Using machine learning as a research tool in experimental psychology

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Tanja Krumpe
aus Ludwigshafen am Rhein

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	29.11.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Volker Franz
2. Berichterstatter:	Prof. Dr. Peter Gerjets

Abstract

This dissertation evaluates how methodologies from machine learning can be applied in experimental psychology to gain new insights from neurophysiological data. Using two examples from memory psychology, the experimentally collected EEG data are evaluated once with classical group-level statistics and once with classification methods from the field of machine learning. The combination of the results of both methods shows that new insights can be gained that will profitably advance research in experimental psychology. The use of new methodologies in this area is necessary because conventional group-level statistics have problems that have spread extensively in science and have had serious consequences, especially in the replication crisis that started in the year 2000. The benefits of machine learning can help to alleviate these problems. In comparison to the use of group-level statistics alone, the combination of both methods allows data to be evaluated equally at both group and single-subject levels in order to obtain a complete picture of the data. Also, the information of individual regions can be compared and evaluated with that of an entire association of sensors collecting data. In this way, underlying patterns can also be considered. The addition of machine learning also enables explorative data analysis, which is not yet feasible in the area of group statistics.

In concrete terms, the application of machine learning techniques has made it possible to refine the characterization of executive functions and to draw up new hypotheses regarding episodic memory. Of great importance were methodologies that make the operation of machine learning processes transparent. This allowed the application to be legitimized and the results to be interpreted for a specific purpose. Furthermore, the comparison of the behavioral accuracy and the accuracy of the machine learning process was particularly valuable. In this comparison it could be shown that there is not necessarily a connection between the visual processing of an image and its active recognition. Both case studies were able to show in a representative manner which possibilities arise from the use of machine learning methods and thus present new findings which would not have been possible without the application of machine learning in this context.

Zusammenfassung

Im Rahmen dieser Dissertation wird evaluiert wie Methodiken aus dem maschinellen Lernen angewendet werden können um in der experimentellen Psychologie neue Erkenntnisse aus neurophysiologischen Daten zu gewinnen. An zwei Beispielen aus der Gedächtnispsychologie, werden die experimentell erhobenen EEG Daten jeweils einmal mit klassischen Gruppen Statistiken ausgewertet und einmal mit Klassifikationsverfahren aus dem Bereich des maschinellen Lernens. Die Kombination aus den Ergebnissen beider Verfahren zeigt, dass neue Erkenntnisse gewonnen werden können, die die Forschung in der experimentellen Psychologie gewinnbringend vorantreiben. Der Einsatz neuer Methodiken in diesem Bereich ist notwendig, da die konventionelle Gruppen-Statistik Probleme aufweist, die sich großflächige in der Wissenschaft ausgebreitet und schwerwiegende Folgen nach sich gezogen haben, die sich insbesondere in der Replikationskrise zeigen, die in den 2000ern begann. Die Vorteile, die sich durch den Einsatz von maschinellem Lernen ergeben, können dazu beitragen diese Probleme maßgeblich zu lindern. Im Vergleich zur Anwendung von Gruppen-Statistik alleine, ermöglicht die Kombination aus beiden Methoden, dass Daten sowohl auf Gruppen als auch auf Versuchspersonenebene gleichermaßen ausgewertet werden können, um ein vollständiges Bild der Daten zu erhalten. Zum anderen können die Informationen einzelner Regionen, mit denen von einem ganzen Verband von Daten sammelnden Sensoren verglichen und ausgewertet werden. So können auch Muster in Betracht gezogen werden. Auch wird mit dem Hinzufügen des maschinellen Lernens die explorative Datenanalyse ermöglicht, welche im Bereich der Gruppen-Statistik bisher nicht durchführbar war.

Konkret konnte durch die Anwendung maschineller Lernverfahren die Charakterisierung von exekutiven Funktionen verfeinert und neue Hypothesen bezüglich des episodischen Gedächtnisses aufgestellt werden. Von großer Wichtigkeit waren Methodiken, die die Arbeitsweise der maschinellen Lernverfahren transparent gestalten. Durch sie konnte die Anwendung legitimiert und eine zweckgebundene Interpretation der Ergebnisse durchgeführt werden. Weiterhin besonders wertvoll war die Gegenüberstellung von der behavioralen Genauigkeit und der Genauigkeit des maschinellen Lernverfahrens. In diesem Vergleich konnte gezeigt werden, dass zwischen der visuellen Verarbeitung eines Bildes und dessen aktiver Wiedererkennung nicht notwendigerweise ein Zusammenhang besteht. Beide Fallbeispiele konnten repräsentativ zeigen, welche Möglichkeiten sich durch den Einsatz von maschinellen Lernverfahren ergeben und dadurch neue Erkenntnisse präsentieren, die ohne die Anwendung des maschinellen Lernens in diesem Kontext nicht möglich gewesen wären.

Acknowledgements

In the following I want to thank everyone that enabled me to get through the last four years which now, finally, result in my degree. My thank you goes to Prof. Rosenstiel to give me the opportunity to write this thesis in the departement of computer engineering. Further I would like to thank Prof. Gerjets for supervising this thesis. He supported me with continous ideas and feedback that lead to extensive and constructive discussions across the boarder of disciplines. Thinking outside the box and across disciplines was indispensable for this work, which would not have been possible without him. My thank you goes also to Dr. Christian Scharinger who supported me greatly in the study designs and conductions, which was an essential part of this thesis. Most importantly, however, I want to thank Dr. Martin Spüler who was the main person throughout my academic career. He accompanied all important milestones such as my bachelor thesis, master thesis as well as my doctoral thesis for which I am very grateful.

Of course my gratitude goes to the full team of the Neural Interfaces and Brain Signal Decoding group including Martin, Sebastian, Katrin, Alex and Christian for their constructive criticism, open ears and minds in all matters, fun conference travels and a great atmosphere to work in. Further I want to extend this to the team of Perception Engineering: Wolfgang, Nora, Benedikt, Thomas, Thiago and David which made the office a better place and ensured lots of fun within and also beyond the gates of Sand.

By this means I also want to thank the LEAD graduate school as a supportive network. Being surrounded by so many other people that experience the same issues and go through the same tough phases in their PhD, independent of the field of research, is incredibly valuable. Thank you for fun retreats, visible fun agendas and the many excellent experiences I made with all of you.

Besides the people that supported me on an academic level, I also want to thank some very important people that supported me on a personal level. The greatest thanks goes to Sophia Steinmetz and her family for all the support throughout the last ten years and more. I would not be standing at this very point in my life without you and your family and I am very grateful for that. I also want to thank Malte Strop, for stepping into my life. He helped me to clear my mind and to focus during the last months of this thesis. He provided valuable input and feedback for the writing of this thesis. Thank you, your devotion, thoughtfulness and care is much appreciated!

Contents

I Introduction **1**

1 Introduction **3**

1.1 Objective and aim of this thesis . . . 4

1.2 Structure of this thesis . . . 5

2 From statistics to machine learning **7**

2.1 Inferential statistics for hypothesis testing . . . 7

2.1.1 T-test . . . 8

2.1.2 Analysis of Variance (ANOVA) . . . 9

2.1.3 Correction for multiple testing . . . 10

2.1.4 Covariance and correlation . . . 10

2.2 Machine Learning . . . 11

2.2.1 Linear Regression . . . 11

2.2.2 Support vector machines . . . 12

2.2.3 Cross-validation . . . 15

3 Brain-Computer Interfaces **17**

3.1 Assessment of brain signals . . . 17

3.1.1 Electroencephalogram (EEG) . . . 17

3.2 Feature extraction . . . 18

3.2.1 Event-related potentials (ERPs) . . . 19

3.2.2 Frequency band power . . . 19

3.3 Filtering of the data . . . 20

3.3.1 Artifact correction . . . 20

3.3.2 Re-referencing . . . 21

3.3.3 Signal-to-noise ratio . . . 21

3.4 Classification . . . 22

3.5 BCIs as a research tool . . . 22

3.5.1 Activation patterns 23

4 Problem statement 27

4.1 Missing analysis of patterns 27

4.2 Missing analysis of single-subject level 28

4.3 Missing identification of latent variables 28

4.4 Strictly explanatory data analysis 29

4.5 Summary: Issues of reproducibility 30

4.6 Proposal for an augmented methodology 30

5 Field of application: Memory 33

5.1 Multi-component models 33

5.2 The focus of this thesis: Two case studies 34

II Working memory load and executive function 37

6 Introduction into working memory and executive functions 39

6.1 Miyake and Friedman's framework of the unity and diversity of EFs 39

6.1.1 Inhibition of information and responses 40

6.1.2 Updating - information updating and monitoring 42

6.1.3 Shifting - mental set shifting 43

6.2 Idea and hypothesis 44

7 Study 1: When Updating meets Inhibition 47

7.1 Task design 47

7.1.1 Participants 49

7.1.2 Technical setup 49

7.2 Data analysis 49

7.2.1 Neurophysiological analysis 49

7.2.2 ML based classification 50

7.2.3 Neural activation patterns 52

7.3 Results 53

7.3.1 Neurophysiological data 53

7.3.2 ML based classification 59

7.4 Discussion 64

7.4.1 Neurophysiological analysis 64

7.4.2 ML based classification 64

7.4.3 Neural activation patterns . . . 65
7.5 Conclusion . . . 65

8 Study 2: When Inhibition meets Shifting 67
8.1 Task design . . . 67
8.1.1 Participants . . . 68
8.1.2 Technical setup . . . 69
8.2 Data analysis . . . 69
8.2.1 Behavioral data analysis . . . 70
8.2.2 Neurophysiological analysis . . . 70
8.2.3 ML based classification . . . 70
8.2.4 Neural activation patterns . . . 72
8.3 Results . . . 72
8.3.1 Behavioral data . . . 72
8.3.2 Neurophysiological data . . . 73
8.3.3 ML based classification . . . 79
8.3.4 Cross-class classification . . . 81
8.4 Discussion . . . 85
8.4.1 Behavioral data . . . 85
8.4.2 Neurophysiological data . . . 86
8.4.3 ML based classification . . . 86
8.4.4 Neural activation patterns . . . 87
8.5 Conclusion . . . 87

9 Study 3: When Inhibition meets Updating II 89
9.1 Task design . . . 90
9.1.1 Participants . . . 91
9.1.2 Technical setup . . . 91
9.2 Data analysis . . . 91
9.2.1 Behavioral data analysis . . . 91
9.2.2 Neurophysiological data analysis . . . 92
9.2.3 ML-based classification . . . 93
9.2.4 Neural activation patterns . . . 94
9.3 Results . . . 94
9.3.1 Behavioral data . . . 94
9.3.2 Neurophysiological analysis . . . 96
9.3.3 ML-based classification . . . 102

9.4 Discussion 108

9.4.1 Behavioral data 108

9.4.2 Neurophysiological data 109

9.4.3 ML-based classification 109

9.4.4 Neural activation patterns 110

9.5 Conclusion 110

10 Study 4: When Shifting meets Inhibition II 113

10.1 Task design 113

10.1.1 Participants 115

10.1.2 Technical setup 115

10.2 Data analysis 115

10.2.1 Behavioral data analysis 116

10.2.2 Neurophysiological analysis 116

10.2.3 ML-based classification 117

10.2.4 Neural activation patterns 118

10.3 Results 118

10.3.1 Behavioral data 118

10.3.2 Neurophysiological data 120

10.3.3 ML-based classification 125

10.3.4 Cross-class classification 127

10.4 Discussion 131

10.4.1 Behavioral data 131

10.4.2 Neurophysiological data 131

10.4.3 ML-based classification 131

10.4.4 Neural activation patterns 132

10.5 Conclusion 132

11 Discussion 133

11.1 Unity and diversity 133

11.2 Block effects 134

11.3 Practical implications 135

11.4 The advantage of the ML approach 136

III Episodic memory processes 139

12 Introduction into episodic memory 141

12.1 Memory encoding: Quality of memorization 142
12.2 Memory retrieval: Familiarity and recognition 142
12.3 Decision making: Confidence about memory 143
12.4 Idea and hypothesis 144

13 Studies I - IV: Task design and data analysis 147

13.1 Task design 147
13.2 Differences between studies 149
13.3 Participants 150
13.3.1 Study I 150
13.3.2 Study II 150
13.3.3 Study III 150
13.3.4 Study IV 150
13.4 Technical setup 150
13.5 Data analysis 151
13.5.1 Behavioral data analysis 153
13.5.2 Neurophysiological data analysis 153
13.5.3 Classification 154
13.5.4 Activation patterns 155

14 Memory encoding 157

14.1 Results 158
14.1.1 Behavioral data 158
14.1.2 Neurophysiological analysis 158
14.1.3 ML-based classification 161
14.1.4 Exploratory analysis approach 161
14.2 Discussion 164
14.2.1 Behavioral and neurophysiological data 164
14.2.2 ML-based classification 165
14.2.3 Exploratory analysis approach 165
14.3 Conclusion 166

15 Stimulus familiarity 167

15.1 Results 167
15.1.1 Behavioral data 167
15.1.2 Neurophysiological analysis 168
15.1.3 ML based classification 169

15.1.4 Exploratory analysis approach . . . 172
15.2 Discussion . . . 175
15.2.1 Behavioral data . . . 175
15.2.2 Neurophysiological data . . . 175
15.2.3 ML-based classification . . . 176
15.2.4 Exploratory analysis approach . . . 176
15.3 Conclusion . . . 177

16 Decision confidence 179
16.1 Results . . . 179
16.1.1 Behavioral data . . . 179
16.1.2 Neurophysiological analysis . . . 180
16.1.3 ML based classification . . . 180
16.1.4 Exploratory analysis approach . . . 183
16.2 Discussion . . . 187
16.2.1 Behavioral data . . . 187
16.2.2 Neurophysiological data . . . 187
16.2.3 ML based classification . . . 187
16.2.4 Exploratory analysis approach . . . 188
16.3 Conclusion . . . 188

17 Discussion 189
17.1 Latent mental processes . . . 189
17.2 Disentanglement of phases . . . 190
17.3 The ratio of known and unknown stimuli . . . 191
17.4 Practical implications . . . 191
17.4.1 Memory encoding . . . 192
17.4.2 Stimulus familiarity . . . 192
17.4.3 Decision confidence . . . 192
17.5 The advantage of using the ML approach . . . 192

IV Benefits of using machine learning in experimental psychology 195

18 Summary 197

19 Discussion 201
19.1 Patterns . . . 201

19.2 Single-subject level 202
19.3 Latent variables 204
19.4 Exploratory and explanatory data analysis 205
19.5 Reproducibility 206

20 Conclusion 209

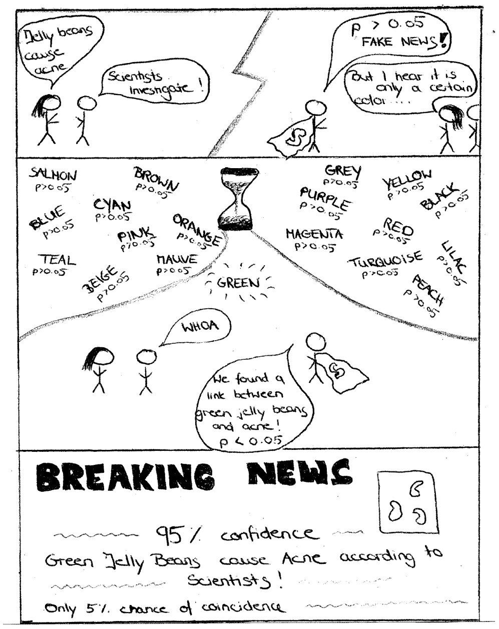
List of Abbreviations 211

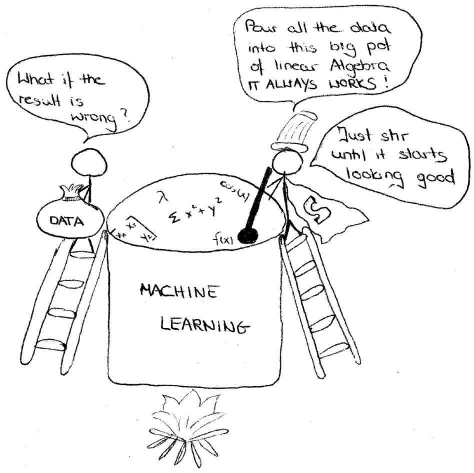
List of publications 212

Bibliography 213

Part I

Introduction





What if the result is wrong?
Pour all the data into this big pot of linear Algebra IT ALWAYS WORKS!
Just stir until it starts looking good
DATA
λ
Σ x² + y²
cos(x)
f(x)
MACHINE LEARNING

Chapter 1

Introduction

Experimental psychology is a methodology that gains scientific knowledge through the conduction and analysis of experiments. Comprehensive knowledge of statistical methods is indispensable for experimental psychology since every finding needs substantiation with tests that asses the statistical significance. However, statistical methods have many limitations that people are not aware of, making the correct use of statistics a challenge. In general, it is easy to unintentionally draw wrong conclusions from statistical tests due to the narrow window each method provides concerning the hypothesis to be tested and properties the data needs to fulfill in order to work correctly.

That the limitations and false interpretations of statistical methods are an issue that has real consequences, shows the current crisis of reproducibility in social sciences, in particular, in the field of psychology. In 2015 researchers created a project (Reproducibility Project) [1] in which they replicated 100 studies from three of the top-ranking journals. As a result, they reported that they failed to reproduce the results of the original publications in 64 % of all cases. One of the consequences of this project is damaged credibility concerning the current state of the art procedures and methodologies in the affected fields of research. Of course, the origin of this crisis is manifold, like malpractices such as p-hacking, issues of power, and mathematical issues that can emerge from flaws in the use and interpretation of statistical methods. If handled with, care group-level statistics can provide reliable results. However, the many pitfalls in standard approaches make it hard to reach reproducible and reliable results.

To this end, it needs to be mentioned that the methods in statistics continuously develop. Some of the progress might, however, be found under the term machine learning instead of statistics. The field has developed in a way that the transition between statistics and machine learning became very fluent, and a distinction between the two fields becomes more and more complex.

To find a solution that makes it easier to avoid the pitfalls of standard approaches, it might be useful to look into an advanced field of statistics, which also includes the field of machine learning. Interestingly there is already a field of research that makes use of machine learning on experimental physiological data, which, however, serves mainly a different purpose. The field of Brain-Computer Interface (BCI) research, which is a small but very interdisciplinary community, develops applications and tools that use brain activity as a control mechanism. The field evolved around patients with severe limitations of muscle control, to enable patients to restore control over body parts or even more fundamental, to regain the ability to communicate with the world. A typical BCI application records brain activity (in many cases with electroencephalography (EEG)) and categorizes the activity into control, and non-control intentions of the subject. The categorization is usually done with the help of machine learning algorithms, that can learn patterns in the brain activity which are associated with the control or non-control intention.

When comparing this application scenario with the analysis of data in experimental psychology, it is possible to draw parallels between the two. In both cases, the focus is on the differentiability of two or more conditions. In BCI research, the success of differentiating the control and non-control intention is measured in classification accuracy, describing in how many cases the correct intention was decoded. The easier it is for the algorithm to separate the data correctly, the higher the classification performance. In addition to the evaluation of the performance of the algorithm and finally of the BCI application, the accuracy can also be used to make a statement about the extent of the differences between the two categories. If, for example, the data is not distinguishable because it originates from the same condition, the performance would be similar to the chance level. If, on the other hand, the data is from two different conditions, performance values significantly above chance level can be expected. Looking now at the described conditions as two or more experimentally manipulated conditions from a psychological study for which the decision should be made whether they originate from the same distribution, the connection between the two disciplines becomes clear. Compared to group-level statistics, machine learning has several advantages that make it attractive concerning an application in experimental psychology. Within the scope of this thesis, these advantages will be worked out to show that machine learning is a suitable tool in experimental psychology to reduce and potentially overcome the current issues.

1.1 Objective and aim of this thesis

The objective of this thesis is to promote the usage of machine learning in the field of experimental psychology. The usage and interpretation of statistical methods have several flaws that need attention since many researchers are not aware of these issues. To simplify interpretation and reduce the potential sources of errors, this thesis proposes the use of machine learning in addition to conventional analysis techniques. Apart from simplification, this thesis aims to show that standard group-level statistics and machine learning techniques are two approaches that work hand in hand and complement each other meaningfully. Both techniques have a different way to approach the available data, and also provide access to different levels of information. The main objective of this thesis is to show which methodology can answer which question, according to the level of information that can be accessed. The hypothesis is that the combination of both methodologies creates a more complete and informative picture of the experimental data, which leads to higher reproducibility and new insights that cannot be gained with conventional methods only.

Overall, this work is highly interdisciplinary, covering aspects of many disciplines and topics ranging from psychology and informatics to neurobiology. Due to this nature, it is not possible to deal with all issues in depth without going beyond the scope of this thesis. Therefore, it has been decided to limit the introductory as well as the discussion chapters to the relevant information that is needed to understand the basic mechanisms of the experiments and the respective effects.

1.2 Structure of this thesis

The thesis is divided into four parts to elaborate on the use of machine learning as a tool in experimental psychology.

Part I starts with an introduction into the relevant fields of statistics and machine learning, including a problem statement that points out the limitations of state of the art group-level statistics. It also includes an introduction to BCI research because this field has an exemplary character for the development of this thesis. The chosen field of application is memory psychology, which will also be presented in its essential parts within this introductory section.

The core of this thesis will be two case studies, which will be used to demonstrate the purpose and benefit of using ML in experimental psychology. Each case study consists of several experimental studies that cover a specific process or research question. The case studies have specifically been chosen to include the core parts of human memory, from short to long-term, and also to show the diversity of the achieved benefits by using the proposed methodology.

Part II includes the first case study and focuses on working memory and associated mental processes. A particular focus will be on the characterization of executive functions (EFs), which are the core processes of working memory. The main aim will be to decode the individual EFs on a neurophysiological level by using ML. In particular, four studies will be performed in which two of the EFs will be combined each to assess their properties in a pairwise comparison.

Part III includes the second case study and focuses on processes related to episodic memory. In particular, the processes of memory encoding, memory retrieval, and decision confidence will be under investigation. Again, four studies have been performed to reveal the characteristics of the respective processes.

Part IV closes the thesis with a summarizing and concluding chapter that elaborates on the benefits that the addition of ML to conventional analysis techniques can achieve in experimental psychology. It also contains an overall discussion to merge the results and findings from the two case studies on an abstract level since the two cases could also be considered as stand-alone research questions.

Figure 1.1 shows an overview of the full structure of the thesis.

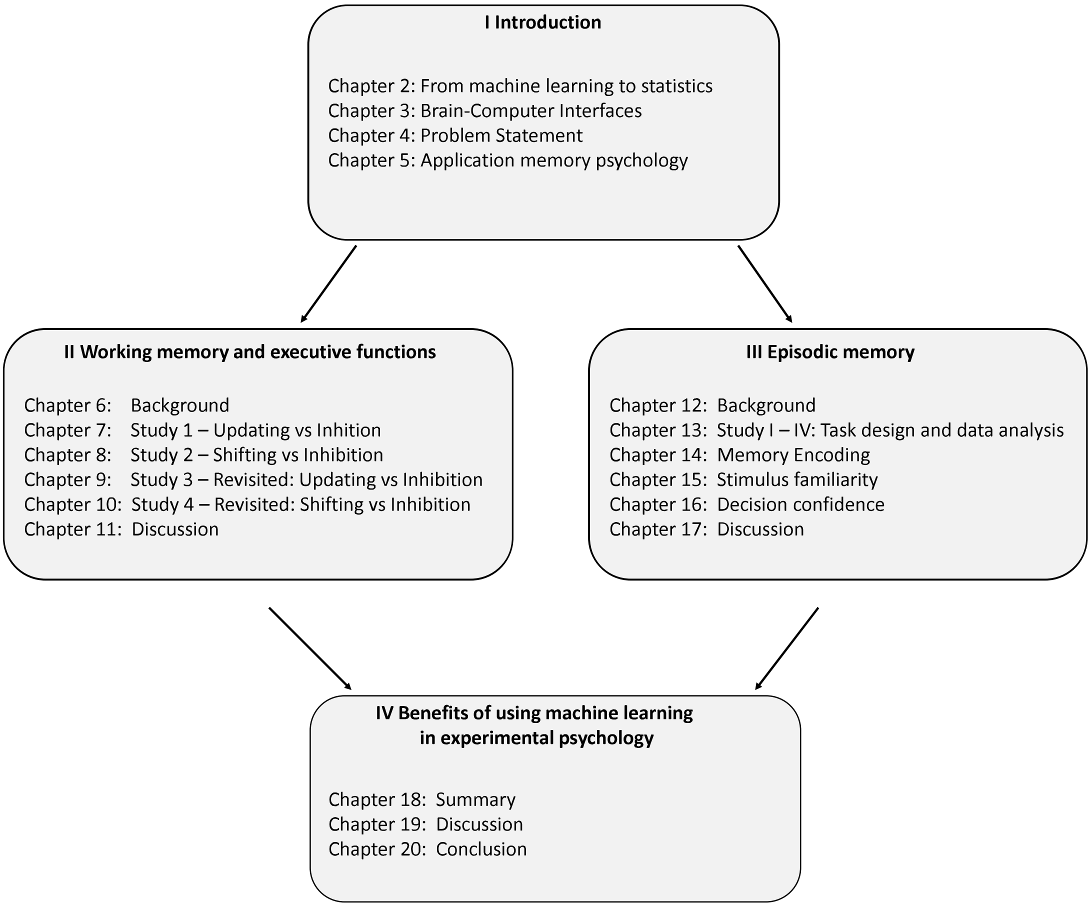

I Introduction
Chapter 2: From machine learning to statistics
Chapter 3: Brain-Computer Interfaces
Chapter 4: Problem Statement
Chapter 5: Application memory psychology
II Working memory and executive functions
Chapter 6: Background
Chapter 7: Study 1 – Updating vs Inhition
Chapter 8: Study 2 – Shifting vs Inhibition
Chapter 9: Study 3 – Revisited: Updating vs Inhibition
Chapter 10: Study 4 – Revisited: Shifting vs Inhibition
Chapter 11: Discussion
III Episodic memory
Chapter 12: Background
Chapter 13: Study I – IV: Task design and data analysis
Chapter 14: Memory Encoding
Chapter 15: Stimulus familiarity
Chapter 16: Decision confidence
Chapter 17: Discussion
IV Benefits of using machine learning in experimental psychology
Chapter 18: Summary
Chapter 19: Discussion
Chapter 20: Conclusion

Figure 1.1: Structure of the thesis - The thesis can be structured in four parts, which consist of one common introductory part, two case studies covering two different fields of memory psychology which can be found in part II and III and lastly a summarizing and concluding part IV that discusses the findings of the two cases studies in an overall way.

Chapter 2

From statistics to machine learning

Statistics as well as machine learning belong to the field of data science. Data science is a discipline that models, organizes and summarizes data to understand its underlying structure. Despite their similarities, Breiman [2] describes the two fields as two cultures, that so far seldom work together. Machine learning, which is often described as predictive modeling, is often used to automatically categorize large amounts of data. It puts a focus on algorithmic methods and model skill to work well on future and so far unseen data. It evolved from statistical methods and is today mostly located in the field of computer science.

Statistics or statistical learning is a mathematical perspective on modeling data with a focus on fitting models as good as possible, to describe a set of data as precise as possible. Overall, the two fields have a lot in common, but do pursue different aims. One way of stating these aims would be to say statistics are hypothesis-driven to find explanations for data, whereas ML is data-driven to find patterns that can be matched on new data. For the experimental psychology, both aims are equally important, but so far only the explanation part is implemented by default. In the following, the basics of both fields that are essential for this thesis will be elaborated, to understand the conceptual differences and to get an idea of the potential gains that can be achieved in combining the two methodologies. For conciseness, this chapter only includes the technical basics. A problem statement that deals with the problems of statistics will be given in Chapter 4. Thus, the focus can be better directed to the relevant points that are dealt with in this thesis, regarding the promotion of using ML in addition to classical statistics in experimental psychology.

Different in aims: Information vs Prediction

2.1 Inferential statistics for hypothesis testing

The field of statistics deals with mathematical descriptions of quantitative data. There is a subdivision into descriptive statistic that describes data with the help of its general tendency and variability and there is inferential statistic that models probabilities and random effects in the generation of data. Standard procedure in any form of statistical data analysis is to propose a hypothesis to be tested, that describes the relationship between the two or more groups of data. For the field of experimental psychology the basis of each research experiment is a theory based hypothesis about a certain cognitive state or form of human behavior. For simplification it can be stated that each hypothesis consists of the following statement: "When X changes in the experiment, then the measurable value Y changes as follows...". Y is representative of the cognitive state Z. Therefore, each theoretical hypothesis needs to be embedded into an experiment that uses the measure(s) Y that is operational for Z to evaluate the hypothesis. To assess whether the measurable change is meaningful, a statistical hypothesis is formulated. This is either corroborated or rejected based on the p-value that indicates the significance of the performed statistical tests. Since only a limited amount of data samples can be collected, it needs to be assured that the samples are representative for the respective population the samples are drawn from. Only then it can be evaluated if the two or more samples originate from the same (no meaningful difference between the samples) or a different (measurable and meaningful difference between the samples) population. A standard assumption is that the population is normally distributed, hence the collection of samples is ideally also normally distributed. Random errors and the influence of chance on the collection of samples though can cause deviations from the normal distribution.Almost every form of statistical data analysis evaluates the likelihood of observing the drawn samples simply by chance, to assess the validity of the results and the respective conclusions that can be drawn [3, 4]. To give a very quick example: When throwing a dice ten times, it is possible to observe ten sixes in a row but not very likely. It would not be correct to conclude that each throw results in a six, except, of course, the dice is biased. Therefore, further analyses are needed to understand and correctly judge the observations of dice throws.

2.1.1 T-test

One of the standard statistical test that is performed, is a T-test. A T-test can answer the question if the means of two collected samples differ. This can be an interesting information to find out if an experimental condition or treatment causes a measurable difference between two samples. As already mentioned, it needs to be assessed if the differences that can be found are meaningful and systematical or if they are due to chance. To find that out, all theoretically possible differences are constructed and sorted according to their probability of being observed in a so-called T-distribution. When looking at two populations X and Y the basic assumption of a T-test is always that the means of X and Y ($\overline{x}$ and $\overline{y}$) do not differ [5]. This assumption is also called the **nullhypothesis** H_0.

$$H_0 : \overline{x} = \overline{y} \tag{2.1}$$

To estimate if the nullhypothesis holds, based on the collected samples x and y, a T-value is calculated, as can be seen in Equation 2.2.

$$t = \frac{\overline{x} - \overline{y}}{\sigma \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{2.2}$$

n and m are the sample sizes of x and y respectively and σ represents the standard deviation of the whole population, which is not known, but can be estimated with the calculated variance of the populations σ_x^2 and σ_y^2.

$$\sigma = \sqrt{\frac{(n-1)\sigma_x^2 + (m-1)\sigma_y^2}{n+m-2}} \tag{2.3}$$

The null hypothesis is rejected if the T-value follows the conditions below

$$|t| > t(1 - \frac{1}{2}\alpha, n + m - 2) \tag{2.4}$$

α represents the level of significance and can be chosen freely. $\alpha = 0.05$ has been used as a standard for years now, but it needs to be stated that this value is arbitrary and can be chosen differently if suitable.

In general, the level of significance expresses a probability value (**p-Value**). It indicates the probability with which the calculated difference of the two samples, or an even more extreme difference, can be observed when the H_0 holds. A value of 0.05 can be translated in a probability of 5 % that this result occurs simply by chance. The p-value is not a measure of how right the assumption is or how important the difference is, but only a measure of how likely it is that the nullhypothesis is falsely rejected.

2.1.2 Analysis of Variance (ANOVA)

If more than two samples are to be tested a different test is required. The analysis of variance (ANOVA) is conceptually similar to a multiple two-sample T-tests, and can therefore be used for more samples. It is more conservative than multiple individual tests and therefore more accurate, because estimation errors do not accumulate over several steps. Again the nullhypothesis H_0 assumes that all group means of the data samples are equal, stating that there is no systematical difference between the investigated populations [6].

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \tag{2.5}$$

A rejected nullhypothesis implies that at least one, but possibly all populations differ in their mean. Once again a distribution is established, that describes the probability with which the individual groups can theoretically vary from each other, a so-called F-distribution. To estimate the probability and therefore the location in the distribution, the variance within SS_W and between SS_B the groups $1, .., k$ is calculated.

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{i,j} - \overline{x})^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_{i,j} - \overline{x_i})^2}_{SS_W} + \underbrace{m\sum_{i=1}^{k}(\overline{x_i} - \overline{x})^2}_{SS_B} \tag{2.6}$$

$$F = \frac{\frac{SS_B}{(k-1)}}{\frac{SS_W}{k(m-1)}} \tag{2.7}$$

k is the number of groups and m the number of samples within the groups. Again α represents the level of significance which is usually determined to be at 0.05. The nullhypothesis is rejected when the following condition holds

$$F < F(k - 1, k(m - 1), \alpha) \tag{2.8}$$

Also in this case, a significance level of $\alpha = 0.05$ is selected as a default value, stating that the chances of achieving this or more extreme results by random are less or equal to 5%.

2.1.3 Correction for multiple testing

As already mentioned, it is possible that differences between two or more data samples are purely coincidental and due to random variations. The bigger the parameter space, the higher the chances of a randomly sampled difference (also known as the look-elsewhere effect). Therefore, it is required to determine the parameter space of interest according to the posed hypothesis, to avoid chance level effects. In some cases, however, testing the hypothesis requires several statistical tests to be performed. To avoid wrong conclusions, a correction of the chance level must be made in such cases.

Bonferroni correction

One of these corrections was proposed by Bonferroni [7] and suggests to simply correct the significance level according to the number of made comparisons m. The criterion according to which the null hypothesis is to be rejected is therefore:

$$p < \frac{\alpha}{m} \tag{2.9}$$

This is known to be a rather conservative method, but overall an accepted measure to avoid misguided findings.

2.1.4 Covariance and correlation

The covariance and the correlation are measures that are used to quantify the relationship between variables of two samples. If two variables co-vary systematically, there is clearly a relation between the two. The covariance can be positive or negative, describing the fact that both variables either vary simultaneously in the same direction (positive) or if they vary in opposite directions (negative). Equation 2.10 describes how the covariance of two samples x and y can be calculated. n represents the size of the samples, whereas $\overline{x}$ and $\overline{y}$ represent the means of the samples, with respect to the investigated variable.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{n-1} \tag{2.10}$$

The covariance can be interpreted linearly. High values (positive as well as negative) describe strong relationships, whereas small values quantify weak relationships. To this end it needs to be mentioned, that the covariance is an unstandardized measure which is hard to compare between different experiments. However, there is a standardized form which is called correlation. It relates the covariance to the maximally achievable covariance, which scales all values to a range between -1 and 1. Equation 2.11 shows how the correlation can be calculated between the two samples x and y. σ represents the standard variation of the sample x or y as indicated by the subscripted coefficient.

$$r_{xy} = \frac{cov_{emp}}{cov_{max}} = \frac{cov(x,y)}{\hat{\sigma_x} \cdot \hat{\sigma_y}} \tag{2.11}$$

The **coefficient of determination** is the squared correlation r^2 which can further be interpreted as the proportion of the variance in the response variable that is predictable from the explanatory variable(s). In other words, r^2 is representative for the percentage of variance that can be explained by the explanatory variables. This can be applied to estimate the variance caused by the experimental condition but also as a qualitative measure to find out how well a statistical or machine learning model fits the real data. Values range from zero to one and the higher the value the higher the explainable variance or the better the fit of the model.

