Hippocampal responses to event boundaries as a predictor of general memory performance: a machine learning approach

Daphne van Dijk
STUDENT NUMBER: 2039660

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. Silvy Collin
dr. Maryam Alimardani

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2020

Preface

This thesis describes the findings of the research I carried out as part of my Master's program in Data Science and Society at Tilburg University.

In recent months I have been busy mastering scientific literature in the field of cognitive neuroscience. Although I had no previous experience with this field, I quickly became fascinated with this scientific discipline. I would therefore like to thank my supervisor Silvy Collin for her help in getting me through the complex papers and her guidance in setting up this research design.

I experienced my thesis project as both challenging and inspiring. It was great to finally be able to put the acquired knowledge into practice. The project showed how much valuable knowledge and skills I have gained during my master's.

I hope you enjoy reading this thesis.

Daphne van Dijk
Tilburg, 9 December 2020

Hippocampal responses to event boundaries as a predictor of general memory performance: a machine learning approach

Daphne van Dijk

This thesis describes a comparative study of commonly used machine learning algorithms in predicting general memory performance. Previous studies have shown that event segmentation is of great importance for event-specific memory performance. However, the link between event segmentation ability, which is reflected by hippocampal activity, and general memory performance has not been addressed. Therefore, the aim of this study was to learn more about the link between event segmentation ability and memory unrelated to that specific event. To achieve this, machine learning models were applied to a dataset from the Cam-CAN project. It was explored whether there were differences between the multilayer perceptron, logistic regression, and the support vector machine in their ability to predict people's performance on a general memory test using their hippocampal time course of continuous movie-viewing. The results show that all models were able to predict memory performance based on hippocampal time courses better than one would expect based on chance alone. However, no significant differences were found between the three different algorithms in terms of classification accuracy. Furthermore, a significant difference was found between people with good versus bad memory performance in their event segmentation consistency. In short, this study has shown that a person's hippocampal time course of ongoing activity provides information about that person's general memory capacity.

1. Introduction

1.1 Context

The importance of event segmentation for memory has been discussed in many scientific studies in the field of cognitive neuroscience. In 1973, Newtson discovered that people tend to agree on where event boundaries should be placed within ongoing activities. So there is a high degree of consistency in the segmentation of events among people. Later studies have shown that consistent event segmentation contributes to better memory for that specific activity (e.g. Kurby and Zacks 2018; Sargent et al. 2013). This means that when people segment events inconsistently, they appear to have worse memory for that particular activity. Furthermore, it is known that there is a relationship between event boundaries and hippocampal activity: when people experience an event boundary, there is increased hippocampal activity (Ben-Yakov and Henson 2018).

Even though a lot of research has already been conducted into this topic, the link between people's event segmentation ability and their overall memory capacity is still unknown. Activity peaks in the hippocampus and event boundaries coincide, but its significance for memory performance on a more general level has never been explored. The purpose of this new research is therefore to extend the findings of previ-

ous studies by gaining more insight into the interaction between event segmentation ability, hippocampal activity, and the performance on a general memory test rather than an event-specific memory test. This will be accomplished by applying machine learning techniques. It will be examined whether hippocampal time courses can be used to make reliable predictions for people's performance on a general memory test, i.e. the Famous Faces task. When people experience event boundaries at timepoints that deviate strongly from the group norm, this is expected to be indicative of a lower score on the Famous Faces task.

As noted above, this new research can provide greater understanding of the relationship between general memory capacity and event segmentation, and how this link is reflected in brain activity. This can lead to useful new insights because the relationship between these three key concepts is still unknown. If the hippocampal responses to event boundaries appear to be a reliable predictor of the performance on the Famous Faces task, it suggests that a reduced event segmentation ability is not only indicative of event-specific memory, but also of memory performance in general. In addition to gaining this kind of fundamental knowledge, the outcomes of this study will also have practical implications for helping people with memory deficits.

For instance, elderly often experience problems with cognitive tasks, especially those that involve memory (Bailey et al. 2013). They may not remember certain things or have difficulty correctly recalling the event structure of recent activities. By gaining more insight into the mechanisms behind event segmentation and its link to general memory performance in a healthy population, we can also start learning more about the mechanisms that may cause memory problems. This might make it possible to develop training methods in the future that can slow down the advancement of these deficits, or that may even improve the memory capacity of people suffering from such deficits. After all, if the hypothesis is correct, improving a person's ability to segment activities could contribute to that person's memory performance.

1.2 Research question

To learn more about the mechanisms behind event segmentation and its link to general memory performance, this study aims to answer the following research question:

To what extent are hippocampal responses to event boundaries in an ongoing activity indicative of general memory performance?

Machine learning algorithms are only able to achieve a good prediction performance when there is a relationship between the input and the targets. Hence, the following sub-question will be examined in order to answer the main research question:

To what extent are there differences between the multilayer perceptron, logistic regression, and the support vector machine in their ability to predict people's performance on the Famous Faces task using their hippocampal time course of continuous movie-viewing?

As an additional test, it will be examined whether there is a significant difference between the level of event segmentation consistency between the *good*-memory group and the *bad*-memory group. This will be accomplished by running an intersubject correlation (ISC) analysis (Chen et al. 2016; Hasson et al. 2004; Nastase et al. 2019).

1.3 Findings

This research showed that all proposed classifiers could predict memory performance based on hippocampal time courses better than one would expect based on chance alone. Moreover, none of these classifiers was found to perform significantly better or worse. Lastly, a significant difference was found between the people in the *good*-memory group and the *bad*-memory group in terms of event segmentation consistency. This suggests that these two groups actually differ in their hippocampal activity and therefore also in their hippocampal responses to event boundaries. In short, this study demonstrates that hippocampal activity is at least to some extent indicative of general memory performance.

2. Related Work

2.1 Event Segmentation Theory (EST)

When people experience everyday activities, they parse this stream of activity into discrete, meaningful events. According to Kurby and Zacks (2008), an 'event' can be described as a segment of time at a certain location where you can indicate a clear beginning and end. The understanding and perception of events is supported by so-called 'event models' that make predictions about what will happen next (Kurby and Zacks 2008). When something happens that is not consistent with the prediction based on the current event model, it will be experienced as a prediction error. Subsequently, the current model gets an update based on the latest perceptual information (Kurby and Zacks 2008). The timepoints at which these updates are made are referred to as 'event boundaries' (Kurby and Zacks 2008).

When people are explicitly asked to mark event boundaries in, for example, a movie, it appears that people strongly agree on the location of these event boundaries (Newtson 1973). This suggests that there is general agreement on the evaluation of event structures in ongoing activities.

However, the fact that people can consistently chunk ongoing activities does not mean that this is a normal and spontaneous mechanism of human information processing. Hence, this phenomenon was also explored in later studies by using functional magnetic resonance imaging (fMRI). For instance, in a study conducted by Zacks et al. (2001), people were shown tapes of everyday events while at the same time their brain activity was measured using fMRI. Participants were unaware of the segmentation task during this stage. When these same people were later explicitly asked to mark event boundaries in these same videos, there appeared to be a significant correlation between transient changes in brain activity and the explicitly self-labeled event boundaries. In other words, around the timepoints where participants later identified event boundaries, there was increased local brain activity. These results suggest that event segmentation is a natural and spontaneous aspect of human information processing.

Ben-Yakov and Henson (2018) have further explored this phenomenon by specifically investigating hippocampal activity as a response to event boundaries in an ongoing activity. Their study showed that increased hippocampal activity is specific and sensitive in its response to subjective event boundaries. Moreover, this activity was larger at those boundaries for which they found high consensus between the participants (Ben-Yakov and Henson 2018). This means that strong, obvious event boundaries also trigger stronger responses in the hippocampus. In this study, they recorded activity peaks in the hippocampus and then analyzed the alignment between the event boundaries and

these activity peaks. These event boundaries were identified by an independent group of participants who did not undergo the fMRI themselves. This analysis demonstrated that increased hippocampal activity was highly correlated with the identified event boundaries (Ben-Yakov and Henson 2018).

