

BACHELOR THESIS
ARTIFICIAL INTELLIGENCE

Radboud University



Decoding fMRI data for color
qualia by using a Multi-Layer
Perceptron

Author:
Kelly Karremans
4802802

First supervisor:
dr. O. Colizoli
Donders Institute for
Brain, Cognition and
Behaviour
o.colizoli@donders.ru.nl

Second supervisor:
dr. L. Geerligts
Donders Centre for
Cognition
l.geerlings@donders.ru.nl



July 7, 2022

Abstract

Participants were scanned with an fMRI while perceiving letter-color combinations. These participants implicitly learned letter-color associations by reading books with colored letters. Afterward, they were scanned again under the same conditions as before. This study tries to determine if a Multi-Layer Perceptron can predict whether a person is looking at a colored or black letter before and after association learning. Suppose color qualia, the conscious experiences of color, are processed in the same brain regions as when visually perceiving colors. In that case, the model is expected to become less accurate when applied to post-training data. It will look at if association learning causes this model to predict black events more often as color events. This is not the case, meaning there is no indication that color qualia might be processed in the same regions in the occipital lobe as visual color processing. It cannot be said if there is any significant positive correlation between the Projector-Associator scores of the subjects (which gives an indication of how strong their associated color has become) and how often their black events are predicted as color events post-training.

Contents

1	Introduction	3
1.1	Previous related work	3
1.2	Research questions	4
1.2.1	Motivation & Scientific relevance	4
1.3	Hypotheses	5
2	Methods	6
2.1	Participants	6
2.2	Procedure	6
2.3	Oddball task	7
2.4	fMRI acquisition	8
2.5	fMRI preprocessing	8
2.6	b0 unwrapping	8
2.7	Motion correction	8
2.8	Temporal High-Pass Filtering	8
2.9	Brain extraction	9
2.10	Brain registration to common space	9
2.11	Regions of Interest	9
2.12	Models	9
2.12.1	Multi-Layer Perceptron	9
2.13	Statistical Analysis	10
2.14	Software & Hardware	11
3	Results	11
3.1	Hypothesis 1	11
3.1.1	Most important voxels	13
3.2	Hypothesis 2	14
3.3	Hypothesis 3	15
4	Discussion	16
5	Conclusions	17
	Appendix A Color Experience Questionnaire	20
	Appendix B Accuracy and loss graphs for all 10 Stratified Folds	22
	Appendix C Voxel Importance	24

1 Introduction

Qualia are the internal conscious experiences of a sentient being. Subsequently, color qualia refer to the conscious experience of seeing or imagining a specific color.

One can wonder where color qualia are processed in the brain and which voxels would be involved. In this study, a Multi-Layer Perceptron will be used to first see if it is possible to predict whether or not a participant is looking at a color or black event. If this is possible, the next step will be to discover if color qualia are perhaps processed in the same brain regions used to process visual color stimuli.

First, an overview of previous related research will be given. Then the research question will be presented with the corresponding hypothesis.

The methods section outlines all the information needed to replicate this study. Next, the results will be presented in the results section, which will be summarized per hypothesis. These results will be discussed in the discussion section. At last, a short conclusion of the study will be given in the conclusion section.

1.1 Previous related work

The participants in this study will undergo association learning to induce a change in color qualia. In previous research, it has been shown that it is possible to induce a change in subjective color experience by using association learning.

In 2012, a study by Colizoli et al. showed that it is possible to train synesthesia to some extent. Synesthesia is the phenomenon where while experiencing one stimulus, a person involuntarily experiences a second sensory stimulus (Cytowic, 2002). For example, tasting specific colors or seeing letters in a particular color. Participants in this study implicitly learned letter-color associations by reading books for an extended time, where letters were assigned a certain color. These participants were tested both before and after training using a modified Stroop task and a crowding task. The results of these tasks showed that it was possible to train synesthesia in those that did not initially experience it, thus showing it is possible to change the subjective color experience of the participants.

In a study by Colizoli et al. (2016) the activity of the visual cortex was used to predict the subjective experience of color perception. This was done by letting participants read a book for several weeks (4-6), where the letters were colored. Before and after these weeks of implicitly learning the color-letter associations, the participants were scanned using an fMRI (functional magnetic resonance imaging). Suppose it is the case that these associations changed some form of visual qualia, for example, a different quality in color experience or a more vivid visual mental image. In that

case, the brain activation while viewing the trained black letters should be able to predict the strength of these associations. The participant was given a Projector-Associator (PA) questionnaire afterward. A PA questionnaire can be used to determine a so-called Projector-Associator score for a subject. This questionnaire can be found in Appendix A. A projector-type of color qualia experience would have a similar experience to normal color vision, while an associator-type experience would be more like experiencing the associated color in the mind’s eye. In this study, the results showed that, indeed, there was a correlation between the reported differences in the color experience and the training-related activation of V4.

Thus, both the study by Colizoli et al. (2012) and Colizoli et al. (2016) showed that it is possible to induce a change in color qualia by the use of association learning.

Using Machine Learning algorithms, such as Multi-Layer Perceptrons, to decode fMRI data is not a new concept. This has been done in numerous previous studies (Feng et al., 2021). Ulloa et al. (2018) used a Multi-Layer perceptron to predict a schizophrenia diagnosis based upon a combination of sMRI and fMRI data. The result was a prediction accuracy of 0.8, which is well above random chance (0.5). In 2020 a study by Wang et al. was done, which used a Multi-Layer Perceptron and ensemble learning to identify if a person had an Autism Spectrum Disorder diagnosis based on fMRI data. This resulted in a classification accuracy of 74.52%. A 96.66% classification accuracy for Alzheimer’s Disease was achieved by using a Multi-Layer Perceptron classifier. However, instead of fMRI data, sMRI data was used, where the features were extracted by using a cascaded 3D CNN (Raju et al., 2021).