2.2 Machine Learning

Machine learning is a field in which algorithms can learn from and make predictions on data without explicitly being told what to learn. There are different forms of algorithms that differ according to the way the knowledge is gained from the data. This can for example be supervised (data is already provided in categories that need to be learned), unsupervised (no additional information is given), or reinforcement learning (good outputs are rewarded to reinforce the model) algorithms. For the application of machine learning, the same statement holds, that has already been formulated for statistical methods: Only a limited amount (samples) of data can be collected, which is why assumptions must be made about the totality of the population to be investigated. Ideally the sample is representative for the population, but chances are that this is not the case. However, machine learning is not concerned with calculating probabilities, but with finding an abstraction level that comprehensively describes the data and at the same time applies to new data as unrestrictedly as possible.

2.2.1 Linear Regression

Linear regression is a method that belongs in both sections, statistics as well as machine learning. It models the relationship of the data by finding a linear function that captures all data points with a minimal error. It can both, predict concrete values for the response variable, but also quantify the relationship between response and explanatory variables. To calculate a linear regression, a standard linear equation, as can be seen in Equation 2.12, is set up in which the gradient a is modeled by the covariance of the two samples x and y and the standard deviation σ^2 of the explanatory variable x (2.13).

$$b = \overline{y} - a \cdot \overline{x} \tag{2.12}$$

$$a = \frac{cov(x, y)}{\sigma_x^2} \tag{2.13}$$

Linear regression is restricted to one predictor or response variable with a linear relation to the explanatory variable. There exist other forms, such as the multiple regression to model more than one predictor or regression that can model non-linear relationships.

Regularization

If it is considered, once again, that a data sample does not necessarily have to be representative for a population, then it becomes intuitively clear why a simple linear equation is often not sufficient to describe a population. The process of creating a model that fits exactly to the available data and loses sight of the generalizability to new data is also known as overfitting. In the field of machine learning, where predictions are the most important, people are aware of the problem of overfitting, which is why various measures have been developed to prevent it.

One of these measures is regularization. For linear regression there are two popular forms of regularization, that modify the objective of the linear function to simplify the model. A simplification of the model is helpful in so far as it disregards special subtleties of the sample and thus reduces the danger of overfitting.

Ridge regression (L2 Norm) is one form of regularization that aims to find a minimal distance of the regression function to the data points by minimizing the sum of errors and the weight vector within the function.

$$min \sum_{i=1}^{k} (y_i - (Ax_i + b))^2 + \lambda ||Ai||^2 \tag{2.14}$$

Lasso regression (L1-Norm) is another form of regularization that, in addition to minimizing the sum of errors, aims to reduce the number of data points that influence the regression function by setting as many weights as possible to zero. By this the model is simplified which helps to view the data on a more abstract level and therefore, to reduce overfitting.

$$min \sum_{i=1}^{k} (y_i - (Ax_i + b))^2 + \lambda ||Ai||^1 \tag{2.15}$$

The only difference between Lasso and Ridge regression is that the regularization term is an absolute value in Lasso.

2.2.2 Support vector machines

Support vector machines (SVMs) have been developed in 1974 by Vapnik *et al.* [8] as a mathematical procedure for pattern recognition. They belong to the group of supervised machine learning algorithms. The data on which this procedure is applied to needs to be represented in a vector space. The aim is to find a hyperplane in this vector space that separates the data into two classes according to the underlying, but not a priori defined pattern. The hyperplane can be described with the following linear equation

$$w = b/x_i \tag{2.16}$$

if the data is available in the following format:

$$(x_1, y_1) \dots (x_n, y_n), x \in \mathbb{R}^m, y \in -1, +1 \tag{2.17}$$

x are the data points in the vector space, whereas y represents the class label of the respective data points. The position of the hyperplane is chosen in a way that the distance between the closest data point and the hyperplane, the so called margin, is maximal to guarantee a clear and optimal separation of the data even if new data points enter the scene. The hyperplane is defined by a normal vector w and a bias b as described in Equation 2.16. The separation of the data can be described mathematically as follows:

$$(w \cdot x_i) - b \geq 1 \ \ \texttt{for} \ \ y_i = 1 \tag{2.18}$$

$$(w \cdot x_i) - b \leq -1 \ \ \texttt{for} \ \ y_i = -1 \tag{2.19}$$

The two Equations 2.18 and 2.19 describe that the data points belonging to the same class will be positioned on the same side of the hyperplane. Data points from the other class will be on the opposite site. The margin has the size of $\frac{2}{||w||}$ which is why minimizing $\frac{1}{2}||w||$ will lead to a maximal margin. To ensure that all data points of one class will end up on the same side of the hyperplane the following constraint needs to be fulfilled:

$$y_i(\langle w, x_i \rangle + b) \geq 1 \tag{2.20}$$

The optimization problem that needs to be solved in order to position the hyperplane with a maximal margin can be described as follows:

$$\min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1] \right\} \tag{2.21}$$

The constraints of the optimization problem are shown in Equation 2.22 and 2.23.

$$\sum_{i-1}^{m} \alpha_i y_i = 0 \tag{2.22}$$

$$w = \sum_{i-1}^{m} \alpha_i y_i x_i \tag{2.23}$$

Plugging Equation 2.22 and 2.23 into the optimization problem displayed in Equation 2.21 will lead to the rephrased optimization problem:

$$\max W(\alpha) = \sum_{i-1}^{m} \alpha_i - \frac{1}{2} \sum_{j,i-1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{2.24}$$

When applying this approach to new and unknown data, the distance of every data point x_i to the hyperplane will be calculated. Due to the sign of the computed distance (positive or negative) a decision can be made on which side of the hyperplane the data point needs to be located. The word machine in the name of the support vector machine is due to the fact that the approach belongs to the class of machine learning algorithms. The addition of support vectors in the name has been chosen as the position of the hyperplane is defined by only a small number of vectors that are closest to the hyperplane. They are used to define and maximize the margin, all other data points are not needed to for this representation. Figure 2.1 visualizes the separation of a set of data points due to their shape with an optimal positioned hyperplane.

Kernel functions

If no linear separation of the data is possible as it was in Fig 2.1, a transformation into a higher dimensional space can be considered in which a separation might become possible. Figure 2.2 provides a simple example for the transformation of data points into a higher dimensional space, enabling a linear separation of the data via the application of a plain quadratic function. The transformation of the data can be computationally very expensive, which is why an application of this might be unfeasible in some cases. The so called kernel trick [9] can be applied, as it manages to solve the problem without actually computing

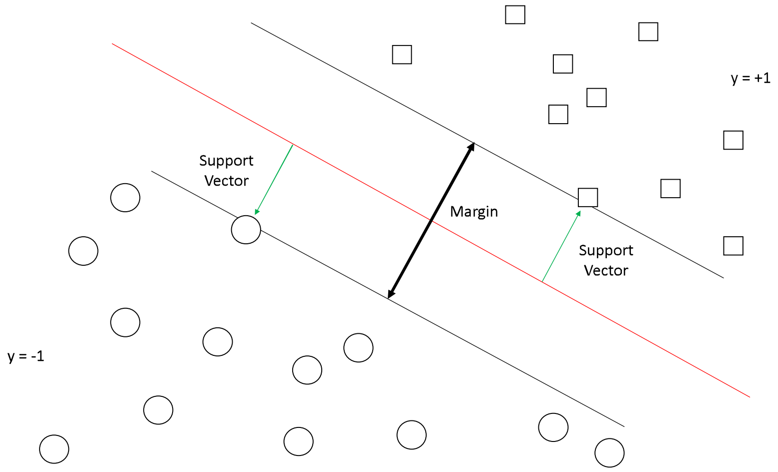


Figure 2.1: Support vector machines (SVMs) - The datapoints x_i of two classes are separated by a line in this 2D example. The two classes are represented by circles and rectangles, which are assigned the labels y = -1 and y = 1 respectively. In a 3D or more dimensional example the line will be represented by a hyperplane. The hyperplane is positioned in space in a way that the two classes (circles and rectangles) are separated and the distance (the margin) between the hyperplane and the closest samples of each class is maximal. The vectors closest to the hyperplane therefore define the position of the hyperplane and are called support vectors. In this example a perfect separation of the two classes is possible.

the transformation. The key insight that is needed for the application is that Equation 2.24 only accesses the training data in terms of their inner products $\langle\varphi(x), \varphi(x)\rangle$. Kernels functions are inner products that can be chosen to match the problem at hand. Therefore, kernel functions are plugged in the equation instead of the 'simple' inner product by which the inner products between the images of all pairs of data are computed in the feature space, without doing the actual transformation.

$$K(x, y) = \langle\varphi(x), \varphi(x)\rangle \tag{2.25}$$

Several types of kernels exist, even for data represented as graphs or strings. In the following, two kernel functions will be named to give an impression of what kind of transformations are possible.
RBF kernel

$$K(x, x') = exp(-\frac{||x - x'||^2}{2\sigma^2}) \tag{2.26}$$

Linear kernel

$$K(x, x') = (x^T x' + c) \tag{2.27}$$

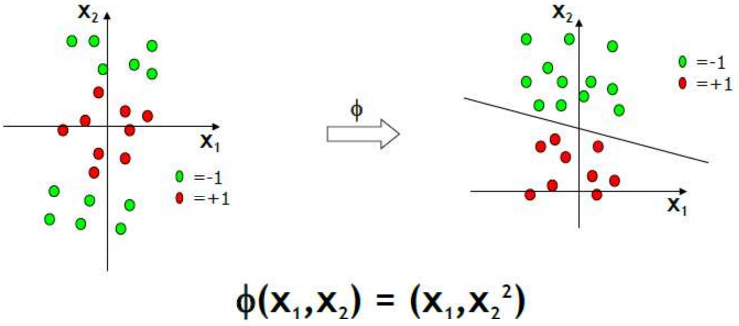


Figure 2.2: Example of a kernel function - Example of a transformation of not linearly separable data into a higher dimensional space. The two classes are displayed with a red and green color. A simple quadratic function makes it possible to find a linear separation of the two classes.

2.2.3 Cross-validation

One possibility to validate the model of the trained classifier of an SVM or any other machine learning algorithm, which can be seen as a prove of its generalizability to new data, is to perform a cross-validation. It enables to test the data model on unseen data, without the need for new data. Typically a k-fold cross-validation is applied which denotes dividing the available data into k subsets. $k-1$ of the subsets are used to train the classifier and the last remaining subset is used to test and evaluate the classifier. This is repeated k times to ensure that every sample was used for both training and testing purposes. The average performance of all k runs is calculated and stated as the overall performance. The most commonly used is the 10-fold cross-validation. Figure 2.3 an example of how a cross validation is performed.

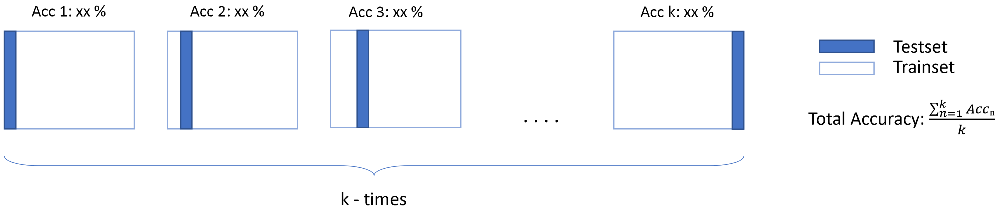


Figure 2.3: Cross-validation - The data is cut into k parts, according to the k specified in k-fold. In k iterations, each time the k-th part is used for testing, all other remaining parts for training the classifier. By this, each part is used for testing once and the full data set is evaluated within the cross-validation. To get an overall measure for the dataset, the average over all k-folds is reported.

Chapter 3

Brain-Computer Interfaces (BCIs)

To understand why the field of BCI research might be a useful example to improve data analysis in experimental psychology, this chapter will cover the fundamentals of the technology. Historically, BCIs are an assisting technology that have been developed for people with severe motoric impairments and disabilities caused by stroke or diseases like amyotrophic lateral sclerosis (ALS) [10],[11]. Their main task is to translate human brain activity, which equates with the users' intent, into machine commands that are executed by a computer. By this, patients get the chance to restore control over body parts or even more fundamental, get the chance to regain the ability to communicate with the world. To this end, research and industry continuously work on the development of possible BCI applications that could be used in an everyday context by both the healthy and the impaired.

Using machine learning on experimental physiological data. An example

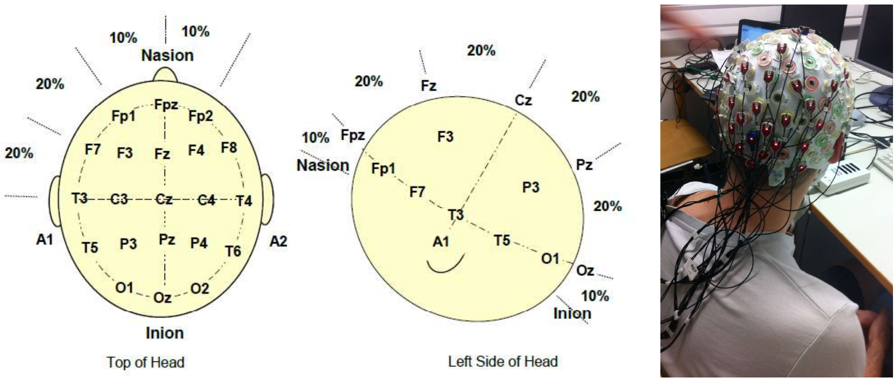
3.1 Assessment of brain signals

As a start a way needs to be found to measure brain activity that can be used to control an application. There exist several possibilities to asses brain signals. However, not all of them are of practical use for a BCI. In general, the techniques can be divided into two categories: invasive and non-invasive. Invasive methods imply that sensors are implanted directly on the brain to derive the brain signals, which is usually only an option for patients in which the benefits outweigh the risks of the implantation. For healthy users, only non-invasive techniques are reasonable. Specifically of interest are near-infrared spectroscopy (NIRS) and electroencephalography (EEG) since they are transportable and comparably cheap, contrasted to magnet-encephalography (MEG), which needs helium cooling or magnet resonance imaging (MRI) that requires a shielded room, due to a high magnetic field. In this thesis, electroencephalography (EEG) was the method of choice. Therefore, the detailed introduction will be limited to this technique.

3.1.1 Electroencephalogram (EEG)

Hans Berger has recorded the first human EEG in 1924 [12]. It measures brain activity via differences in voltage on the surface of the scalp by placing electrodes on the surface of a subjects' head. The placement of the electrodes follows a pattern that was designed

to standardize the setup. Most commonly used is the 10-20 system [13] that refers to a pattern that distributes the electrodes relative to the head size, all over the head. Herbert Jasper invented the system in 1958, and the numbers 10 and 20 refer to a percentage of a distance reaching from nasion (approximately between the eyebrows) to inion (a tangible point where skull ends) and from ear to ear. The two distances describe a horizontal and a vertical axis on the head on which the electrodes will be positioned. This system has been introduced to be able to produce comparable and reproducible results. In Figure 3.1 a) an exemplary placement of electrodes according to the 10-20 system can be seen.
A typical EEG recording consists of a time series of voltage values for each electrode. Commonly used sampling frequencies of the signal are 256 or 512 Hz (resulting in 256 or 512 values per second). Hence, the EEG has a very high temporal resolution in the range of milliseconds, which is a desirable property for the research of signal processing in the brain. Unfortunately, it has a low spatial resolution. One reason for this is that the measured signal represents a sum activity of many brain cells. The signal is passively conducted through water and the top of the skull, making it challenging to locate the exact origin of the signal. On the other hand, the skullcap in particular, but also other layers located between the electrode and the brain, have an extremely dampening effect. Only signals that originate from an outer cortex of the brain can be measured, or those that are so strong that they can still be detected at a great distance despite the signal loss. Despite this deficit, EEG recordings have clinical use for diagnostics of epilepsy, brain death, or various kinds of sleeping disorders. It has no side effects and can be used without restriction on any subject or patient. Therefore, it became one of the most used techniques for the assessment of brain activity in BCIs.



(a) 10-20 placement (b) EEG setup

Figure 3.1: EEG setup and placement - a) Schematic overview of the electrode placement according to 10-20 system [13]. The numbers 10 and 20 refer to a percentage distance in relation to the size of the head. This ensures a standardized placement of the electrodes independent of the concrete head size. b) EEG cap placed on subject. The electrodes are placed in a cap and are connected to the amplifier via cable.

3.2 Feature extraction

As a second step, in the development and application of a BCI, it is necessary to identify properties of interest within the recorded brain signal. Intentions and thoughts which are used to control the application are usually limited to short and specific time frames. Relevant properties of the signal regarding time or sensor space, are usually called features. A certain set of features can qualify as a BCI control signal if it is characteristic and can reliably be produced by a subject. In this thesis, two categories of features have been used that will be introduced in the following.

3.2.1 Event-related potentials (ERPs)

One common class of features are event-related potentials (ERPs). They represent the activity that is caused by a certain event, in the form of the time series of recorded voltage values. Event-related potentials can either be elicited by sensory perception or cognitive processes. The standard approach to extract an ERP is, therefore, to look at the onset of the event and follow the signal for a predefined window of time. To get a clean and meaningful form of an ERP, it is reasonable and advisable to look at more than one event of the same type since EEG recordings vary a lot and are prone to record artifacts. Averaging over all events reduces the variation of the signal, especially the variation that is not due to the event itself. To give one specific example, the P300 will be introduced as the most commonly used ERP in BCI research. The P300 is an ERP that is elicited due to evaluation and categorization of a presented stimulus [14]. When time-locked to the event it appears strongest at electrodes over the parietal cortex, with a latency of 250-500 ms and an intensity of 10 - 20 μV [15]. Typically, the peak of the potential is centered around 300 ms, and since it is a positive deflection, the name P300 can be explained. It does not matter if the stimulus is visual or auditory, in both cases P300s can be elicited. A general observation is, the less frequent the stimulus, the higher the amplitude of the P300. A typical task for the examination of the P300 signal is the oddball paradigm [16]. Two different stimuli are presented in a sequence of stimuli. One of the stimuli is less frequent than the other and therefore called the oddball stimulus. The waveform of a P300 can be seen in Figure 3.2 among other typical and well known ERPs such as the error-related negativity (ERN) which associated with the monitoring of errors [17], the N400, a component that can be observed as a response to words or other meaningful stimuli and [18] the P600 which is a language relevant component that can be associated with syntactical processing of sentences [19].

3.2.2 Frequency band power

A second common class of features is the frequency band power. They represent the activity that is caused by oscillations and rhythmic activity of groups of neurons. To visualize and identify this kind of activity, it is easier to look at the recorded EEG signal in the frequency domain instead of the time domain. Synchronized activity of many groups of neurons is specified as event-related synchronization (ERS), whereas a desynchronized activity is specified as event-related desynchronization (ERD). This transformation from time to frequency domain can be performed with methods like the Fast Fourier Transformation (FFT) or Burgs maximum entropy method [20]. Regarding nomenclature, it has been established to group ranges of frequencies into so-called bands. For brain activity, a grouping into five bands has been proposed, which are enumerated in the following: Delta 1-4 Hz slow wave sleep, Theta 5-7 Hz drowsiness or arousal, Alpha 8-12 Hz relaxation to attenuates in mental exertion, Beta 13-31 Hz attenuated during active movement, and Gamma $>$ 32 Hz. The presence or absence of waves in individual frequency bands can

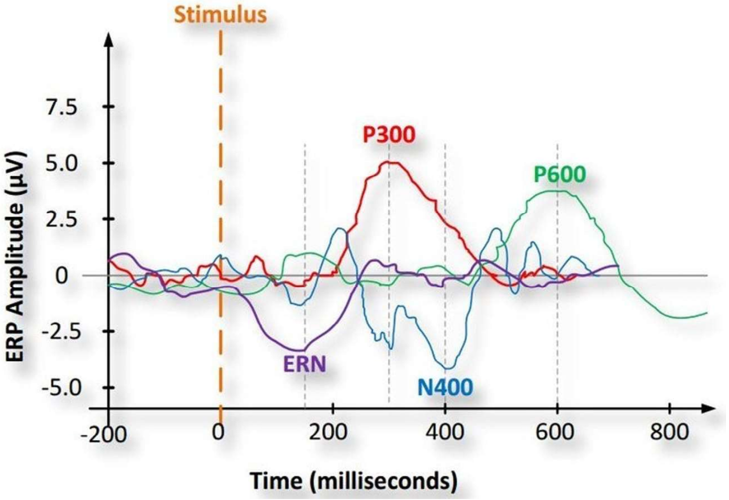


Figure 3.2: ERPs - Example of waveforms of commonly known ERPs, such as the ERN, P300, N400 and P600. They are all stimulus-locked which means that they can only be observed and identified in a specific time frame after the stimulus has occurred [1].

be associated with cognitive states and pathological disorders. Alpha waves, for example, can be found as a steady rhythm during relaxation which will get attenuated while doing mental exercises, especially over the occipital lobe [21]. Beta waves, as another example can usually be found over the motor cortex and can be associated with active muscle contraction and movement [22].

3.3 Filtering of the data

In addition, to extract relevant time frames from the signal, it is also important to ensure a certain level of quality of the signal. Signals that can be measured from the EEG electrodes via the amplifiers are in the range of μV and, unfortunately, prone to pick up noise from many different kinds of sources. To ensure that only signals related to brain activity remain, components related to something else need to be eliminated. In the following, filter methods for three kinds of errors or issues will be described that improve the data quality.

3.3.1 Artifact correction

A common source of artifacts that is not related to brain activity is movement. Changes in the signal arise when the subject starts moving during the measurement, which includes even minor movements such as blinking, yawning, crossing legs, or finger stretching. Subjects are instructed not to move if possible. However, some movements occur naturally that cannot be restricted (e.g., blinking). For that reason, algorithms have been developed that filter out eye movement artifacts via independent component analysis (ICA) or regression methods [23].

[1] Picture taken from https://www.inverse.com/article/32958-ethics-brain-computer-interface-technology-concerns

3.3.2 Re-referencing

Voltage is a relative measure, that refers to a difference in potential between two sites. Hence, a reference electrode needs to be placed in addition to the scalp electrodes. Depending on the position, the reference has a larger or smaller effect on all other electrode sites. The most frequently used sites of reference, to minimize the effect of the reference on all other electrodes, is the nose or the mastoids (point behind the ears). Despite minimizing the effect, it is still present, which is why it can be recommended to remove it. The **common average referencing** filter (CAR) [24] is a filter that subtracts the average value of all electrodes. The average across all electrodes is considered to be an estimate of the activity at the reference site that is equally present at all sites. Subtracting the mean, therefore, removes the influence of reference. Equation 3.1 demonstrates how the CAR is calculated and applied to the data.

$$V_i = V_i^{ER} - \frac{1}{n} \sum_{j=1}^{n} V_j^{ER} \tag{3.1}$$

n represents the number of used electrodes and V^{ER} is the potential of electrode i with respect to the previously used reference.

3.3.3 Signal-to-noise ratio

In addition to that, there are also methods that improve the signal-to-noise ratio, which can be useful to better bring out the control signal, in contrast to the non-control signal. One method is, for example, the **canonical correlation analysis** (CCA) [25], [26]. It is an approach for feature extraction that finds the maximal correlation between two sets of data X and S by applying the linear transformations W_x and W_s (see Equation 3.2).

$$\max_{W_x W_s} \frac{W_X^T X S^T W_S}{\sqrt{W_X^T X X^T W_X \cdot W_S^t S S^T W_S}} \tag{3.2}$$

In practical terms, this means for the example of ERP detection, that the algorithm aims to maximize the correlation between the averaged ERP signal and the single-trial recordings. Figure 3.3 visualizes how the CCA is used during a k-fold-cross validation in a typical classification example for BCIs.

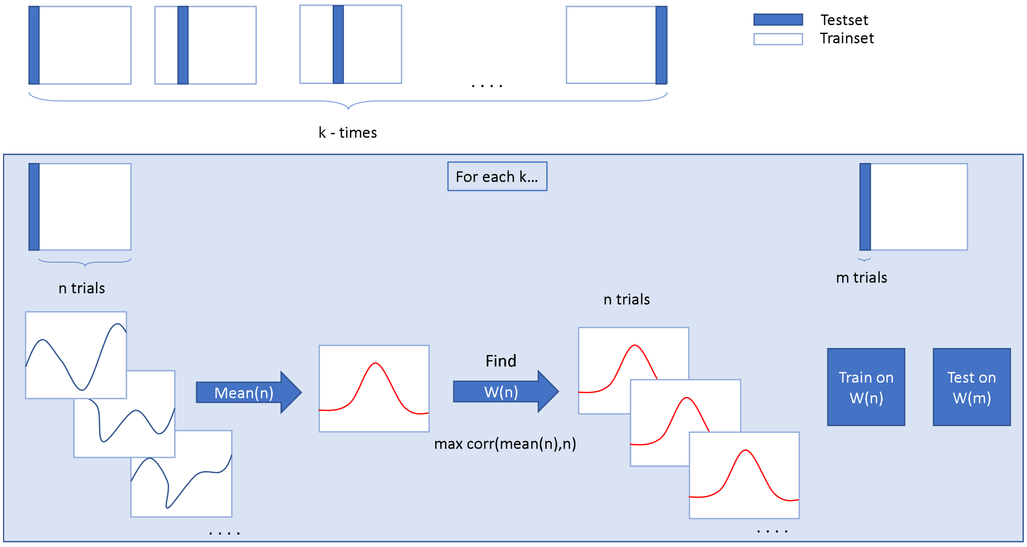


Figure 3.3: Example for CCA in ERP detection - Simplified visualization of the CCA which is classically applied within a cross-validation during the classification. The average ERP of the condition is used as a target to which the correlation shall be maximized within all individual trials. The found transformation W is then applied to the test data.

3.4 Classification

As the last step, it is necessary to identify the control signal when it occurs in the recording. This is necessary to identify which intent the user has and, therefore, to decide which command needs to be sent to the application. To achieve that, supervised learning approaches are most commonly used, which implement a specified training in which the user creates control and non-control signals to collect training data. The training data is then used to train a machine learning classifier that learns to detect and distinguish the two types of signals from each other. For many BCIs, SVMs are used, but also other algorithms are possible and equally appropriate [27]. The classifier is used in the application to automatically detect the users' intent, which is then sent to the application to execute a command. Mostly, feedback is provided to the user, which then closes the loop of the application. Figure 3.4 shows an example of a BCI setup depicting the workflow of a BCI application, to make the closed loop scenario more comprehensible.

3.5 BCIs as a research tool

In general, there is a movement in the field of BCI research that aims to utilize BCIs as a research tool. So far, this topic has not made much progress, but few examples show the potential of data analysis within the closed-loop of an application. To sketch the idea, one example will be shown in the following.
Schultze-Kraft and colleagues were interested in the so-called readiness potential (RP) and its significance during the execution of movements [28]. The RP is associated with the preparation of movement and can be found in the EEG shortly (around 1000 ms) before the onset of movement. The authors were interested in the question if the RP is the

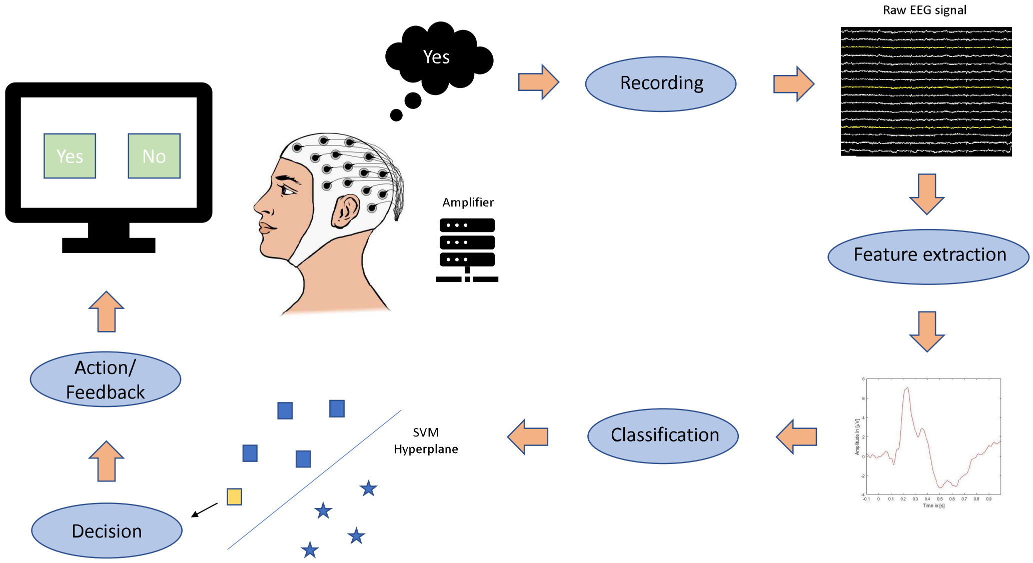


Figure 3.4: Workflow of a BCI - The application is visually presented on a computer screen to a subject that is connected to a EEG cap. The recorded signals go through an amplifier and are filtered to remove artifacts, and possibly to improve the signal to noise ratio. Features are extracted and fed into a classification algorithm that tries to identify the control command. The result of the classification is passed on to the application layer which implements the command. The usage of the application can be seen as a closed-loop.

trigger that initiates the movement or if it is rather an indicator for the decision to move. In practical terms, they aimed to find out: Is cancellation of the movement still possible after the RP has been elicited or not? A closed-loop application was built to identify the appearance of the RP in real-time and to send a stop signal to the application to instruct the user not to perform the planned movement (if possible). The authors found that it is possible to cancel the movement if the stop signal appeared at least 200 ms before the onset of movement, and therefore at least 800 ms after the onset of the RP. After 800 ms it is no longer possible to cancel the movement, indicating that a point of no return has been passed.

In this example, the use of feedback, customized to the real-time brain processes, enabled to determine the point in time after which a movement can no longer be canceled. The study showed that using a real-time classification of brain signals can be a powerful tool to answer open research questions.

3.5.1 Activation patterns

When taking one step back from making use of the full loop of a BCI application, it might come to mind that also intermediate results from processes within the loop could be of value to answer research questions. For example, it could be of interest based on which criteria the decision boundary between the two classes was build. In terms of an SVM, this decision boundary is build by the support vectors. Mathematically, each feature of the input data gets assigned a weight, which determines its importance with respect to this decision boundary. It is possible to interpret these weights in terms of activation patterns that represent the distinct properties of the categories that are separated by the SVM. A method developed by Haufe and colleagues [29] can be used to transform the weights of the SVM classifier into neurophysiological interpretable values. This transformation is a necessary processing step since multivariate methods like SVMs combine information from several channels to improve the signal to noise ratio, thereby preventing the possibility of interpreting the involved parameters directly. The authors describe the step that is done in an SVM as a backward model that transforms data $x(n)$ to the optimized and separable form $s(n)$ by multiplying a transformation matrix on the data (Eq. 3.3).

$$W^T x(n) = s(n) \tag{3.3}$$

The transformation matrix represents the weights of the SVM, which are mathematically optimized but cannot be interpreted regarding the neurophysiological importance of the features that are used for the distinction of classes. The so-called activation pattern A is calculated by multiplying the covariance matrices of the data with the weights of the SVM to reveal the individual importance of the features (see Eq. 3.4).

$$A = \Sigma_X W \Sigma_S^{-1} \tag{3.4}$$

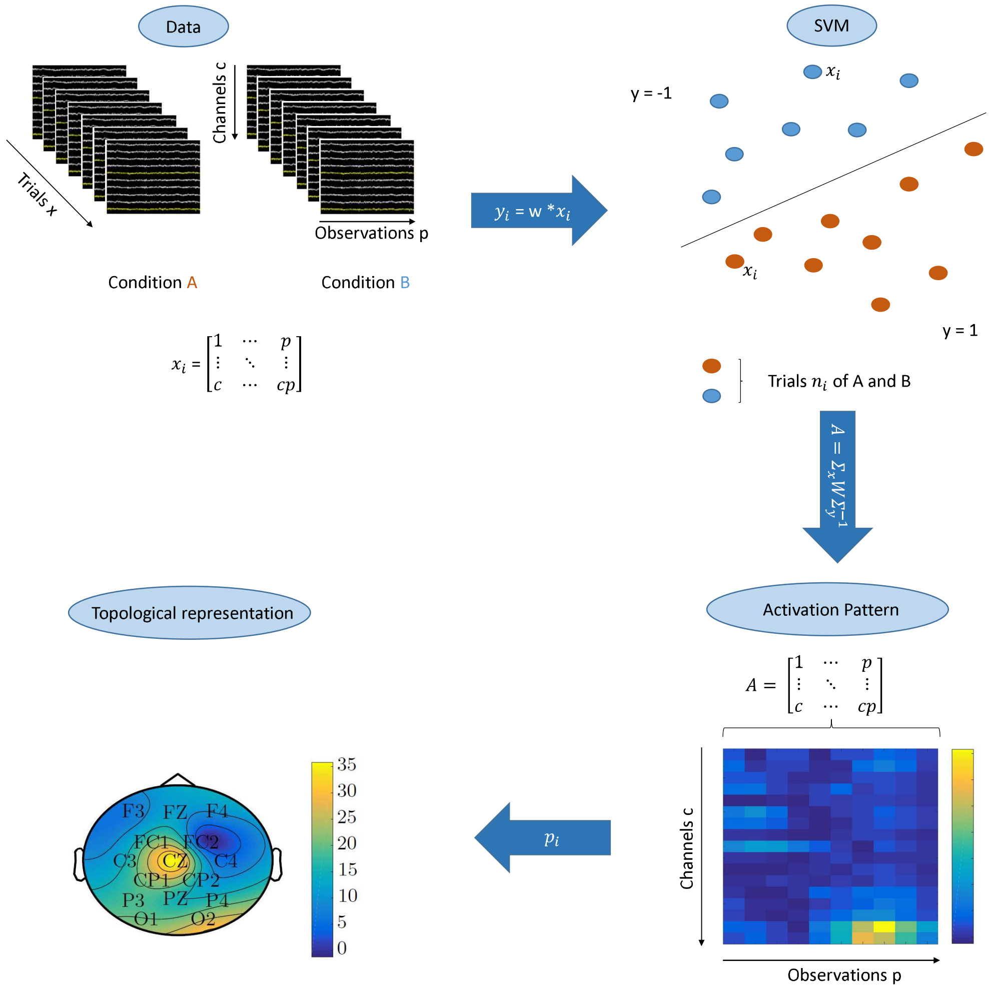

Data
Channels c
Trials x
Observations p
Condition A
Condition B
$x_i = \begin{bmatrix} 1 & \cdots & p \\ \vdots & \ddots & \vdots \\ c & \cdots & cp \end{bmatrix}$
y_i = w *x_i
SVM
y = -1
x_i
x_i
y = 1
Trials n_i of A and B
$A = \Sigma_x W \Sigma_y^{-1}$
Activation Pattern
$A = \begin{bmatrix} 1 & \cdots & p \\ \vdots & \ddots & \vdots \\ c & \cdots & cp \end{bmatrix}$
Channels c
Observations p
p_i
Topological representation
F3 FZ F4
FC1 FC2
C3 CZ C4
CP1 CP2
P3 PZ P4
O1 O2
35
30
25
20
15
10
5
0

Figure 3.5: Workflow for the calculation of activation patterns - Schematic workflow of the calculation of activation pattern A from EEG data. EEG data is cut into trials x_i and assigned a label (A/B) which represents the respective experimental condition. The features of each trial are represented by the number of observations p and channels c which are used to categorize and separate each trials x_i by mathematical optimization in an SVM. The separable form of the data is now called y. Parameter w of the SVM and the covariance matrices of the data x and the transformed version of the data y is used to calculate the activation pattern A. Each value in A can be seen as a rank of importance for the respective feature during the separation. The topological representation of activation pattern A for a specific observation p is optional, but can be used to make it easier to interpret the pattern.