In summary, these studies have shown that brain activity seems to be modulated by event structures and that there is general agreement on those structures among people (Ben-Yakov and Henson 2018; Newtson 1973; Zacks et al. 2001). Moreover, it has become clear that event boundaries are reflected by peaks in hippocampal activity (Ben-Yakov and Henson 2018; Zacks et al. 2001).

2.2 Event segmentation and memory

Numerous previous studies indicated that event segmentation of ongoing activity plays an important role in people's ability to remember and understand things (e.g. Baldassano et al. 2017; Kurby and Zacks 2018; Sargent et al. 2013).

Baldassano et al. (2017) discovered a relationship between event boundaries and hippocampal encoding in a movie-viewing experiment. Their research showed that the hippocampus was triggered at the end of an event, i.e. at event boundaries, to encode this new information about this event into memory. This implies that the segmentation of events contributes to how memories are organized in memory. The encoding seemed to be most powerful when the activity in the hippocampus was relatively low during an event, but considerably high at an event boundary (Baldassano et al. 2017).

Kurby and Zacks (2018) also note that the event segmentation process is used to update human working memory and regulate encoding in people's long-term memory. Worse event memory can therefore be explained by bad event segmentation ability. Event segmentation ability can be defined as the level at which someone agrees with a larger sample regarding the location of the event boundaries in ongoing activities (Sargent et al. 2013). According to Kurby and Zacks (2018), if people cannot segment events properly, these events will not be properly encoded and this in turn has a negative effect on memory recall.

In a study conducted by Sargent et al. (2013) it was assessed whether the event segmentation ability can predict subsequent memory. This study showed that the ability to segment a continuous experience can accurately predict memory related to this specific activity. In this study, people who were better at segmenting a movie were able to remember more actions from this movie afterwards. Thus, this study also proposes that poor memory about a specific activity can be explained by poor segmentation of this activity while experiencing it.

In conclusion, the ability to properly segment ongoing activity is crucial for memory, but so far only event-specific memory has been explored. It is not known yet whether a reduced event segmentation ability, which is reflected by hippocampal activity, is also indicative of people's memory performance on a more general level.

2.3 Machine learning and fMRI data

In recent years, interest in using machine learning models to analyze fMRI data has grown significantly (Arbabshirani et al. 2017). The advantage of fMRI research is that it is a non-invasive method with which it is possible to measure local brain activity very accurately (Vemuri, Jones, and Jack 2012). Not only does this make it possible to investigate differences in brain activity across different groups, but it also offers possibilities for diagnostic or prognostic tools using machine learning methods.

A classifier is a machine learning function that takes a number of features as input and then tries to predict the corresponding class for each set of features (Müller and Guido 2017). Thus, a classifier is basically a model that represents the relationship between the features and the class labels in a given dataset. When there is no relationship between the input and the target labels, an algorithm will fail to achieve good classification performance.

These type of methods have been widely used in recent years to classify neurological brain disorders such as schizophrenia, mild cognitive impairment and Alzheimer's disease (Arbabshirani et al. 2017). In most cases, the goal of these studies was to correctly classify healthy control subjects and people with a certain neurological disorder (e.g. Er et al. 2017; Mourao-Miranda et al. 2005).

When dealing with neuroimaging data, a time course of brain activity measurements can be used as input features (Pereira, Mitchell, and Botvinick 2009). In such a time course, each data point / feature represents a brain activity measurement at a specific timepoint. Many previous studies applied machine learning models to fMRI data to classify different patterns of brain activity (e.g. Challis et al. 2015; de Vos et al. 2018). In these studies, fMRI data was acquired while participants were exposed to a certain stimulus. Next, patterns of brain activity for different voxels were extracted from these data. These patterns were then used as input for a classifier that tries to predict what a participant was experiencing. This means that classification models were trained to learn the mapping between brain activity patterns and certain predefined stimulus categories. The classification performance for such methods appeared to be very good in most cases (Arbabshirani et al. 2017).

One of the most commonly used algorithms in such studies is the Support Vector Machine (SVM), but the Logistic Regression (LR) model is also used regularly (see Arbabshirani et al. 2017). Furthermore, in recent years there has been an increase in the number of studies that applied deep learning methods for neuroimaging classification (Arbabshirani et al. 2017). Deep learning is a branch of machine learning that uses algorithms inspired by the neural network architecture of the human brain (Sarraf and Tofighi 2016). The use of these neural network algorithms in the study of cognitive neuroscience is still relatively new, but various studies have already shown promising results (e.g. Güçlü and van Gerven 2017; Wen et al. 2018).

2.4 Current study

The purpose of this study is to fill the research gap described in the previous section. Thus, this study aims to find out whether reduced event segmentation ability is indicative of general memory performance, and not only event-specific memory performance. It will therefore be examined whether the event segmentation ability, which is reflected in people's hippocampal time course, can be used to predict the performance on a general memory task. This will be accomplished by applying several machine learning algorithms to the Cam-CAN dataset. It is expected that people with impaired event segmentation ability score worse on a general memory test.

Functional brain imaging data from a movie-viewing experiment will be used to obtain hippocampal responses to event boundaries. The results of the Famous Faces task will be used as the measure of general memory performance, i.e. the target labels. By using fMRI data as input in algorithms to predict general memory performance, a relationship between hippocampal activity and general memory performance can be established. After all, the performance on the Famous Faces task is entirely unrelated

to the movie-viewing experiment and an algorithm cannot achieve good classification performance when there is no relationship between the input and the target labels.

3. Methods

Various types of classification algorithms can be implemented for the task of predicting memory performance. In this study, the performance of the most commonly used algorithms in this research area will be compared with each other: logistic regression and the support vector machine. In addition, these algorithms will be compared with a basic artificial neural network: the multilayer perceptron. The background of these three different models will be briefly explained in the following section.

Furthermore, the intersubject correlation (ISC) can be computed to assess the consistency of hippocampal time courses among participants. The background and application of this analysis will also be explained in this section.

As a last point, the statistical technique used to compare the performance of the different algorithms will be discussed.

3.1 Algorithms

3.1.1 Logistic Regression. Logistic regression (LR) is a supervised classification algorithm that predicts the probability that an observation belongs to one out of two possible classes (Bonaccorso 2017). The outcome variable in logistic regression is dichotomous or binary (Hosmer Jr, Lemeshow, and Sturdivant 2013). In logistic regression, the sigmoid function is applied to the output of a linear function x to get a probability value within the range of 0 to 1 (Bonaccorso 2017). This function is defined as:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

Thus, the sigmoid of x is 1 over 1 plus the exponential of negative x. We set a certain threshold on the probability value (p), e.g. 0.5, to make the logistic regression a linear classifier. For each training data point x, the predicted class is y. This results in predicting y = 1 when $p \geq 0.5$ and y = 0 when $p < 0.5$.

3.1.2 Support Vector Machine. A support vector machine (SVM) is a supervised learning algorithm that can be used for classification tasks. The idea behind it is that there is some unknown dependency between input and target labels. Its goal is to find a hyperplane in an N-dimensional space that has the maximal margin of separation between the two classes (Daumé III 2017). This means that the distance between the data points of the different classes should be as large as possible to reduce the risk of misclassification. The support vectors are the data points that are closer to the hyperplane and therefore directly affect the position of the hyperplane (Daumé III 2017). After all, the margin is computed as the distance from the line to only the closest points.

By applying a kernel trick, the SVM is also capable of fitting nonlinear decision boundaries (Mourao-Miranda et al. 2005). Class labels in SVM are set to -1 and 1 and this results in a reinforcement range that functions as boundary margins (Bonaccorso 2017). For the hyperplane there is a weight vector w^p and an offset b such that:

$$y_i\Big((w^p)^T v_i^p + b\Big) > 0$$

Here: y_i is the class label (+1 for the class A and -1 for the class B), and v_i^p are the training examples. The hypothesis for this classification model is defined as:

$$h\left(x_i\right) = \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases}$$

3.1.3 Multilayer Perceptron. The multilayer perceptron (MLP) is an example of an artificial neural network that is composed of an input layer, one or more hidden layers and an output layer. Each layer consists of one or more neurons. The more hidden nodes, the more complex the model is. It is often used for pattern classification because it is capable of learning complex patterns that are not linearly separable (Patterson and Gibson 2017).