Overall, there seems to be a precedent of successfully using Multi-Layer Perceptron classifiers in combination with brain data acquired by fMRI.

1.2 Research questions

The research question that this study will try to answer is: "Can a Multi-Layer Perceptron correctly classify observed colors by using fMRI data?".

1.2.1 Motivation & Scientific relevance

In previous research by Colizoli et al. (2012) and Colizoli et al. (2016) it was shown that subjective color experiences can be trained by association learning. This study will build upon that finding by using data from pre- and post-trained participants, assuming that the color qualia of these subjects will change due to the association learning.

This data can be used to see to what extent a Multi-Layer Perceptron can classify a subject’s visual color experience. It might be possible to get an insight into which voxels mainly contribute to this classification,

implicating that these brain regions are used for processing colors in vision. After constructing this model and using it on the data after association learning, it becomes possible to get a clue about if color qualia are processed in the same brain regions as the ones used to perceive colors before training.

Assuming that after association learning, the color qualia are experienced in the same brain regions, then by looking at the Projector-Associator scores and how they relate to predictions made after association learning, an insight into how subjectively intense color qualia experiences translate to brain activation can be made as well.

In a broader sense, a better understanding of the brain and its neural basis concerning color qualia can be obtained. A better understanding of the neural basis and how it relates to subjective experiences can provide insight into our still very limited understanding of consciousness.

1.3 Hypotheses

The hypotheses will be split up into three parts.

1. Considering that previous studies have shown that a Multi-Layer Perceptron can be used to make accurate predictions based on fMRI data, the first hypothesis becomes:

A Multi-Layer Perceptron is able to predict if a participant is looking at a color or black event, before association learning, with an accuracy above chance level.

2. Assuming that processing color qualia uses the same brain regions as visually perceiving colors before any association training, one would expect previous black events to be classified as color events. This would be because if it used the same brain regions, it should not just process the presented black color but also the now associated color qualia. Thus the second hypothesis will be:

After association learning, the model should predict black events more often as a color event.

3. Assuming that color qualia are processed in the same brain regions as visually perceiving colors before any association training, then if a participant experiences the new associated color experience more intensely, one would expect the brain activation to mirror that of the experienced color qualia, especially in the projector-type experience cases, where the experience is similar to that of visually perceiving the color stimulus. Therefore, it could be expected that a higher Projector-Associator score of a subject would mean that their black events are now more often predicted as a color event to mirror their color qualia experience. The third hypothesis becomes:

After association learning, there should be a positive correlation between the Projector-Associator scores of the subjects and how often their black events are classified as a color event.

2 Methods

2.1 Participants

The research experiment described below was approved by the ethical committee of the Donders Centre for Cognitive Neuroimaging. Every participant was informed that they had the option to withdraw their participation at any given moment. Before participating in the research, each participant has signed an informed consent form.

A group of $N = 46$ participants were selected. The demographic consisted of 34 women and 12 men, with an age range of 18 to 34 years. The mean age of the participants was $M \approx 24.85$ years with a standard deviation $SD \approx 4.24$.

All the subjects were screened before participating based on an interview and questionnaire. This was done to rule out participants that experienced synesthesia. Participants were also ruled out if they had color blindness, attention deficit disorder (ADD), or dyslexia.

2.2 Procedure

The experimental design consists of three parts. The first part is the pre-test phase, the second is the training phase, and the third is the post-test phase. The participant's brain activity was measured using an fMRI in both the pre-test and post-test phases.

In the pre-test, the participants were asked to read the letters presented to them. These consisted of the 26 letters of the Roman alphabet. These letters could either have a color or be black. There were a total of 13 different colors (not counting black). Every letter was presented to the participant twice, once in color and once in black. The data collected in the pre-test will be referred to later in the study as either session 1 data or data before association learning.

The participants were split into two groups before moving on to the training phase. One group was allowed to choose which color they wished to correspond to each letter. The other group was not allowed to select these color-letter combinations for themselves; instead, each participant in this group was assigned a color-mapping corresponding to the chosen colors of a subject in the first group. These letter color combinations were counterbalanced across the participants. In total, there were 13 trained letters with 13 unique colors.

The participants were asked to start with the association learning in the training phase. This training was done by letting the participants read books where each letter was consistently assigned a certain color. This was the only requirement for the books. Besides this, the participants were free to choose whichever book(s) they wanted to read. This training phase lasted between 4 to 6 weeks, and the participants were asked to read at least 100.000 words. Each book had a word count to make it easy for the participants to keep track of.

In Figure 1, an example of the letter-color mapping can be seen.

bcdefghijklma (Arial Black)

**Sunset is the time of day when
our sky meets the outer space
solar winds. There are blue, pink,
and purple swirls, jogging, spinning
and twisting, like clouds of balloons
caught in a whirlwind. A jellyfish
started joking all the time.**

Figure 1: Example of a possible letter-color mapping

In the post-test phase, the participants were again asked to read the letters presented to them. The goal was to observe the impact of the association learning. Thus, it is desired to have minimum interference of other variables because then it is possible to measure more precisely the influence of the association learning. Therefore, the post-test was set up identically to the pre-test. The data collected in the post-test will be referred to later in the study as either session 2 data or data after association learning.

The participants did not only look at the presented graphemes during the pre-test and post-test. To make sure the participants were paying attention, the letters were presented with the use of an oddball task. The oddball task is described in further detail in the section below.

The data collected from the pre-test and post-test sessions is the data that will be used for this study.