Chapter 4

Problem statement: Limitations of statistics in experimental psychology

As already stated in the beginning, the use of statistical methods is not straightforward. Especially in the field of experimental psychology, statistics have a crucial role, making it indispensable for researchers to understand the implications of the applied methodology. In addition to men-made difficulties that need to be considered, there are also limitations of classical group-level statistics which are currently state-of-the-art in experimental psychology. In the following, the biggest limitations within statistical methods will be explained to create an understanding of the necessity to introduce ways that prevent possible pitfalls and narrow down limitations. Since this thesis focuses on one predominant type of experimental data (EEG), the issues will be pinpointed to this data type.

Why do inferential statistics need to be augmented for data analysis in experimental psychology ?

4.1 Missing analysis of patterns

EEG recordings are an example of a highly multidimensional dataset in experimental psychology, because setups consist of 16 up to 128 electrodes from which each measures hundreds of values per second. The first step in the analysis and processing of multidimensional data is usually to narrow the data down to time frames and sensors of interest because much of the signal might be irrelevant or redundant. Interestingly, whenever experimental psychology uses EEG as a measure, it is usually broken down to a minimal number of dimensions, to test the posed hypothesis. The standard approach is to choose the electrode(s) for analysis that is most likely to be affected by the experimental condition according to the literature on which the hypothesis is based. The motivation behind this is to reduce error accumulation that can be caused by multiple significance tests. Choosing more sites and performing an appropriate correction for multiple comparisons, can be done, but greatly reduces the chance of finding small to medium effects. The minority of studies has medium-sized effects, let alone big effects, hence making it clear why most researchers have suspended the idea of choosing more electrodes in the analysis. Testing only sites of relevance, which are supported by the literature, seems to be a good solution to avoid that problem.

However, the approach of testing only individual sites omits the relation between sensors. When considering that whole networks of structures control most processes in the human brain, it seems plausible and even necessary to take more sensors and especially the relation between sensors into account. The analysis of this relationship could reveal meaningful patterns that help to understand the underlying processes better. Consequently, finding a way that avoids choosing a loss of rich and meaningful information over a statistical technicality, would be an excellent benefit for the analysis of high dimensional data (such as EEG data) in experimental psychology.

4.2 Missing analysis of single-subject level

State-of-the-art analysis techniques comprise mostly of group-level statistics. They aim to assess and quantify effects that generalize over whole populations. Averages and standard deviations involved in T-tests or ANOVAs are the method of choice to achieve that. By looking at the dataset as one construct, the information of a single subject does not weigh in as an individual piece but as part of a whole. Apart from identifying individuals as outliers in the analysis, which often occurs due to technical difficulties, the information of a single subject is not important. By looking at the variance and confidence intervals of the data, the information on individual subjects can, however, be captured to a certain extent. It might be of interest though, in which subjects specifically the found effect is present or not, and with which intensity. Counter-intuitively, this question gets especially relevant when no overall significant effect can be found in the population at all. Not finding a significant effect does not necessarily mean that there is no effect to be found. Investigating the data on a single-subject level might reveal that an effect can be found but only within a subgroup of the population. Or it might reveal that an effect can be found within each subject but with a systematical variation that cannot be captured by the group-level methods. Accordingly, the importance of patterns in brain data is once more emphasized, but not only on a group but also a single-subject level. Finding methods that also cover the level of single subjects comprehensively would be a great gain for the acquisition of knowledge in experimental psychology. The relevance of this issue gets clearer when turning the focus back to EEG data. The low spatial resolution of the technology, in combination with individual differences in brain anatomy, can lead to different measurable outcomes in the EEG signal even though the underlying processes are the same in all subjects. Again, not finding an effect on the group-level does not necessarily imply that there is no effect. In terms of EEG data, it could merely mean that the phenotype of the effect is different between subgroups, if not between all subjects altogether. Inter-subject variability in EEG data is highly present and even can be used for the authentication of individuals without specific feature engineering or filtering [30]. Therefore, adding a single-subject level to the standard analysis seems to be a useful idea.

4.3 Missing identification of latent variables

Latent variables are variables that are not or cannot be measured directly but do have an influence on the measurable variables in the experiment. This can for example be external factors such as the room temperature or the light situation that can cause differences in the subjects behavior that are not caused by the experimental manipulations. The same applies to internal factors such as fatigue, pain or emotions. They might not uniformly be present within the measured population, but they likely account for a considerable proportion of variance within the signal. Some methods can assess and quantify the presence of latent variables. However, it is not common to check data for latent variables. Since the general aim of statistical models is to fit the data as good as possible, latent variables might have a major influence on the construction of the models. Depending on which latent variable(s) are involved, they may be specific to a sample or a particular property of the experiment, so the gain in knowledge that can be achieved by the model may be limited. Since latent variables are neither qualitatively nor standardized recorded, this limitation is very far-reaching and affects all experimentally obtained data. Therefore, adding a methodology that accounts for the identification of latent variables by default seems to be a useful idea.

4.4 Strictly explanatory data analysis

Statistical data analysis is hypothesis-driven and allows only conclusions concerning the proposed hypothesis. The hypothesis makes statements about the to be expected effects that are caused by the controlled manipulation of experimental conditions. In almost every case, the theory cannot be measured directly but needs to be assessed by intermediate variables, such as reaction time or task accuracy, which are then used to corroborate or rejected the hypothesis. Due to this, very strict design and testing are inevitable to guarantee findings that can be associated with the variable of interest. Therefore, all details from design to testing methodology needs to be well thought out and defined even before data collection. In the end, this means if errors in the design are detected or the hypothesis cannot be confirmed, the collected data often becomes useless. Explorative data analysis, which would mean testing several hypotheses on the same set of data, is not suitable for this methodology. Due to the sampling of data from bigger and unknown populations, effects can always be found that stand out significantly from chance, with pronounced multiple testing. By correcting the chance level according to the number of performed tests, this issue can be contained. Still, other problems such as overfitting of statistical models play a crucial role in this, further preventing excessive testing. As already stated, most statistical models aim to model the data as well as possible. By modeling data as precisely as possible to a specific research question other essential aspects on which the data is based get neglected. This leads to ignoring relevant information while putting an emphasize on potentially irrelevant information, which finally results in a distortion of facts. Naturally, this is not desired and would lead only to insufficiently reliable scientific findings. Therefore, it is always necessary to perform several consecutive experiments, each specifically designed for one hypothesis, to explore more than one research question. This is methodically flawless but very time and money consuming and, therefore, often not feasible. Finding methodologies that allow exploratory data analysis within one experiment by overcoming the issue of multiple testing and overfitting would be a desirable goal to make the most out of existing data.

4.5 Summary: Issues of reproducibility

The listing above might already suffice to grasp how it can easily happen that experimental results might occur once but not necessarily again in a replication of a study. Of course, there are other reasons, such as power issues caused by small sample sizes and malpractices, which, however, will not be addressed here. This thesis wants to create awareness that there is an urgent need for a methodology that can take multidimensional data into account, without causing issues of multiple comparisons and extensive overfitting, while taking patterns and the single-subject level into account. Understanding that patterns could describe the measured effect and its underlying theory comprehensively or at least more accurate than individual measurement points is crucial for decoding cognitive processes further. Including pattern analysis is, therefore, a useful and necessary extension. Especially for the example of EEG data, but also for other data types, it is also necessary to take the single-subject level into account. Patterns may differ from subject to subject regarding their particular location and strength of expression. But they can show a correlation that cannot be described by averages. Especially due to the poor spatial resolution of the measurement technology and the complexity of the human brain, a single-subject view of the data would be advantageous. Overall, the focus should be on generalizability and not only on the goodness of the fit. This ensures that the generated models also work on new data. Thus, the data needs to be considered at a different level of abstraction, which may do better justice to the target variables and capture them better than a model that limits itself to the goodness of the fit. All of these issues would help to keep issues of reproducibility down because. Mainly because the diversity of the data is no more regarded as a disadvantage but integrated advantageously into the analyses. Last but not least, the ability to perform exploratory data analysis without committing data dredging and abusing the flaws of statistical testing would be desirable to make more and steady progress in experimental psychology.

4.6 Proposal for an augmented methodology

This thesis represents a proposal for the introduction and promotion of machine learning within the field of experimental psychology. In this thesis, SVMs are chosen to classify data that correspond to a priori determined variables of interest. The classification process, which is literally a process of separation, can be regarded as a practical implication of the research question: "Are those two samples that have been obtained under different experimental conditions from the same population?"
If the data can be separated, it is highly likely that the data arises from two populations. If it cannot be separated, the data originates from the same distribution. The separation is quantified by accuracy values that state in how many cases the categorization into one of the two determined variables is correct. Therefore, if the process of classification is successful, which is expressed in high classification accuracies, the proposed null hypothesis that both samples are from the same population can be rejected. Classification accuracies around the chance level, on the other hand, would be indicators to corroborate the null hypothesis and, therefore, to assume that the samples are from the same population.

In addition to delivering an equivalent function as classical hypothesis tests, machine learning has the prospect of solving the above-stated issues. The field of machine learning has its origin to a large extent in pattern recognition and the recognition of regularities in large amounts of data. Therefore, it can take multiple levels of the data into account, without the necessity of multiple hypothesis testing. Due to the high focus on real-world applications and especially the application to new and unseen data, the issue of overfitting is addressed, and solutions have been provided in many forms (e.g., cross-validation and regularization). In many cases, the form of application has an exploratory character. This helps to understand what the available data, and therefore, the problem at hand is about. By exploring all properties of the data, it is possible to find general models that can describe the data in its entirety and on an appropriate level of abstraction. Last but not least, it can also be seen in Chapter 3 that using machine learning methodology on single subjects only is not a concern, as long as a representative amount of data has been collected for each subject individually. It is a standard approach in BCI research and by that means, is also used for the research of mental states [31]. Further, the Chapter shows that ML does not need to be opaque. Methods exist that make the approach more transparent and traceable. By this the results get interpretable and provide deeper insights into the data. It has been shown that the approach provides meaningful results in those fields. Overall, this short sketch of the idea of proposing the application of machine learning in experimental psychology gives already a promising estimate of how the existing issues could be overcome. This thesis will implement the idea and provide results for creating a proof of concept.

Chapter 5

Field of application: Memory

To test the proposed methodology of using machine learning in experimental psychology, the field of learning and memory was chosen as an exemplary field. One of the reasons for this choice is that it is a considerably broad topic that is still relatively poorly understood. The second and more important reason is that it is a field that heavily relies on brain data to make further progress to comprehensively identify processes during the learning, storage, and retrieval of information. In the problem statement, it was made clear that especially high-dimensional data, such as brain recordings via EEG, can lead to problems when using standard techniques for statistical analysis. Accordingly, learning and memory seem like a suitable field of application to establish a proof of concept of the proposed methodology. In the following, an overview of the standard models of human memory will be given to get an insight into the complexity of the field and the connection between the individual components that will later be the focus of this thesis. Chapter 6 and 12 will then give a more detailed introduction to cover the basics and standard theories that will be needed to understand the paradigms of the two case studies performed in this thesis.

Choosing a stage for a proof of concept

5.1 Multi-component models

Human memory is a fascinating construct on which we highly depend on a moment to moment basis. It can capture facts and episodes of daily life, conscious or unconsciously, with almost limitless capacity. Regarding the issue of memory capacity, however, it quickly comes to mind, that many things are easily forgotten or can only be remembered with a high effort of learning. The proposition of a fragmenting of our memory system into several components that differ in their capacity and the duration of how long information can be stored there was therefore readily accepted. It was made by Atkinson and Shiffrin in 1968 and can be found in the literature under the term multi-store model [32]. It consists of three types of memory components that all have in common that they maintain information for a certain amount of time after the source of information has vanished. The three components are usually referred to as **sensory memory**, **short-term memory**, and **long-term memory**.

Sensory memory allows retaining impressions of sensory information, such as touch, vision, and sound, with a high capacity for a very short amount of time (under a second). For many, this is rather a perception than a form of memory. Short-term memory deals with information that can be received and processed at the same time and is hence, limited in capacity and can store information only for a short period (under a minute). Long-term memory is theoretically unlimited in capacity and can store information for years or even a lifetime. It represents knowledge that we are not consciously aware of all times, but that can be actively retrieved into consciousness. All in all, this is a very comprehensive model that still holds today, but it is certainly too simple to capture the complex reality.

Short-term memory

Only a few years after Atkinson and Shiffrin proposed their model, Baddeley and colleagues presented an extension of that model. It further fractions short-term memory into a system of several components that are not only able to store, but also to manipulate information, while still holding it within memory. The authors engraved the term working memory for the system and describe it with three different components: **central executive**, **phonological loop** and **visio-spatial sketchpad** [33, 34]. The central executive is an attention-related control structure, which coordinates the storage, rehearsal, and transformation of information in memory through the so-called executive functions (EFs). Since the EFs will be the center of the first case study, they will be introduced in detail in Chapter 6. The phonological loop and the visio-spatial sketchpad can be described as memory-related storage components for verbal and visual information. In 2000 the authors refined the model by adding the **episodic buffer** as an intermediate component that links working memory with long-term memory [35]. It transfers information from and to long-term memory while consciously processing it and, therefore, closes the loop between the two major memory components.

Long-term memory

Naturally, long-term memory is also an extensive construct of several components that can store vast amounts of information. Current models define long-term memory as a system of at least two components that distinguish between **explicit** (conscious) and **implicit** (unconscious) knowledge [36]. A more coherent definition of the two types of memory is the distinction of ”knowing what” (facts, the experience of time, places, and feelings) and ”knowing how” (skills and conditioning), which can be assigned to explicit and implicit memory respectively. For the component of explicit memory, there was a further subdivision into **episodic** and **semantic** memory, categorizing information into personal experience and factual knowledge. More subdivisions exist, but the so far given level of detail will suffice for a general overview of the core constructs in memory that will be needed in this thesis.

5.2 The focus of this thesis: Two case studies

To sum everything up, Figure 5.1 visualizes the connection and arrangement of all relevant memory components that have been described within this chapter. For this thesis, the focus is put on two central aspects within the construct of human memory, to show representatively what can be achieved with the use of machine learning within experimental psychology. The first core topic deals with current questions regarding the central executive of working memory and its underlying constructs. The central executive

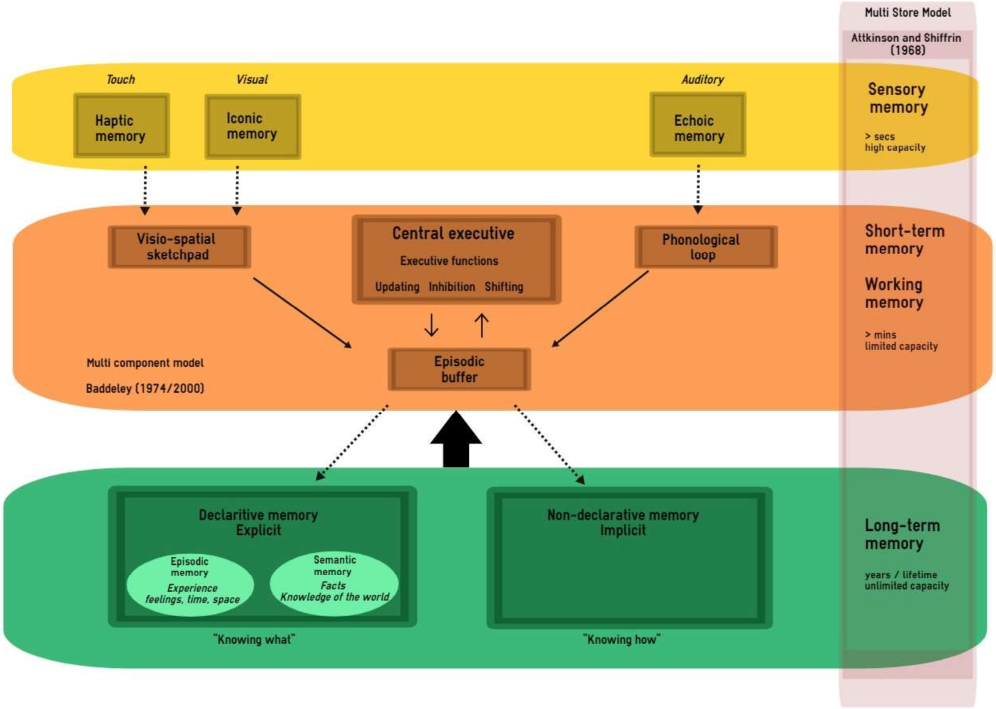


Figure 5.1: Memory models - The three main components of human memory, sensory, short-term and long-term memory are separate but interdependent constructs. Information that enters human memory always starts in the sensory memory system and then either travels from short to long-term memory or leaves the human memory system in either of the three components. Especially the long and short-term memory system maintain a permanent connection. This connection is used to exchanged information for the storage, retrieval, or maintenance of information. Different subcomponents of each memory system are known, which are in parts depicted within this figure.

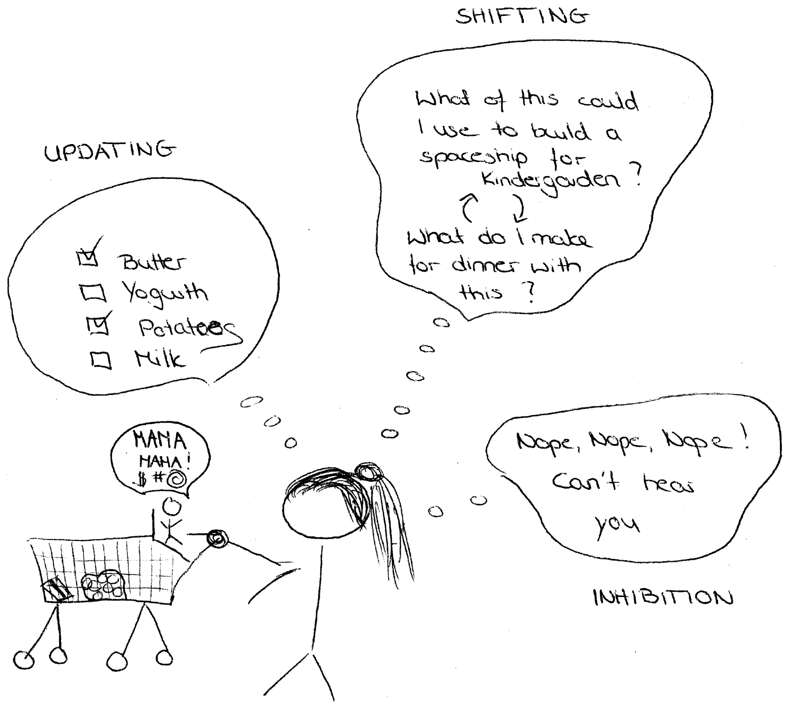
is the key player in working memory and serves as a gateway and controlling instance regarding the information that will be passed on to long-term memory. It is, therefore, considered to be of particular importance. Due to its importance and the fact that it is still not entirely understood, EFs the basis of the central executive, have been chosen as the stage for the first case study.

As a second topic, the core functions of episodic memory, or rather long-term memory in general, are under investigation. The central chain of processes that every information needs to pass to be captured in long-term memory includes the encoding, storage, and retrieval from long-term memory components. Here, too, there are still many unanswered questions that clarification, especially when the processes are viewed in interaction and not just as individual components.

The selected topics serve as a model for an exemplary proof of concept, which at the same time should clarify central questions in memory psychology to significantly advance research in this field.

Part II

Working memory load and executive function


SHIFTING
What of this could I use to build a spaceship for Kindergarden ?
What do I make for dinner with this ?
UPDATING
☑ Butter
□ Yoguth
☑ Potatees
□ Milk
MAMA MAHA! $#@
Nope, Nope, Nope! Can't hear you
INHIBITION

Chapter 6

Introduction into working memory and executive functions

The first case study will deal with working memory (WM), and its core construct executive functions (EFs). To complete the overview of Chapter 5, a more detailed description of working memory will be given, with a major focus on EFs. The chapter is supposed to give a basic overview over standard models and theories to provide the basic knowledge that is needed for the studies described in the following.

As has been described in Chapter 5, WM can be seen as an extension of the concept of short-term memory. It is as a system that allows temporary maintenance and manipulation of information, which is helpful and indispensable to perform complex tasks. It is a construct of several components, one of which is the central executive, an attention-related control structure. Miyake and colleagues tied the concept of EFs to this central executive. EFs are cognitive processes that are essential in the planning and control of goal-directed behavior [37]. Together with Baddeley's model of working memory, Miyake's concept of EFs describes the dominant view of the standard literature concerning working memory [38].

6.1 Miyake and Friedman's framework of the unity and diversity of EFs

Based on a latent-variable analysis of a bundle of executive tasks, Miyake and colleagues [39] developed one of the best-known models describing the structure of different EFs. According to this model, the three core EFs are named *updating*, *inhibition*, and *shifting.* EFs play a pivotal role in many other and even recent WM theories (e.g., [40, 41, 42, 43, 44] and [45]). According to these theories, the attentional control processes required to pursue WM tasks are responsible for the severe limitation of human WM. Accordingly, individual differences in WM capacity trace back to individual differences concerning these EFs [42, 44]. Miyake and colleagues [39, 46] describe the relationship between the three core functions in terms of unity and diversity, as they share many properties but also have distinct characteristics as individual functions (see Figure 6.1). They state the EFs are "separable but moderately correlated constructs" [39]. The unity and diversity of EFs as individual components of WM, describing shared and unique properties of the

EFs, has so far mainly been investigated through statistical analyses of behavioral data in either healthy subjects or patients with frontal lobe impairments, see e.g., [47, 48] and [49]. Until now, it remains unclear whether the shared and unshared variances of performance in different EF tasks can be mapped either on a common attentional and limited resource in the brain [50], a shared network [51] or, on EF-specific brain functions. In 2012 [46] Miyake and colleagues have adapted their model based on studies performed by Friedman [52, 53], stating that there is no unique variance for inhibition as it perfectly correlates with the common properties of EFs. Overall this is a complicated question to address. It has been found that tasks that reliably induce load on the same EF often have low correlations with each other. Burgess and colleagues describe this is as an issue of task impurity [48] because many non-executive processes are required to solve a task, which can vary greatly even though the same core EF processes are present [54].

Collette et al. performed one example of studies that aimed to further investigate the unity and diversity of EFs on healthy subjects with neuroimaging techniques ([55, 56]). The authors explored neural substrates for the EFs updating, inhibition, and shifting using positron emission tomography (PET). Via conjunction and interaction analysis, they compare a battery of tasks for each EF to reveal both unity and diversity aspects within the collected brain data.

An assessment of differences in neural correlates of EFs using EEG data, that allows a within-subject and not only between-subject comparison between EFs has so far only been done by Scharinger and colleagues [57]. In their study, they manipulated demands on the two EFs updating and inhibition independently. They analyzed indicators of working memory load (WML), both at the behavioral level (reaction times (RTs) and accuracies) and at the physiological level (pupil diameter, event-related potentials in the EEG (ERPs), and EEG power spectra). The aim was to develop a detailed and brain-related account of how load on different executive functions is interrelated. Their study will be the starting point of this case study, but before I introduce the studies of this thesis, the three core EFs will be presented to get a better understanding of the issue.

6.1.1 Inhibition of information and responses

The EF inhibition describes the ability to suppress irrelevant information and to prevent impulsive behavior in inadequate situations. For both scenarios, a high level of focused attention is required to perform accordingly. To give a real-world example, the ability to focus on one out of many simultaneous things comes in handy, for example, during conversations in crowded and noisy environments. But also the avoidance of reflex-like behavior is useful in daily life as in keeping information to oneself. Either literally by not expressing an opinion or by keeping a straight face, which can be challenging in many situations.

Many tasks have been developed to assess and test for the EF inhibition experimentally. The so-called Stroop task [14] [58] [59] is one of them, in which color words are presented on the screen and subjects are instructed to name the color of the font out loud. The written words either match the color font or form a different color word. In

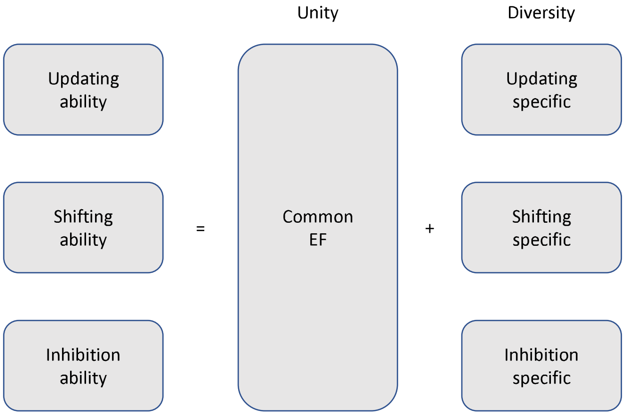


Figure 6.1: Miyakes model of unity and diversity - The ability to perform executive tasks is based on three core EFs: updating, inhibition and shifting. Each EF consists of properties that are common for all three EFs (unity) and a individual component that is different for each EF (diversity).

case of a mismatch, the information creates interference, and the dominant dimension (in this case reading the word) needs to be suppressed to perform adequately. A mismatch between the dimensions results in longer reaction times and also in more errors. The same task also exists with number stimuli, in which the numerical and the physical size of the numbers interfere. In this case dominant dimension is the numerical size of the stimuli which needs to be suppressed.
The Eriksen Flanker Task is also a well known and standard tool in psychology to measure and induce response inhibition [60], [61], [62]. In its original form, it consists of arrows horizontally aligned on a screen. One of them is in the center. The subjects' task is to indicate by means of a button press whether the arrow is pointing to the left or the right. Three arrows accompany the central arrow to its left and right, which either all point in the same direction, which can be equal or different from the direction of the central arrow. Independent of the concrete task, it has been established to describe matching dimensions as congruent and mismatching dimensions as incongruent from each other. In some experiments, a third and neutral dimension is introduced to create a visual distraction but without interference in the target dimension. Figure 6.2 shows a short outline of the three tasks to get a better understanding. Since the Flanker task will be the method of choice for the induction of inhibition demands within this thesis, some more details about the task will be provided.

A general observation is that the reaction time (RT) is higher and the accuracy lower for trials in which the flanker is incongruent (different) from the central item in contrast to trials with congruent (equal) flankers. This effect has been described as the Flanker congruency effect (FCE) [63]. Different ratios of congruent and incongruent flanker items in- or decrease this congruency effect [64], [65]. A possible explanation is the level of paid attention to the flanker items. The smaller the ratio of incongruent flanker items, the lesser attention is paid to the irrelevant and peripheral stimuli. This

phenomenon can also be described as the utility principle. The processing of the flanker items is optimized by allocating only a specific amount of attention to the flanker items depending on their utility [65]. In general, this observation can be described as an interaction between the trial type (congruent or incongruent) of the current trial and the previous trial, which is also known under the name Gratton effect. Seeing this as modulation of cognitive control it is the main effect supporting the conflict adaptation theory by Botvinick [66].
This phenomenon of interaction can be explained by an adjustment of cognitive control due to the experience of conflict, mainly when conflict is aversive [67], [68], to be able to manage the subsequent conflict better. Therefore, this effect is also known as a so-called conflict adaptation effect.

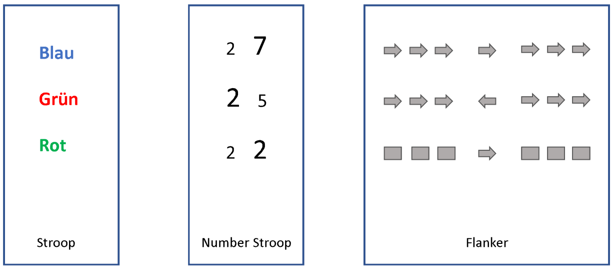


Figure 6.2: Examples inhibition task - The Stroop task requires the naming of the color font while inference is created by displaying written color words in matching or mismatching color fonts. A version of the task with numbers creates inference by a mismatch of physical and numerical size of the displayed numbers. The flanker task requires to identify the direction of the central arrow while the flanking arrows might interfere by pointing to the opposite direction. Generally, a mismatch of stimulus dimensions creates inference that influences task performance. A mismatch of stimuli dimensions is also described with as an incongruent, a match of dimensions as a congruent condition.

6.1.2 Updating - information updating and monitoring

The EF updating describes the ability to monitor and store information into working memory and to manipulate, change, or rearrange information within memory when necessary [69]. A real-world example that requires updating is, e.g., a mental shopping list that is adapted while walking through the grocery store and adding items to the cart. The list can be altered, extended, or shortened at any time. Despite the active control of what is entering and leaving WM, this process can also be passive. It has been found that there is a temporal tagging of information to decide what is no longer needed [70] to clear the storage space.

Again many tasks have been developed to assess the EF updating experimentally. One example is the letter memory task [69] in which a strict recall of the last four letters of a list is required, which can in theory, be arbitrarily long. Another example is the Sternberg task [71] in which items need to be recognized as already seen or not out of a list of up to 8 items. It has been found that the reaction time for the categorical answer (yes or no) correlates positively with the length of the list. The method of choice for this thesis to induce load on updating demands is the so-called n-back task [72]. In a typical n-back task paradigm, participants see a sequence of stimuli, presented one after another. The task of the participants is to indicate via key press whether the current stimulus matches or mismatches the reference stimulus they saw n-steps back concerning a particular feature dimension (e.g., location, size, color, or identity). Depending on the n-back level, the n-back task is supposed to impose increased updating demands with increased n-level. The higher n-level the higher the reaction time as well as the error rate [73], [74]. The n-back task is a widely used WM task in neurophysiological research for studying effects of WM updating (e.g., [75], [76], [77], [78], [79], [80], [81], [82]), which is why it seemed to be an appropriate choice.

6.1.3 Shifting - mental set shifting

The EF shifting is often described as cognitive flexibility. A key concept in cognitive flexibility is a task set, representing a set of rules that need to be collected, understood, and linked to the situation to perform the correct actions. Shifting between task sets is essential to be able to solve new and daily situations. An example of this abstract definition is as simple as trying to imagine what purpose an item can have. A knife can be used to cut food but can also be used as a weapon. A t-shirt is an item of clothing for the torso, but can also be used as a protection from the sun on other body parts. It can be used to dry something when using it instead of a towel, when more than one are tied together they could work as a rope or a blanket. Possibilities are manifold but require to think out of the box and to use the knowledge that is usually applied when using other items. Overall it can be stated that this way of thinking is essential to fulfilling goal-directed actions [83].

For the assessment of requirements on the EF shifting, many tasks have been developed. One typical task is the Wisconsin card sorting test (WCST), in which subjects need to find the appropriate rule of sorting cards into categories [84]. The rule stays the same for several trials and can be worked out with the help of categorical feedback from the experimenter for each trial. Other typical tasks are plus and minus calculus operations or to evaluate numbers and letters concerning a specific question (consonant or a vowel, bigger than 5) in an alternating order. A general concept for the assessment of shifting is the alternation between two tasks, requiring the application of two different rules to solve the task correctly. Therefore, the appropriate response varies as a function of the task. Jersild [85] did pioneer work in the area of task switching, investigating differences in error rate and RT between blocks in which subjects only performed one task and blocks in which they had to complete two different tasks. The result was that subjects are much slower in blocks with two tasks than with one task. It was also found that the effect is present between trials of different tasks within a block in which two tasks are presented. Trials can be defined as switch trials if, in the previous trial, a different task needed to be performed than in the current one, and as repeat trials when the previous trial presented the same task as the current one. Switch trials have been shown to be answered slower and with a higher error rate than repeat trials. The difference, especially in RT, has been defined as a cost that is due to task switching [86]. Several subtypes of costs have been identified in task switching, each caused by different processes that need to be implemented during task switching in general. One major part of shifting costs is aggregated in the term alternation cost, which can be further divided into switch costs (immediate shift costs) and mixing costs (costs due to having two tasks in general) [87]. Shifting costs reduce when getting a chance to prepare before a task switch. Showing cues to indicate what is going to happen next allows preparation, which can reduce costs significantly. Biedermann states that showing a cue speeds up the process of answering a task in different calculus operations [88]. Still, it has also been demonstrated for a task in which needs to be decided if the stimulus is odd or even[89].

6.2 Idea and hypothesis

The following sections will present a total of four studies that aim to characterize and distinguish the three executive functions updating, shifting, and inhibition based on their EEG correlates. The main question is if the theories and models by Miyake and colleagues can be supported and interconnected with neural components. The hypothesis is that the difficulty of finding individual variance of EFs arises from the incapability of current state of the art methods of capturing the variance within EEG data. It seems likely that EFs are based on patterns distributed all over the brain, which need to be adequately assessed. It is also likely that the variance between subjects is rather high, which is why an investigation on single-subject level is advisable to make progress with this question. In particular the combination of updating with inhibition and inhibition with shifting will be investigated in the following studies (for a schematic overview see Figure 6.3).

To this end, research has established that executive functions have commonalities but are individual processes that can be distinguished. The detailed characterization of the commonalities and differences of the three EFs (unity and diversity), however, is still an open question. Four studies have been designed in this thesis to realize the aim of finding evidence for a better characterization of the EFs. Two crucial factors for solving this problem will be a careful and unique task design combined with a machine learning approach. The task design is unique as it combines two EFs in one task while keeping non-EF variance to a minimum. It enables a comparison between two executive functions within the same experimental setting but more importantly, within the same subject. This design reduces the non-EF variance to a minimum and ensures that if a difference occurs, the variance of the two EFs is the cause of it. To make use of this within subject comparison prepared by the task design, the machine learning approach is needed. It will extend the classical explanatory group-level approach also to the single-subject level. Individual differences can therefore be investigated per subject as well as per group. In addition, it enables to consider patterns in the rich and high dimensional EEG data instead of single electrodes only without hitting statistical limitations.

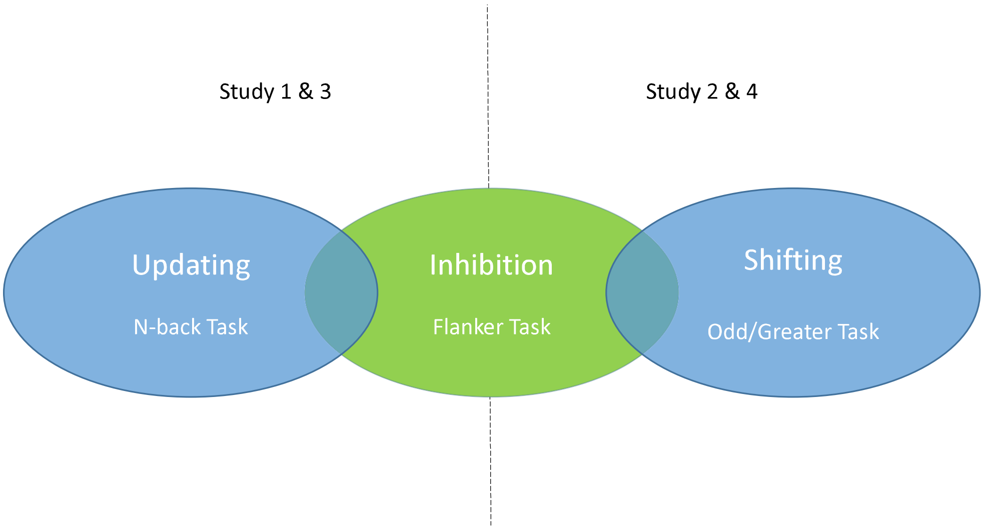


Figure 6.3: Idea and implementation - Content of working memory studies at a glance

Chapter 7

Study 1: When Updating meets Inhibition

Study 1 combines the EFs updating and inhibition to investigate their properties with concerning unity and diversity aspects. Conceptually, Study 1 is a reanalysis of a study initially performed by Scharinger and colleagues [57], with which I actively collaborated. The results of Study 1 have in part been published in [90] and in [91] and will now be presented in detail, from task design and experimental setup to the achieved results and the implications of the latter.