The input consists of a set of features that are pushed into the neurons of the input layer: $x_i|x_1, x_2, ..., x_n$. There are as many neurons in the input layer as there are features. As a result, there is a neuron with an associated weight value for each feature. The dot product of these input values with the associated weights yields a value at the hidden layer like this: $w_1x_1 + w_2x_2 + ... + w_nx_n$ followed by pushing this value along with a bias term through one of the many activation functions.

The output of this procedure is used as the input for the neurons in the next layer. In the next layer, we again take the dot product of this input with the corresponding weights of the neurons in that new layer. This outcome is also pushed through an activation function and passed on to the next layer. These steps are repeated until the output layer is reached. The output in this output layer is used to make the final classification decision. What kind of output this output layer has depends on the chosen activation function (Patterson and Gibson 2017).

The training process involves gradually fine-tuning the randomly initialized weights in order to minimize the error of the network. Backpropagation is used to make these weight changes relative to the error, and the error itself is measured using the Cross-Entropy loss function (Patterson and Gibson 2017).

3.2 Intersubject correlation (ISC)

Intersubject correlation (ISC) analysis of functional brain imaging data offers great insight into how brain activity is correlated across different participants when they are exposed to the same ongoing activity (e.g. movie-watching). This means that ISC quantifies the consistency of neural responses to these kind of naturalistic stimuli among people (Hasson et al. 2004; Nastase et al. 2019). The ISCs can be established by computing the correlation coefficients between all possible pairs of participants. This is known as the pairwise approach. This approach results in $N(N-1)/2$ correlation values for each possible pair of N participants.

According to Nastase et al. (2019), statistical inference for ISC analysis is complicated as each participant contributes to the calculation of the ISC of all other participants. For that reason, the assumption of independence is violated and standard parametric tests (e.g. T-tests) are not suitable for this type of analysis.

Chen et al. (2016) demonstrated that subject-wise permutation tests are actually well suited for comparing two groups that are expected to have different ISC values. This non-parametric permutation test makes it possible to statistically evaluate two-sample tests while disrupting the correlation structure among pairs (Chen et al. 2016).

For this test, a list needs to be created containing the group label for each indi-

vidual participant. First, the observed test statistic is determined. This is the observed difference in the median ISC value between the two groups. Subsequently, the group labels are randomly reassigned at each iteration, and the test statistic is then calculated at all possible rearrangements of the group labels (Chen et al. 2016; Nastase et al. 2019). This generates a null distribution against which the observed test statistic can be statistically evaluated. This procedure results in a reliable permutation-based *p*-value corresponding to the two-sided test.

3.3 Cochran's Q Test

Cochran's Q test is a non-parametric statistical technique to compare the performance of multiple machine learning models (Raschka 2018). It tests the null hypothesis that there is no significant difference between multiple classification models in their accuracy on the same test set.

For this method, $\{M_1, \ldots, M_L\}$ is defined as the list of classification models that were trained and tested on the same dataset. If these L models indeed perform equally well, then the Q statistic will be roughly distributed as χ^2 where the degrees of freedom equals $L - 1$.

$$Q_C = (L-1)\frac{L\sum_{i=1}^{L} G_i^2 - T^2}{LT - \sum_{j=1}^{N_{ts}} (L_j)^2}$$

To perform Cochran's Q test, several steps must be taken that were described by Raschka (2020). In this equation, G_i defines the number of items out of N_{ts} that were correctly classified by model $M_i = 1, \ldots L$. Furthermore, L_j can be defined as the number of models out of L that did correctly classify item $\mathbf{z}_j \in \mathbf{Z}_{ts}$. Here, $\mathbf{Z}_{ts} = \{\mathbf{z}_1, ...\mathbf{z}_{N_{ts}}\}$ represents the test dataset that was used to test the different models. Lastly, the total number of right number of votes among the L models is indicated by T:

$$T = \sum_{i=1}^{L} G_i = \sum_{j=1}^{N_{ts}} L_j.$$

In machine learning, model predictions are usually organized in a binary $N_{ts} \times L$ matrix. In such a matrix, a 0 indicates that the model M_j incorrectly predicted data example $\mathbf{z}_i$, and a 1 indicates a correct prediction.
The Q-value can be obtained by plugging in the right values into the first equation. This Q-value allows us to find the corresponding *p*-value, assuming a χ^2 distribution with $L - 1$ degrees of freedom. When this *p*-value is smaller than the significance level of α = 0.05, the null hypothesis (i.e. that all models perform equally well), will be rejected.

4. Experimental Setup

The following section provides detailed information about the dataset, the preprocessing steps, the experimental procedures, the evaluation metrics, and the software used in this study.

4.1 Dataset

The dataset used in this study was retrieved from the Cambridge Center for Aging and Neuroscience (University of Cambridge 2010; Shafto et al. 2014; Taylor et al. 2017). This Cam-CAN project uses cognitive, neuroimaging and epidemiological data to learn more about how elderly can best maintain cognitive capacities. Only healthy people participated in this study. The following material was used from this dataset: the imaging data (fMRI) from the movie-viewing experiment (N = 649), and the cognitive/behavioral data with the results of the Famous Faces task (N = 660). In total there were 631 participants who took part in both experiments. One observation had to be deleted because the behavioral data for this participant was incomplete and therefore useless. Thus, the final dataset consisted of 630 observations. Among the participants were 315 men and 315 women with a mean age of 54.88 (SD = 18.31). The age of the participants ranged from 18.5 to 88.92 years old.

4.2 Data description

For this current study, we worked with two types of data: Functional Magnetic Resonance Imaging data (fMRI) data and behavioral / cognitive data. Background information on these two types of data is briefly discussed in the following sections.

4.2.1 Functional Magnetic Resonance Imaging data (fMRI). The fMRI data from a movie-viewing experiment was used to obtain the hippocampal responses to event boundaries. Participants who took part in this movie-viewing experiment underwent fMRI to measure brain activity while they were watching a condensed 8-minute version of Alfred Hitchcock's *"Bang! You're Dead"*. Even though the full 25-minute video was considerably shortened, the narrative of the video was preserved in this process.

The fMRI data were collected using a 3T Siemens TIM Trio equipped with a 32-channel head coil. High-resolution T1-weighted structural images were acquired for each participant to depict structural properties of the brain and enable cross-correlational functional analyses. For the functional scan, T2*-weighted echo planar images (EPIs) were obtained using a multi-echo sequence (whole brain coverage; TR = 2470 ms).

The acquired functional and structural images were preprocessed using SPM12 (see `http://www.fil.ion.ucl.ac.uk/spm)`, as implemented in the Automatic Analysis pipeline system described in the paper by Cusack et al. (2015). In short, the obtained functional images were corrected for slice-timing differences and motion. Next, the T1 and T2 images were combined to make it possible to distinguish different tissue classes. In order to optimize the alignment among the participants, an anatomical group template was created using the DARTEL procedure and then transformed into Montreal Neurological Institute (MNI) space. Lastly, the EPI images were coregistered to the T1 image and normalized into MNI space.

In this current study, the starting point is the preprocessed data from the Cam-CAN project that comes in a MATLAB format. The preprocessing of the raw fMRI data does not fall within the scope of this thesis project. Further details about this imaging data, the MRI scanner, and the preprocessing pipeline can therefore be found in the paper written by Taylor et al. (2017).

In the dataset that is used in this study, a MATLAB-file is available for each participant who took part in the movie-viewing experiment. Each file contains 116 brain regions-of-interest (ROIs) that were measured by MRI scanner. Only the hippocampus

of the left and right hemisphere are of interest for this current study. The MRI scanner stored 193 fMRI-scans over the course of the movie. For each ROI, there is a sequence in the MATLAB-file that shows the mean activity of that brain region at each of these 193 timepoints. This means that a hippocampal time course can be obtained for each participant: a sequence consisting of 193 timepoints at which the brain activity in the hippocampus was quantified.