2.3 Oddball task

During the oddball task, the participants were asked to press a button whenever they were presented with a so-called oddball stimulus. An oddball stimulus, in this case, was any stimulus that was not part of the intended graphemes conditions discussed in the section above (letters in either black or any of the other 13 colors). The oddball stimulus was either a letter in

a grey color or a number (irrelevant of which color it was presented in). By using these oddballs, it could be made sure that the participants were actively paying attention to both the shape and color of the grapheme.

There are a total of 52 events. There are 26 events where each letter is presented in black and 26 events where each letter is presented in one of the 13 colors. There are two sessions, the pre-test and post-test sessions. This makes a total of 104 events for every participant. This study aims to build a Multi-Layer Perceptron model which can predict whether a participant is experiencing a black event or a color event based on the fMRI data.

A General Linear Model was created by dr.O. Colizoli. Z-scores were used for inference on the event parameters. The voxels and their corresponding Z-scores were used for the datasets (pre-test and post-test data) mentioned in this study.

2.4 fMRI acquisition

The MRI used was a 3 Tesla (3T) MAGNETOM PrismaFit MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) equipped with a 32-channel head coil. The MRI was located in Nijmegen (The Netherlands). A multiband acceleration factor of 4 was used. The repetition time (RT) is 1.5s, and the echo time (ET) is 39.6ms. The voxel resolution is 2 x 2 x 2 mm. The fMRI acquisitions used the interleaved scheme, and there are 68 slices.

2.5 fMRI preprocessing

2.6 b0 unwrapping

When participants are in the scanner, they mess up the uniform magnetic field. The proper position can be recovered when the signal has become displaced. To correct these distortions, b0 unwrapping was used (Veer et al., 2017).

2.7 Motion correction

Motion correction was used to correct data influenced by the participants' movements. When a participant moved their head too much, the data was removed from the dataset.

2.8 Temporal High-Pass Filtering

Physiological and physical noise could cause low-frequency drifts. These drifts can result in statistical analysis losing its power in data analysis. These drifts also have a tendency to invalidate event-related averaging. Temporal High-Pass Filtering was used to remove this drift (Woolrich et al., 2001).

2.9 Brain extraction

Voxels that are not part of the brain were removed. For example, voxels that were part of the skull were removed.

2.10 Brain registration to common space

The data is spatially normalized to account for differences between the participants' brains. This was done by matching their brains to a standard brain template. The Montreal Neurological Institute 152 (MNI-152) was the template used (Chau & McIntosh, 2005).

2.11 Regions of Interest

For this study, there is only one region of interest, namely the occipital lobe. The brain's visual areas are located in the occipital lobe (Rehman & Al Khalili, 2019). In this study, the focus lies on building a Multi-Layer Perceptron that is able to predict whether or not a participant is seeing a color or black event. Thus, the areas related to visual processing are needed. The brain voxels were extracted using the Harvard Oxford brain atlas. (Makris et al., 2006)(Frazier et al., 2005)(Desikan et al., 2006)(Goldstein et al., 2007)

2.12 Models

2.12.1 Multi-Layer Perceptron

The Multi-Layer Perceptron consists of an input, hidden, and output layer. The Multi-Layer Perceptron in this study consists of an input layer that takes an input dimension of (2392, 10000). Where 2392 comes from the 46 participants, that each have 52 events for a session ($46 * 52 = 2392$). To reduce the problem of overfitting, only a single hidden layer is used with 25 nodes and the ReLu activation function. In addition, one Dropout layer with a rate of 0.4 is added to combat overfitting. The Dropout layer sets a fraction of 0.4 of the features to 0 during each training step. The output layer has one single node and uses the Sigmoid activation function. The loss function will be binary cross-entropy because it will try to classify binary events (either black or color). Early stopping is used to prevent the Multi-Layer Perceptron from overtraining, which can result in overfitting.

2.12.1.1 Dataset preprocessing The dataset has 29642 voxels. It is possible to reduce the number of voxels used in the model using F-scores. An F-score is determined by looking at the variance a feature has between other features and between itself. A high F-score indicates that a feature is more likely to be important for correct classification. The 10000 voxels with the highest F-scores are used for the training and validation sets, and the

others are dropped. This results in a lower need for computational power and memory size.

See Figure 2 to see the F-score distribution among all voxels.

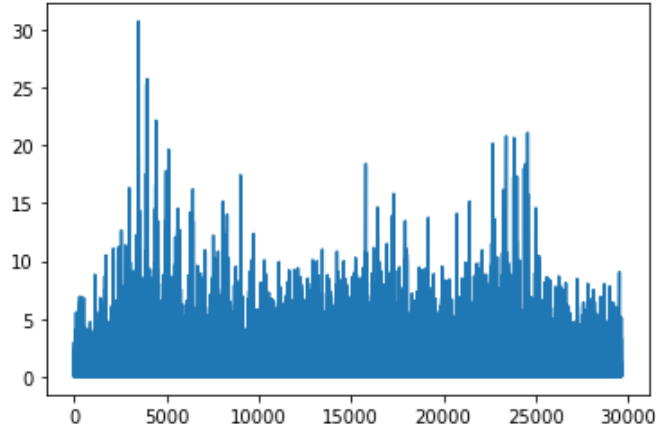


Figure 2: F-score distribution among all 29642 voxels

2.13 Statistical Analysis

For the first hypothesis, only the mean prediction accuracy is used because it only needs to be established that the model’s accuracy is above the chance level, i.e., an accuracy above 50%. Then, Permutation Importance can be used to find out which voxels contributed the most to the classification. A feature is shuffled, after which the decrease in the model accuracy is noted. The more the accuracy declines, the more important a feature is considered to be for classification. This will be done for every feature.