Can correlates of unity and diversity for the two EFs updating and inhibition be found in EEG data?

7.1 Task design

The experiment uses an integrated n-back flanker task to study interactions between the two EFs updating and inhibition. The n-back task is used to induce demands on the EF updating, whereas the flanker task is used to induce demands on the EF inhibition. The simultaneous presentation of the two tasks was realized by showing seven items at once, from which one was positioned centrally, the other six on a flanking position to the left and right of the central item (see Figure 7.1). The n-back task was performed on the central item and used as the primary task in the experiment. The flanker was only used as a secondary task, to which no action was required. The six flanking items, where therefore congruent (identical) or incongruent (different) to the central item and fulfilled a distracting purpose.

The list of stimuli of the task consisted of the four letters, S, H, C, and F. For each trial, one out of these four letters was randomly chosen and presented centrally on the screen either flanked by the same letters (e.g., HHH H HHH) or by randomly chosen different letters (e.g., FFF H FFF). All letters were presented in gray on black backgrounds in Arial at 25-point font size. Each stimulus was shown for 500 ms, followed by a black screen for 1500 ms. Thus, each trial lasted 2000 ms. For a schematic overview of the task see Figure 7.1. In the experiment, three levels of updating demands were implemented (n = 0, 1, or 2) in a block design. For each trial in a block subjects indicated via a key press (yes/no key) whether the central letter of the current trial was identical to

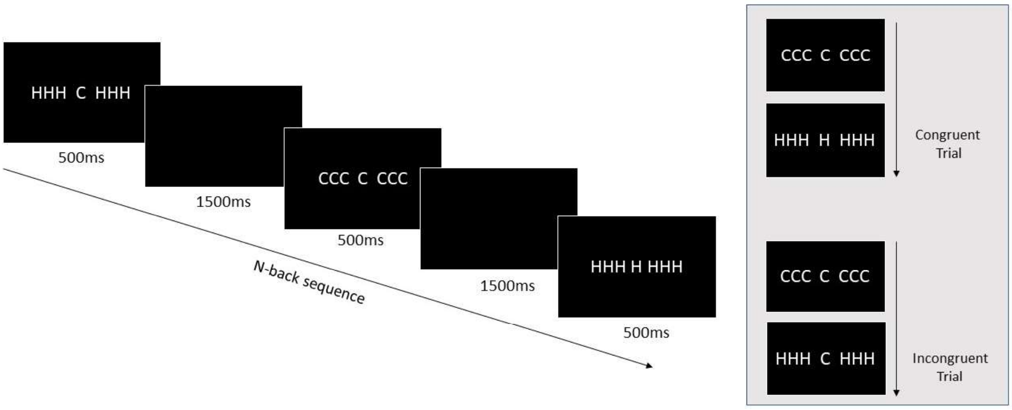


Figure 7.1: Experimental design - The task was presented on a black screen with white letters as stimuli and flanker items. Stimuli were presented for 500 ms, followed by a blank black screen in which the subject needed to answer the n-back task on the central item with yes or no by pushing the respective key on the keyboard. The box on the right shows exemplary which trials were used in the analysis as congruent or incongruent trials.

(target) or was different from (nontarget) the central letter they had seen in the sequence n steps back. In the 0-back condition, before the stimulus sequence started a randomly chosen letter (S, H, C, or F) was displayed as the n-back target letter for the whole block (no updating required). During the following stimulus sequence, each time this letter occurred as the central letter, subjects had to press the yes key, in all other cases, the no key. Subjects answers and reaction times were recorded. Each n-back level was presented twice. Thus, subjects performed a total of six blocks with 154 trials each. The sequence of blocks was randomly assigned for each subject, with the constraint that each n-back level was presented once before an n-back level was presented for the second time. One block lasted about 5 min. Within each block, half of the trials were targets, half of the trials were nontargets. Concerning the flanking items, the experiment was designed in a way that about one-third of the stimuli of each response category was incongruent and two-thirds were congruent. The first four trials of each block were always congruent nontargets. The trial sequences within the blocks were pseudorandomized, to avoid attenuation of the interference effect for incongruent stimuli due to conflict adaptation processes (i.e., the so-called Gratton effect; [66], [92], [65]), incongruent-incongruent stimuli sequences were excluded in advance during construction of the stimuli lists. In addition to that, in each block 20 randomly chosen stimuli a sequence of 10 targets and non-targets was replaced by stimuli without a central letter (i.e., ten targets and ten nontargets per block consisted only of the flankers on both sides of a gap). In this cases, the subjects were instructed to remember the flanker letters of the current trial for the following comparison and to base their current target/nontarget judgment on a comparison of the flanker letters with the previous central letter according to the n-back level. This was done to avoid that the subjects become increasingly unaware of the flankers during a block.

Stimuli were presented using E-Prime presentation software (E-Prime 2 Professional, Psychology Software Tools, Inc.). At the beginning of the study, subjects performed training blocks for each n-back level. Training was repeated until subjects reached an accuracy of at least 60 percent correct responses. During training, subjects' accuracy was displayed at the end of a block to give them feedback regarding their performance. No feedback was given during the actual task presentation.

7.1.1 Participants

22 subjects (22 right-handed, 12 females) with an average age of 22.64 (±4.31) participated in the study, for which they were reimbursed with 8 euro per hour. All subjects had normal or corrected to normal vision and no reported neurological disorders. Written and informed consent was given by each subject, and the study was approved by the ethical committee of the University of Tübingen. All subjects were right-handed, which was validated with a standardized questionnaire.

7.1.2 Technical setup

EEG was assessed with 32 electrodes (Acticap Brain Products) at a sampling rate of 500 Hz. The electrodes were placed according to the international 10/20 system [93] (Fp1, Fp2, F7, F3, Fz, F4, F8, FC5,FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, O2) with the reference at right mastoid and the ground electrode at AFZ. Three additional electrodes were placed at the outer canthi of the and below the left eye, to record an electrooculogram (EOG). Further details considering the technical setup can be found in the original publication [57].

7.2 Data analysis

The reanalysis of the data was limited to the physiological data, as the behavioral data has already been sufficiently evaluated in the original study. Therefore, only the procedure of the physiological data will be explained in the following. The data was cut into epochs of 0-1000 ms after stimulus onset with respect to the six categories as can be seen in Figure 7.2. All in all, only artifact-free trials with correct responses were used for data analysis, with additional exclusion of trials that might yield to any Gratton-like effect. This includes congruent trials following incongruent trials, and all trials in which the central item was missing (plus the two preceding trials each). Additionally, the first four trials of each block were excluded. Two types of physiological features were evaluated: ERPs and power spectra.

7.2.1 Neurophysiological analysis

The EEG data was bandpass filtered between 0.4 - 40 Hz and re-referenced to common average. To remove artifacts, a threshold of 100 μV was chosen, and all trials exceeding this level were discarded. Trials including eye movement artifacts were corrected using independent component analysis (ICA) (rejection by visual inspection). For the investigation of ERPs, the grand average over all subjects and trials was computed for each of the six categories separately. Based on the results of [57], the electrode positions FZ, CZ, PZ were chosen to be of major interest. Apart from the ERP analysis, the data was also analyzed in the frequency domain to get insights into the spectral properties of the two executive functions. For the calculation of the power spectra Burgs maximum entropy method [20] was used with a model order of 32 and a bin size of 1. Again the grand average will be calculated to visualize the differences between the categories over all subjects.

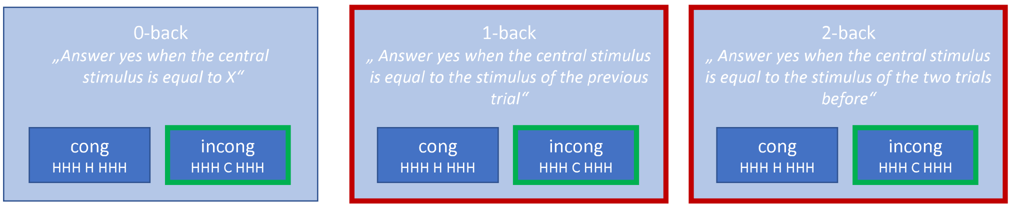

0-back
„Answer yes when the central stimulus is equal to X"
cong
HHH H HHH
incong
HHH C HHH
1-back
„ Answer yes when the central stimulus is equal to the stimulus of the previous trial"
cong
HHH H HHH
incong
HHH C HHH
2-back
„ Answer yes when the central stimulus is equal to the stimulus of the two trials before"
cong
HHH H HHH
incong
HHH C HHH

Figure 7.2: Datastructure - The data can be divided into three categories regarding the n-back level, which represent different levels of updating demands. Each category can further be divided into two subcategories regarding the Flanker condition, which represent different levels of inhibition demands. In total a set of six categories can be distinguished and will be used for analysis. 🟩: inhibition, 🟥: updating

To test whether the differences in the ERPs and power spectra between the categories are statistically significant a Wilcoxon rank sum test was conducted over all subjects and trials. The resulting p-values were Bonferroni corrected (according to the number of time points or frequency bins), and the significance level was set to $p < 0.05$

7.2.2 ML based classification

For the investigation of the EFs by means of machine learning, support vector machine (SVM) classification was chosen. A SVM with a linear kernel ($C = 1$) [94], [8] was applied to differentiate between the six categories introduced above using the libsvm implementation for Matlab [95], [96]. The classification between categories was conducted for the following pairs for each subject individually:

- Inhibition: Cong vs Incong
- Updating: Zero vs One, Zero vs Two and One vs Two
- EFs: Updating vs Inhibition

The aim is to separate EF demands from baseline demands, but also different EF demands from each other. As the baseline demand within this study, congruent trials from the zero back condition have been chosen, because neither updating nor inhibition demands should be induced during those trials. For each data pair a subset of the data is used to train a classifier, to learn the characteristics of each category. The remaining data is used to evaluate the success of the learning and hence the skill of the classifier. This is done on a single-trial and single-subject level. To ensure stable results a 10-fold cross validation was performed for each classification. The datasets (training set as well as the test set) were balanced for each comparison, by removing all spare trials if one of the classes had more trials than the other, to ensure that the distribution of examples per class does not have an influence on the result. For classification, again two different types of features were used: CCA filtered ERP features [97] and power spectra. For each of the features, an individual SVM was trained and evaluated. The performance of the classification approach is measured in accuracy, stating in how many cases the classifier categorized a trial correctly. Statistical significance of the classification results was determined by using permutation tests on the data. The statistical significance of the results was determined by permutation tests with 1000 iterations [98, 99]. The classification performance achieved in the permutations establishes an empirical null distribution on random observations, which can be used to determine significance boundaries. Therefore, in each iteration classification was performed in a 10-fold cross validation, but with randomly assigned class labels in the training set instead of the correct class labels. The achieved accuracy values were compared with the ones determined in the standard 10-fold cross validation. Significance level was determined to be at $p < 0.05$, stating that the original classification performance is significant when the performance values are higher than the 95th percentile of the calculated empirical distribution.

ERPs

The ERP features upon which the SVM classifies are composed as follows. The 0-1000 ms epochs that have been prepared in the beginning will be used from 17 EEG channels (FP1,FP2,F3,FZ,F4,FC1,FC2,C3,CZ,C4,CP1,CP2,P3,PZ,P4,O1,O2). All other channels of the setup are discarded to reduce the influence of noise and artifacts in the data. Due to the sampling rate of 500 Hz one trial of ERP data is represented by 500×17 features. As a way to improve the signal-to-noise ratio of the data, a spatial filtering method based on canonical correlation analysis (CCA) was applied [26] with a filter size of 27×17. The filter aims to minimize the variance within a class and to maximize the variance between classes to improve the separability.

Power spectrum

Classification using features from the frequency domain was conducted on the power spectra between 1-20 Hz, calculated on the same epochs as described above (0-1000 ms after stimulus onset) with Burgs maximum entropy method [20] with a bin size of 1. The same 17 channels were used resulting in 19×17 features.

Cross-Class classification

Since the question of separability aims to answer the diversity aspect of Miyake's model of executive functions, another approach needs to be introduced to answer the unity aspect of the EFs. For this, a cross-class classification was performed. Cross-class in this case means using a classifier across classes and not only for the classes it was trained for. General aims of this approach are, for example, to evaluate how well the trained classifier generalizes on data that shares the same underlying mental process but has been collected from different tasks. In BCI research this approach has practical reasons, regarding a minimization of training time of a classifier. If the classifier generalizes well across tasks, no or only little additional training is necessary on the new task to get the application running efficiently. Overall, this can be used to evaluate how much of the measured effects are task specific and how much can be accounted to the underlying shared mental process. Within this study, this approach was used to measure the shared and overlapping properties of the two EFs updating and inhibition. To simplify the explanation of the procedure, the two EFs are named EF1 and EF2. The placeholders EF1 and EF2 need to be seen as an abstract description, a concrete assignment of values or processes is not important for now. For this study, a classifier was trained on the distinction of EF1 vs. BL, and then tested on EF2. Theoretically, this means that all answers the classifier will give are wrong. However, if the two functions EF1 and EF2 have significant overlaps in their properties, EF2 should be classified as EF1 with above-average frequency. If there is no such overlap, the classification accuracy should be at a more random level, because no commonalities can be found and the criteria the classifier is based on are also based on random choices. This evaluation is done in both directions. Hence each EF is part of the train and the test set once. Cross-class classification is performed on ERP as well as on power spectral features. Again, for each classification the number of trials per class are balanced to avoid the overrepresentation of one of the classes.

7.2.3 Neural activation patterns

To inspect the features used for the distinction in the classification approach, a method developed by Haufe and colleagues [29] was used that transforms the weights of the SVM classifier into neurophysiological interpretable values, in so-called neural activation patterns. One classifier model is trained for each subject on the data of the respective categories. Applying the method on the model results is one activation value for each feature that was used in the classification. To create a comprehensive picture of the resulting neural activation pattern, the values are averaged within and according to the two frequency bands alpha (8-12 Hz) and theta (4-7 Hz). This is done for each subject individually, but the median values across subjects will be depicted in a color-coded topological distribution, to visualize the results. By calculating the activation patterns the underlying neurophysiological patterns that are responsible for the distinction can be inspected, which can provide valuable information analyzing the unity and diversity of different EFs.

7.3 Results

7.3.1 Neurophysiological data

ERPs

Figure 7.3 shows the grand average ERPs for the comparison of congruent and incongruent flankers, the comparison of inhibition and baseline demands, under three different n-back levels. It can be seen that statistically significant differences between the two flanker conditions can be found across all three n-back levels. However the amount of statistically significant different time points seems to decline with increasing n-back level. Figure 7.4 shows the grand average ERPs of the pairwise comparison of the three n-back levels, hence the comparison of updating and baseline demands, as well as the comparison of two different magnitudes of updating demands. Again it can be seen that all comparisons provide statistical significant differences, across the displayed electrode positions. Differences are largest for the Zero vs Two comparison (see Figure 7.4 b), followed by One vs Two and Zero vs One. The ERP at position PZ can be identified as a P300, equally for all comparisons.

Power spectra

Again, Figure 7.5 shows the comparison of congruent and incongruent flankers, hence the comparison of inhibition and baseline demands, under three different n-back levels. It can be seen that the comparison shows little to no statistical significant differences throughout the three n-levels in the power spectra. At n-level zero, differences can be found within occipital alpha (9-12 Hz) and within central theta (5-7 Hz). At n-level two, no statistical significant differences can be found. When assessing properties of the EF updating in the power spectra, Figure 7.6 shows that statistical significant differences occur in all made distinctions. Again differences are biggest for the comparison of Zero vs Two, representing the comparison baseline vs high updating demands.

To get a better overview of all available levels of the EFs in the data, the grand average ERPs and power spectra for four out of the six categories are shown in Figure 7.7. It can be seen that the four displayed conditions differ in amplitude, but the waveform remains rather constant. The amplitude of the identified P300 at position Pz decreases continuously with an increasing amount of load on the EFs. Also, the parietal/occipital change in alpha power can be identified as an ERD as the power decreases with an increasing amount of load, whereas the observed change in frontal theta power can be identified as an ERS since the power increases with the amount of load.

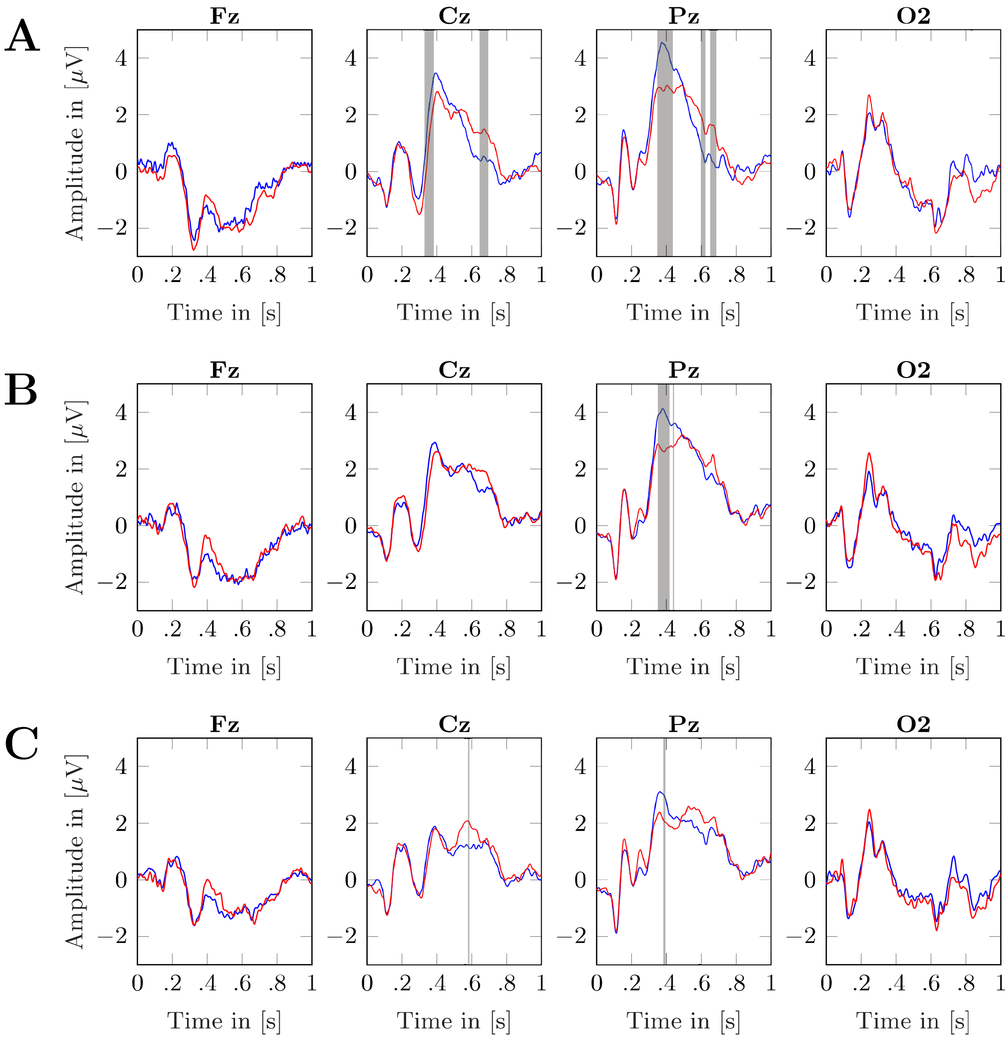


Figure 7.3: Grand average ERPs for Inhibition demands: Displayed are the electrode positions Fz, Cz, Pz and O2 during all three n-back levels. Each subfigure represents one of the n-back levels and displays a comparison of trials with congruent and incongruent flankers. The grand average has been calculated over all 22 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Zero, **B**: One, **C**: Two —: cong, —: incong

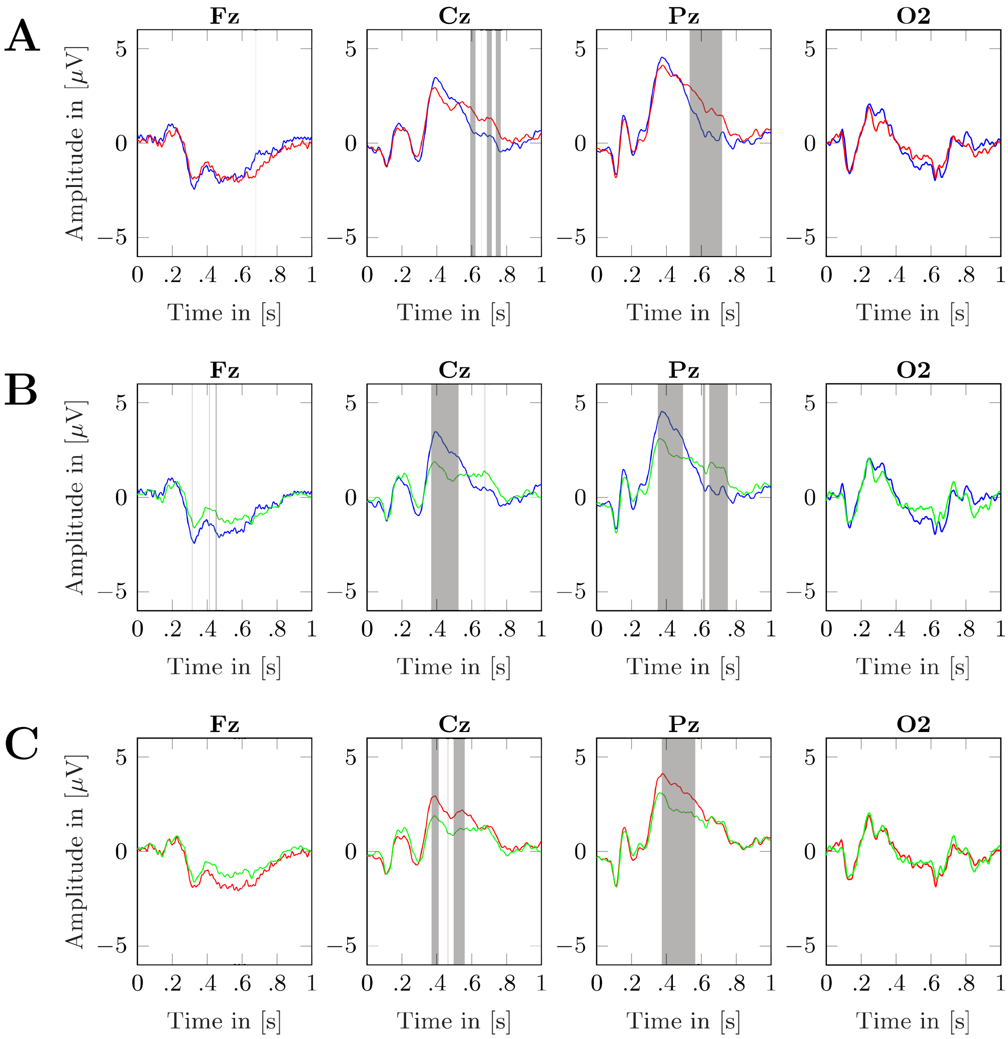


Figure 7.4: Grand average ERPs for Updating demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during a congruent Flanker. A pairwise comparison of trials between the n-back levels can be seen in the three subfigures. The grand average has been calculated over all 22 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). The three n-back levels are depicted as follows **A**: Zero vs One, **B**: Zero vs Two, **C**: One vs Two, —: zero, —: one, —: two

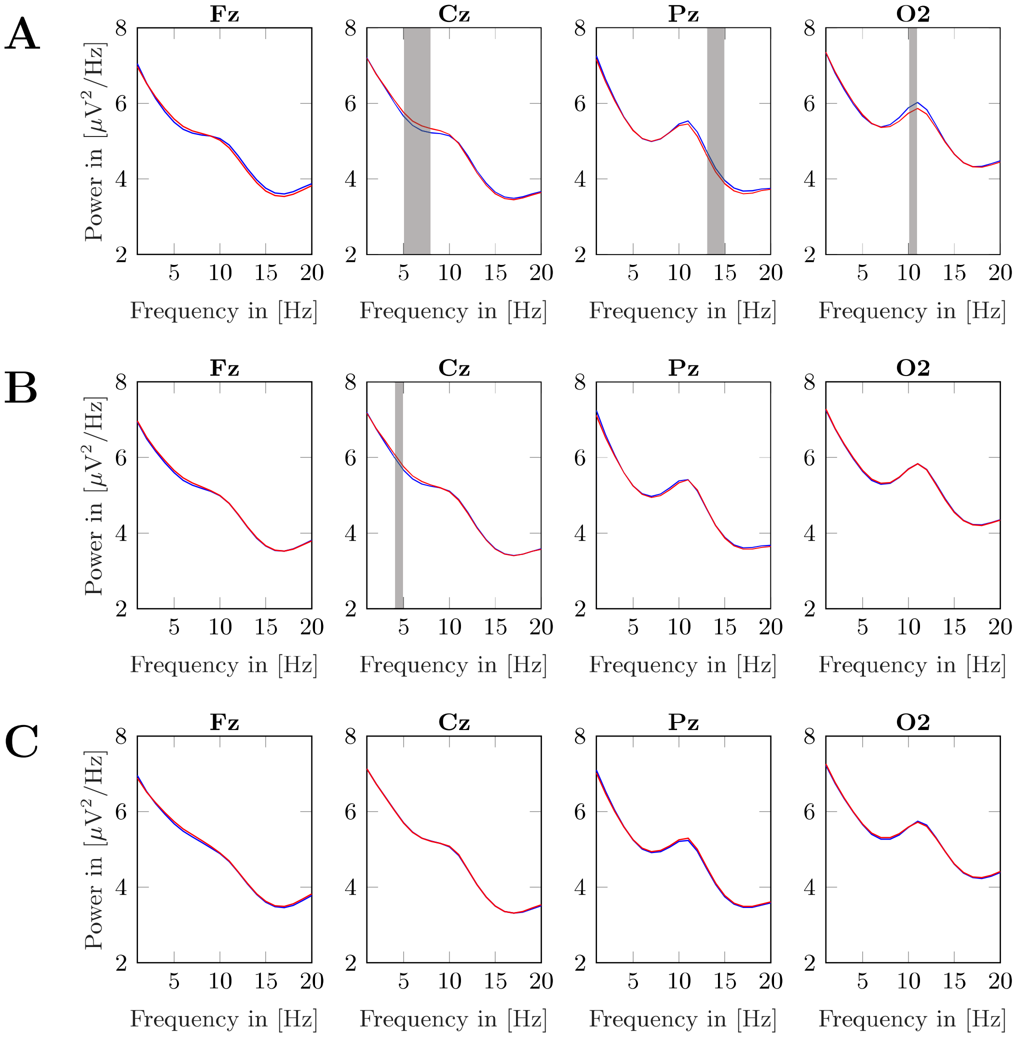


Figure 7.5: Grand average power spectra for Inhibition demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during all three n-back levels. Each subfigure represents one of the n-back levels and displays a comparison of trials with congruent and incongruent flankers. The grand average has been calculated over all 22 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points). The three n-back levels are depicted as follows **A**: Zero, **B**: One, **C**: Two, —: cong, —: incong

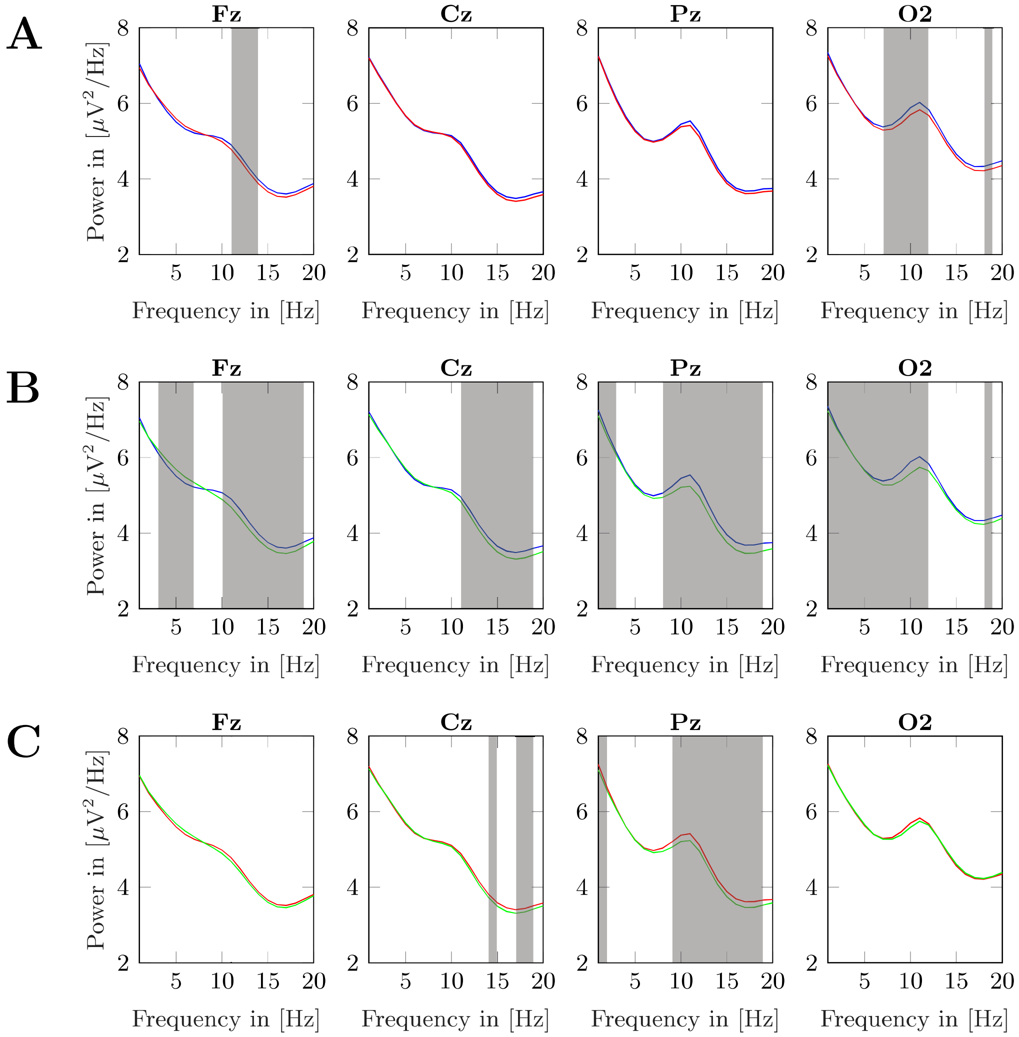


Figure 7.6: Grand average Power spectra for Updating demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during a congruent Flanker. A pairwise comparison of trials between the n-back levels can be seen in the three subfigures. The grand average has been calculated over all 22 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Zero vs One, **B**: Zero vs Two, **C**: One vs Two, —: zero, —: one, —: two

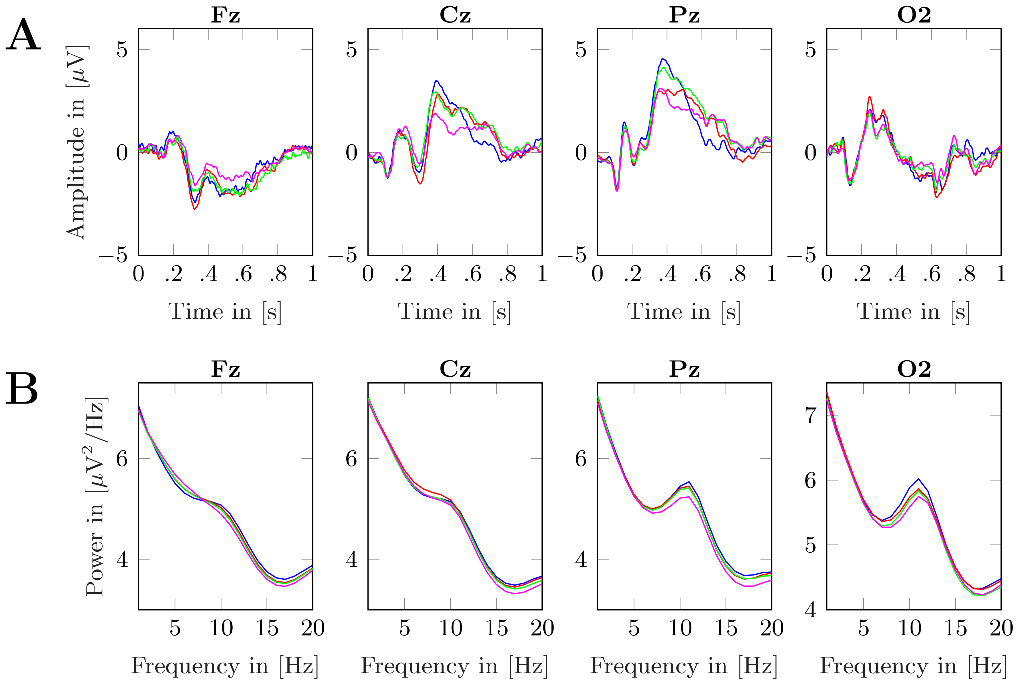


Figure 7.7: Grand average All in One: Displayed are the electrode positions Fz, Cz, Pz and O2. The grand average has been calculated over all 22 subjects. **A**: ERP, **B**: Spectra, —: zero cong, —: zero incong, —: one cong, —: one incong

7.3.2 ML based classification

The results for separating the EFs from baseline demands and from each other with the help of an SVM based ML classification approach can be seen in Table 7.1 - Table 7.3. The comparison of updating demands reaches values between 59.77 % and 69.75 % for ERP features and between 52.68 % and 66.32 % for spectral features. With minor exceptions, the comparison works equally well under a congruent and incongruent flanker, as can be seen in Table 7.1. The comparison between inhibition demands reaches accuracy values between 55 and 60 % for ERP features, and around 50 % for the power spectral features, which can be seen in Table 7.2. To evaluate if the EFs can be separated from each other, categories that each only induce demands on one of the two EFs, were compared to each other. Table 7.3 shows that the two EFs can be distinguished with up to 75 % accuracy on the basis of their ERP features and with up to 64 % on the basis of their spectral features.

Table 7.1: Classification updating demands - Classification accuracies achieved for the distinction of updating demands, from baseline and from each other, are shown with (incong) and without (cong) additional load on the EF inhibition. Results are the average performance of a SVM in a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. Significance level was determined to be at $p < 0.05$.