4.2.2 Behavioral data. The results of the Famous Faces task have been selected to serve as the measure of general memory performance. It is a good measure because the performance on this task is entirely unrelated to the movie-viewing experiment and this task is sensitive to brain disorders such as Alzheimer's (Estévez-González et al. 2004). Even if people are only in the early stages of this disease, they already score significantly worse on this test (García et al. 2020). Thus, this test is able to identify mild memory deficits and this is desirable since only healthy people participated in the experiment.

This semantic memory test measures participants' ability to recognize famous people from photos. Participants were shown photos of 30 celebrities and 10 random unknown people. Participants were asked if they recognized the person in the photo. If so, they were asked if they knew this person's name, and if they could provide occupational information about this person. Records were kept of how often participants gave a correct answer in terms of recognition, the provided name and the provided occupation.

4.3 Preprocessing

Before the data could be used for this research, some preprocessing steps had to be performed. The preprocessing steps for both types of data and the reasoning behind them are explained in detail in the following section.

4.3.1 Functional Magnetic Resonance Imaging data (fMRI). As mentioned earlier, a sequence is available for each ROI that shows the mean activity of that brain region at 193 different timepoints. For each participant in this study, only the brain activity in the hippocampus of the left and right hemisphere was of interest. Since no difference was expected between the left and right hemisphere, the average of these two sequences was used to represent people's hippocampal time course. In short, for each participant two sequences were extracted from the MATLAB-file and the average of these two was then calculated to represent hippocampal brain activity over the course of the movie.

When analyzing these types of time courses, it is important to realize that the level of the signal can differ substantially across people. For that reason, the obtained hippocampal time courses were normalized (z-scored) within subjects to enable better comparison across participants. Due to this transformation, each individual sequence has fluctuating values within the same range. This forces the algorithms to classify based on the pattern of the activity fluctuations rather than the signal level.

4.3.2 Behavioral data. The file with the results of the Famous Faces task shows multiple memory aspects that were measured in the experiment. In the paper by Shafto et al. (2014), a distinction was made between four components: (1) number of famous faces recognized, (2) number of faces for which occupational information was given, (3) number of faces whose full names were given, and (4) the number of unfamiliar faces that were correctly identified as unfamiliar.

To gain a more reliable insight into people's ability to recognize faces, the scores of components 1 and 4 were combined into a new score. This new score takes the false

recognitions into account. The term "*false recognitions*" refers to the phenomenon that people wrongly indicate that they recognize an non-famous face. In theory, it is possible to achieve a perfect score by always answering *"yes"* to the recognition questions. The overall recognition score must therefore be compensated for the situations where people answer *"yes"* when this was in fact impossible because the face shown was a non-famous face. This compensation was accomplished by subtracting the percentage of false recognitions (which can be deduced from component 4) from the percentage of correct answers on component 1.

The outcome of the procedure described above, together with items 2 and 3, form the sub-components of the final overall memory score. This means that this final overall memory score is the sum score of these three separate components. The descriptive statistics of these components can be found in table 1.

Table 1
The descriptive statistics per memory component.

component	min	max	mean	SD	median
true recognition	20	100	87.56	14.64	90
naming	4	30	23.58	5.79	25
occupation	6	30	26.86	4.20	28
final score	39	160	138.00	21.79	144.67

Numerical scores must be discretized to enable the use of classification algorithms. A threshold was therefore determined to differentiate between *"good"* and *"bad"* memory performance. Thus, this study deals with a binary classification problem. It has been decided to set this threshold at the median score to minimize the class imbalance. This same discretization technique was also applied to the scores of the three individual memory components. This allows the classification performance of the different algorithms to be compared on multiple memory aspects. The distribution of the participants over the two classes per memory aspect is shown in figure 1.

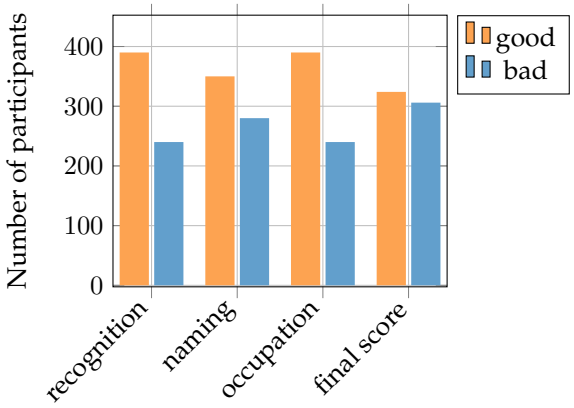

Figure 1
The distribution of the participants over the two classes ("good" and "bad") per memory component.

For a second experiment, a subset of the data was created in which only the best 10% and worst 10% of the participants were included. Which participants belonged to the best and worst 10% was determined based on people's original final memory score (numerical). Duplicates were also added to this data subset. This means that all participants who had the same score as the participant at the cutoff point, i.e. the score of the 63rd and 567th participant in the sorted list, were also added to the subset. The 63rd because this concerns the top 10% of the total of 630 participants and the 567th because this concerns the bottom 10%.

This procedure resulted in a subset consisting of 159 observations. This subset included 82 men and 77 women with a mean age of 52.06 (SD = 19.24). The range of the age values remained the same for this subset (18.5 to 88.92). The distribution of the participants in this subset over these two classes is shown in figure 2. With considerably more duplicates added on the top 10% side, class imbalance got worse for this subset compared to the original dataset. This should be taken into account when interpreting the results.

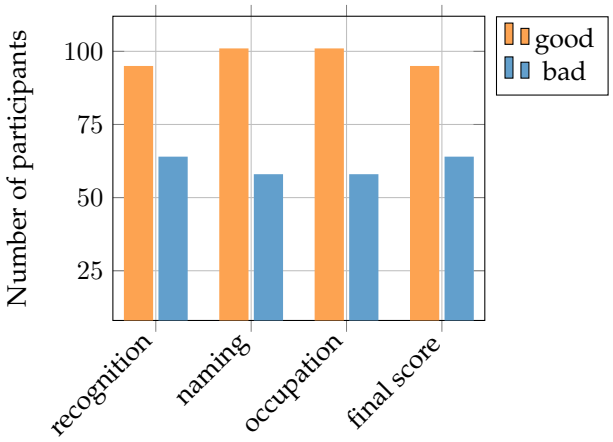

Figure 2
The distribution of the participants in this subset over the two classes "good" and "bad" per memory component.

In this subset, the mean final score of the *good*-memory group (M = 159.56) and *bad*-memory group (M = 88.38) were further apart than in the complete dataset (154.27 and 120.77 respectively). It was expected that this would lead to clearer differences in the hippocampal time courses between those two groups, which in turn would benefit the classification ability of the algorithms. Figure 3 and 4 show that the mean hippocampal time course of the two groups do indeed differ more from each other in the subset data than in the complete dataset.

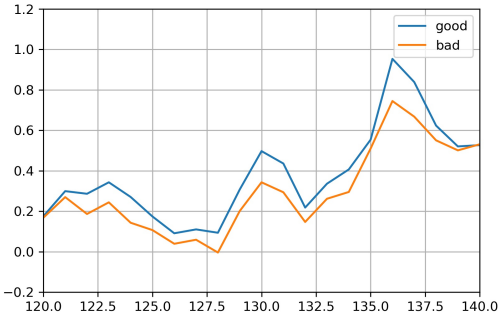

Figure 3
Part of the mean z-scored hippocampal time course in the complete dataset per memory group (timepoint 120-140)

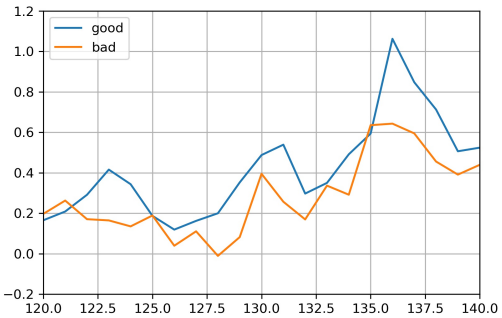

Figure 4
Part of the mean z-scored hippocampal time course in the subset of the data per memory group (timepoint 120-140)

4.4 Experimental procedure

To be able to properly answer the research question of this study, a set of experiments was conducted. The following section explains in detail which experiments were carried out, how they were carried out and why these specific methods were chosen.