For the second hypothesis, if the accuracy decreases when using data from session 2, it will be necessary to determine if this decrease was caused because the color qualia result in black events being predicted as color. This can only be possible in cases where the event is trained.

To get a good overview of how often, for a subject, out of all presented black events, a black event ended up being predicted as color, a variable called `BlackEventScore` is introduced. This variable is calculated by counting the number of times a participant saw a black event, but the prediction was a color event. This number is then divided by the number of events per subject.

To determine if the decrease is caused by the color qualia, a t-test needs to be performed on the means of the trained `BlackEventScores` and the untrained `BlackEventScores`. The mean of the trained `BlackEventScores` should be higher than that of the untrained `BlackEventScores` if the decrease was caused by the trained black events.

For the final hypothesis, a Pearson correlation coefficient is needed to

determine if there is a positive correlation between the Projector-Associator scores of the subjects and their BlackEventScores.

2.14 Software & Hardware

The software used is Python version 3.9.7. To perform as previously discussed the paired t-test and Pearson correlation coefficient, the package `scipy.stats` is used. The machine learning was done on the GPU. The computations were performed on a computer with the following hardware specifications:

- **Processor:** Intel(R) Core(TM) i7-4770K CPU @ 3.50GHz 3.50 GHz
- **Installed RAM:** 16,0 GB
- **GPU:** NVIDIA GeForce GTX 1060 6GB

3 Results

3.1 Hypothesis 1

The first hypothesis states that a Multi-Layer Perceptron is able to predict if a participant is looking at a color or black event, before association learning, with an accuracy above chance level. (With H_0 : *the Multi-Layer Perceptron predicts if a participant is looking at a color or black event, before association learning, with the same accuracy as chance level.*)

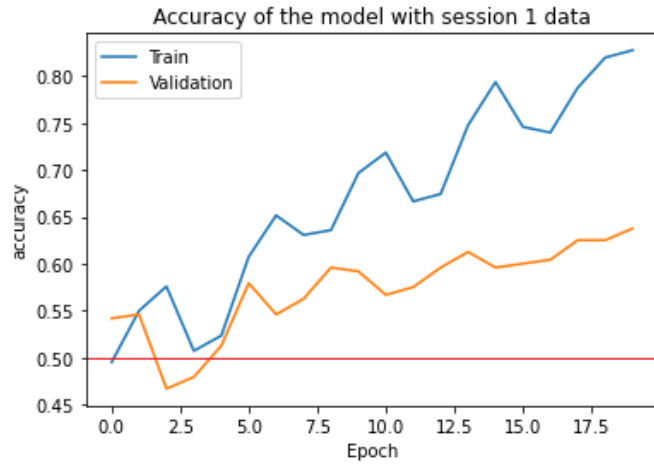
For this hypothesis, only the data from the first session, which consists entirely of untrained events, was needed. When using the data from the first session, the Multi-Layer Perceptron had a mean accuracy of 90.62% for the training data, with a standard deviation of 0.05 and a mean loss of 0.27.

For the validation set, the model had a 60.79% mean accuracy, with a standard deviation of 0.04 and a mean loss of 0.87. The model's accuracy and loss have been plotted for both the training and validation sets. In Figure 3a and Figure 3b below, the plots of one stratified fold can be seen. The plots of all ten folds can be found in Appendix B.

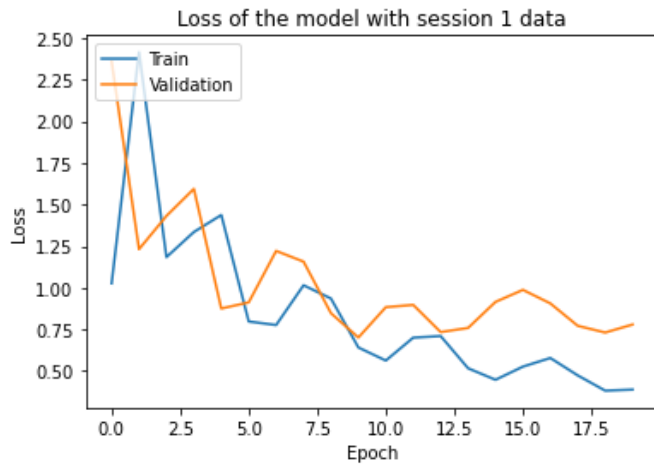
Suppose an accuracy value in these accuracy plots is above the red line. In that case, it indicates that the accuracy at that epoch is better than the chance level. The accuracy plots show that the accuracy for both the training and validation sets is above the chance level for most epochs across all folds.

When looking at the loss values, it can be seen that the average training loss is lower than the average validation loss ($0.27 < 0.87$). This means that the model is a better fit for the training set than for the validation set, which was expected considering the differences in accuracy. With the differences in

accuracy and loss values, it can be said that the model is overfitting despite the measures that were taken to prevent this.



(a) On the x-axis the epochs are shown and on the y-axis the accuracy values are presented. The red line indicates an accuracy of 0.5, which is the chance level.



(b) The epochs are shown on the x-axis and the loss values on the y-axis.

Figure 3: The accuracy (a) and loss (b) plot of the first fold.

For every fold, the validation set's correctly and wrongfully predicted events were stored. This resulted in a total of 1196 black events. The model predicts 694 times correctly that a participant was experiencing a black event, and 502 times the model predicts it as a color event. There was a total of 1196 color events as well. Here the model correctly predicts 760 times that the event a participant experienced was indeed a color event. However, it wrongfully predicts that the other 436 color events are black events. In Figure 4, a confusion matrix can be seen, which gives an overview of these

predictions.