Flanker	Features	Zero vs One	Zero vs Two	One vs Two
cong	ERP (CCA)	61.20 %*	69.75 %*	63.59 %*
	Power (1-20)	52.68 %	63.17 %*	64.00 %*
incong	ERP (CCA)	59.77 %*	69.56 %*	63.58 %*
	Power (1-20)	57.31 %*	66.32 %*	63.23 %*

Table 7.2: Classification inhibition demands - Classification accuracies achieved for the distinction of inhibition demands, from baseline, are shown with (One, Two) and without (Zero) additional load on the EF updating. Results are the average performance of a SVM in a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. Significance level was determined to be at $p < 0.05$.

Features	cong vs incong	cong vs incong	cong vs incong
	Zero	One	Two
ERP(CCA)	60.56 %*	60.62 %*	55.01 %*
Power(1-20)	50.75 %	51.22 %	49.91 %

Table 7.3: Classification Inhibition vs Updating - Classification accuracies achieved for the distinction between updating and inhibition demands are shown. Results are the average performance of a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. Significance level was determined to be at $p < 0.05$

Features	Up1 vs Inh	Up2 vs Inh
ERP (CCA)	68.23 %*	74.73 %*
Power (1-20)	55.46 %*	64.13 %*

Neural activation patterns

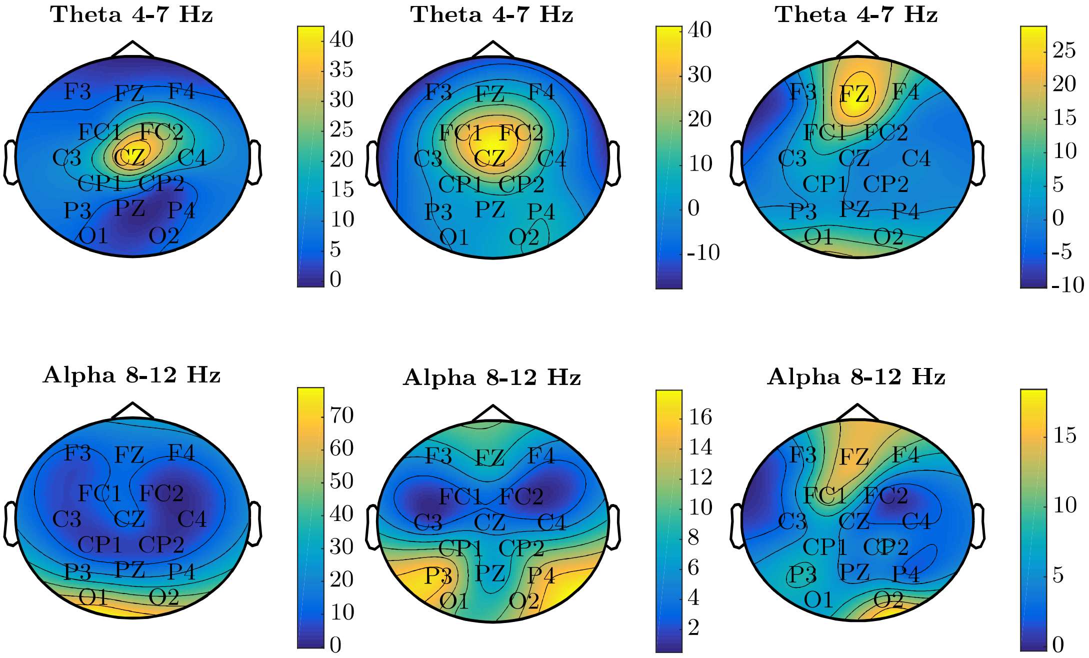
In addition to the classification performance, the weights of the used SVM classification approaches were also evaluated. After the transformation, they can be interpreted as neural activation patterns that describe the relevant differences between the tested conditions. First, the distinction of the EFs against baseline demands is evaluated. Figure 7.8 shows the neural activation pattern for the distinction of inhibition from baseline demands. It can be seen that theta power at CZ seems to discriminative for this distinction. For the distinction of updating from baseline demands, especially theta power at Fz seems to be a discriminating factor, as can be seen in Figure 7.9. In direct comparison, when trying to separate updating from inhibition demands, it can be seen that Theta power at Cz and Alpha power at Cz and Pz seem to be discriminating between the two EFs. Figure 7.10

Cross-class classification

The results obtained by cross-class classification, to reveal if joint feature characteristics are shared in large proportions by the two EFs, can be seen in Table 7.4. All results are close to random except the classifications performed with the model trained on Up2 vs. BL. The statistical analysis revealed that significantly more Inh trials had been assigned to the class BL and not to Up2. So the classifier trained on one EF does not recognize the demands imposed onto another EF. All other results do not provide statistical significance. These results provide evidence that the neurophysiological signatures of the two EFs are substantially different from each other, with a small number of joint features rendering a cross-class classification impossible.

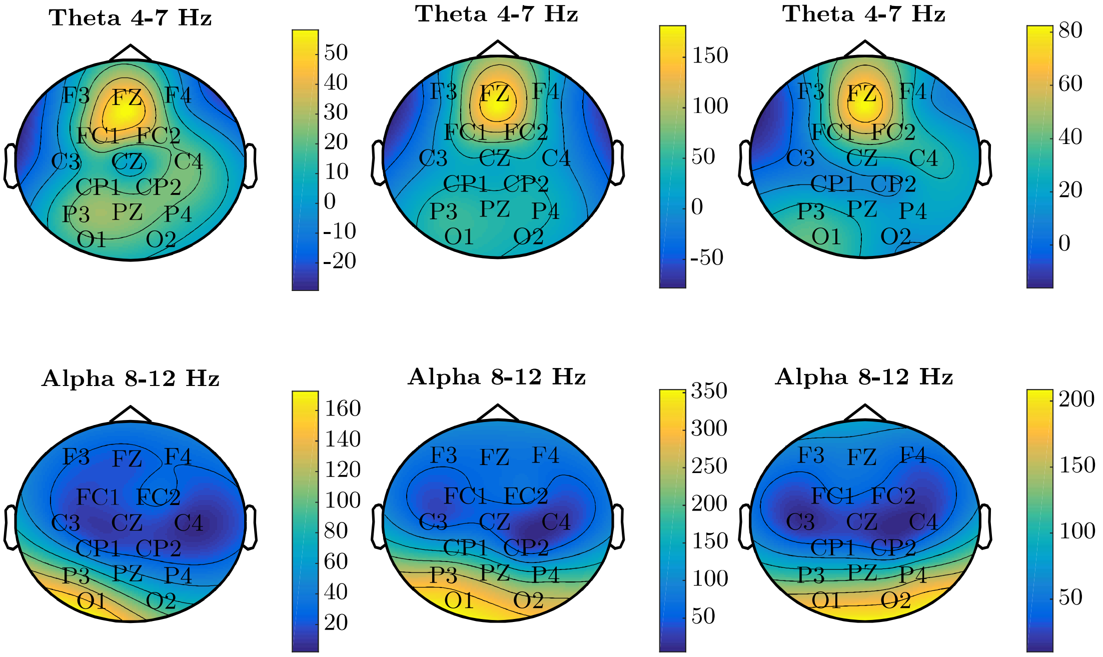
Table 7.4: Cross-class classification - The table provides cross-class classification with a SVM for ERP features as well as for the power spectra. The classifier was trained on Demand 1 vs BL and tested on trials belonging to Demand 2 only. Therefore, the here presented accuracies represent the percentage of trials classified as BL and 100 - the here displayed percentage reveals the share of trials classified as Demand 1 respectively. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. Significance level was determined to be at $p < 0.05$

Features	Trainset	Inh vs BL	Inh vs BL	Up1 vs BL	Up2 vs BL
	Testset	Up1	Up2	Inh	Inh
ERP		52.93 %	51.22 %	49.81 %	57.92 %*
Power (1-20)		51.55 %	47.86 %	52.84 %	61.26 %*



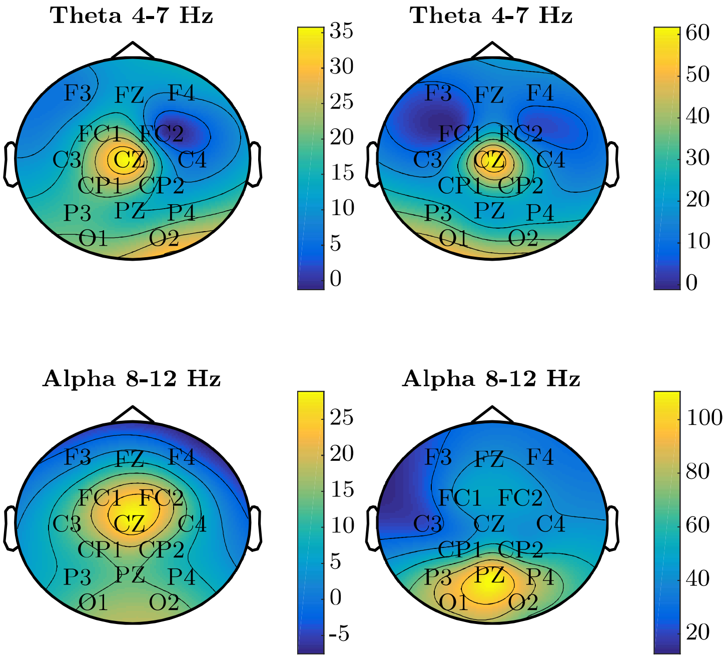
(a) Zero (b) One (c) Two

Figure 7.8: Neural activation pattern of inhibition demands - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of conflict conditions congruent (cong) and incongruent (incong) is shown for the three n-back levels. The resulting values are an average over the individual patterns of all 22 subjects. The three n-back levels are depicted as follows a) Zero, b) One , c) Two



(a) Zero vs One (b) Zero vs Two (c) One vs Two

Figure 7.9: Neural activation pattern of updating demands - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of n-back levels is shown for the congruent flanker condition. The resulting values are an average over the individual patterns of all 22 subjects. a) Zero vs One, b) Zero vs Two, c) One vs Two



(a) Up1 vs Inh (b) Up2 vs Inh

Figure 7.10: Neural activation pattern of EFs - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of updating (in two levels) and inhibition demands is shown in the following. The resulting values are an average over the individual patterns of all 22 subjects. a) Up1 vs Inh (One vs Incong), b) Up2 vs Inh (Two vs Incong)

7.4 Discussion

Can neural correlates for the EFs updating and inhibition be found in EEG data ? Are there indicators for the unity and diversity of the EFs ?

7.4.1 Neurophysiological analysis

The standard analysis of the ERPs averaged over the individual categories, reveals an elicited P300 which decreases in amplitude throughout the conditions. A similar pattern of results was found in the power spectra. The power spectra yield well-known indicators for WML, namely alpha desynchronization and theta synchronization cf., [100, 101]. For both feature sets, power spectra and ERPs, a statistically significant difference can be found for tested levels of load on each of the EFs. However, it is likely that the significant differences are due to the general amount of WML that is present throughout the task, and not necessarily due to the demands of the individual EFs. Since updating is induced by the main task and inhibition only by a secondary stimulus presentation to which no explicit response was necessary, it can be assumed that the induced load is likely to be higher for updating than for inhibition. Moreover, it is uncontroversial that updating demands elicited by the 2-back task are more challenging than the ones elicited by the 1-back task, as more letters need to be updated continuously in WM. A decrease in P300 amplitude, as is present in the data, is also well in line with previous findings from the literature concerning an increase of overall WML [100, 73]. Standard analysis techniques, therefore, reveal differences between the six categories, but they can only be reliably linked to the general amount of WML and not to specific properties of the individual EFs. To this end, the neurophysiological data analysis does not provide specific correlates for the EFs and also no indicators for unity or diversity of the two EFs, despite the found statistically significant differences.

7.4.2 ML based classification

It could be shown that a significantly better than random distinction with an SVM classification approach is possible between the EF conditions by single-trial ERPs and also single-trial power spectra. Therefore, it can be concluded that the statistically significant differences are not only present on group-level, but also a single-trial and single-subject level. A general observation that can be made is that the accuracy values mirror the gradient that was found in the measured physiological signals. The bigger the difference in induced WML between the conditions, the higher the accuracy for the respective distinction by SVM classification for both feature sets. The reliable generation of WML by both executive functions, which seem to differ mainly in the amount, is an indicator for the unity between the two EFs.

Concerning the comparably weak results for the EF inhibition, two reasons may be given. In a more recent version of model Miyake and colleagues [46] state that the inhibition ability is the core property of all EFs. Friedman and colleagues emphasize this statement by postulating that there is no unique variance describing inhibition [53, 52]. Potentially, this hypothesis of Miyake and Friedman could provide a theoretical explanation of why inhibition trials provided the overall weakest results, neurophysiologically as well as in terms of classification accuracy. However, it could also be argued that the secondary nature of the flanker stimulus inducing the inhibition demands is the reason for that. No explicit reaction to this stimulus was demanded in contrast to the central stimulus used for inducing updating demands. Designing a task that puts the inhibition demands more into focus might resolve this issue and reveal clearer and stronger neural correlates for inhibition.

To rule out that the differences that can be classified rely only on the amount of induced WML and not on individual variance caused by the EFs, as well as the fact that the EF inhibition does not have a unique variance, a cross-class classification was performed. The cross-class classification was used to indicate how much the characteristics of the EFs overlap. If signal strength caused by the overall WML, would be the only difference between the conditions, the overlap of characteristics should be large and cross-class classification accuracies should be significantly above chance level. Three out of four performed tests provided accuracies around the chance level for both types of features (ERPs and Power spectra). The only exception was a classifier trained on Up2 vs. BL and tested on Inh trials. It turned out that Inh is rather classified as BL than as Up2, revealing that Inh seems to be closer to BL demands than to Up2 demands. These results provide evidence for the hypothesis that there are larger differences in the signals reflecting individual variance and therefore the diversity of the EFs.

7.4.3 Neural activation patterns

In addition to cross-class classification, the neural activation patterns were inspected to find out which features are prominently used in afore performed classifications. Overall, the resulting values indicate that especially features in the occipital/parietal alpha and the frontal theta yielded the highest weights. These features are known from the literature to correlate with WML strongly. Apart from these WML related features, no other features seem to play a prominent role. This is an important insight as it makes the ML approach transparent and legitimizes the interpretation of the classification results concerning unity and diversity of the EFs. The neural activation patterns revealed that the main difference between the two EFs that can be assessed in the EEG signal is located in the theta band power. Inhibition correlates with central theta, whereas updating with a more frontal theta band power synchronization. This difference in feature characteristics renders the two EFs differentiable by their neural signatures, thereby, accounting for the diversity of the two EFs. Yet, the same values reveal a common correlation with occipital/parietal alpha desynchronization in both EF, hence accounting for the unity of these functions. Since the experimental design was chosen very carefully reducing all non-EF related variance to a minimum, it seems legitimate to assume that the discovered differences in the patterns can be traced back to the EFs and not to any confounds from external stimuli. Overall, it can be concluded that both aspects of the theoretical model put forward by Miyake [39], in which the unity and diversity of all three EFs is described, can be confirmed with the data of this study.

7.5 Conclusion

The data collected in this study and analyzed with conventional group-level statistics and machine learning techniques provide insights that support the theoretical model of Miyake and colleagues [39] describing the unity and diversity of EFs. It can be shown that the two executive functions updating and inhibition, which both induce WML, can be separated by single-trial ERPs and power spectra. Using power spectra yielded less accurate results but allowed to reveal patterns in the spectra that can be extracted and linked to the two individual executive functions. Inhibition is characterized by increased frontal activity in the theta band, whereas updating demands are characterized by increased central activity in the theta band. The results, substantiate the hypothesis that the two executive functions should be considered as two separable processes in WM. Applying machine learning techniques supplements the classical approaches, by taking the single subject level and the full pattern of the available data into account. In the example of this study it enabled to extract more knowledge out of existing data, then standard analysis techniques could have provided.

Chapter 8

Study 2: When Inhibition meets Shifting

Study 2 combines the EFs inhibition and shifting to investigate their properties concerning unity and diversity aspects. The design of the task is very closely related to the paradigm of Study 1, to keep the current and the respective study as comparable as possible. So far the results are unpublished and are therefore presented in full detail in the following sections.

Can correlates of unity and diversity for the two EFs shifting and inhibition be found in EEG data?

8.1 Task design

The experiment uses an integrated odd/greater and flanker task to study interactions between the two EFs shifting and inhibition. The odd/greater task is used to induce demands on the EF shifting, whereas the flanker task is used to induce demands on the EF inhibition. The simultaneous presentation of the two tasks was realized by showing seven items at once, from which one was positioned centrally, the other six on a flanking position to the left and right of the central item. The odd/greater task was the primary task in the experiment and performed on the central item. The flanker did not require any action from the subject and was only presented as a secondary task. The six flanking items, where congruent (identical) or incongruent (different) to the central item and fulfilled a distracting purpose. Therefore, the visual presentation is precisely the same as in study 1.

The stimuli were no longer letters but numbers in the range of 1 to 9, excluding 5. The instruction is to answer either of the two questions with yes or no by a button press on a standard keyboard (D and L): ”Is the central item greater than 5 ?” or ”Is the central item odd ?” Which of the two questions needs to be answered is defined and signaled by a cue which is presented shortly before each trial. Again a block design is used, presenting eight blocks in total. The eight blocks were presented divided into two parts, enabling a break after half of the experiment. Two blocks consisted of the odd task only, two of the greater than five task only and four blocks presented both tasks in an alternated and random but balanced order. In each block, 120 trials are shown

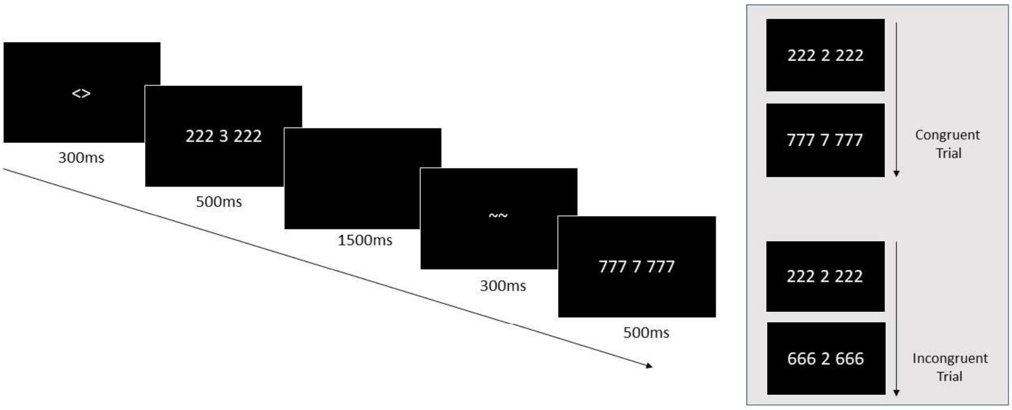


Figure 8.1: Experimental design - The task was presented on a black screen with white numbers as stimuli and flanker items. Stimuli were presented for 500 ms, followed by a blank black screen in which the subject needed to answer the question indicated by the cue with yes or no by pushing the respective key on the keyboard. The box on the right shows exemplary which trials were used in the analysis as congruent or incongruent trials. The cue '<>' posed the greater than 5 question, the '∼∼' cue posed the is the central item odd question.

consisting of a cue (<> for greater, ∼∼ for odd) presented for 300 ms, the stimulus presentation for 500 ms and a blank screen (1500 ms). Hence, experimental time accounts approximately for 40 minutes plus individual brake time in between blocks. The ratio for the congruency of the flanker items was on third congruent and two thirds incongruent. Four trials per block were randomly chosen to be presented without a central stimulus in which the subjects were asked to answer the greater or odd question with the flanking items. Each number was presented equally often as a central item, leading to a balanced amount of even and odd items as well as a balanced number of smaller and greater than 5 items. In addition, it was ensured that exactly half of the trials have to be answered with yes, to counterbalance the key presses. The assignment of the yes and no key (yes on 'D' or 'L') was counterbalanced across all subjects to avoid any bias due to handedness.

Before the start of the experiment, a short training phase was presented to familiarize the subjects with the task. Three blocks consisting of 24 trials each are presented to show each possible block once. During training feedback based on the performance is provided to the subject after each block to indicate to the subject and instructor that the task was fully understood and can be executed with sufficient accuracy. The subjects are also asked to rate the effort after each training block that is needed and how successful the subjects perceived their performance. During the real task, no feedback was provided.

8.1.1 Participants

21 subjects (18 females) participated in the study, for which they were reimbursed with 8 euro per hour. All subjects had normal or corrected to normal vision and no reported neurological disorders. The participants gave written and informed consent, and the study was approved by the local ethics committee. On average the subjects were 23.0 (±3.52) years old and all right-handed.

8.1.2 Technical setup

The subjects are seated in front of a computer screen (19 inches) on which the experiment was presented by the software E-Prime (Version 2.0.10.356). A standard keyboard was used for entering the answers, by which the correctness of an answer and the reaction time are assessed. For recording EEG, a Brain Products Acticap system with 32 electrodes was used and one Brain Products actiChamp amplifier which was sampled at 500 Hz (PyCorder). The integrated high pass filter was set to 0.1 Hz and the built-in low pass filter to 100 Hz. Additionally, a notch filter between 48-52 Hz was applied to eliminate power line noise. 28 electrodes were used for the recording and placed according to the extended 10-20 system [93] (FP1, FP2, F7, F3, FZ, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, PZ, P4, P8, O1, O2). The ground and reference electrodes were placed on the right and left mastoid respectively, and impedances were kept below 10 kΩ.

8.2 Data analysis

Data will be analyzed with respect to behavioral data, which includes reaction time and task accuracy as well as concerning physiological measures, including the EEG signal. Trials are categorized according to the presented flanker condition (cong and incong), the task (odd and greater) and the Shift which can occur trial wise (Repeat and Switch) or blockwise (Single and Mixed). Shifting demands are supposed to be induced when more than one set of rules needs to be applied. Within a series of shifting trials it can further be differentiated between trials in which the prior trial was from the same task (same rule set), which is then called a Repetition trial and trials in which the preceding trial is from a different task (different rule set) and is therefore called a Switch trial. Inhibition demands are supposed to be induced when the flanker is incongruent to the central stimulus. To

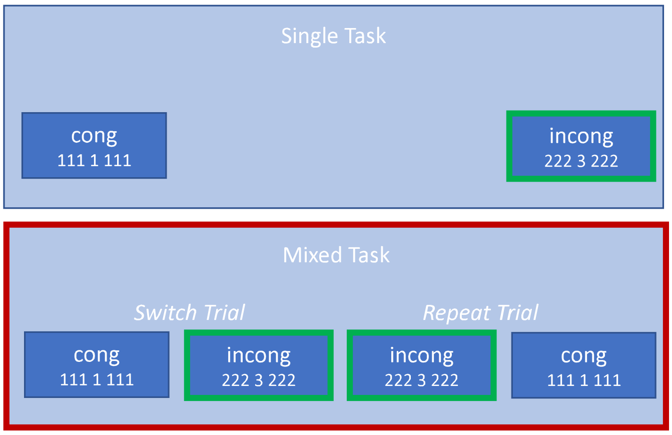


Figure 8.2: Datastructure - The data can be divided into two categories regarding the number of tasks within one block, which represent different levels of shifting demands. Each category can further be divided into two subcategories regarding the flanker condition, which represent different levels of inhibition demands. In total a set of six categories can be distinguished and will be used for analysis. 🟩: inhibition, 🟥: shifting

visualize the possible categories, Figure 8.2 shows the data structure of the experiment. To investigate the data concerning properties of the two demands trials are sorted and cut into epochs of 0-1300 ms after stimulus onset, based on six categories concerning the shifting and inhibition demands that are present during that trial. In this task design a blockwise manipulation of shifting demands, but also trial wise manipulation of shifting demands can be investigated. Therefore, blocks with a single task are compared to blocks with mixed tasks, as well as Switch vs. Repeat trials from mixed blocks. But baseline demands are also compared to pure inhibition and pure shifting demands.

8.2.1 Behavioral data analysis

In terms of behavioral data, RT and task accuracy can be investigated an compared to reveal differences and commonalities between the properties of the two executive functions shifting and inhibition. Statistical significant differences between the categories have been evaluated with an ANOVA, calculated on a linear regression model either on the RT or task accuracy.

8.2.2 Neurophysiological analysis

For the physiological data, only artifact-free trials with correct responses were used for data analysis, with additional exclusion of trials that might yield to any Gratton-like effect. The data was bandpass filtered between 0.4 - 40 Hz and re-referenced to common average. To remove artifacts, a threshold of 100 μV was chosen and all trials exceeding this level were discarded. Trials including eye movement artifacts were corrected using a regression method by Schloegl and colleagues [23]. A baseline correction was performed with 100 ms pre-cue onset.

In addition to this restriction, only trials that were preceded by another congruent trial have been selected. Incong trials are correctly answered trials with incongruent flankers from blocks with mixed flankers, again with the restriction that only trials which are preceded by a congruent trial are used for the analysis. In terms of physiological data, the grand average ERPs, as well as the grand average spectra can be computed for each of the six categories separately. Based on the results of [57], the ERPs at electrode positions FZ, CZ, PZ were of major interest. For the calculation of the power spectra Burgs maximum entropy method was used with a model order of 32 and a bin size of 1. To test whether the differences in the ERPs and power spectra between the factor levels are statistically significant a Wilcoxon rank sum test [102] was conducted over all subjects and trials. The resulting p-values were Bonferroni corrected [103] and the significance level was set to $p < 0.05$.

8.2.3 ML based classification

For the investigation of the EFs by means of machine learning, SVM classification was chosen. A SVM with a linear kernel (C = 1) [94], [8] was applied to differentiate between the six categories introduced above using the libsvm implementation for Matlab [95], [96]. The classification between categories was conducted for the following pairs for each subject individually:

- Inhibition: Cong vs. Incong
- Shifting: Single vs. Mixed, Switch vs. Repeat
- EFs: Shifting vs. Inhibition

The aim is to separate EF demands from baseline demands, but also different EF demands from each other. As the baseline demand within this study, congruent trials from the single blocks have been chosen, because neither shifting nor inhibition demands should be induced during those trials. For each data pair, a subset of the data is used to train a classifier, to learn the characteristics of each category. The remaining data is used to evaluate the success of the learning and hence the skill of the classifier. This is done on a single-trial and single-subject level. To ensure stable results 10-fold cross-validation was performed for each classification. The datasets (training set as well as the test set) were balanced for each comparison, by removing all spare trials if one of the classes had more trials than the other, to ensure that the distribution of examples per class does not influence the result. For classification, again two different types of features were used: CCA filtered ERP features [97] and power spectra. The performance of the classification approach is measured in accuracy, stating in how many cases the classifier categorized a trial correctly. Statistical significance of the classification results was determined by using permutation tests on the data.

The statistical significance of the results was determined by permutation tests with 1000 iterations [98, 99]. The classification performance achieved in the permutations establishes an empirical null distribution on random observations, which can be used to determine significance boundaries. Therefore, in each iteration classification was performed in a 10-fold cross validation, but with randomly assigned class labels in the training set instead of the correct class labels. The achieved accuracy values were compared with the ones determined in the standard 10-fold cross validation. Significance level was determined to be at $p < 0.05$, stating that the original classification performance is significant when the performance values are higher than the 95th percentile of the calculated empirical distribution.

ERPs

The ERP features upon which the SVM classifies are composed as follows. The 0-1000 ms epochs that have been prepared in the beginning will be used from 17 EEG channels (FP1,FP2,F3,FZ,F4,FC1,FC2,C3,CZ,C4,CP1,CP2,P3,PZ,P4,O1,O2). All other channels of the setup are discarded to reduce the influence of noise and artifacts in the data. Due to the sampling rate of 500 Hz one trial of ERP data is represented by 500×17 features. As a way to improve the signal-to-noise ratio of the data, a spatial filtering method based on canonical correlation analysis (CCA) was applied [26] with a filter size of 27×17. The filter aims to minimize the variance within a class and to maximize the variance between classes to improve the separability.

Power spectrum

Classification using features from the frequency domain was conducted on the power spectra between 1-20 Hz, calculated on the same epochs as described above (0-1000 ms after stimulus onset) with Burgs maximum entropy method [20]. The same 17 channels were used resulting in 19 × 17 features.

Cross-class classification

Since the question of separability aims to answer the diversity aspect of Miyake's model of executive functions, another approach needs to be introduced to answer the unity aspect of the EFs. For this, a cross-class classification was performed. In this approach, a classifier was trained on the distinction of EF1 vs. BL and tested on EF2. The reason for this is as follows. If the functions have significant overlaps in their properties, EF2 should be classified as EF1 with above-average frequency. If there is no such overlap, the classification accuracy should be at a more random level. This evaluation is done in both directions. Hence each EF is part of the train and the test set once. Cross-class classification is performed on ERP as well as on power spectral features.

8.2.4 Neural activation patterns

To inspect the features used for the distinction in the classification approach, a method developed by Haufe and colleagues [29] was used that transforms the weights of the SVM classifier into neurophysiological interpretable values, in so-called neural activation patterns. One classifier model is trained for each subject on the data of the respective categories. Applying the method on the model results is one activation value for each feature that was used in the classification. To create a comprehensive picture of the resulting neural activation pattern, the values are averaged within and according to the two frequency bands alpha (8-12 Hz) and theta (4-7 Hz). This is done for each subject individually, but the median values across subjects will be depicted in a color-coded topological distribution, to visualize the results. By calculating the activation patterns the underlying neurophysiological patterns that are responsible for the distinction can be inspected, which can provide valuable information analyzing the unity and diversity of different EFs.

8.3 Results

8.3.1 Behavioral data

In Table 8.1 the average accuracy and reaction time of the subjects can be seen, separated according to the different levels of shifting and inhibition demands. Subjects were significantly slower for trials in mixed blocks (blocks in which both tasks were performed alternately) compared to single blocks (only one task needed to be performed). Subjects were also much faster for repetition trials, compared to switch trials. The difference in reaction time yielded a p-value of 0.056 and is therefore not statistically significant. However, it is very close to the chosen threshold of $p = 0.05$ that it can be hypothesized that a non-random effect might be present. Regarding the inhibition demands it can be stated that subjects are faster for congruent trials compared to incongruent trials, but in this study the difference is not significant. Task accuracy reveals the same tendencies, showing statistically significant differences for the block wise shifting demands (Single vs Mixed), an almost significant difference for the trial wise manipulated shifting demands, but no significant effect for the inhibition demands. The p-values for all performed tests can be seen in Table 8.2.

Table 8.1: Behavioral accuracy and reaction time - Average accuracy (Acc) and reaction time (RT) of the subjects categorized according to the flanker condition (cong, incong), shift condition (Single, Mixed, Switch or Repeat) and task (odd or greater).

	Task	Flanker	Single	Mixed	Repeat	Switch
Avg Acc [%]	Greater5	cong	94	90	93	88
	Greater5	incong	92	90	89	91
	OddEven	cong	92	87	92	84
	OddEven	incong	89	83	85	81
	Both	cong	93	89	92	86
	Both	incong	91	86	87	85
Avg RT [ms]	Greater5	cong	504.93	653.12	627.38	669.67
	Greater5	incong	526.54	651.78	631.00	670.59
	OddEven	cong	557.20	681.28	632.43	709.68
	OddEven	incong	590.18	721.93	684.60	753.69
	Both	cong	531.06	667.20	629.90	689.67
	Both	incong	558.36	686.85	657.80	712.14

Table 8.2: ANOVA on behavioral data - P-Values calculated for the average reaction times (RT) and accuracies (ACC) of all subjects per condition. An ANOVA was performed on a linear regression model, taking the Task, flanker condition and the Shift into account. Significance level has been determined to be at $p < 0.05$.

	Shift (Block)	Shift (Trial)	Task	Flanker
RT	< 0.05	0.06	< 0.05	0.23
Acc	< 0.05	0.07	< 0.05	0.08

8.3.2 Neurophysiological data

ERPs

Figure 8.3 shows the ERPs for the comparison of inhibition and baseline demands, whereas Figure 8.4 shows the ERPs for the comparison of the shifting and baseline demands. For the EF inhibition it can be seen that small differences exist between the ERPs at the four presented electrode positions. The differences are more pronounced when performing a single task, compared to multiple tasks and they are mainly present at electrode position Cz. With regard to the EF shifting it can be seen that major differences exist for the Single vs. Mixed comparison at several points in time at all electrode positions. The differences seem to be again, most pronounced at position Cz. In contrast to that, the Switch vs. Repeat comparison does not provide any statistically significant differences at the four examined electrode positions. This mirrors the results that have been found in the analysis of the behavioral data.

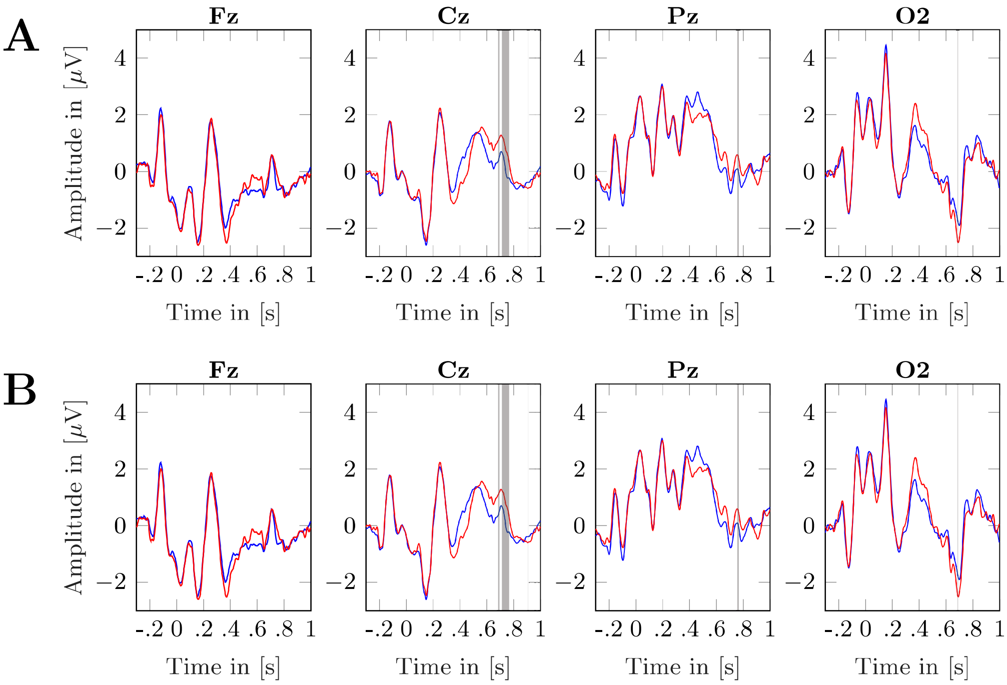


Figure 8.3: Grand average ERPs for inhibition demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during Single and Mixed blocks. A pairwise comparison of trials with congruent and incongruent flankers can be seen in the two subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Single, **B**: Mixed, —: cong, —: incong

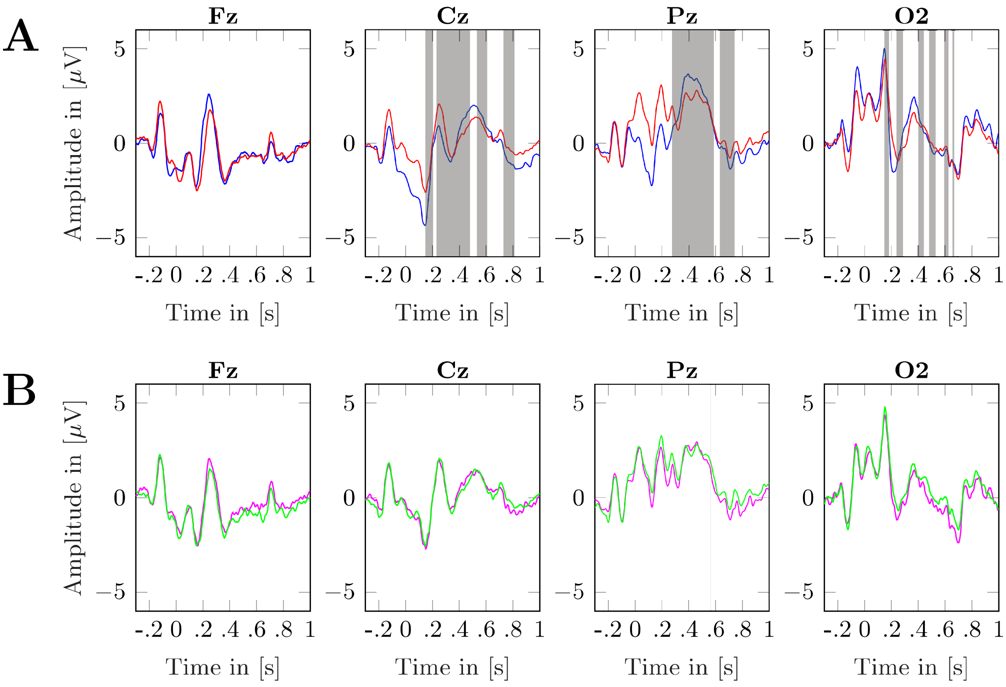


Figure 8.4: Grand average ERPs for Shifting demands - Displayed are the electrode positions Fz, Cz, Pz and O2. A pairwise comparison of trials with different shifting levels all with congruent flanker can be seen in the two subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points).
A: Switch vs Repeat, **B**: Single vs Mixed, —: single, —: mixed, —: repeat, —: switch

Power spectra

The same comparisons have also been done for the signal in the power spectra. Figure 8.5 shows the comparison of inhibition and baseline demands. It can be seen that the two conditions vary in the alpha and beta range, at all displayed electrode positions. Figure 8.6 shows the comparison of shifting and baseline demands. In the blockwise comparison (Single vs Mixed), which can be seen in subfigure A, statistically significant differences can be seen throughout all displayed electrode positions, especially in the alpha and theta range. For the trial wise comparison (Switch vs Repeat) no differences can be found. Again this mirrors the results found in the ERPs and also in the behavioral data.