4.4.1 Model comparison for for predicting memory performance. The z-scored mean array of the activity in the hippocampus of the left and right hemisphere was used as input in three different algorithms. For these three algorithms, it was examined whether they can be used to correctly predict four components of people's performance on the Famous Faces task. The following algorithms were compared: logistic regression, the support vector machine and the multilayer perceptron.

To avoid overfitting, part of the available data was hold out as a test dataset. This means that the dataset was divided into 80% training data and 20% test data in a stratified fashion.

First, all classifiers were trained on the entire 80% training dataset. The models were trained using k-fold cross-validation (k = 5) to make optimal use of the available data.

This means that for each testing fold, 4 other folds were used to train. Secondly, to find the optimal hyperparameters, a grid search was performed per algorithm during the training on the binarized final memory score. The selected hyperparameters based on these grid searches were then applied to all models for the individual memory components. As a result, the same hyperparameters were used for each memory component per algorithm. Thus, the hyperparameters only vary between algorithms and not across memory components. This makes it possible to interpret the classification performance on the various memory components more unambiguously. The hyperparameters that gave the best results can be found in Appendix A.

Lastly, the classification performance of the algorithms was evaluated on the 20% test dataset. This allowed the performance on the training set to be compared with the performance on the test set, resulting in insight into possible overfitting or underfitting. Furthermore, to test whether there were significant differences between the algorithms in their classification performance, Cochran's Q Tests were performed for each memory component using the accuracy scores achieved on the test dataset.

4.4.2 Model comparison on a subset of the data. The experimental procedure described above was repeated for the second experiment in this study. The difference between the previous experiment is that this time only a subset of the available dataset was used.

Because this subset was substantially smaller (N = 159) than the complete data set (N = 630), the distribution between training and test data was adjusted for this experiment. It was decided to train on 70% of the data and to use 30% test data for evaluation. This larger test set allows for a more reliable indication of classification performance on unseen data. This modification made it even more important to make efficient use of the available training data. Therefore, during k-fold cross-validation (k = 5) was changed to (k = 10). Thus, for each testing fold, 9 other folds were used to train.

For this separate experiment, new grid searches were performed for each algorithm to find the best hyperparameters. These were also found by training on the binarized final memory score, and then these hyperparameters were copied for the models for the individual memory components. The hyperparameters that yielded the best classification performances are listed in Appendix A.

Finally, Cochran's Q Tests were also performed for this experiment to find out if there were differences between the algorithms in their classification accuracies. These tests were performed for each memory component separately on the 30% test data. In addition, the test accuracy was compared to the training accuracy to gain insight into possible overfitting or underfitting.

4.4.3 Analysis of the inter-subject correlation of hippocampal time courses. As an additional test, it was examined whether there was a significant difference between the level of event segmentation consistency between the *good*-memory group and the *bad*-memory group. This was established by performing an intersubject correlation (ISC) analysis on the hippocampal time courses (complete dataset). The ISCs were computed using the pairwise approach. Next, a two-sample Monte Carlo approximate permutation test was performed on these ISCs using 1000 iterations. The labels belonging to the binarized final memory score were used as the group labels. In this way, differences were computed between the median ISC-score for within-group correlations while the between-group correlations were excluded. Monte Carlo resampling had to be applied because an exact test would result in an infinitely long list of possible permutations. After all, in a two-sample test the number of possible permutations equals the factorial of N (N = 630 in this case).

4.5 Evaluation criteria

Several kind of evaluation metrics were implemented to evaluate and compare the classification performance of the different machine learning models. The models were trained using accuracy, but other metrics are preferred for evaluating classification performance with unbalanced sample sizes (Arbabshirani et al. 2017). For example, when 90 percent of the observations belong to the positive class, a model can have an accuracy of 90% by simply always predicting *"positive"*. Thus, the accuracy score is not always a reliable metric. For that reason, model performance was not only quantified using accuracy, but also using precision, recall, F1-score and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). The baseline performance was set at an accuracy score of 50% since this is the score expected by chance in such a binary classification task. The different evaluation metrics will be briefly explained below, with the following definitions applicable:

An outcome is labeled as a true positive (TP) when the classification model correctly predicts the positive class, and a true negative (TN) if the negative class is correctly predicted. On the other hand, an outcome is considered a false positive (FP) when the classification model wrongly predicts the positive class, and a false negative (FN) when it wrongly predicts the negative class. In this study we treat the *"bad"* category as the postive class and the *"good"* category as the negative class. After all, the objective of the study is to be able to detect memory problems.

Accuracy: This metric computes the fraction of predictions that the model classified right. The equation is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

Precision: The precision score indicates the capability of a classifier to return only the relevant instances. For this study that means the following: how many of the observations that were labeled as belonging to the *"bad"* class, actually belong to the *"bad"* class? This is expressed in the following equation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Recall is the fraction of the total amount of relevant instances that were actually retrieved. For this study that means the following: how many of the observations belonging to the *"bad"* class were correctly classified as *"bad"*? This is expressed in this equation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score: This metric can be described as the weighted average of the precision and recall score. The most optimal score is 1, and the worst value is 0. The relative contribution of recall and precision to this F1-score are equal. The equation for the F1-score is:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

ROC AUC: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of binary classification models. It is built by plotting the true positive rate (TPR) against the false positive rate (FPR) at several thresholds. The Area Under an ROC Curve is a summary measure that can be computed from prediction scores. It describes the performance of the model across all decision thresholds and this can be interpreted as the degree of separability. This measure provides a complete picture of the classification performance. The higher the AUC-score, the better the model is at making a distinction between two classes. The at chance-level is an AUC-score of 0.5 and a perfect classification score would be an AUC-score of 1.

4.6 Software

For this study, all (pre)processing steps were executed by using Python (version 3.7.3) in Jupyter Notebooks (version 6.1.1). The MATLAB-files from the Cam-CAN dataset were converted to Python format using the SciPy library (version 1.5.2). The algorithms were implemented and defined using the scikit-learn library (version 0.23.2). Furthermore, NumPy (version 1.19.1), pandas (version 1.1.1), glob2 (version 0.7) and Matplotlib (version 3.3.1) were used for data preprocessing, data visualization, and the implementation of the algorithms. The Cochran's Q Tests were implemented using the Mlxtend library for Python (version 0.17.3). The intersubject correlation analysis was performend using the Brain Imaging Analysis Kit for Python (see `http://brainiak.org`). The implementation of this analysis kit is based on the ISC tutorial of Nastase (2019).

5. Results

In this section, the results of all experiments in this study will be presented. First, the classification performances of the different models will be given, followed by information on overfitting / underfitting and finally the results of the Cochran's Q Tests. Subsequently, these same elements will also be discussed for the analysis on the subset. As a final point, the results of the intersubject correlation analysis will be described.

5.1 Model comparison for the classification tasks

For all three algorithms, four different models have been built, each trying to predict the performance on a different memory component. The models were trained on the entire dataset. The classification performances found in this experiment are shown in table 2. The best score is shown in bold for each memory component and evaluation metric.

Table 2
The classification performance on the test dataset per algorithm for the different memory components. Test performance was evaluated based on accuracy, recall, precision, F1-score and the ROC AUC score. The label of the positive class is "bad".