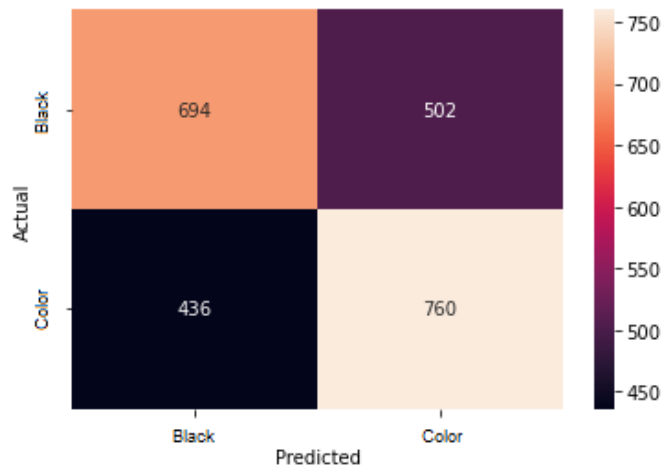


Figure 4: A confusion matrix of the predictions made with session 1 data. The top left: correctly predicts black. Top right: incorrectly predicts color (instead of black). Bottom left: incorrectly predicts black (instead of color). Bottom right: correctly predicts color.

With these results, we can say that the first hypothesis is correct and thus reject the null hypothesis. With a mean accuracy of 60.79%, it is indeed true that a Multi-Layer Perceptron can predict better than chance whether or not a participant is looking at either a colored or black letter.

3.1.1 Most important voxels

Because the first hypothesis has been accepted, it is possible to look at which voxels were the most important in determining whether a participant was looking at either a black or colored event. To find these voxels, permutation importance was used. In Figure 5, all the voxels with a positive influence on the prediction are shown. See Appendix C for all the permutation importance values of all the other voxels, excluding those with zero values. It is noticeable that there are no significant clusters of voxels. The Discussion section discusses the possible reasons for the lack of clusters.

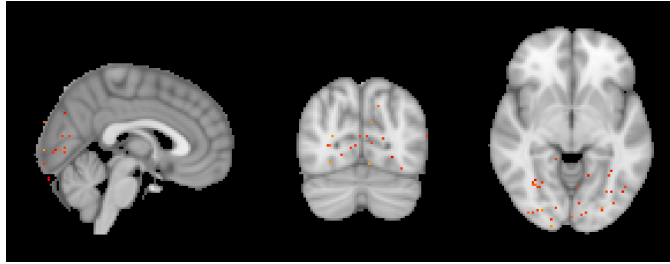


Figure 5: All the voxels within the occipital lobe which have a positive influence on the event predictions are highlighted.

3.2 Hypothesis 2

The second hypothesis states that the model should predict black events more often as color events after association learning. (With H_0 : *after association learning, there is no effect on the number of black events being predicted as color events.*)

For this hypothesis, the entire dataset of session 2 is needed, and a subset of this data with only the trained events of session 2. After applying the model to the session 2 data, an accuracy of 54% is achieved and a loss of 1.07. Thus, the accuracy of the model decreases. This decrease can indicate that, indeed, black colors are now more often wrongly predicted as color. More analysis is needed to confirm the second hypothesis. It is important to ensure that decrease is due to the trained data having a lower accuracy and not due to other factors, such as insignificant data variation.

First, an overview of the specific predictions of this second session. Of the 1196 black events, 624 are correctly predicted. The other 572 black events were predicted as color. Comparing this with the session 1 data shows that $572 - 502 = 70$ more black events are classified as color in the session 2 data. For the 1196 color events, it is 679 times correctly predicted to be color. The remaining 517 color events are predicted to be black. Compared to the session 1 data, there are $517 - 436 = 81$ more cases where color events are incorrectly predicted.

In Figure 6 below, a confusion matrix can be seen, which gives an overview of the predictions with session 2 data.

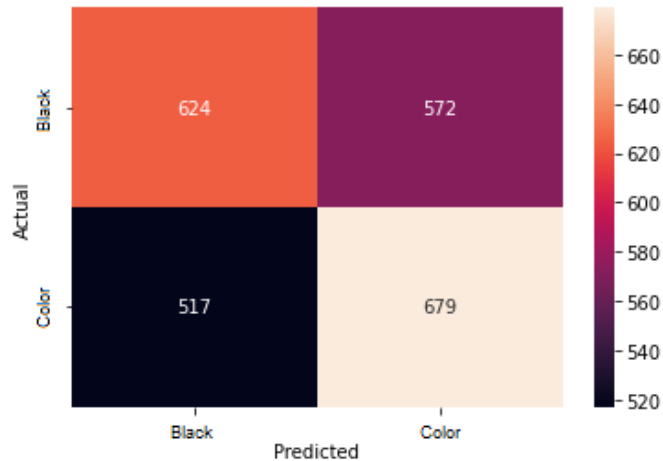


Figure 6: A confusion matrix of the predictions made with session 2 data. The top left: correctly predicts black. Top right: incorrectly predicts color (instead of black). Bottom left: incorrectly predicts black (instead of color). Bottom right: correctly predicts color.

As presented in the methods section, a variable called `BlackEventScore` was used to get a good overview of how often, out of all the black events presented to a subject, a black event ended up being predicted as color. The `BlackEventScores` were taken for both the trained and untrained data of session 2. One would expect the mean of the trained `BlackEventScores` to be higher than the mean of untrained `BlackEventScores` if the decrease in accuracy was caused by the trained black events. In other words, the two means should not be equal. To test if the accuracy has decreased due to the trained black events being more often incorrectly predicted as color events, a two-sided paired t-test, with a significance level of $\alpha = 0.05$ was used on the means of the `BlackEventScores` of the trained data and the untrained data. If the p-value is smaller than α , then the null hypothesis (meaning both means are equal) can be rejected.