For the overall comparison and to get a complete picture, the grand average ERPs and power spectra for four out of the six categories are shown in Figure 8.7. It can be seen that the four conditions differ in amplitude, but the waveform remains rather constant. Mixed ERPs are more positive in amplitude at Cz and Pz. The parietal/occipital change in alpha power can be identified as an ERD as the power decreases with an increasing amount of load. The change in frontal theta power can be identified as an ERS since the power increases with the amount of load.

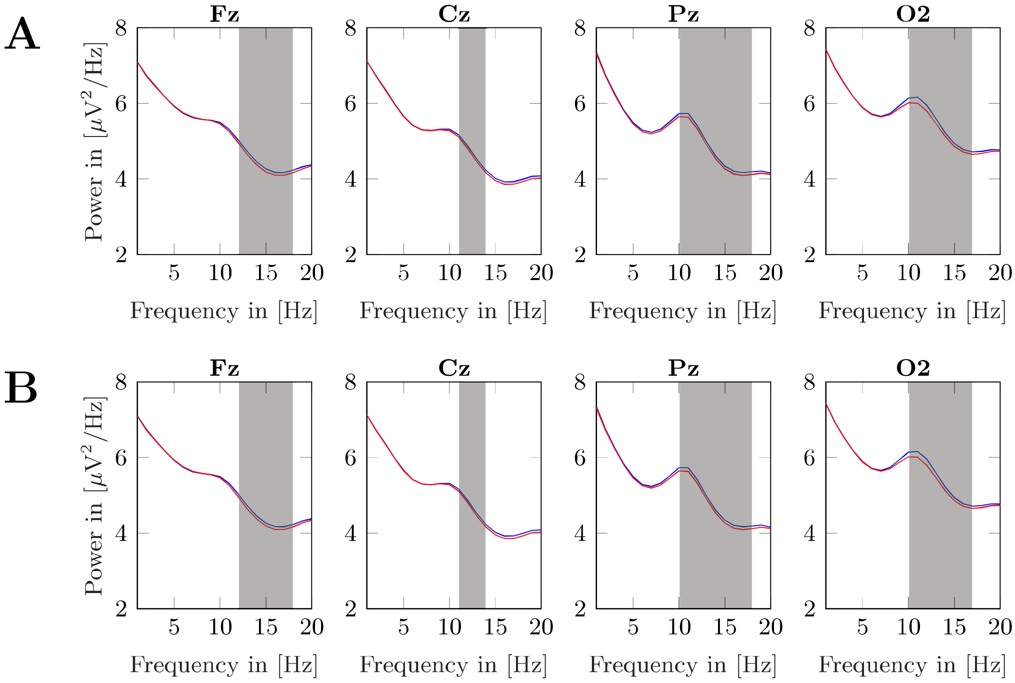


Figure 8.5: Grand average Power spectra for inhibition demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during Single and Mixed blocks. A pairwise comparison of trials with congruent and incongruent flankers can be seen in the two subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Single, **B**: Mixed, —: cong, —: incong

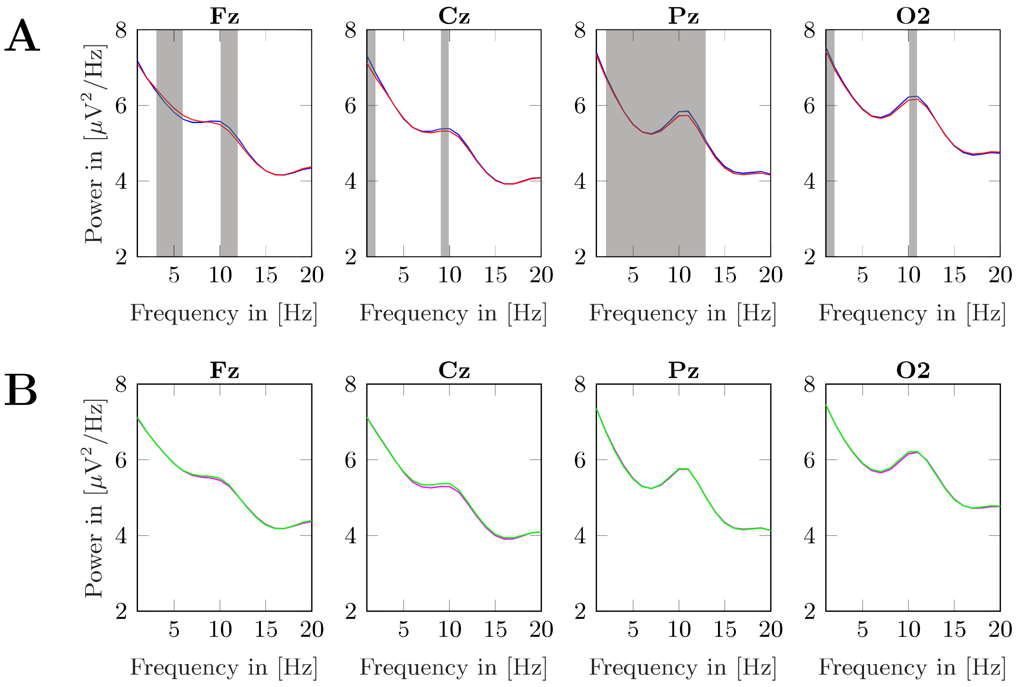


Figure 8.6: Grand average Power spectra for shifting demands - Displayed are the electrode positions Fz, Cz, Pz and O2. A pairwise comparison of trials with different shifting levels can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points).
A: Switch vs Repeat, **B**: Single vs Mixed, —: single, —: mixed, —: repeat, —: switch

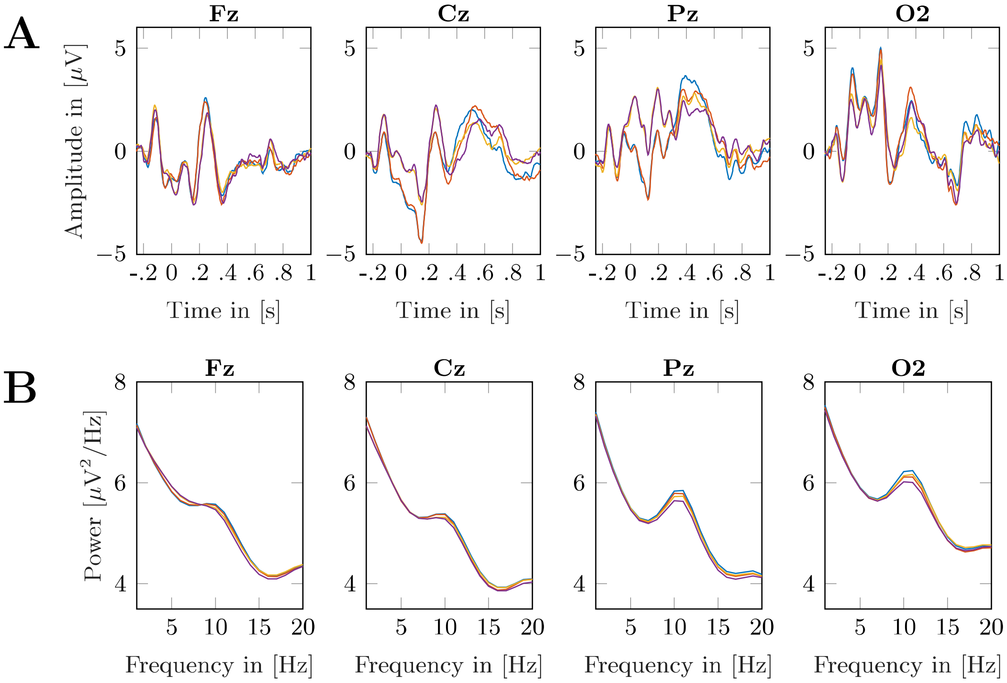


Figure 8.7: Grand average All in One - Displayed are the electrode positions Fz, Cz, Pz and O2. The grand average has been calculated over all 21 subjects.
A: ERP, **B**: Spectra, 🟦: single cong, 🟥: single incong, 🟨: mixed cong, 🟪: mixed incong

8.3.3 ML based classification

Classification results describe how well data could be discriminated on a single trial and single subject basis. Table 8.3 - Table 8.5 provide the accuracy values for three distinctions concerning the level of shifting and inhibition demands. First, again the inhibition demands are compared to baseline demands which can be seen in Table 8.3. It can be seen that all results have low accuracy values below 60 %. A general observation is that accuracy tends to be higher for ERP than for the power spectral features. Secondly, it was evaluated how well shifting demands can be separated from baseline demands. Table 8.5 shows the trial wise manipulation of shifting demands, the Switch vs. Repeat trials. Independent of present or absent inhibition demands none of the performed classifications reaches chance level, revealing that a distinction on single trial level is not possible. The results for the block wise manipulation of shifting demands compared to baseline demands can be seen in Table 8.4. Interestingly, in this case the classification exceeds chance level in all cases and reaches maximum values of 75.22 %. As before, ERP features can be distinguished with higher accuracy rates than power spectral features, independent of present or absent inhibition demands. When trying to separate shifting from inhibition demands, once while using all congruent trials from mixed blocks and once while using only congruent switch trials it can be seen that the distinction is possible with accuracy values up to 77 % (see Table 8.6). Both cases work equally well, stating that the neural correlates of the two EFs can be separated.

Table 8.3: Classification inhibition demands - Classification accuracies achieved with ML approach with (Mixed) and without (Single) an additional load factor of shifting. Displayed is the classification accuracy achieved with a SVM and a linear kernel during a 10 fold cross-validation. Accuracy is reported for the two tasks (OddEven and Greater5) individually and combined (both). The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	cong vs incong Single	cong vs incong Mixed	cong vs incong Both
OddEven	ERP (CCA)	55.84 %*	53.21 %	60.68 %*
	Power (1-20)	49.93 %	54.77 %*	53.56 %
Greater5	ERP (CCA)	57.04 %*	55.44 %*	57.86 %*
	Power (1-20)	49.83 %	53.31 %	52.77 %
Both	ERP (CCA)	57.89 %*	56.78 %*	59.67 %*
	Power (1-20)	51.37 %	52.02 %	52.53 %

Table 8.4: Classification shifting demands (block) - Classification accuracies achieved with ML approach with (incong) and without (cong) an additional load factor of inhibition. Displayed is the classification accuracy achieved with a SVM and a linear kernel during a 10 fold cross-validation. Accuracy is reported for the two tasks (OddEven and Greater5) individually and combined (both). The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	Single vs Mixed cong	Single vs Mixed incong	Single vs Mixed both
OddEven	ERP (CCA)	72.22 %*	71.24 %*	74.70 %*
	Power (1-20)	59.10 %*	59.08 %*	60.53 %*
Greater5	ERP (CCA)	73.13 %*	70.34 %*	75.66 %*
	Power (1-20)	61.85 %*	56.67 %*	61.88 %*
Both	ERP (CCA)	75.22 %*	72.41 %*	76.98 %*
	Power (1-20)	61.14 %*	58.30 %*	62.63 %*

Table 8.5: Classification shifting demands (trial) - Classification accuracies achieved with ML approach with (flanker = incong) and without (flanker = cong) an additional load factor of inhibition. Displayed is the classification accuracy achieved with a SVM and a linear kernel during a 10 fold cross-validation. Accuracy is reported for the two tasks (OddEven and Greater5) combined only due to a insufficient number of trials. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$

Task	Features	Repeat vs Switch cong	Repeat vs Switch incong	Repeat vs Switch both
Both	ERP (CCA)	53.26 %	52.40 %	54.88 %
	Power (1-20)	49.37 %	49.28 %	50.08 %

Table 8.6: Classification Inhibition vs Shifting - Classification accuracies achieved with ML approach. Displayed is the classification accuracy achieved with a SVM and a linear kernel during a 10 fold cross-validation. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	Inh vs Switch	Inh vs Mixed
Both	ERP (CCA)	77.56 % *	76.44 % *
	Power (1-20)	59.39 % *	60.68 % *

Neural activation patterns

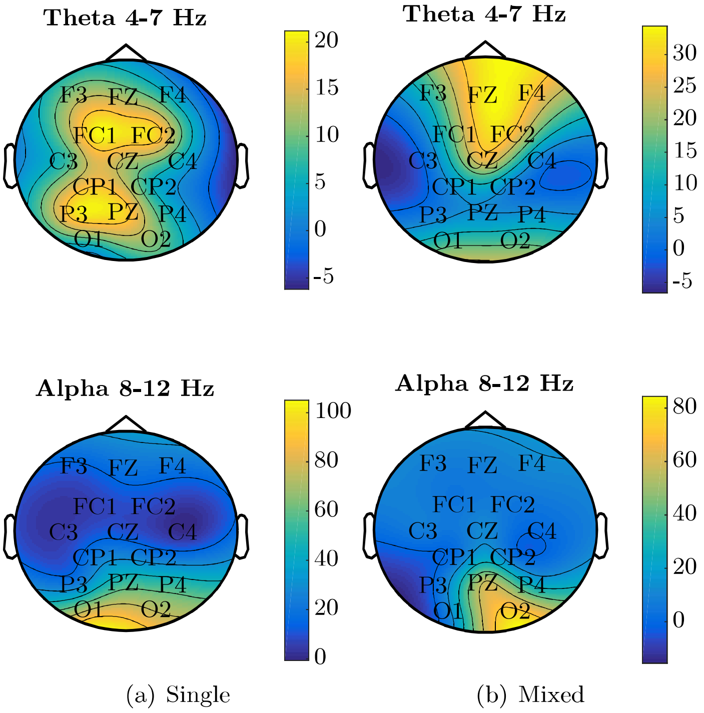
The neural activation patterns that correspond to the SVM classifications can be seen in the following. When trying to separate congruent from incongruent trials (inhibition form baseline demands) it can be seen in Figure 8.8 that frontal but also parietal theta, and occipital alpha seem to provide crucial and discriminative properties. For the distinction of shifting vs. baseline demands it can be seen in Figure 8.9 that parietal theta and occipital alpha seem to play an important role. In direct comparison it can be seen that especially central theta is discriminative between the two EF (see Figure 8.10).

8.3.4 Cross-class classification

Last but not least the results for the cross-class classification are shown in Table 8.7. They provide the last piece of evidence regarding the unity and diversity of inhibition and shifting for the performed study. It can be seen that all results are close to random, which make it seem like there is no significant amount of joint features, that could lead to a confusion of the two EFs.

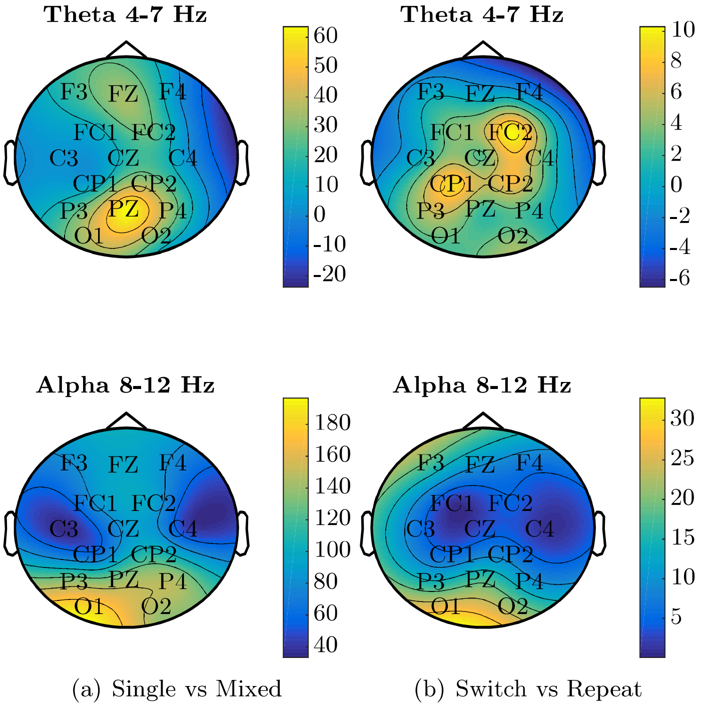
Table 8.7: Cross-class classification - The table provides cross-class classification with a SVM for ERP features as well as for the power spectra. The classifier was trained on Demand 1 vs BL and tested on trials belonging to Demand 2 only. Therefore, the here presented accuracies represent the percentage of trials classified as BL. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Features	Trainset	Inh vs BL	Inh vs BL	Single vs Mixed	Repeat vs Switch
	Testset	Mixed	Switch	Inh	Inh
ERP (CCA)		48.85 %	50.69 %	47.84 %	50.70 %
Power (1-20 Hz)		54.27 %	48.97 %	49.85 %	47.72 %



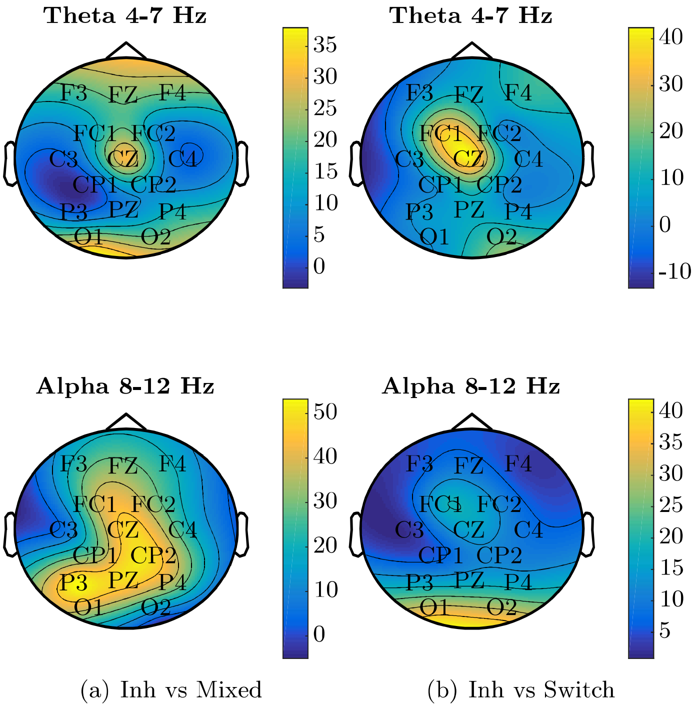
(a) Single (b) Mixed

Figure 8.8: Neural activation pattern inhibition demands - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of conflict conditions congruent (cong) and incongruent (incong) is shown for the shifting levels Single and Mixed. The resulting values are an average over the individual patterns of all 21 subjects. a) Single , b) Mixed



(a) Single vs Mixed (b) Switch vs Repeat

Figure 8.9: Neural activation pattern shifting demands - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of shifting conditions Single and Mixed, as well as Switch vs Repeat is shown for the congruent flanker condition. The resulting values are an average over the individual patterns of all 21 subjects. a) Single vs Mixed, b) Switch vs Repeat



(a) Inh vs Mixed

(b) Inh vs Switch

Figure 8.10: Neural activation pattern Inhibition vs Shifting - Displayed is the color coded activation pattern A calculated from the weights of the SVM, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of shifting and inhibition demands . The resulting values are an average over the individual patterns of all 21 subjects. a) Mixed vs Incong, b) Switch vs Incong

8.4 Discussion

Can neural correlates for the EFs shifting and inhibition be found in EEG data ?
Are there indicators for the unity and diversity of the EFs ?
Can the results be compared to study 1?

8.4.1 Behavioral data

The analysis of the behavioral data showed statistically significant effects for shifting demands compared to baseline demands when manipulated block wise, and almost but no statistically significant effects for trial wise manipulation of shifting demands. The manipulation of inhibition demands did not show significant effects. The lack of statistical significance for the manipulation of inhibition demands can be due to the secondary nature of the flanker task. No explicit reaction is required, therefore the focus is clearly on the primary task odd or greater. Since the odd or greater task, especially when the two are alternating, is already challenging it could also be hypothesized that a limit is reached regarding the capacity of working memory. This would be in line with a study from Sorqvist et al. which showed that task-irrelevant auditory stimuli caused less interference under high as compared to low visual WM updating load [104].

When looking at the shifting demands, it can be seen that differences exist for both types of manipulations (block and trial wise) when analyzing behavioral data. This is well in line with the literature [85], which states the differences between blocks can be observed, but also that the immediate shift of task also results in a measurable differences regarding reaction time. The two manipulations, block and trial wise, can be summarized under the term alternation cost, which is composed of switching and mixing costs. The block wise manipulation and the resulting differences in RT between single and mixed blocks can be described as mixing costs. Switch costs account for the differences in performance due to immediate and trial wise switches of tasks and rule sets. That the switch costs are not statistically significant might be due to the task ambiguity. The used stimuli are exact the same in both tasks which can lead to so-called cross-talk effects. Both task sets are active at the same time and only the cue indicates which needs to be used to solve the task correctly. It has been shown that there are significant mixing cost when mixing two unambiguous tasks [87].

Another effect independent of shifting and inhibition demands is the differences between the odd and greater task. The data suggests that effect is significant for odd but not for the greater task. Answers for the greater task were given much faster than for the odd task. There is one possibility that can be the cause of this, which can be found under the term spatial-numerical association of response code (SNARC). The SNARC effect states that when dealing with numbers a mental number line is active in the background. Due to this reactions to bigger numbers are faster with the right hand, whereas reactions to smaller numbers are faster with the left hand [105]. Overall, it has been found that the judgment of number magnitude is activated even when it is not relevant for the task. Therefore, it is easier to blend out a flanking item if the distance (difference) between the target and the flanker item is big compared to when it is small [106],[107]. A closely related effect is the linguistic markedness association of response codes (MARC), which relates the parity of numbers with the response hand. It states that even numbers are more associated with the right and are therefore answered faster with the right hand compared to odd numbers which are answered faster with the left hand [108]. Since this aspect of the data was not of relevance for the posed research question, no tests that would confirm the presence or absence of these effects were performed. For future studies, this relations should be kept in mind for a thorough and more refined task design that controls for those effects.

8.4.2 Neurophysiological data

The neurophysiological data creates a similar impression compared to the behavioral data. For the EF inhibition almost no differences can be found in the ERPs but some differences in the power spectra when comparing it to baseline demands. Since similar results have already been found in Study 1, this is not surprising. Here, too, the secondary nature of the flanker task could play an important role. When comparing the demands of the EF shifting, differences for single vs. mixed comparisons (blockwise manipulation) can be found in the ERP as well as in the spectra. Interestingly for the switch vs. repeat comparison (trialwise manipulation) no differences can be found at all. As already stated in the discussion of the behavioral data, the task ambiguity could play a major role for the not existing difference for trialwise manipulations. Concerning the very pronounced difference in the blockwise manipulation the factor of task uncertainty can be named. In the mixed blocks the two tasks were both presented equally often but in a random manner, which task will be presented in the next trial was therefore always uncertain. In single blocks in which only one task was presented, no task uncertainty prevailed. Rubin and Meiran [87] found indicators that especially task uncertainty and differential general control effort might be the main trigger for mixing costs, which is also reflected in neurophysiological signals. The general control effort is well reflected in the ERPs, but also in the spectra, resulting in an overall higher WL for the EF shifting compared to inhibition. Classical ERS/ERD but also decreased P300 amplitudes could be found showing again that the induction of WML is a common trait of EFs. Overall the hypothesis arises that block wise manipulation compared to trial wise manipulation creates different patterns of behavior and neurophysiological signals. The different anticipation and expectation of task demands seems to control the behavior and therefore also the level of attention that is paid towards to potentially conflicting stimulus.

8.4.3 ML based classification

Using ML based classification as an analysis tool, provided in many aspects the same results as the classical group-level analysis. For the distinction of the EF inhibition from baseline demands chance level performance has been found in the most cases, likewise for the switch vs. repeat comparison of the EF shifting. In some cases the performance exceeds the significance threshold but still, the values are close to 50 %. It can therefore be assumed, that the addition of the single-subject level is not helpful for this particular task. For the distinction of shifting vs baseline demands accuracies above 70 % have been reached for ERP features, stating that this is a stable and pronounced difference. Again it needs to be hypothesized that it is mainly the general amount of load that is classified and distinguished in this case and not necessarily the specific properties of the EFs.

When classifying the two EFs against each other, however, high performance values above

75 % were achieved. This indicates that the two EFs can be distinguished. The cross-class classification results support this finding, since there only performance values around chance level have been found. This is a crucial indicator that the two EFs do not share a significant amount of joint features, supporting the diversity of the two EFs inhibition and shifting.

8.4.4 Neural activation patterns

The neural activation patterns that have been established from the weights of the SVMs suggest that again only features are of relevance in the distinction of the EFs that highly correlate with WML. This corroborates the results from study 1 and further legitimizes this approach and interpreting the patterns as characteristic properties of the EFs. Again it seems that for the EF inhibition frontal theta plays a major role, whereas for shifting rather a central or parietal theta are of relevance. Especially when comparing the two EFs against each other, a main difference can be seen at central theta. Doing this comparison with switch and repeat trials reveals also a discriminative component in the central/parietal alpha range, which is a lot less pronounced when using switch trials only.

8.5 Conclusion

The data collected in this study and analyzed with conventional group-level statistics and machine learning techniques provides insights that support the theoretical model of Miyake and colleagues describing the unity and diversity of EFs. Similar but not identical results have been achieved compared to study 1. Inhibition is characterized by frontal theta activation which cannot be located as precisely as in study 1. Shifting is characterized by parietal theta. No significant cross-class classification has been found suggesting that no joint features between the two EFs exist that make them interchangeable. Again the induction of WML has been found as a common trait between EFs that modulates the neural correlates (unity) but nevertheless a clear distinction between the EFs was possible (diversity). More general, it was found that a blockwise manipulation of demands has a greater effect than a trial wise manipulation. Due to this finding, it seems relevant to reevaluate the task design concerning this factor, to avoid misguided hypothesis.

Chapter 9

Study 3: When Inhibition meets Updating II (Between vs Within Block effects)

Study 3 again investigates the unity and diversity of the EFs updating and inhibition. Since the results from Study 2 brought a new perspective on the so far performed studies, a follow-up study was designed. Study 2 revealed that a blockwise manipulation of EF demands, in this case shifting, works well, whereas a within block manipulation does not. Due to task design, an equivalent comparison for inhibition demands from Study 1 or 2 cannot be made. Neither can they for updating demands of study 2. Due to the nature of an n-back task, updating demands need to be manipulated blockwise, else the task cannot be solved. Flanker items, however, are constantly varied within each block, but a variation between blocks can easily be realized. The results achieved in Study 2 give rise to the assumption that a between blockwise manipulation of inhibition demands might be a useful extension of the analysis. Therefore, study 1 was extended by one more condition that included blocks in which the flanker items simply did not alternate between congruent and incongruent. A blockwise comparison of inhibition demands, thereby gets feasible. For simplicity and to avoid an extensive duration of the experiment, only blocks in which no inhibition demands are induced are presented. Hence, the flanker items remained congruent throughout the full block. The results of this Study have in part been published in [109], but will now be presented in full detail.

Is there a block effect for inhibition demands in the modified flanker task ?

9.1 Task design

The experiment is based on the task design of Study 1, which was originally reported in [57]. It combines two tasks: the n-back and the Eriksen-Flanker task. The simultaneous presentation of the two tasks was realized by showing seven items at once, from which one was positioned centrally, the other six on a flanking position to the left and right of the central item. The n-back task was performed on the central item and used as the primary task in the experiment. The flanker was only used as a secondary task, to which no action was required. The six flanking items, where therefore congruent (identical) or incongruent (different) to the central item and fulfilled a distracting purpose.

Again the experiment was presented blockwise, this time with six blocks in total. The number of n-back levels was reduced from three to two (0 and 1), where each level was presented three times. For both n-levels, one block consisted of congruent flankers only, and the two other blocks consisted of alternating congruent and incongruent (mixed) flankers. To quickly repeat the task of the subjects: In each trial, subjects were asked to decide whether the central item is equal to the one presented n-steps before. Therefore, in each trial, an answer of yes or no was required by button press (keys D and L on a standard keyboard). Yes and no answers were randomly distributed over each block with a ratio of 1:1 and given with the index finger of either the right or left hand. Which key represented the yes answer was counterbalanced throughout all subjects. Each block included 120 trials, of two seconds length. One trial consisted of 500 ms stimulus presentation and a 1500 ms long blank screen, hence one block was 4 minutes long. In study 3 trials with no central item were omitted. All other details regarding the implementation have been kept equal, therefore, for more detail see the original publication [57] or [90]. To familiarize the subjects with the experiment, each subject had to perform training before starting the task. The training consisted of 2 short blocks (24 trials), one for each n-back level (0 and 1). The training blocks had to be repeated if the

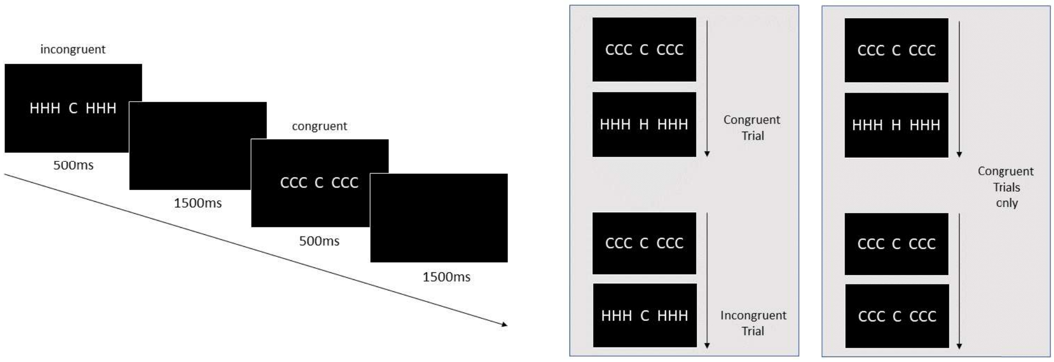


Figure 9.1: Experimental design - The task was presented on a black screen with white letters as stimuli and flanker items. Stimuli were presented for 500 ms, followed by a blank black screen in which the subject needed to answer the n-back task on the central item with yes or no by pushing the respective key on the keyboard. The box on the right shows exemplary which trials were used in the analysis as congruent or incongruent trials in the mixed flanker block.

accuracy was below 60 % to ensure that the subject was able to solve the task correctly.

9.1.1 Participants

A new set of 21 subjects (18 females) was recruited to participate in the study, for which they were reimbursed with 8 euro per hour. All subjects had normal or corrected to normal vision and no reported neurological disorders. The participants gave written consent, and the study was approved by the local ethics committee, and the study was performed in accordance with the declaration of Helsinki. On average, the subjects were 22.95 (±3.23) years old.

9.1.2 Technical setup

The subjects were seated in front of a computer screen (19 inches) on which the experiment was presented by the software E-Prime (Version 2.0.10.356). A standard keyboard was used for entering the answers, by which the correctness of an answer and the reaction time were assessed. For the EEG recording, a Brainproducts Acticap system with 32 electrodes was used and one BrainProducts actiChamp amplifier, which was sampled at 500 Hz (PyCorder). The integrated high pass filter was set to 0.1 Hz and the integrated low pass filter to 100 Hz. Additionally, a notch filter between 48-52 Hz was applied to eliminate power line noise. 28 electrodes were used for the recording and placed according to the extended 10-20 system [93] (FP1, FP2, F7, F3, FZ, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, PZ, P4, P8, O1, O2). The ground and reference electrodes were placed on the right and left mastoid respectively and impedances were kept below 10 kΩ.

9.2 Data analysis

Data will be analyzed with respect to behavioral data, which includes reaction time and task accuracy as well as with respect to physiological measures, including the EEG signal. Trials are categorized according to the presented flanker condition (congonly, cong and incong) and according to the n-back level (0 and 1). A total of six categories can be distinguished according to these criteria. Congonly trials represent all correct trials from a block in which only congruent flankers were presented. Trials from the category cong, are correct trials with a congruent flanker originating from blocks in which mixed flankers (alternating between congruent and incongruent) were presented. Lastly, incong trials are correct trials with incongruent flanker, originating from blocks with mixed flankers as well. In addition to this restriction, only trials that were preceded by another congruent trial have been selected. Incong trials are correctly answered trials with incongruent flankers from blocks with mixed flankers, again with the restriction that only trials which are preceded by a congruent trial are used for the analysis. For a better overview of the chosen categories, they have been visualized in Figure 9.2.

9.2.1 Behavioral data analysis

The behavioral data will be analyzed with respect to reaction time and task accuracy. To get an estimate of the group-level performance the averages will be calculated. After assessing the task accuracy for each category and subject, only trials with correct responses

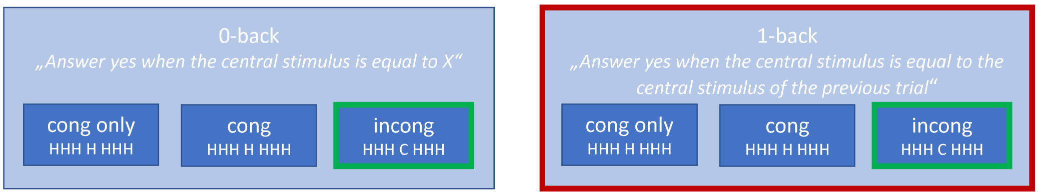

0-back
„Answer yes when the central stimulus is equal to X"
cong only
HHH H HHH
cong
HHH H HHH
incong
HHH C HHH
1-back
„Answer yes when the central stimulus is equal to the central stimulus of the previous trial"
cong only
HHH H HHH
cong
HHH H HHH
incong
HHH C HHH

Figure 9.2: Data structure - The data can be divided into two categories regarding the n-back level, which represent different levels of updating demands. Each category can further be divided into three subcategories regarding the Flanker condition, which represent different levels of inhibition demands. In total a set of six categories can be distinguished and will be used for analysis. 🟩: inhibition, 🟥: updating

have been used to calculate the average reaction times again per category and subject. As an additional constraint, the first four trials per block were left out. Statistical significant differences between the categories have been evaluated with an ANOVA, calculated on a linear mixed effect model either on the RT or task accuracy. To reveal the specific level on which significant differences are present, a pairwise t-test was performed as well on RT and accuracy.