		Classification report (classes = 2)				
Component	Models	Accuracy	Recall	Precision	F1-score	ROC
recognition	LR	0.57	0.46	**0.37**	**0.41**	0.57
	SVM	**0.64**	**0.73**	0.16	0.26	**0.62**
	MLP	0.61	0.60	0.12	0.20	0.61
naming	LR	0.59	0.51	**0.51**	**0.51**	0.59
	SVM	**0.60**	0.54	0.38	0.44	**0.60**
	MLP	**0.60**	**0.57**	0.23	0.32	0.57
occupation	LR	0.62	0.50	**0.46**	**0.48**	0.58
	SVM	**0.63**	**0.60**	0.12	0.21	**0.62**
	MLP	0.56	0.40	0.29	0.34	0.58
final score	LR	0.53	0.53	**0.51**	0.52	0.51
	SVM	**0.62**	**0.66**	0.49	**0.56**	**0.66**
	MLP	0.56	0.59	0.35	0.44	0.60

All models appeared to score better than the baseline performance of 50% accuracy. It is noticeable that the SVM model seemed to perform best in most cases. However, when only precision scores were considered, the SVM and MLP appeared to perform very poorly. Thus, the LR model turned out to be the best classifier in terms of precision scores. Furthermore, no major differences were observed between the different memory components.

As a second step, it was examined whether there was overfitting or underfitting in this experiment. The differences between training and test accuracy for the final memory score can be found in figure 5. This figure shows that there was no remarkable overfitting or underfitting on the training data since the test accuracy was almost exactly the same as the training accuracy. The differences between training and test accuracy for the other memory components were quite similar and can be found in Appendix B.

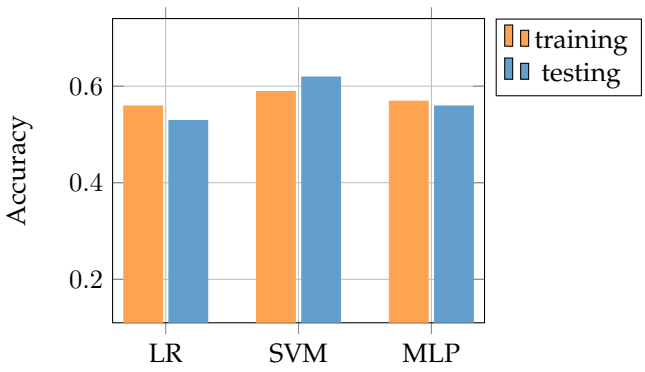

Figure 5
The difference between training and test accuracy on the final score per classifier.

The next step was to find out whether there were significant differences between the three algorithms in their classification accuracies. This was accomplished by running Cochran's Q Tests for each memory component using the accuracies on the test dataset. The results of these tests can be found in table 3. Differences were considered significant when the p-value was smaller than 0.05. Thus, no significant difference was found between the classifiers for any memory component.

Table 3
Results of Cochran's Q Tests: the Q-value and associated p-value per memory component.

component	Q-value	p-value
recognition	2.450	0.295
naming	0.195	0.907
occupation	2.577	0.276
final score	4.217	0.121

5.2 Model comparison on the data subset

The second experiment was performed on the earlier described subset of the dataset. Again, twelve classification models were trained in this experiment: three algorithms that each try to predict four different memory components. The classification performances of these twelve different models can be found in table 4. Again, the best score is shown in bold for each memory component and evaluation method.

Table 4
The classification performance for the subset data per algorithm for the different memory components. Test performance (on the 30% unseen test data) was evaluated based on accuracy, recall, precision, F1-score and the ROC AUC score. The label of the positive class is "bad".

		Classification report (classes = 2)				
Component	Models	Accuracy	Recall	Precision	F1-score	ROC
recognition	LR	0.61	0.53	0.53	0.53	0.65
	SVM	**0.63**	**0.55**	**0.58**	**0.56**	**0.71**
	MLP	0.61	0.53	0.47	0.50	0.59
naming	LR	**0.67**	**0.64**	0.47	0.55	**0.70**
	SVM	**0.67**	0.62	**0.53**	**0.57**	0.68
	MLP	0.46	0.25	0.16	0.19	0.48
occupation	LR	**0.67**	0.60	0.50	**0.55**	**0.65**
	SVM	0.63	0.53	**0.56**	0.54	0.63
	MLP	**0.67**	**0.62**	0.44	0.52	0.58
final score	LR	0.61	0.53	0.53	0.53	0.65
	SVM	**0.63**	**0.55**	**0.58**	**0.56**	**0.71**
	MLP	0.61	0.53	0.47	0.50	0.59

In general, it can be concluded that the classification performance on the subset was better than on the complete dataset. Again, all models appeared to score better than the baseline performance of 50% accuracy. The differences between the models were even smaller than in the previous experiment. It was therefore not possible to indicate a winning model unambiguously. It is noticeable that the big difference between precision scores and the other types of scores was no longer present. Furthermore, no major differences were found between the memory components.

As a next step, it was examined whether there was overfitting or underfitting in this second experiment. The differences between training and test accuracy for the final memory score can be found in figure 6. This figure shows that there was overfitting on the training data for all classifiers since the test accuracy was substantially lower than the training accuracy. The differences between training and test accuracy for the other memory components show a similar pattern and are listed in Appendix B.

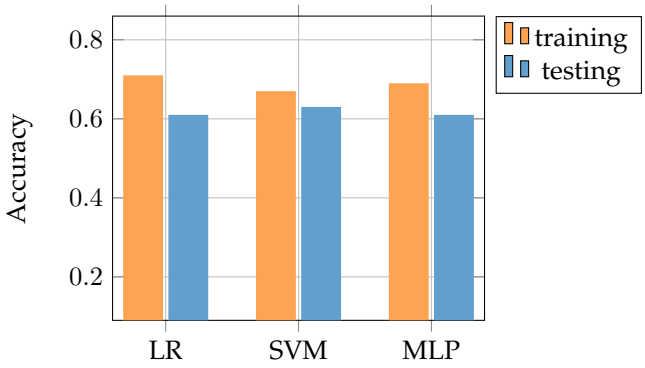

Figure 6
The difference between training and test accuracy on the final memory score for the three different classifiers on the subset data.

In order to test if there were significant differences between the three classifiers in their accuracies, Cochran's Q Tests were performed for each memory component using the accuracies achieved on the test dataset. An overview of the results of these tests can be found in table 5. Scores were considered significant if $p < 0.05$. This means that again no significant difference was found between the classifiers for any component.

Table 5
Results of Cochran's Q Tests: The Q-value and associated p-value per memory component.

component	Q-value	p-value
recognition	0.250	0.882
naming	0.400	0.819
occupation	1.000	0.607
final score	0.250	0.882

5.3 Intersubject correlation (ISC) analysis of hippocampal time courses

The intersubject correlation analysis revealed a median correlation of 0.439 across all values. A two-sample Monte Carlo approximate permutation test was conducted to compare the ISCs in the *bad*-memory group with the *good*-memory group. The actual observed group difference in terms of the median ISC values turned out to be -0.088. The non-parametric test showed that there was a significant difference (p = 0.014) in the median ISC values for the *bad*-memory group (0.393) and the *good*-memory group (0.480). Thus, the time courses correlated significantly stronger in the *good*-memory group compared to the *bad*-memory group. This also means that there is more consistency in the time courses in the *good*-memory group than in the *bad*-memory group.

6. Discussion

The purpose of this current study was to find out whether hippocampal responses to event boundaries are indicative of general memory performance. This was examined by comparing three different algorithms in their ability to predict people's performance on the Famous Faces task using hippocampal time courses of continuous movie-viewing. As an additional test, an intersubject correlation (ISC) analysis was carried out to gain insight into the group differences in the hippocampal time courses between people with good versus bad memory performance.

Several classification models were trained to predict the performance on several memory components of the Famous Faces task. As a first step, twelve models were trained on the entire dataset: three different algorithms each on four different memory components. All models scored slightly better than the baseline performance of 50% accuracy, and no major differences have been observed between the different memory components. In general, it can be concluded that the SVM model performs slightly better on all memory components in comparison to the LR model and the MLP model. However, the differences were extremely small and non-significant. Furthermore, completely different conclusions could be drawn if only the precision score and the F1-score were taken into account.

The precision scores and F1-scores for the SVM models were in fact considerably low compared to the scores on the other evaluation metrics. The MLP models also suffered from this issue, but the LR models did not. It does make sense that this problem arises when using the *bad*-memory group as the positive class rather than the *good*-memory group. After all, only healthy people took part in this experiment. This means that there were probably relatively few good representative examples in the dataset of people with poor memory capacity. It has been decided to set the threshold at the median final score to minimize class imbalance, but the disadvantage of this approach is that some people labeled with *"bad"* memory may in fact have quite normal / average memory performance. As a consequence, the hippocampal time course of these people will presumably not deviate much from many people in the *good*-memory group. This probably made it more difficult for algorithms to learn the relationship between the *"bad"* memory label and the hippocampal time courses.