The t-test gave a p-value of around 0.50 and a t-statistic of around 0.41. Since the p-value is larger than the significance level ($0.50 > 0.05$), the null hypothesis cannot be rejected. Therefore it cannot be said that the decrease in accuracy is caused by the trained black events being incorrectly predicted as color events when using the model.

3.3 Hypothesis 3

The last hypothesis states that there should be a positive correlation between the Projector-Associator scores of the subjects and their `BlackEventScores` of the trained session 2 data. (With H_0 : *there is no correlation between the Projector-Associator scores of the subjects and their `BlackEventScores`*

of the trained session 2 data.)

The Pearson correlation coefficient was used to determine if there is a correlation between the Projector-Associator scores and BlackEventScores. Again we use a significance level of $\alpha = 0.05$. The Pearson correlation coefficient between these two scores was around 0.15, with a p-value of around 0.3. Thus, there is a very weak positive correlation between the scores. In Figure 7, a graph is shown that visualizes the correlation between the scores.



Figure 7: The positive correlation between the Projector-Associator scores and BlackEventScores. With the Projector-Associator score on the x-axis and the BlackEventScores on the y-axis.

However, the p-value is larger than the significance level ($0.32 > 0.05$). Thus it is not possible to reject the null hypothesis.

4 Discussion

When looking at the results, it becomes clear that the Multi-Layer Perceptron can predict with an accuracy above change level whether or not a participant, before association learning, is looking at either a colored or black letter. When looking at the most essential voxels in regards to predicting the color and black events, it can be seen in Figure 5 that there are no apparent clusters. This could be due to the pre-processing, where only the 10000 most likely to be relevant features were selected. This could have filtered out other nearby voxels because if the Z-scores were very much related to nearby voxels, the variance between these features would be lower, thus resulting in a lower F-score and subsequently being filtered out. A suggestion for future research would be to see if there would be clusters of voxels when feature selection is not used. However, this would increase the dataset's size,

resulting in a need for better hardware with more computational power and memory than was used in this study.

When looking at how the dataset is created, there could be a possible problem with how the current model predicts these events. Namely the problem of different events, but of the same subject, being in the training and testing data. When creating the session 1 dataset (before association learning), all the events of all the subjects were put together in the hopes of the Multi-Layer Perceptron being able to find a general link between the activation of the voxels in the participant’s brains that could be used to predict the events. However, this could mean that the Multi-Layer Perceptron could predict events in the validation set solely because there was similar activation in the voxels present in the training set from the same subject instead of predicting based on the generalization of the activation from the events.

Thus, splitting the dataset into a training and validation set could result in a training set with many events from one subject. This might mean that the model could better predict the smaller occurring amount of different events from that same subject in the validation set. With Stratified K-Fold, this problem is reduced but could still occur. Thus, it might be possible that the current model is less accurate than presented in the result section. For future research, it is suggested to rule out this problem by ensuring the training and validation data are split so that the same subject’s events do not occur in both the training and validation sets.

A common issue when decoding fMRI data with a Multi-Layer Perceptron is the limited amount of data because a Multi-Layer Perceptron can better generalize when there is more available data. When there is insufficient data, it might result in overfitting (Raschka, 2015). As seen in the Results section, the current model tends to overfit the training data. This issue could possibly be resolved with the use of Transfer Learning. With Transfer Learning, a model has first trained on a different but somewhat similar dataset before being used with the actual data to improve accuracy (Zhang et al., 2018). Another option could be data augmentation (Perez & Wang, 2017). For example, using a synthetic data generator that can create synthetic fMRI data. Thus enlarging the dataset without the need for more participants.

Therefore, it is suggested that in future research, either Transfer Learning or data augmentation is used to try and improve the model’s accuracy.

5 Conclusions

Overall, only the first hypothesis can be accepted. This means that the Multi-layer Perceptron has a better than chance accuracy when predicting whether or not a participant was shown a black or color event. Before asso-

ciation learning (session 1), the mean accuracy was 60.76% with a standard deviation of 0.04. However, due to data of subjects being split across both training and validation sets, this reported accuracy could be higher than it would have been if the events of the subjects had been assigned to only one of the sets.

It was impossible to reject the other two hypotheses' null hypotheses. Therefore, it cannot be said if a decrease in accuracy when applying the model on the session 2 data can be explained by an increase in trained black events now being predicted as color events. This means there is currently no indication that for the experience of color qualia, the brain might use the same regions as when perceiving color visually. It is also not possible to say if a higher amount of trained black events being predicted as a color event indicates a higher Projector-Associator score for a subject.

An improvement of the current model could be splitting the training and validation data differently, namely, by making sure that all events of the same subject are either in the training set or in the validation set. This should be done to prevent the model from learning color and black events solely from already present data of that subject in the training set instead of it being able to make predictions based on the generalization of the events. In future research, Transfer Learning and data augmentation could be used to prevent overfitting and increase accuracy of the model.

References

- Chau, W., & McIntosh, A. R. (2005). The talairach coordinate of a point in the mni space: How to interpret it. *Neuroimage*, *25*(2), 408–416.
- Colizoli, O., Murre, J. M., & Rouw, R. (2012). Pseudo-synesthesia through reading books with colored letters. *PloS one*, *7*(6), e39799.
- Colizoli, O., Murre, J. M., Scholte, H. S., van Es, D. M., Knapen, T., & Rouw, R. (2016). Visual cortex activity predicts subjective experience after reading books with colored letters. *Neuropsychologia*, *88*, 15–27.
- Cytowic, R. E. (2002). *Synesthesia: A union of the senses*. MIT press.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.
- Feng, W., Liu, G., Zeng, K., Zeng, M., & Liu, Y. (2021). A review of methods for classification and recognition of asd using fmri data. *Journal of neuroscience methods*, 109456.
- Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange, N., Kennedy, D. N., Herbert, M. R., Bent, E. K., Koneru, V. K., Dieterich, M. E., et al.