9.2.2 Neurophysiological data analysis

For the physiological data, only artifact-free trials were used for data analysis. The data was bandpass filtered between 0.4 - 40 Hz and re-referenced to the common average. To remove artifacts, a threshold of ± 80 μV was chosen, and all trials exceeding this level were discarded. EOG artifact correction was performed with a regression method by Schloegl and colleagues [23]. A pre-stimulus baseline (-100 and 0 ms) was chosen to perform a baseline correction for every trial. Stimulus onset starts with stimulus presentation in the n-back task. For the calculation of the power spectra Burgs maximum entropy method [20] was used with a model order of 32 and a bin size of 1.

After choosing Fz, Cz, Pz, and O2 as representative channels for the evaluation, the statistical significance of the grand average of the ERPs and spectra is investigated. To reveal statistically significant differences in the signal, a Wilcoxon Ranksum Test [102] is used to calculate the p-value. A Bonferroni correction according to the number of used tests [103] is applied to correct for multiple comparisons. Significance level was determined to be at $p < 0.05$.

9.2.3 ML-based classification

For the investigation of the EFs by means of machine learning, support vector machine (SVM) classification was chosen. A SVM with a linear kernel (C = 1) [94], [8] was applied to differentiate between the six categories introduced above using the libsvm implementation for Matlab [95], [96]. The classification between categories was conducted for the following pairs for each subject individually:

- Inhibition: Cong vs. Incong vs. Congonly
- Updating: Zero vs. One
- EFs: Inhibition vs. Updating

The aim is to separate EF demands from baseline demands, but also different EF demands from each other. As the baseline demand within this study, congruent or congruent only trials from the zero back condition have been chosen, because neither updating nor inhibition demands should be induced during those trials. For each data pair, a subset of the data is used to train a classifier, to learn the characteristics of each category. The remaining data is used to evaluate the success of the learning and hence the skill of the classifier. This is done on a single-trial and single-subject level. To ensure stable results a 10-fold cross-validation was performed for each classification. The datasets (training set as well as the test set) were balanced for each comparison, by removing all spare trials if one of the classes had more trials than the other, to ensure that the distribution of examples per class does not have an influence on the result.

For classification, again two different types of features were used: CCA filtered ERP features [97] and power spectra. The performance of the classification approach is measured in accuracy, stating in how many cases the classifier categorized a trial correctly. Statistical significance of the classification results was determined by using permutation tests on the data.

The statistical significance of the results was determined by permutation tests with 1000 iterations [98, 99]. The classification performance achieved in the permutations establishes an empirical null distribution on random observations, which can be used to determine significance boundaries. Therefore, in each iteration classification was performed in a 10-fold cross-validation, but with randomly assigned class labels in the training set instead of the correct class labels. The achieved accuracy values were compared with the ones determined in the standard 10-fold cross-validation. Significance level was determined to be at $p < 0.05$, stating that the original classification performance is significant when the performance values are higher than the 95th percentile of the calculated empirical distribution.

ERPs

The ERP features upon which the SVM classifies are composed as follows. The 0-1000 ms epochs that have been prepared in the beginning will be used from 17 EEG channels (FP1,FP2,F3,FZ,F4,FC1,FC2,C3,CZ,C4,CP1,CP2,P3,PZ,P4,O1,O2). All other channels of the setup are discarded to reduce the influence of noise and artifacts in the data. Due to the sampling rate of 500 Hz, one trial of ERP data is represented by 500 × 17 features. As a way to improve the signal-to-noise ratio of the data, a spatial filtering method based on canonical correlation analysis (CCA) was applied [26] with a filter size of 27 × 17. The filter aims to minimize the variance within a class and to maximize the variance between classes to improve the separability.

Power spectrum

Classification using features from the frequency domain was conducted on the power spectra between 1-20, calculated on the same epochs as described above (0-1000 ms after stimulus onset) with Burgs maximum entropy method. The same 17 channels were used, resulting in 19 × 17 features respectively.

Cross-class classification

Since the question of separability aims to answer the diversity aspect of Miyake's model of executive functions, another approach needs to be introduced to answer the unity aspect of the EFs. For this, a cross-class classification was performed. In this approach, a classifier was trained on the distinction of EF1 vs. BL and tested on EF2. The reason for this is as follows. If the functions have significant overlaps in their properties, EF2 should be classified as EF1 with above-average frequency. If there is no such overlap, the classification accuracy should be at a more random level. This evaluation is done in both directions. Hence, each EF is part of the train and test set once. Cross-class classification is performed on ERP as well as on power spectral features.

9.2.4 Neural activation patterns

To inspect the features used for the distinction in the classification approach, a method developed by Haufe and colleagues [29] was used that transforms the weights of the SVM classifier into neurophysiological interpretable values, in so-called neural activation patterns. One classifier model is trained for each subject on the data of the respective categories. Applying the method on the model results is one activation value for each feature that was used in the classification. To create a comprehensive picture of the resulting neural activation pattern, the values are averaged within and according to the two frequency bands alpha (8-12 Hz) and theta (4-7 Hz). This is done for each subject individually, but the median values across subjects will be depicted in a color-coded topological distribution, to visualize the results. By calculating the activation patterns, the underlying neurophysiological patterns that are responsible for the distinction can be inspected, which can provide valuable information analyzing the unity and diversity of different EFs.

9.3 Results

9.3.1 Behavioral data

Table 9.1 shows the reaction time averaged over all participants and the overall accuracy of the participants' responses. To display differences between the experimental conditions, the results are sorted and averaged individually for each flanker condition (congonly, cong and incong) and n-back level (zero and one). Irrespective of the task conditions,

Table 9.1: Behavioral accuracy and reaction time - Average accuracy (acc) and reaction time (RT) of the subjects categorized according to the flanker condition (congonly, cong, incong) and the n-back level (zero, one)

		Flanker		
n-back Level		congonly	cong	incong
Zero	RT [ms]	465.91	464.46	500.12
	Acc [%]	94.9	94.6	90.4
One	RT [ms]	519.41	543.10	541.29
	Acc [%]	91.9	91.6	92.7

Table 9.2: ANOVA on behavioral data - P-Values calculated for the average reaction times (RT) and accuracies (ACC) of all subjects per condition. An ANOVA was performed on a linear regression model, taking the participant, n-Level, flanker condition and the interaction between flanker condition and n-Level into account. Significance level has been determined to be at $p < 0.05$.

	Participant	Flanker	n-Level	n-Level:Flanker
RT	< 0.05	< 0.05	< 0.05	< 0.05
Acc	< 0.05	0.16	0.15	< 0.05

participants were able to achieve an overall task accuracy of more than 90 %. Regarding the n-back level, it can be seen that participants were consistently slower in trials from blocks with n-back level one than in blocks with n-back level zero. With one minor exception, it can also be said that the task accuracy is lower during n-back level one than for level zero. When looking at the flanker conditions, it can be seen that congonly and cong trials are answered equally fast and correct, whereas incong trials are slower and less correctly answered by the participants. Interestingly, during 1-back cong and incong answers are almost equally fast, while congonly answers have been given much faster compared to the other two categories. An ANOVA revealed that reaction time is significantly influenced by all tested factors, including participant, flanker congruency, and the n-back level. It could also be shown that there is a significant interaction between the flanker and the n-back level (see Table 9.2). When comparing the flanker conditions pairwise to reveal all levels of the effect, it can be seen that there are differences between the two available n-back levels (see Table 9.3). During n-back level zero, we find a significant difference in RT between cong and incong trials as well as between congonly and incong trials. During n-back level one we find significant differences in RT between congonly and cong trials as well as between congonly and incong trials.

Interestingly, no significant difference in accuracy can be found for the flanker conditions during n-back level one. During the 0-back condition, the congonly vs cong comparison is not significantly different, whereas the other two comparisons are. Even though the differences between the three levels are not consistently found across all comparisons, the analysis of behavioral data provides first clues about three different levels of inhibitory control and the interaction of the two EFs updating and inhibition.

Table 9.3: Pairwise T-test on behavioral data - More detailed analysis of the two individual factors flanker condition and n-Level. P-Values have been calculated for the average reaction times (RT) and accuracies (ACC) of all subjects per condition with a paired t-test. Significance level has been determined to be at $p < 0.05$.

		Flanker		
n-back Level		congonly vs cong	congonly vs incong	cong vs incong
Zero	RT	0.67	< 0.05	< 0.05
Zero	Acc	0.70	< 0.05	< 0.05
One	RT	< 0.05	< 0.05	0.90
One	Acc	0.66	0.28	0.24

9.3.2 Neurophysiological analysis

ERPs

The grand average event-related potentials (ERPs) and spectra were calculated for the four electrode positions Fz, Cz, Pz, and O2, which were chosen as representative positions since it was shown by Scharinger and colleagues [57] as well as in study 1 [90], that those positions are of particular interest. While evaluating the neurophysiological signals to reveal if three levels of inhibition can be distinguished, three possible comparisons were investigated (cong vs. incong, congonly vs. cong, congonly vs. incong), separately for each of the two n-back levels. The results for the ERPs for n-back level zero can be seen in Figure 9.3. It can be seen that at all four electrode positions almost no statistically significant differences can be found. Figure 9.4 shows the ERPs for the comparison of the n-levels.

Power spectra

In Figure 9.5 the comparisons of Flanker categories can be seen during n-back level zero. For the power spectra, there are some indicators for differences in the congonly vs. incong comparison, but none for the other two comparisons. In contrast to that, the comparison of n-back levels shows statistically significant differences throughout all flanker conditions as can be seen in Figure 9.6. The differences seem to be most pronounced during a congruent flanker. The grand average ERPs and power spectra for four out of the six categories are shown in Figure 9.7 to make the waveforms comparable between all conditions. It can be seen that the four conditions differ in amplitude, but the waveform remains rather constant. The change in frontal theta power can be identified as an ERS since the power increases with the amount of load (zero congonly < zero cong < zero incong < one congonly).

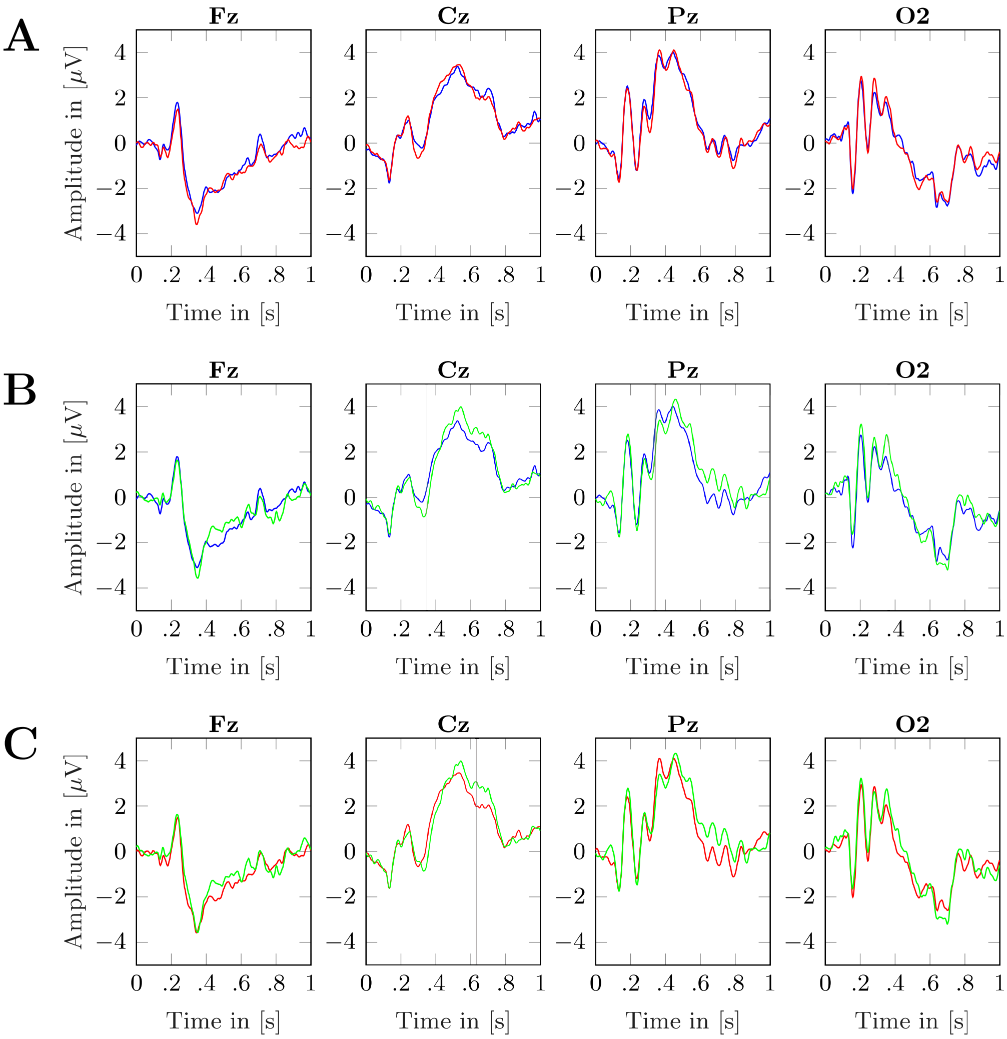


Figure 9.3: Grand average ERPs for Flanker conditions - Displayed are the electrode positions Fz, Cz, Pz and O2 during n level 0. A pairwise comparison of trials with congruent, incongruent and congruent only flankers can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Cong vs Congonly, **B**: Congonly vs Incong, **C**: Cong vs Incong, ▬: congonly, ▬: cong, ▬: incong

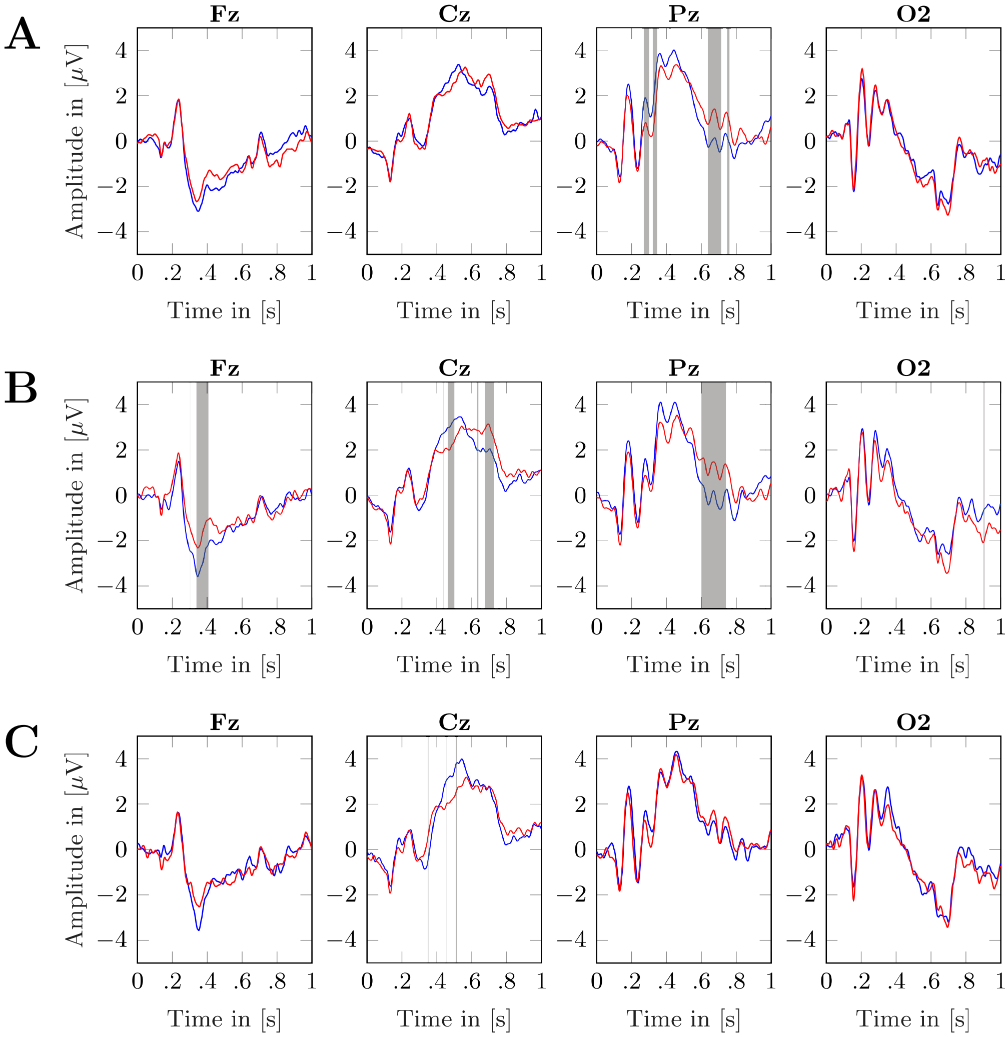


Figure 9.4: Grand average ERPs n-level one vs. zero - Displayed are the electrode positions Fz, Cz, Pz and O2 during the three Flanker conditions. A comparison of trials between n-back level one and zero can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Congonly, **B**: Cong, **C**: Incong, —: zero, —: one

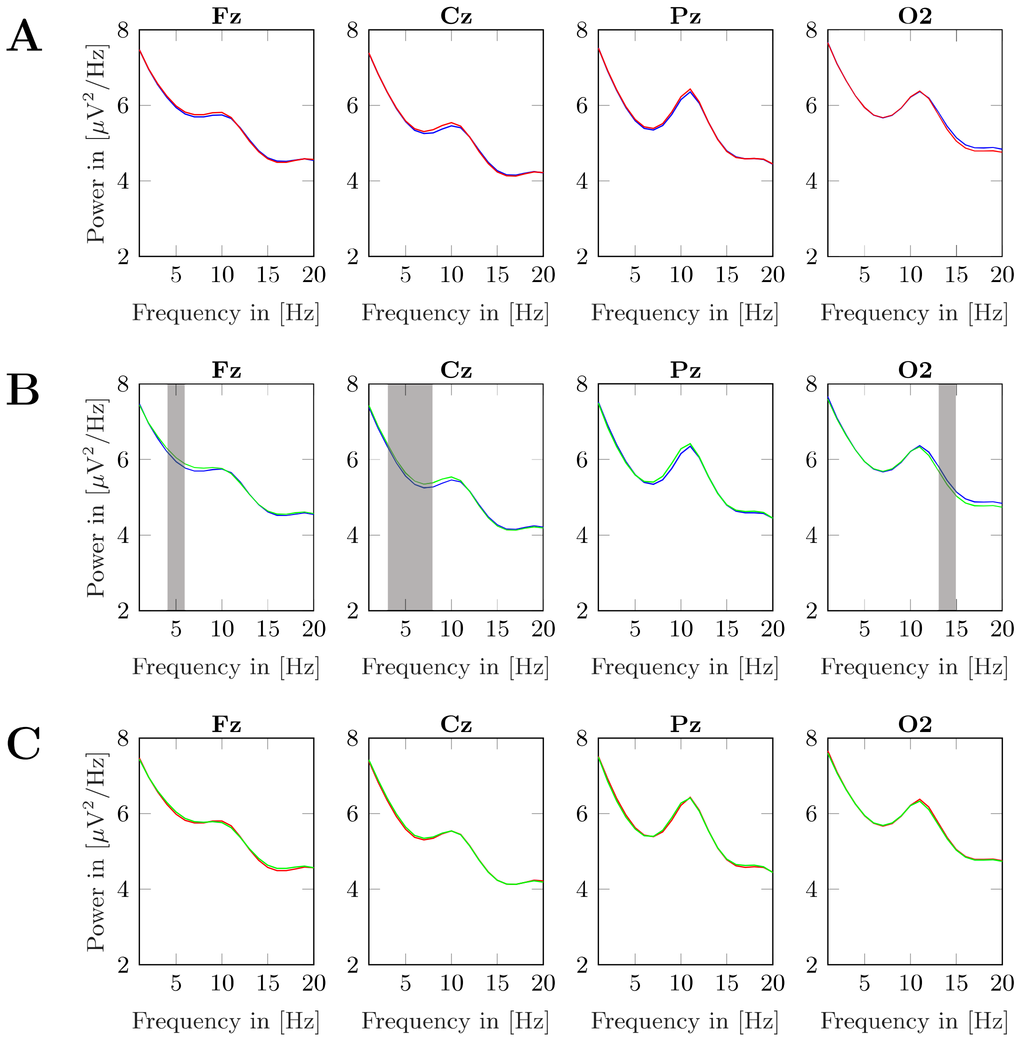


Figure 9.5: Grand average power spectra for n-back level 0 - Displayed are the electrode positions Fz, Cz, Pz and O2 during n level 0. A pairwise comparison of trials with congruent, incongruent and congruent only flankers can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of frequency bins). **A**: Cong vs Congonly, **B**: Congonly vs Incong,
C: Cong vs Incong, —: congonly, —: cong, —: incong

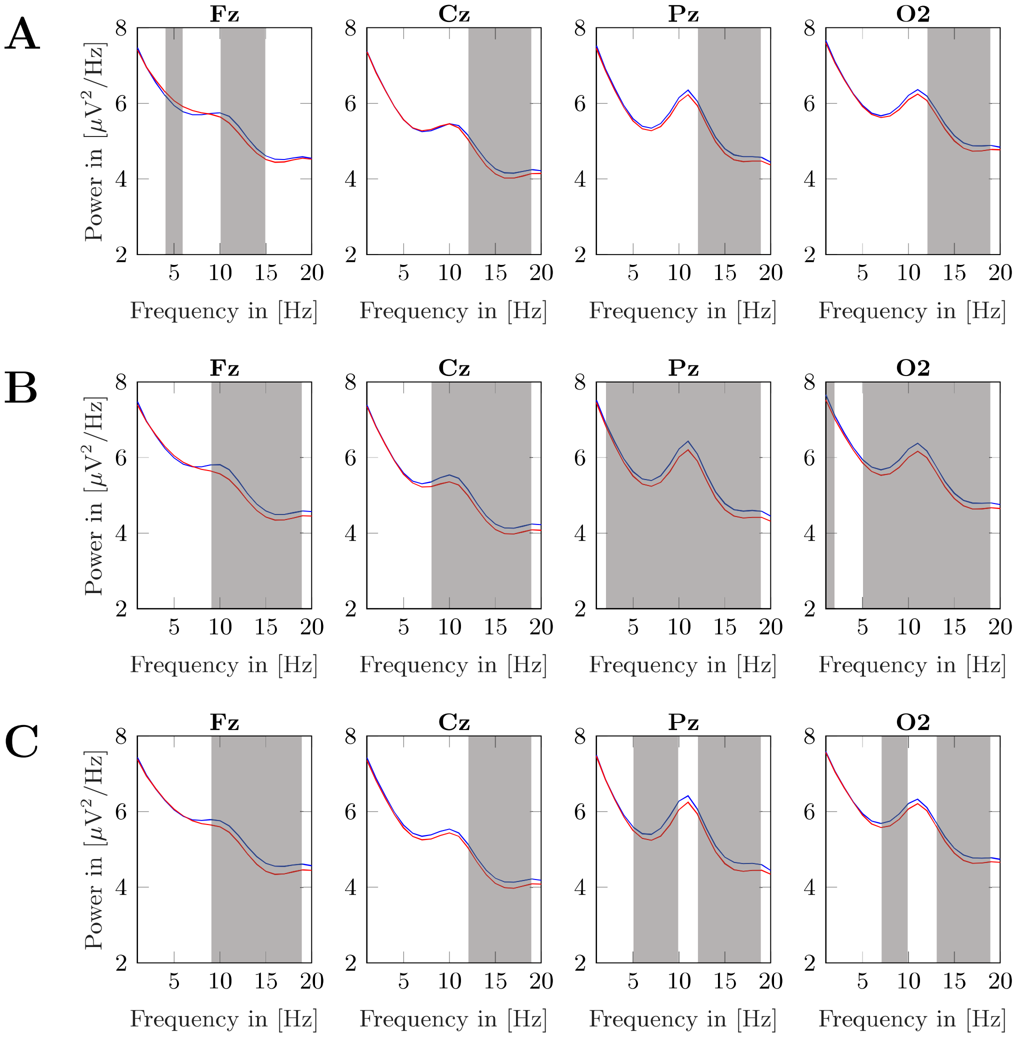


Figure 9.6: Grand average Power spectra n-level one vs. zero - Displayed are the electrode positions Fz, Cz, Pz and O2 during the three Flanker conditions. A comparison of trials between n-back level one and zero can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points). **A**: Congonly, **B**: Cong, **C**: Incong, —: zero, —: one

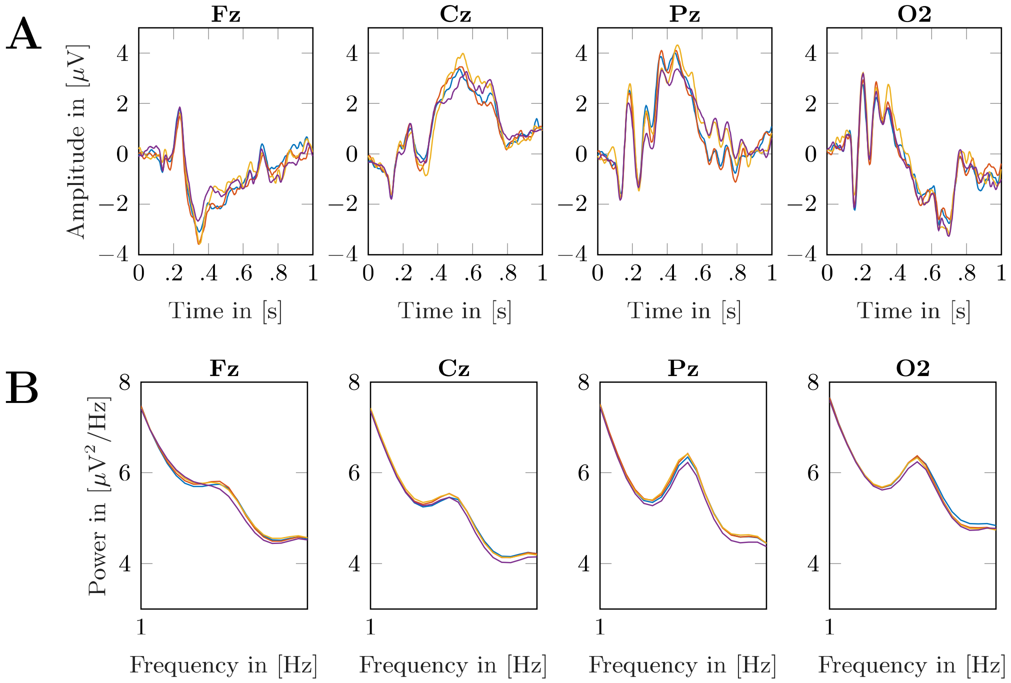


Figure 9.7: Grand average All in One - Displayed are the electrode positions Fz, Cz, Pz and O2. The grand average has been calculated over all 21 subjects. A: ERP, B: Spectra, —: zero congonly, —: zero cong, —: zero incong, —: one congonly

9.3.3 ML-based classification

In the classification approach, the same comparisons as in the previous analysis steps were made. Table 9.4 shows the classification accuracies for the pairwise comparisons. Depending on the used feature set (ERP or power spectra) the values range from 51.03 % up to 63.07 % for the distinctions. The distinction worked best for the congonly vs. incong comparison at n-back level zero and was least successful for cong vs. incong at n-back level one. The null distribution determined by permutation tests for each distinction individually revealed that most of the classification accuracies are statistically significant above chance level. The only distinction in which no statistical significance was reached with any of the used feature sets is the cong vs. incong comparison during n-back level one.

In Table 9.5 the performance values for the distinction of the two n-back levels can be seen. Accuracy values between 60 - 68 % have been reached. Interestingly, the performance is best for the distinction under the incongruent flanker condition for the ERP feature set, while the performance is worst also during the incongruent flanker condition, but for the power spectral feature set. The other two comparisons are almost equal in performance and values do not vary between the feature sets.

Table 9.6 shows that when trying to distinguish the two EFs on the basis of their ERPs or power spectra classification accuracies between 59 % and 62 % can be reached. The performance values do not indicate that there is a difference between cong and congonly trials while separating them from inhibition trials.

Table 9.4: Classification inhibition demands - Classification accuracies achieved for the distinction of inhibition demands, from baseline, are shown with (One) and without (Zero) additional load on the EF updating. Results are the average performance of a SVM in a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

n-back level	Features	congonly vs. cong	congonly vs. incong	cong vs. incong
Zero	ERP (CCA)	56.97 %	63.07 %*	57.44 %*
	Power	59.90 %*	60.54 %*	51.03 %
One	ERP (CCA)	54.56 %	55.23 %	50.11 %
	Power	59.85 %*	59.96 %*	51.27 %

Table 9.6: Classification Updating (Up) vs Inhibition (Inh) - Classification accuracies achieved for the distinction between updating and inhibition demands are shown. Results are the average performance of a SVM in a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Features	Up vs Inh (cong only)	Up vs Inh (cong)
ERP (CCA)	59.75 %*	62.20 %*
Power	59.31 %*	60.25 %*

Table 9.5: Classification updating demands - Classification accuracies achieved for the distinction of updating demands, from baseline and from each other, are shown with (incong) and without (cong, congonly) additional load on the EF inhibition. Results are the average performance of a SVM in a 10-fold cross-validation with a linear kernel on 0-1000 ms time frame. Used were the CCA filtered ERPs of the 17 channels as one feature set and the power spectra between 1-20 Hz for the same 17 channels as a second feature set. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Features	Zero vs One (congonly)	Zero vs One (cong)	Zero vs One (incong)
ERP (CCA)	65.97 %*	62.95 %*	68.34 %*
Power	66.23 %*	62.39 %*	60.46 %*

Neural activation patterns

To ensure the reliability of the ML approach, the features that have been used in the classification approach are analyzed. This way it can be controlled that only features related to the experimental condition and not caused by artifacts or features unrelated to executive control are factored in.

Figure 9.8 shows the neural activation patterns for the spectral features during n-back level zero. In particular for the alpha and theta frequency band which are known to correlate highly with working memory load. It can be seen that the patterns that are formed are similar in all three distinctions and only include frontal/central theta and parietal alpha components. Therefore, it is clear that the distinction is not based on noise but on features that are known for their correlation with working memory load and executive control. The ability to interpret the process of classification neurophysiologically legitimizes its use and also legitimizes the interpretation of the classification accuracies in the context of mental state characterization.

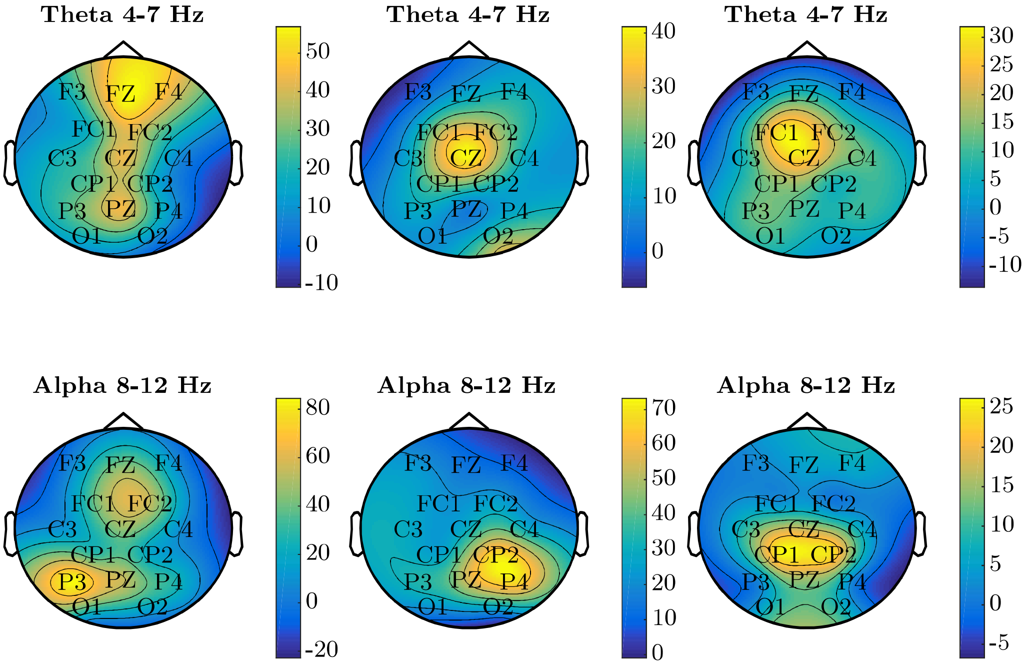
In Figure 9.9, the neural activation patterns for the distinction of updating levels can be found under different flanker conditions. In contrast to study 1, no clear frontal theta pattern can be found, but it still can be seen, that frontal theta features play an important role. Interestingly, not only occipital alpha but also occipital theta are further of importance. In direct comparison of the two EFs, it can be seen in Figure 9.10 that the pattern does not differ much when using congruent flanker only compared to congruent flankers from the mixed flanker condition.

Cross-class classification

Table 9.7 shows the results of the cross-class classification. It can be seen that performance values close to chance level have been achieved. The values are slightly above chance level, but not distinctly. Therefore, it can be assumed that a small overlap in properties can be found for the tested classes.