This issue may also explain why the SVM and MLP model show a large difference between the precision scores and the scores on the other evaluation metrics. After all, these models are nonlinear and probably created complex decision boundaries, with many *"average"* observations ending up in the *bad*-group because those were almost identical to the observations that truly belonged to the *bad*-group. The LR model uses a less complex decision boundary since it is a linear classifier. With the LR model,

observations were classified almost at chance-level. The difficulty the model has in finding the right location for a linear decision boundary could possibly explain this. In this situation, some observations will incorrectly end up in the *good*-group and some incorrectly in the *bad*-group. While in a nonlinear model there will be a tendency to classify too many observations as *"bad"*, and this will result in poor precision scores. However, this explanation is based on intuition only and must be further investigated in follow-up research.

In the second experiment, twelve new models were trained on a subset of the data consisting of the best 10% and worst 10% of the observations. This experiment showed that the overall classification performance on this subset was slightly better than on the complete data set, especially in terms of accuracy and the ROC AUC scores. This may be explained by the greater difference between the two groups in terms of memory capacity. After all, people with an average memory performance are difficult to classify in a binary classification task. By excluding those observations, more strictly demarcated groups emerged.

In contrast to the previous experiment, there were no exceptionally big differences found between the precision scores and the scores on the other evaluation metrics. This can be explained by the more strictly demarcated groups. Thus, it can be concluded that removing the observations with an average memory performance has probably made it a bit easier for algorithms to learn the characteristics of the time courses in the *bad*-memory group. For the SVM model, it may also be explained by the fact that a linear kernel was applied in this second experiment.

In this second experiment it is not possible to unambiguously identify the best algorithm. The performance scores were again very close to each other and each component had a different winner for each evaluation metric. Only the MLP model appeared to perform relatively poorly. However, there was no significant difference in the accuracy scores between the different models for any memory component.

The ISC analysis showed that there was a significant difference in the ISCs values between the *bad*-memory group and the *good*-memory group. People with good overall memory performance appeared to be more consistent in segmenting activities than people with poor memory performance. This is reflected by the fact that the median ISC is significantly higher in the *good*-memory group compared to the *bad*-memory group.

All these results together make it plausible that the hippocampal responses to event boundaries are related to general memory capacity. First of all, algorithms were capable of classifying memory performance using hippocampal time courses better than one would expect based on chance alone. Secondly, there appeared to be a significant difference in event segmentation consistency between the two groups. Thus, what someone's hippocampal time course of ongoing activity looks like provides information about this person's overall memory capacity. However, given the research approach in this study, it can never be established with certainty that it is actually only the event boundaries that are indicative of the memory performance. The indicative factors in the time course could involve other things as well. However, since the study by Ben-Yakov and Henson (2018) indicated that activity peaks in this exact same dataset are specific to event boundaries, it is extremely likely that the event boundaries are indeed indicative factors.

Despite the significant difference in hippocampal activity between the two groups, classification performance turned out to be only slightly better than the baseline. A weak classification rate like this is not uncommon in this type of research. According to Arbabshirani et al. (2017), this can be explained by the fact that the time courses of the different groups usually overlap to a great extent. As a consequence, a significant group difference does not necessarily guarantee a strong classification performance.

Especially in this study, in which only healthy subjects took part, it is very likely that the overlap in time courses caused this rather weak classification rate. In addition, MRI data is characterized by a lot of noise. When preprocessing the data, attempts are made to remove this noise as much as possible. However, it is never possible to completely remove the noise in MRI data. For instance, when people are distracted during the scan, this will result in brain activity that is unrelated to the stimulus.

What is also striking in these results is that the MLP model only yields the best results in a very few cases. In previous studies, the use of neural networks has proven to be very effective for neuroimaging classification (e.g. Güçlü and van Gerven 2017; Wen et al. 2018). However, most deep learning studies have used more complex neural network architectures such as Convolutional Neural Networks (ConvNet / CNN) or Recurrent Neural Networks (RNN). The MLP is one of the most basic versions of a neural network. It is possible that the classification of complex fMRI data may require these type of more complex neural networks. In addition, neural networks perform especially well when a large amount of training data is available (Sun et al. 2017; Ulloa, Plis, and Calhoun 2018). It is therefore likely that the disappointing results of the MLP model are (partly) caused by the relatively small size of the dataset.

Another big disadvantage of a small dataset is the risk of overfitting. This study also showed overfitting when training the models on the subset, especially when applying the MLP model. Furthermore, previous studies have indicated that the use of small datasets results in rather poor classification performance (e.g. Kahou et al. 2016; Ulloa, Plis, and Calhoun 2018). The fact that the improvement in classification performance from the complete dataset to the subset is quite minimal, can therefore be explained by the small size of the subset. However, test performance has actually improved with this subset, so having strictly demarcated groups almost certainly contributes to better classification performance. These results suggest that if more data were available, the models could potentially yield very good classification results.

In summary, only significant group differences in ISC-values and models that score slightly above baseline, is insufficient evidence to establish a clear association between event segmentation ability and general memory performance. However, this study provides strong indications that such a relationship exists, but additional research is needed to be able to make solid statements about it. It can be expected that clearer insights could be acquired by taking the limitations described above into account in future studies.

6.1 Future research

For future studies, it would be good to make use of larger, strictly demarcated groups. By also collecting data from people suffering from memory deficits, it is easier to investigate differences in the hippocampal time courses since this is expected to result in less overlap in the time courses. This will allow stronger conclusions to be drawn about the relationship between hippocampal activity and overall memory capacity. Moreover, the amount of data in this dataset is not sufficient to achieve great success with deep learning methods. More data could improve this, and it would also reduce the risk of overfitting. Furthermore, it could also be very interesting to experiment with more complex neural network architectures.

A major problem here is that it is often difficult to acquire a large (labeled) medical imaging dataset due to the high costs (Sun et al. 2017). In follow-up research it may therefore be interesting to explore ensemble learning strategies in which models are used with low precision performance but relatively higher recall performance. If these

models are diverse enough, it most likely means that the false positives will be diverse as well. This then makes it possible to cancel those false positives out by averaging the models (Ma et al. 2020). The study conducted by Ma et al. (2020) showed that this approach can yield great results when dealing with small classes. It has also been shown in other research areas that combining different classification models, including deep learning methods, can yield very good results (e.g. Kahou et al. 2016; Sun et al. 2017).

Lastly, for future research it could also be interesting not to use the entire time course as input. For example, it is possible to choose to only investigate brain activity at the event boundaries that have a high intersubject agreement. This will make it possible to determine with more certainty that the responses to event boundaries are indicative of memory performance because only boundary-evoked responses are studied, meaning that there will be less random noise in the data.

7. Conclusion

The results of this study show that all proposed classifiers were capable of predicting memory performance based on hippocampal time courses better than one would expect based on chance alone. Accuracy scores were found to be around 60%. These are not exceptionally high scores, but this is expected to be partly caused by various limitations associated with the dataset used in this study. These results suggest that hippocampal responses to event boundaries are indeed indicative of general memory performance.

Furthermore, no significant differences were found between the performance of the three different classifiers. This means that, based on this study, there is no good reason to prefer one model over the others. As a final point, significant difference were found between people with good versus bad memory in their event segmentation consistency: people with good overall memory performance appeared to be more consistent in hippocampal activity than people with poor memory performance. This suggests that these groups actually differ in their hippocampal activity pattern and therefore in their hippocampal responses to event boundaries. These results support the conclusions drawn from the model comparison experiment.

References

Arbabshirani, Mohammad R, Sergey Plis, Jing Sui, and Vince D Calhoun. 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165.