- (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, *162*(7), 1256–1265.
- Goldstein, J. M., Seidman, L. J., Makris, N., Ahern, T., O'Brien, L. M., Caviness Jr, V. S., Kennedy, D. N., Faraone, S. V., & Tsuang, M. T. (2007). Hypothalamic abnormalities in schizophrenia: Sex effects and genetic vulnerability. *Biological psychiatry*, *61*(8), 935–945.
- Makris, N., Goldstein, J. M., Kennedy, D., Hodge, S. M., Caviness, V. S., Faraone, S. V., Tsuang, M. T., & Seidman, L. J. (2006). Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia research*, *83*(2-3), 155–171.
- Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.
- Raju, M., Gopi, V. P., & Anitha, V. (2021). Multi-class classification of alzheimer's disease using 3dcnn features and multilayer perceptron. *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 368–373.
- Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.
- Rehman, A., & Al Khalili, Y. (2019). Neuroanatomy, occipital lobe.
- Ulloa, A., Plis, S., & Calhoun, V. (2018). Improving classification rate of schizophrenia using a multimodal multi-layer perceptron model with structural and functional mr. *arXiv preprint arXiv:1804.04591*.
- Veer, A., Hafkemeijer, A., Steenbergen, H., & Bas-Hoogendam, J. M. (2017). Philips b0 unwarping in fsl feat on shark (or locally).
- Wang, Y., Wang, J., Wu, F.-X., Hayrat, R., & Liu, J. (2020). Aimafe: Autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning. *Journal of neuroscience methods*, *343*, 108840.
- Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, *14*(6), 1370–1386.
- Zhang, H., Chen, P.-H., & Ramadge, P. (2018). Transfer learning on fmri datasets. *International conference on artificial intelligence and statistics*, 595–603.

Appendix A Color Experience Questionnaire

Subject ID:

Date:

Questionnaire 'Color Experience'

We are interested in your experience at the time that you finished reading the colored book(s).

The following questions refer to your experience when seeing or thinking about black (or colorless) letters. For example, you may use the (black) letters that are part of this questionnaire to guide you in your answers.

Please indicate to what extent you agree or disagree with these statements about your experience.

1 = strongly disagree, 3 = neutral, 5 = strongly agree

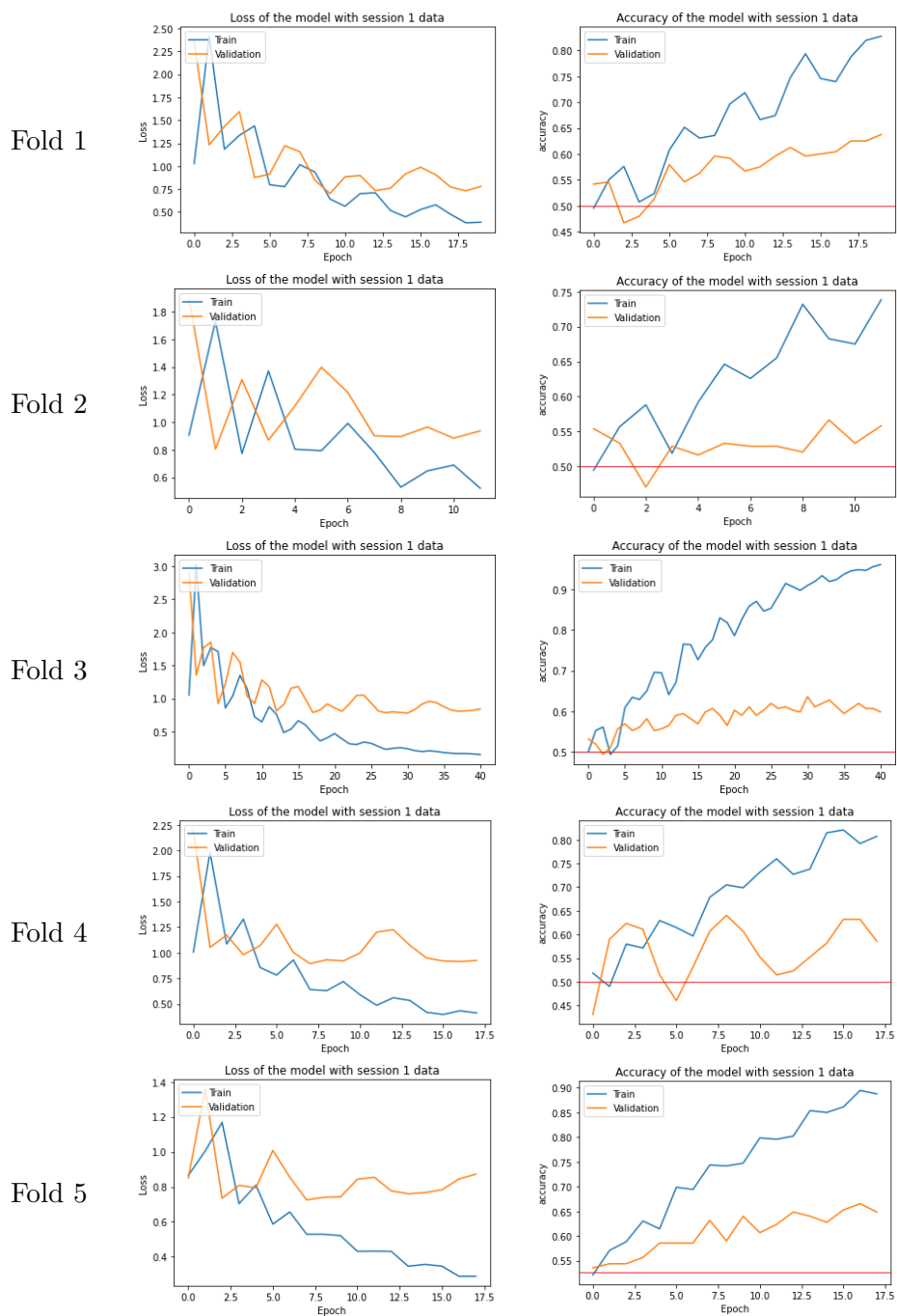
1	I 'see' the colors of certain letters in my mind's eye.	1	2	3	4	5
2	When looking at certain letters, the colors have the same shape as the letters themselves.	1	2	3	4	5
3	When I look at certain letters, the accompanying colors appear in my mind's eye.	1	2	3	4	5
4	When I look at certain letters, I see specific colors.	1	2	3	4	5
5	When I look at letters, I do not see any colors.	1	2	3	4	5
6	When I think about letters, I do not have any mental color experience.	1	2	3	4	5
7	I don't have a strong feeling about which colors 'belong' to the letters.	1	2	3	4	5
8	I am not aware that letters are associated with colors.	1	2	3	4	5
9	For certain letters, their colors have a clear location in the outside world.	1	2	3	4	5

