** Segmentation ***
Segmentation relative to reference marker positions
Reference markers:
    Stimulus    S 22
    Stimulus    S 21
Advanced Boolean Expression:
    ---
Segment size and position relative to reference markers:
Start: -1000.00 ms, End: 1500.00 ms, Length: 2500.00 ms
Allow overlapped segments? Yes
Skip bad intervals?        Yes

Data was not stored but will be calculated on demand.

*** Ocular Correction (Gratton & Coles) ***
Name of HEOG channel: HEOG
Common reference
Name of VEOG channel: VEOG
Common reference

The following channels have been selected for correction:
Fp1    AF7    AF3    F1
F3     F5     F7     FT7
FC5    FC3    FC1    C1
C3     C5     T7     TP7
CP5    CP3    CP1    P1
P3     P5     P7     P9
PO7    PO3    O1     Iz
Oz     POz    Pz     CPz
Fpz    Fp2    AF8    AF4
AFz    Fz     F2     F4
F6     F8     FT8    FC6
FC4    FC2    FCz    Cz
C2     C4     C6     T8
TP8    CP6    CP4    CP2
P2     P4     P6     P8
P10    PO8    PO4    O2

*** Edit Channels ***

The following channels have been disabled:
VEOG   HEOG

*** Artifact Rejection - Automatic Inspection ***

Used Channels: 64
AF3    AF4    AF7    AF8    AFz    C1     C2     C3     C4     C5     C6     CP1    CP2
       CP3    CP4    CP5    CP6    CPz    Cz     F1     F2     F3     F4     F5     F6
       F7     F8     FC1    FC2    FC3    FC4    FC5    FC6    FCz    Fp1    Fp2    Fpz
       FT7    FT8    Fz     Iz     O1     O2     P10    P3     P4     P5     P6     P7
       P8     P9     PO4    PO7    PO8    POz    Pz     T7     T8     TP7    TP8    Oz
       P1     P2     PO3
Check Gradient:
Maximal allowed voltage step: 50 µV/ms
Mark as Bad:  Before Event: 200 ms      After Event: 200 ms
Check Difference (Max-Min):
Maximal allowed difference of values in intervals: 100 µV
Interval Length: 100 ms
Mark as Bad:  Before Event: 200 ms      After Event: 200 ms
Check Amplitude:
```

```
Minimal allowed amplitude: -200 µV
Maximal allowed amplitude: 200 µV
Mark as Bad:  Before Event: 200 ms      After Event: 200 ms
Check Low Activity:
Lowest allowed activity in intervals: 0.5 µV
Interval Length: 100 ms
Mark as Bad:  Before Event: 200 ms      After Event: 200 ms

*** Segmentation ***
Segmentation relative to reference marker positions
Reference markers:
    Stimulus    S 22
    Stimulus    S 21
Advanced Boolean Expression:
    ---
Segment size and position relative to reference markers:
Start: -1000.00 ms, End: 1500.00 ms, Length: 2500.00 ms
Allow overlapped segments? Yes
Skip bad intervals?        Yes

Data was not stored but will be calculated on demand.

*** Segmentation ***
Segmentation relative to reference marker positions
Reference markers:
    Stimulus    S 22
    Stimulus    S 21
Advanced Boolean Expression:
    S1 (0, 1000) or S2(0, 1000)
Segment size and position relative to reference markers:
Start: -1000.00 ms, End: 0.00 ms, Length: 1000.00 ms
Allow overlapped segments? Yes
Skip bad intervals?        No

Data was not stored but will be calculated on demand.

*** Fast Fourier Transformation (FFT) ***
Maximum Resolution
Power
Non-Complex Output
Half Spectrum used

Data Window:
Hanning Window
Length = 10 %
Variance Correction used
Periodic

*** Generic Data Export ***
File name parameter: failed$h
File extension: .dat
Write header file: yes
Write marker file: no
Format: ASCII
Orientation: VECTORIZED
Line Delimiter: CRLF (PC style)
Add channel names: yes
Overwrite default decimal symbol: no
Export all channels: yes

*** Segmentation ***
Segmentation relative to reference marker positions
Reference markers:
    Stimulus    S 22
    Stimulus    S 21
Advanced Boolean Expression:
    not S1 (0, 1000) and not S2(0, 1000)
Segment size and position relative to reference markers:
```

```
Start: -1000.00 ms, End: 0.00 ms, Length: 1000.00 ms
Allow overlapped segments? Yes
Skip bad intervals?        No

Data was not stored but will be calculated on demand.

*** Fast Fourier Transformation (FFT) ***
Maximum Resolution
Power
Non-Complex Output
Half Spectrum used

Data Window:
Hanning Window
Length = 10 %
Variance Correction used
Periodic

*** Generic Data Export ***
File name parameter: successful$h
File extension: .dat
Write header file: yes
Write marker file: no
Format: ASCII
Orientation: VECTORIZED
Line Delimiter: CRLF (PC style)
Add channel names: yes
Overwrite default decimal symbol: no
Export all channels: yes
```

Appendix B

Per participant removed electrodes and the amount of remaining trials.

Ptp	Removed electrodes	Failed trials	Success trials	Total	Balanced total
Pp01	P1, PO3	80	92	172	160
Pp02	FPz, F7, F8, FT7	93	36	129	72
Pp03	P2	80	69	149	138
Pp04	-	86	55	141	110
Pp05	-	102	72	174	144
Pp06	-	96	82	178	164
Pp07	-	101	70	171	140
Pp09	-	93	70	163	140
Pp10	P1	85	75	160	150
Pp101	Iz	135	112	247	224
Pp102	-	115	124	239	230
Pp103	-	102	135	237	204
Pp104	Excluded from further analysis. Only 6 successful trials even before artifact rejection.				
Pp106	-	86	101	187	172
Pp107	-	143	135	278	270
Pp108	-	125	126	251	250
Pp109	-	103	180	283	206
Pp11	T7	141	29	170	58
Pp110	-	84	137	221	168
Pp111	-	122	120	242	240
Pp112	P2	71	142	213	142
Pp12	-	75	92	167	150
Pp13	FPz, FP1, FP2, AFz, AF3, AF4, AF7, AF8, F8	60	77	137	120
Pp14	FPz, FP1, FP2, AFz, AF7, AF8, F1, F3, F4, F5, F7, F8, FC5, POz, P9, T7	66	71	137	132
Pp15	-	64	86	150	128
Pp16	P1, PO3	76	85	161	152
Pp17	POz, P03, Pz, P1, P2, CPz, Oz	62	91	153	124
Pp18	Excluded from further analysis. All electrodes have artifacts for 100% of the trials.				
Pp19	P1, PO3, PO7, PO8	55	89	144	110
Pp20	Pz, P1, P8, PO3	64	97	161	128
Pp21	FP1, Pz, P1, P3, P5, PO3, PO7	77	72	149	144
Pp22	Pz, PO3	105	78	183	156

Appendix C

The code for the classifiers as well as for the statistical tests can be found here:

https://github.com/laravancappelle/classification_response_inhibition.git