Bailey, Heather R., Jeffrey M. Zacks, David Z. Hambrick, Rose T. Zacks, Denise Head, Christopher A. Kurby, and Jesse Q. Sargent. 2013. Medial temporal lobe volume predicts elders' everyday memory. *Psychological Science*, 24(7):1113–1122.

Baldassano, Christopher, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. 2017. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.

Ben-Yakov, Aya and Richard N. Henson. 2018. The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *The Journal of Neuroscience*, 38(47):10057–10068.

Bonaccorso, Giuseppe. 2017. *Machine Learning Algorithms*. Packt Publishing.

Challis, Edward, Peter Hurley, Laura Serra, Marco Bozzali, Seb Oliver, and Mara Cercignani. 2015. Gaussian process classification of alzheimer's disease and mild cognitive impairment from resting-state fmri. *NeuroImage*, 112:232–243.

Chen, Gang, Yong-Wook Shin, Paul A Taylor, Daniel R Glen, Richard C Reynolds, Robert B Israel, and Robert W Cox. 2016. Untangling the relatedness among correlations, part i: nonparametric approaches to inter-subject correlation analysis at the group level. *NeuroImage*, 142:248–259.

Cusack, Rhodri, Alejandro Vicente-Grabovetsky, Daniel J Mitchell, Conor J Wild, Tibor Auer, Annika C Linke, and Jonathan E Peelle. 2015. Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using matlab and xml. *Frontiers in neuroinformatics*, 8:90.

Daumé III, H. 2017. *A Course in Machine Learning*, 2 edition. self-published.

Er, Füsun, Pınar Iscen, Sevki Sahin, Nilgun Çinar, Sibel Karsidag, and Dionysis Goularas. 2017. Distinguishing age-related cognitive decline from dementias: A study based on machine learning algorithms. *Journal of Clinical Neuroscience*, 42:186 – 192.

Estévez-González, Armando, Carmen García-Sánchez, Anunciación Boltes, Pilar Otermín, Berta Pascual-Sedano, Alex Gironell, and Jaime Kulisevsky. 2004. Semantic knowledge of famous people in mild cognitive impairment and progression to alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 17(3):188–195.

García, Sara, Fernando Cuetos, Antonello Novelli, and Carmen Martínez. 2020. Famous faces naming test predicts conversion from mild cognitive impairment to alzheimer's disease. *Acta Neurologica Belgica*, page 1–7.

Güçlü, Umut and Marcel AJ van Gerven. 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336.

Hasson, Uri, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. 2004. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*, volume 398. John Wiley & Sons.

Kahou, Samira Ebrahimi, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Kurby, Christopher A. and Jeffrey M. Zacks. 2008. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2):72–79.

Kurby, Christopher A. and Jeffrey M. Zacks. 2018. Preserved neural event segmentation in healthy older adults. *Psychology and Aging*, 33(2):232–245.

Ma, Tianyu, Hang Zhang, Hanley Ong, Amar Vora, Thanh D. Nguyen, Ajay Gupta, Yi Wang, and Mert Sabuncu. 2020. Ensembling low precision models for binary biomedical image segmentation.

Mourao-Miranda, Janaina, Arun LW Bokde, Christine Born, Harald Hampel, and Martin Stetter. 2005. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional mri data. *NeuroImage*, 28(4):980–995.

Müller, Andreas C. and Sarah Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 1 edition. O'Reilly Media, Inc.

Nastase, Samuel. 2019. Intersubject correlation tutorial (github). https://github.com/snastase/isc-tutorial.

Nastase, Samuel A, Valeria Gazzola, Uri Hasson, and Christian Keysers. 2019. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, 14(6):667–685.

Newtson, Darren. 1973. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28.

Patterson, Josh and Adam Gibson. 2017. *Deep learning: A practitioner's approach*, first edition. O'Reilly Media, Sebastopol, CA.

Pereira, Francisco, Tom Mitchell, and Matthew Botvinick. 2009. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199 – S209. Mathematics in Brain Imaging.

Raschka, Sebastian. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808.

Raschka, Sebastian. 2020. Cochran's q test. http://rasbt.github.io/mlxtend/user_guide/evaluate/cochrans_q/.

Sargent, Jesse Q., Jeffrey M. Zacks, David Z. Hambrick, Rose T. Zacks, Christopher A. Kurby, Heather R. Bailey, Michelle L. Eisenberg, and Taylor M. Beck. 2013. Event segmentation ability uniquely predicts event memory. *Cognition*, 129(2):241–255.

Sarraf, Saman and Ghassem Tofighi. 2016. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*.

Shafto, Meredith A, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, and et al. 2014. The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14(1).

Sun, Wenqing, Tzu-Liang Bill Tseng, Jianying Zhang, and Wei Qian. 2017. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, 57:4–9.

Taylor, Jason R., Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-Can, and Richard N. Henson. 2017. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269.

Ulloa, Alvaro, Sergey Plis, and Vince Calhoun. 2018. Improving classification rate of schizophrenia using a multimodal multi-layer perceptron model with structural and functional mr.

University of Cambridge. 2010. Cam-can. https://www.cam-can.org/.

Vemuri, Prashanthi, David T Jones, and Clifford R Jack. 2012. Resting state functional mri in alzheimer's disease. *Alzheimer's research & therapy*, 4(1):1–9.

de Vos, Frank, Marisa Koini, Tijn M Schouten, Stephan Seiler, Jeroen van der Grond, Anita Lechner, Reinhold Schmidt, Mark de Rooij, and Serge ARB Rombouts. 2018. A comprehensive analysis of resting state fmri measures to classify individual patients with alzheimer's disease. *Neuroimage*, 167:62–72.

Wen, Dong, Zhenhao Wei, Yanhong Zhou, Guolin Li, Xu Zhang, and Wei Han. 2018. Deep learning methods to process fmri data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion. *Frontiers in neuroinformatics*, 12:23.

Zacks, Jeffrey M., Todd S. Braver, Margaret A. Sheridan, David I. Donaldson, Abraham Z. Snyder, John M. Ollinger, Randy L. Buckner, and Marcus E. Raichle. 2001. Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4(6):651–655.

Appendix A: Selected hyperparameters based on grid searches

Table 1
The selected hyperparameters per algorithm based on performed the grid searches. For the hyperparameters not mentioned, the default settings of the scikit-learn library (version 0.23.2) have been applied.

	algorithm	hyperparameters
Complete dataset	LR	{'C': 1, 'penalty': 'l1', 'solver': 'saga'}
	SVM	{'C': 1, 'cache_size': 100, 'kernel': 'rbf'}
	MLP	{'activation': 'logistic', 'hidden_layer_sizes': 1000, 'solver': 'adam', 'alpha': '0.001', 'learning_rate': 'adaptive'}
Subset	LR	{'C': 0.1, 'penalty': 'l2', 'solver': 'saga', 'class_weight': 'balanced'}
	SVM	{'C': 0.25, 'cache_size': 100, 'kernel': 'linear', 'class_weight': 'balanced'}
	MLP	{'activation': 'logistic', 'hidden_layer_sizes': 500, 'solver': 'lbfgs', 'learning_rate': 'adaptive'}

Appendix B: The difference between training and test accuracy

Table 1
The differences between train and test accuracy for the different classifiers and components (complete dataset).

model	component	training	testing
LR	true recognition	0.57	0.57
	naming	0.55	0.59
	occupation	0.57	0.62
	final score	0.56	0.53
SVM	true recognition	0.64	0.64
	naming	0.61	0.60
	occupation	0.61	0.63
	final score	0.59	0.62
MLP	true recognition	0.62	0.61
	naming	0.58	0.60
	occupation	0.62	0.56
	final score	0.57	0.56

Table 2
The differences between train and test accuracy for the different classifiers and components (subset).

model	component	training	testing
LR	true recognition	0.71	0.61
	naming	0.74	0.67
	occupation	0.70	0.67
	final score	0.71	0.61
SVM	true recognition	0.67	0.63
	naming	0.70	0.67
	occupation	0.62	0.63
	final score	0.67	0.63
MLP	true recognition	0.69	0.61
	naming	0.70	0.46
	occupation	0.66	0.67
	final score	0.69	0.61