10	I am aware that certain letters are associated with colors.	1	2	3	4	5
11	When I look at certain letters, the colors appear somewhere 'outside' of my head.	1	2	3	4	5
12	I have a strong feeling about which colors belong to certain letters.	1	2	3	4	5
13	When I look at certain letters, it seems as if the letter's color really appears on the paper/screen that the letter is printed on.	1	2	3	4	5
14	I have a mental color experience when I think about certain letters.	1	2	3	4	5
15	I don't associate colors with any of the letters.	1	2	3	4	5
16	I do not know what the colors of the letters are.	1	2	3	4	5
17	When looking at certain letters, it seems as if their colors are projected onto them.	1	2	3	4	5
18	The colors of certain letters are located 'inside' my head.	1	2	3	4	5

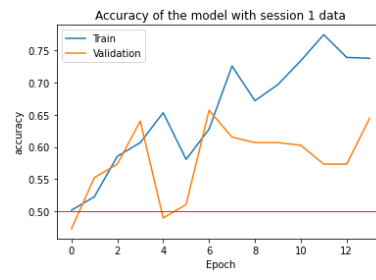
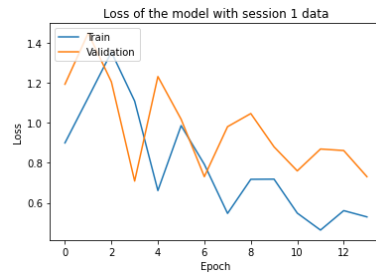
1 = strongly disagree, 3 = neutral, 5 = strongly agree

Thank you for your time!

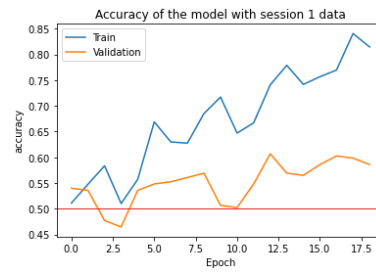
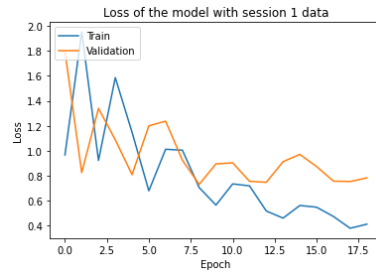
Appendix B Accuracy and loss graphs for all 10 Stratified Folds



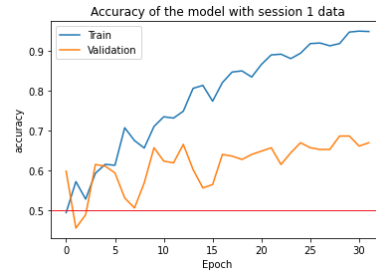
Fold 6



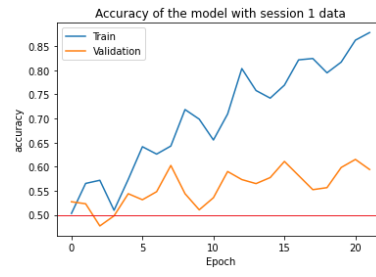
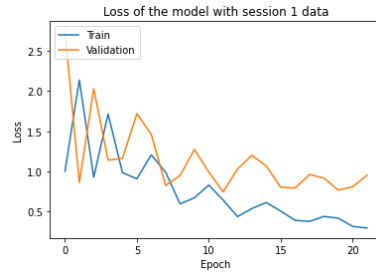
Fold 7



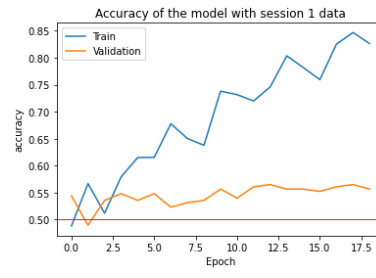
Fold 8



Fold 9



Fold 10



Appendix C Voxel Importance

Excluding voxels where the permutation importance equals zero

Voxel Permutation Importance

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 1

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 2

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 3

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 4

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 5

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 6

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 7

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 8

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 9

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

Page 10

Table with 2 columns: Voxel ID and Permutation Importance. Contains 100 rows of data.