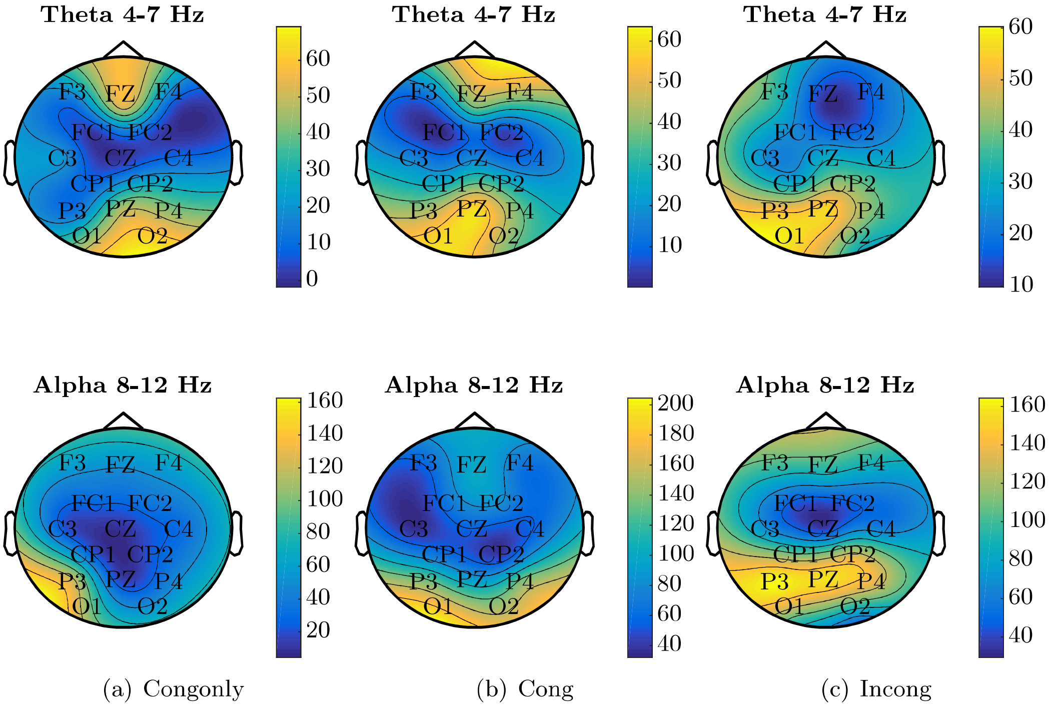
Table 9.7: Cross-class classification - The table provides cross-class classification with an SVM for ERP features as well as for the power spectra. The classifier was trained on Demand 1 vs. BL and tested on trials belonging to Demand 2 only. Therefore, the here presented accuracies represent the percentage of trials classified as BL. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Trainset	Zero congonly vs. incong	Zero vs. One congonly
Testset	One congonly	Zero incong
ERP	46.72 %*	46.12 %*
Power (1-20)	39.99 % *	48.12 %



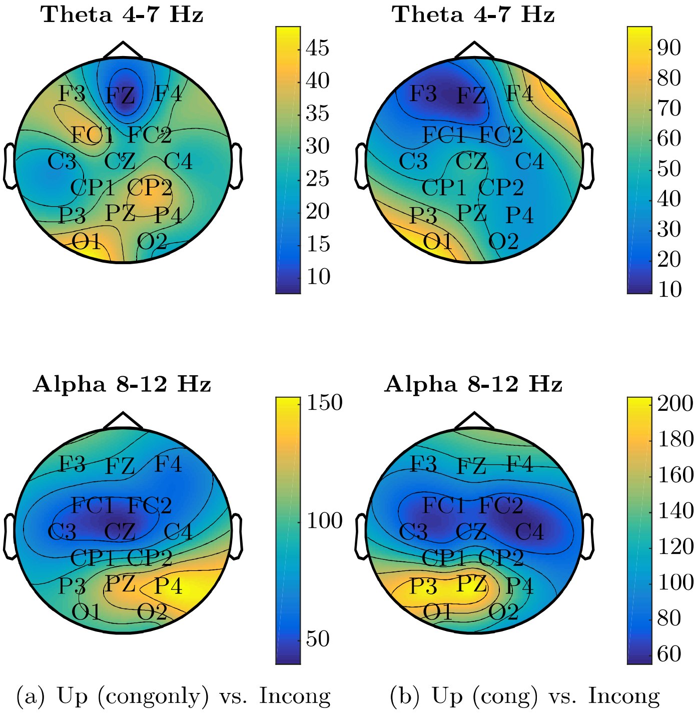
(a) Congonly vs. Cong (b) Congonly vs. Incong (c) Cong vs. Incong

Figure 9.8: Neural activation pattern inhibition demands - Displayed is the color-coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A pairwise comparison of conflict conditions congruent (cong), incongruent (incong) and congruent only (congonly) is shown for n-back level zero. The resulting values are an average over the individual patterns of all 21 subjects. a) Congonly vs. Cong, b) Congonly vs. Incong, c) Cong vs. Incong



(a) Congonly (b) Cong (c) Incong

Figure 9.9: Neural activation pattern updating demands - Displayed is the color-coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of n-back levels zero and one is shown for all flanker conditions. The resulting values are an average over the individual patterns of all 21 subjects. a) Congonly, b) Cong, c) Incong



(a) Up (congonly) vs. Incong (b) Up (cong) vs. Incong

Figure 9.10: Neural activation pattern Updating vs. Inhibition - Displayed is the color-coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of shifting and inhibition demands . The resulting values are an average over the individual patterns of all 21 subjects. a) Up (cong only) vs. Incong, b) Up (cong) vs. Incong

9.4 Discussion

Can the results from study 1 be reproduced ?
Is there a block effect for the EF inhibition in the modified flanker task ?

9.4.1 Behavioral data

The general level of accuracy was high with more than 90 % in all conditions, stating that the difficulty level was adequate and all subjects were able to perform the task with sufficient accuracy. The ANOVA revealed that the n-back level and the flanker congruency have a significant interaction, as well as a significant effect on the RT. Both effects are expected according to literature [60, 110, 111]. Significant differences between congruent and incongruent flanker trials have been found during zero, but not during one back with respect to reaction time as well as to accuracy. The opposite result has been found for congruent trials from mixed blocks compared to congruent trials without flanker variation. RT is significantly different during one back, but not during zero back. This reaction time effects that have been found to differ slightly from the original study [57]. Scharinger and colleagues found an underadditive flanker effect on updating load, but only at n-back level 2 not already at level 1. An underadditive flanker effect describes a decrease in interference due to a general increase in attentional processes. During n-back level one, no differences in RT between cong and incong trials were found. This is an effect that can be due to a general increase in the attentional process has already been found by other authors [110]. The effect underlines the unity aspect of EFs (c.f. [39]) since it can be explained by relying on a common attentional resource, based on the general activation of attentional processes that are shared. This means that based on a high attentional level no more/additional capacities can be used for the flanker processing.

It can be argued that the cognitive load is lower in zero back condition and during low cognitive load enough attentional resources are available to take the flanker into account and also to solve the task sufficiently well. Hence, this could be an explanation of why no difference in accuracy and RT can be found at 0-back between cong and congonly trials. It can be hypothesized that the influence of the readiness to inhibit (cong trials) on the performance seems to be minor but gets visible when additional load is present. This would also lead to the assumption that no other confounds are introduced by using congruent flankers only if no differences between the two experimental conditions can be found during 0-back.

The approach using a block design to asses differences between congruent trials in mixed flanker blocks and blocks without flanker variation is conceptually similar to other studies [85], [112] in which the authors aimed to asses switch cost. They designed blocks in which one task was performed purely and a block in which shifting between tasks was necessary. The difference in time that was needed per block was called switch cost. A general observation that was made and can be transferred to the results of this study is that the reaction time is slower when mixed blocks are performed in contrast to pure blocks. Theoretical validation that allows comparing congruent trials between the two blocks (congruent only vs. mixed flanker block) can be found in a study by Rogers *et al.* [113]. They showed that accuracy and reaction time improve immediately after a switch trial, but no further improvement can be seen in trial three and four after the switch. Since we filter all congruent trials out that are not proceeded by a congruent trial, differences in behavioral data should not reflect an acute shifting process that influences the current congruent trial.

Using neutral flankers only would be interesting to compare with, since they do not represent competing items, but still, the use of inhibitory control might be necessary to focus solely on the central item. In addition to that, using incongruent flankers only would also be interesting. This would complete the analysis regarding the question of how much influence the preparedness to inhibit, actual response inhibition and distractors without conflict have. From the behavioral results of this study, it can be concluded that two different levels of inhibitory control can be distinguished with statistical significance, but not three.

9.4.2 Neurophysiological data

EEG signals have been investigated on a grand-average basis at four channels of interest, covering frontal and parietal sites which are of interest concerning working memory load. No nameable significant differences have been found, suggesting that either potential differences between the experimental conditions cannot be assessed with EEG or that existing difference vanish and lose statistical significance due to averaging over 21 subjects. It could be argued that showing the results of more channels could reveal different results. However, choosing to display more channels would result in a more strict correction for multiple comparisons, making it even harder to find significant differences between the experimental conditions. As already mentioned, the four chosen channels are the most important on which effects would be expected. For the averaged power spectra, the situation looks similar. Finding no significant effects there leads to the assumption that no neurophysiological differences can be found. Due to this, it needs to be stated that the classical standard analysis of ERPs and power spectra were not sufficient to draw conclusions out of the data for distinguishing three levels of inhibitory control.

9.4.3 ML-based classification

With one minor exception, classification accuracy is highest for congonly vs. incong, followed by congonly vs. cong trials. Accuracy is lowest for cong vs. incong comparisons, independent of the n-back level. For each level distinction, statistically significant results exist, indicating that three different levels of inhibitory control can be distinguished. No inhibition, readiness to inhibit, inhibition. This fact underlines the initial assumption that, even though behavioral data does not give rise to any difference between no inhibition and readiness to inhibit at n-back level zero, statistically significant differences exist in neurophysiological data.

Classification accuracy is less accurate when the n-back level is one, compared to n-back level zero, emphasizing the interaction between the two conditions, which also suits the flanker effect we found relying on shared attentional resources [110]. Due to this, the non-significant performance for distinguishing cong vs. incong trials during n-back level one can be explained. Therefore, it can be assumed that the higher the current cognitive load, the less accurate is the classification accuracy. Congonly trials only require the focus on one task without distraction, cong trials in mixed flanker blocks already the preparedness to inhibit potentially conflicting flankers and incong trials require the actual execution of inhibitory control. This difference can be made visible with a classification approach on EEG signals.

In general, the classification approach is interesting for differentiating between different mental states or experimental conditions as it works on a single trial basis of each subject individually. Between-subject variability does not compromise the overall results as much as this is the case in standard ERP analysis. No differences in behavioral data do not mean that the experimental conditions do not induce different mental states. The minor overlap of properties of updating and inhibition, specifically tested with the new congonly condition that has been revealed in the cross-class classification might be related to a unity aspect of the two EFs. Since the classification accuracy only deviates minimally from the chance level, it can be assumed that it is a very small overlap of properties. Therefore, it can also be assumed that a large amount of properties is distinct and speaks for the diversity aspect.

9.4.4 Neural activation patterns

Apart from classifying the signals, it can be evaluated which neurophysiological features led to the distinction in the classification approach. This knowledge can be a useful extension to the prevailing analysis, as it gives more insight into the origin of the difference in signals between conditions. The revealed results show that only features related to working memory load have been denominated as important during the distinction. The features with the greatest importance are in the theta range, at frontal electrode positions, and in the alpha range at parietal electrode positions which are known to correlate with working memory load [101, 114, 115]. No other features have been ranked important. Therefore, it has been ensured that no artifacts or non-task related features have been factored in. For this reason, the usage of this approach can be seen as valid.

Another general conclusion that can be drawn is that the same pattern, that has been found in Krumpe et al. [90] could be replicated in this study for the EF inhibition. The shift of alpha activity to PZ rather than OZ compared to the original study can possibly be explained by using a different approach to correct for eye movement artifacts. But in general, it needs to be noted that the study at hand was no exact replication. Therefore, minor differences can be expected. Despite the differences and potential high inter-subject variability, the reproducibility of the pattern reveals that this must be a rather robust pattern that has been made visible by this technique.

9.5 Conclusion

The here presented results could show that three levels of inhibitory control can be distinguished with the help of ML approaches in a modified Flanker task. Classical analysis approaches, including the analysis of behavioral and physiological data, did not create a consistent picture of the presence and differentiability of three different levels of inhibitory control. Behavioral data only gave rise to the presence of two levels, whereas the group-based averages of EEG signals were not meaningful at all. When additionally looking at the classification accuracies, it could be found that three levels of inhibitory control can be distinguished with statistical significance. The classification approach and its validity are supported by investigating the neurophysiological interpretation of the underlying patterns that make the data distinguishable. This reveals how important it can be to also integrate single subject analysis steps, to not lose meaningful inter-subject variability. It could also be shown that the results from study 1 can in great parts be reproduced, and it could additionally be shown that blockwise manipulation of inhibition demands does differ from trialwise manipulation as was hypothesized. The effects are not as major as in study 2 but nevertheless measurable.

Chapter 10

Study 4: When Shifting meets Inhibition II (Between vs. within block effects)

Study 4 again investigates the unity and diversity aspects of the EFs Inhibition and Shifting. Since the results from Study 2 brought a new perspective on the so far performed studies, another follow-up study was designed. To achieve consistency in the results, it has been decided to repeatedly perform study 2 with the extension of one more condition in which there was no manipulation of inhibition demands by varying flanker items. Conceptually, study 4 equals study 3. However, it is based on the task design of study 2. The resulting design, therefore, allows the block- and trialwise manipulation of inhibition as well as shifting demands. Hence, study 4 completes the analysis regarding the properties of the EF inhibition by evaluating it in two different tasks (n-back and greater/odd) and two different manipulation settings (block and trialwise)

Is there a block effect for inhibition demands in the modified flanker task ?

10.1 Task design

As already stated, the experiment is based on the task design of study 2 with the addition of one condition which has conceptually been introduced in study 3. The design of the study comprises of two tasks: the odd and greater 5 task and the Eriksen-Flanker task. The odd/greater task is used to induce demands on the EF shifting, whereas the flanker task is used to induce demands on the EF inhibition. The simultaneous presentation of the two tasks was realized by showing seven items at once, from which one was positioned centrally, the other six on a flanking position to the left and right of the central item. The odd/greater 5 task was performed on the central item and used as the primary task in the experiment. The flanker was only used as a secondary task, to which no action was required. The six flanking items, where therefore congruent (identical) or incongruent (different) to the central item and fulfilled a distracting purpose.

Again the experiment was presented blockwise, this time with six blocks in total. As in study 2, the levels of shifting demands can be categorized as follows: single (only

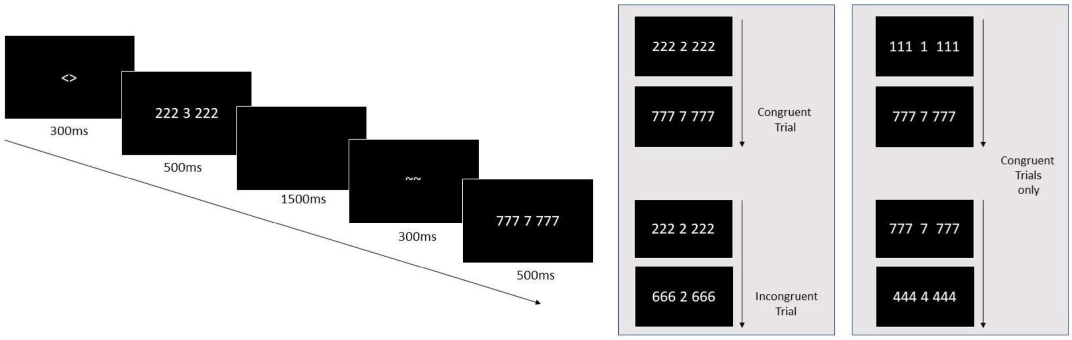


Figure 10.1: Experimental design - The task was presented on a black screen with white letters as stimuli and flanker items. Stimuli were presented for 500 ms, followed by a blank black screen in which the subject needed to answer the odd-even/smaller-greater task on the central item with yes or no by pushing the respective key on the keyboard. The box on the right shows exemplary which trials were used in the analysis as congruent or incongruent trials in the mixed flanker block.

one task during the entire block, odd or greater), mixed (both tasks during the block in an randomly interleaved order), switch (current trial is from different task as the previous one), repeat (current trial is from the same task as the previous one). Single and mixed categories manipulate shifting demands blockwise, whereas switch and repeat manipulate shifting demands trialwise. For a better overview of the categories and the resulting data structure see Figure 10.2. To quickly repeat the task of the subject: In each trial, subjects were asked to decide whether the central item is greater than 5 or odd. Which of the two tasks needed to be performed was indicated by a queue which was presented for 300 ms before each trial (<> for greater, ∼∼ for odd). Therefore, in each trial, an answer of yes or no was required by button press (keys D and L on a standard keyboard). Yes and no answers were randomly distributed over each block with a ratio of 1:1 and given with the index finger of either the right or left hand. Which key represented the yes answer was counterbalanced throughout all subjects. Each block included 120 trials, of 2.3 seconds length. One trial consisted of 300 ms queue, 500 ms stimulus presentation and a 1500 ms long blank screen. For a schematic overview of the experimental design, see Figure 10.1.

The stimuli were again numbers in the range of 1 to 9, excluding 5. Each number was presented equally often as a central item, leading to a balanced amount of even and odd items as well as a balanced number of smaller and greater than 5 items. The ratio for the congruency of the flanker items was on third congruent and two thirds incongruent. The six blocks were presented divided into two parts, enabling a break after half of the experiment. Two blocks consisted of the odd task only (single), one of the greater than five task only (single) and three blocks presented both tasks in an alternated and random but balanced order (mixed). One of the single blocks was designated to show congruent flanker items only, as well as one of the mixed blocks, whereas all other four blocks showed congruent and incongruent flanker items. For consistency throughout all participants, it was chosen to use one of the odd blocks in the single condition to present congonly flankers.

Before the start of the experiment, a short training phase was presented to familiarize the subjects with the task. The training consisted of 3 short blocks (24 trials), one for each task level (odd (single) and greater 5 (single) and both tasks (mixed)). The training blocks had to be repeated if the accuracy was below 60 % to ensure that the subject was able to solve the task correctly. During training feedback based on the performance is provided to the subject after each block to indicate to the subject and instructor that the task was fully understood and can be executed with sufficient accuracy. The subjects are also asked to rate the effort after each training block that is needed and how successful the subjects perceived their performance. During the real task, no feedback was provided.

10.1.1 Participants

21 subjects (18 females) participated in the study, for which they were reimbursed with 8 euro per hour. All subjects had normal or corrected to normal vision and no reported neurological disorders. The participants gave written consent and the study was approved by the local ethics committee. On average, the subjects were 22.95 (±3.23) years old.

10.1.2 Technical setup

The subjects were seated in front of a computer screen (19 inches) on which the experiment was presented by the software E-Prime (Version 2.0.10.356). A standard keyboard was used for entering the answers, by which the correctness of an answer and the reaction time were assessed. For recording EEG, a Brain Products Acticap system with 32 electrodes was used and one Brain Products actiChamp amplifier which was sampled at 500 Hz (PyCorder). The integrated high pass filter was set to 0.1 Hz and the integrated low pass filter to 100 Hz. Additionally, a notch filter between 48-52 Hz was applied to eliminate power line noise. 28 electrodes were used for the recording and placed according to the extended 10-20 system [93] (FP1, FP2, F7, F3, FZ, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, PZ, P4, P8, O1, O2). The ground and reference electrodes were placed on the right and left mastoid respectively and impedances were kept below 10 kΩ.

10.2 Data analysis

Data was analyzed with respect to behavioral data, which includes reaction time and task accuracy as well as with respect to physiological measures, including the EEG signal. Trials are categorized according to the presented flanker condition (Congonly, Cong and Incong), the task (Odd and Greater) and the shift which can occur trialwise (Repeat and Switch) or blockwise (Single and Mixed). Figure 10.2 again shows an overview of all categories that were investigated.

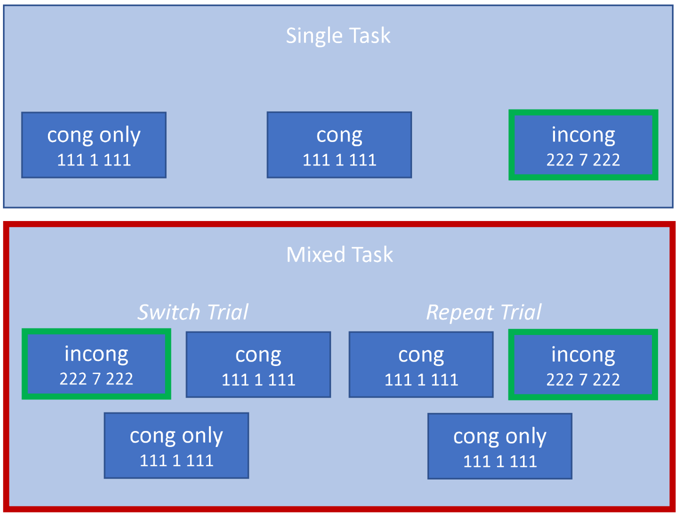


Figure 10.2: Datastructure - The data can be divided into three categories regarding the number of tasks within one block, which represent different levels of shifting demands. Each category can further be divided into three subcategories regarding the Flanker condition, which represent different levels of inhibition demands. In total a set of nine categories can be distinguished and will be used for analysis. 🟩: inhibition, 🟥: shifting.

10.2.1 Behavioral data analysis

In terms of behavioral data, RT and task accuracy can be investigated an compared to reveal differences and commonalities between the properties of the two executive functions shifting and inhibition. Statistical significant differences between the categories have been evaluated with an ANOVA, calculated on a linear regression model either on the RT or task accuracy.

10.2.2 Neurophysiological analysis

For the physiological data, only artifact-free trials with correct responses were used for data analysis, with an additional exclusion of trials that might yield to any Gratton-like effect. The data was bandpass filtered between 0.4 - 40 Hz and re-referenced to the common average. To remove artifacts a threshold of 100 μV was chosen and all trials exceeding this level were discarded. Trials including eye movement artifacts were corrected using a regression method by Schloegl and colleagues [23]. A baseline correction was performed with 100 ms pre-cue onset.

In addition to this restriction, only trials that were preceded by another congruent trial have been selected. Incong trials are correctly answered trials with incongruent flankers from blocks with mixed flankers, again with the restriction that only trials which are preceded by a congruent trial are used for the analysis. In terms of physiological data, the grand average ERPs, as well as the grand average spectra, can be computed for each of the six categories separately. Based on the results of [57], the ERPs at electrode positions FZ, CZ, PZ were of major interest. For the calculation of the power spectra

Burgs maximum entropy method was used with a model order of 32 and a bin size of 1. To test whether the differences in the ERPs and power spectra between the factor levels are statistically significant, a Wilcoxon ranksum test [102] was conducted over all subjects and trials. The resulting p-values were Bonferroni corrected [103] and the significance level was set to $p < 0.05$.

10.2.3 ML-based classification

For the investigation of the EFs by means of machine learning, SVM classification was chosen. A SVM with a linear kernel ($C = 1$) [94], [8] was applied to differentiate between the nine categories introduced above using the libsvm implementation for Matlab [95], [96]. The classification between categories was conducted for the following pairs for each subject individually:

- Inhibition: Cong vs. Incong vs. Congonly
- Shifting: Single vs. Mixed, Switch vs. Repeat
- EFs: Shifting vs. Inhibition

The aim is to separate EF demands from baseline demands, but also different EF demands from each other. As the baseline demand within this study, congruent or congruent only trials from the single blocks have been chosen, because neither shifting nor inhibition demands should be induced during those trials. For each data pair, a subset of the data is used to train a classifier, to learn the characteristics of each category. The remaining data is used to evaluate the success of the learning and hence, the skill of the classifier. This is done on a single-trial and single-subject level. To ensure stable results a 10-fold cross-validation was performed for each classification. The datasets (training set as well as the test set) were balanced for each comparison, by removing all spare trials if one of the classes had more trials than the other, to ensure that the distribution of examples per class does not have an influence on the result.

For classification, again two different types of features were used: CCA filtered ERP features [97] and power spectra. The performance of the classification approach is measured in accuracy, stating in how many cases the classifier categorized a trial correctly. Statistical significance of the classification results was determined by using permutation tests on the data.

The statistical significance of the results was determined by permutation tests with 1000 iterations [98, 99]. The classification performance achieved in the permutations establishes an empirical null distribution on random observations, which can be used to determine significance boundaries. Therefore, in each iteration classification was performed in a 10-fold cross-validation, but with randomly assigned class labels in the training set instead of the correct class labels. The achieved accuracy values were compared with the ones determined in the standard 10-fold cross-validation. Significance level was determined to be at $p < 0.05$, stating that the original classification performance is significant when the performance values are higher than the 95th percentile of the calculated empirical distribution.

Cross-Class classification

Since the question of separability aims to answer the diversity aspect of Miyake's model of executive functions, another approach needs to be introduced to answer the unity aspect of the EFs. For this, a cross-class classification was performed. In this approach, a classifier was trained on the distinction of EF1 vs. BL and tested on EF2. The reason for this is as follows. If the functions have significant overlaps in their properties, EF2 should be classified as EF1 with above-average frequency. If there is no such overlap, the classification accuracy should be at a more random level. This evaluation is done in both directions. Hence, each EF is part of the train and the test set once. Cross-class classification is performed on ERP as well as on power spectral features.

10.2.4 Neural activation patterns

To inspect the features used for the distinction in the classification approach, a method developed by Haufe and colleagues [29] was used that transforms the weights of the SVM classifier into neurophysiological interpretable values, in so-called neural activation patterns. One classifier model is trained for each subject on the data of the respective categories. Applying the method on the model results is one activation value for each feature that was used in the classification. To create a comprehensive picture of the resulting neural activation pattern, the values are averaged within and according to the two frequency bands alpha (8-12 Hz) and theta (4-7 Hz). This is done for each subject individually, but the median values across subjects will be depicted in a color-coded topological distribution, to visualize the results. By calculating the activation patterns the underlying neurophysiological patterns that are responsible for the distinction can be inspected, which can provide valuable information analyzing the unity and diversity of different EFs.

10.3 Results

10.3.1 Behavioral data

Table 10.1 shows the average accuracy and reaction time of the subjects, summarized according to the above defined categories. Regarding accuracy, it can be stated that accuracies are higher for single than for mixed blocks, which is statistically significant as can be seen in Table 10.2. When looking at RTs, it can be seen that answers have been given significantly faster in single than in mixed blocks, as well as in repeat trials compared to switch trials. Cong only trials are overall not more accurate than cong trials but they are faster in the mixed condition. When comparing the three flanker conditions in a pairwise t-test to each other, it can be seen in Table 10.3, that there seems to be no statistically significant difference in neither of the tested combinations.

Table 10.1: Behavioral accuracy and reaction time - Average accuracy (Acc) and reaction time (RT) of the subjects categorized according to the flanker condition (cong, incong), shift condition (Single, Mixed, Switch or Repeat) and task (Odd or Greater).

	SubTask	Flanker	Single	Mixed	Repeat	Switch
Avg Acc [%]	Greater5	cong	97	90	91	90
	Greater5	incong	96	91	89	89
	Greater5	cong only	-	90	92	89
	OddEven	cong	93	86	90	84
	OddEven	incong	92	88	89	87
	OddEven	cong only	93	86	86	87
	Both	cong	95	88	90	87
	Both	incong	94	89	89	88
	Both	cong only	-	88	89	88
Avg RT [ms]	Greater5	cong	517.19	716.7	712.55	720.65
	Greater5	incong	529.13	726.34	715.24	742.24
	Greater5	cong only	-	680.73	669.70	689.16
	OddEven	cong	577.81	762.46	729.16	784.11
	OddEven	incong	601.90	783.75	726.65	836.30
	OddEven	cong only	593.11	727.86	692.44	753.43
	Both	cong	547.50	739.58	720.85	753.28
	Both	incong	565.51	755.04	720.94	789.27
	Both	cong only	-	704.29	681.07	721.29

Table 10.2: ANOVA on behavioral data - P-Values calculated for the average reaction times (RT) and accuracies (ACC) of all subjects per condition. An ANOVA was performed on a linear regression model, taking the task, flanker condition and the Shift into account. Significance level has been determined to be at $p < 0.05$.

	Shift (Block)	Task	Flanker
RT	< 0.05	< 0.05	0.38
Acc	< 0.05	< 0.05	0.99

Table 10.3: Pairwise T-test on behavioral data: More detailed analysis of the two individual factors flanker condition and shifting on block level. P-Values have been calculated for the average reaction times (RT) and accuracies (ACC) of all subjects per condition with a paired t-test. Significance level has been determined to be at $p < 0.05$.

	Flanker		
Shift block	congonly vs. cong	congonly vs. incong	cong vs. incong
Single RT	0.61	0.79	0.43
Single Acc	0.88	0.58	0.33
Mixed RT	0.39	0.21	0.68
Mixed Acc	0.93	0.72	0.61

10.3.2 Neurophysiological data

ERPs

The analysis of the ERPs shows, as in the previous studies, that there is no significant difference for the EF inhibition compared to baseline conditions (see Figure 10.3). Figure 10.4 shows the ERPs for the shifting vs. baseline conditions and in this case, it can also be seen what was found in study 2. There is no visible difference for the switch vs. repeat comparison, but several areas at more than one electrode position that differ significantly between single and mixed blocks. The grand average ERPs and power spectra for four out of the six categories are shown in Figure 10.7 to make the waveforms comparable between all conditions. It can be seen that the four conditions differ in amplitude, but the waveform remains rather constant. The ERPs of the mixed block are more positive in amplitude at Cz and Pz at several points in time compared to ERPs from the single blocks. As in the previous studies, the parietal/occipital change in alpha power can be identified as an ERD as the power decreases with an increasing amount of load and the change in frontal theta power can be identified as an ERS since the power increases with the amount of load.

Power spectra

Figure 10.5 shows the power spectra for the comparison of the inhibition demands in Single Blocks. Only in the congonly vs. incong comparison provides statistically significant differences. In Figure 10.6 the power spectra for the comparison of shifting demands can be seen. The results are in line with the ones form the ERP analysis. The trialwise manipulation (switch vs. repeat) does not show statistical significance, the single vs. mixed comparison (blockwise manipulation) however, provides significantly different areas in the signal at electrode position Cz.

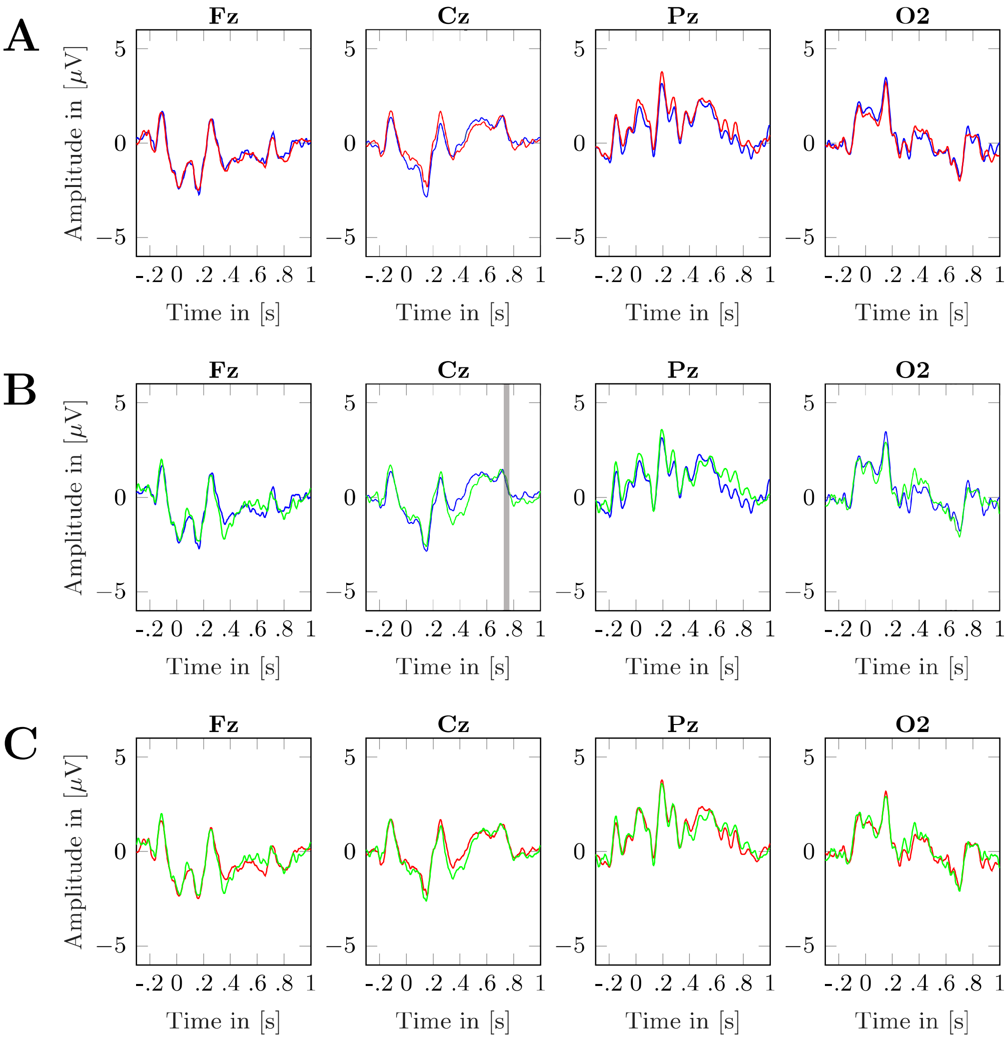


Figure 10.3: Grand average ERPs for inhibition demands Displayed are the electrode positions Fz, Cz, Pz and O2 during Single blocks. A pairwise comparison of trials with congruent only, congruent and incongruent flankers can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points). **A**: Congonly vs Cong, **B**: Congonly vs Incong, **C**: Cong vs Cong, —: cong, —: congonly, —: incong.

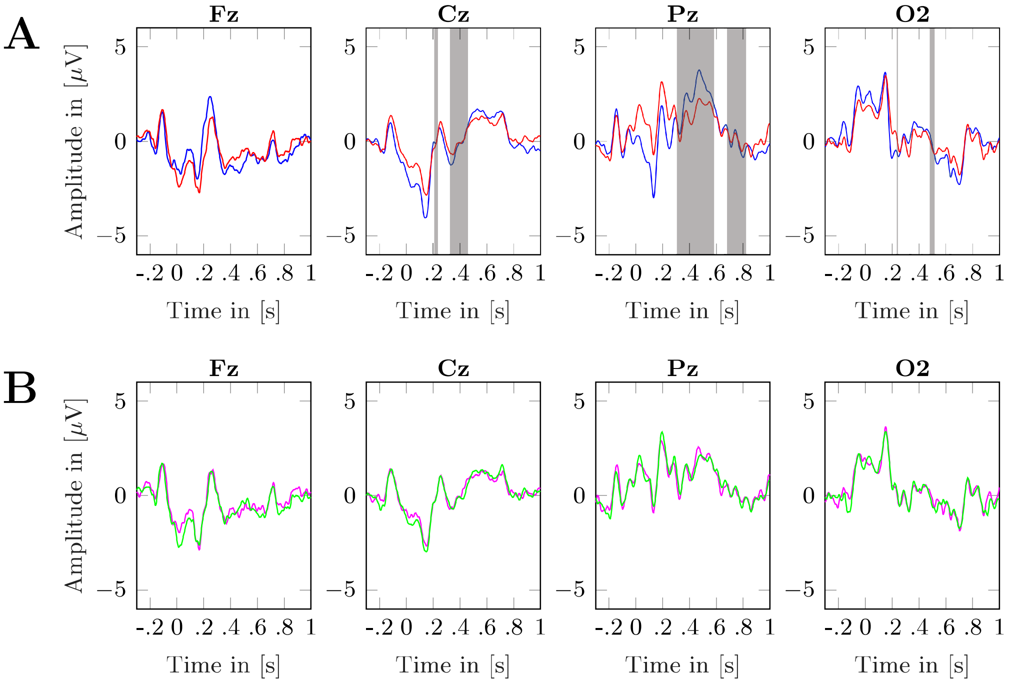


Figure 10.4: Grand average ERPs for shifting demands - Displayed are the electrode positions Fz, Cz, Pz and O2. A pairwise comparison of trials with different shifting levels during a congruent Flanker can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points). **A**: Single vs Mixed, **B**: Switch vs Repeat, —: mixed, —: single, —: repeat, —: switch.

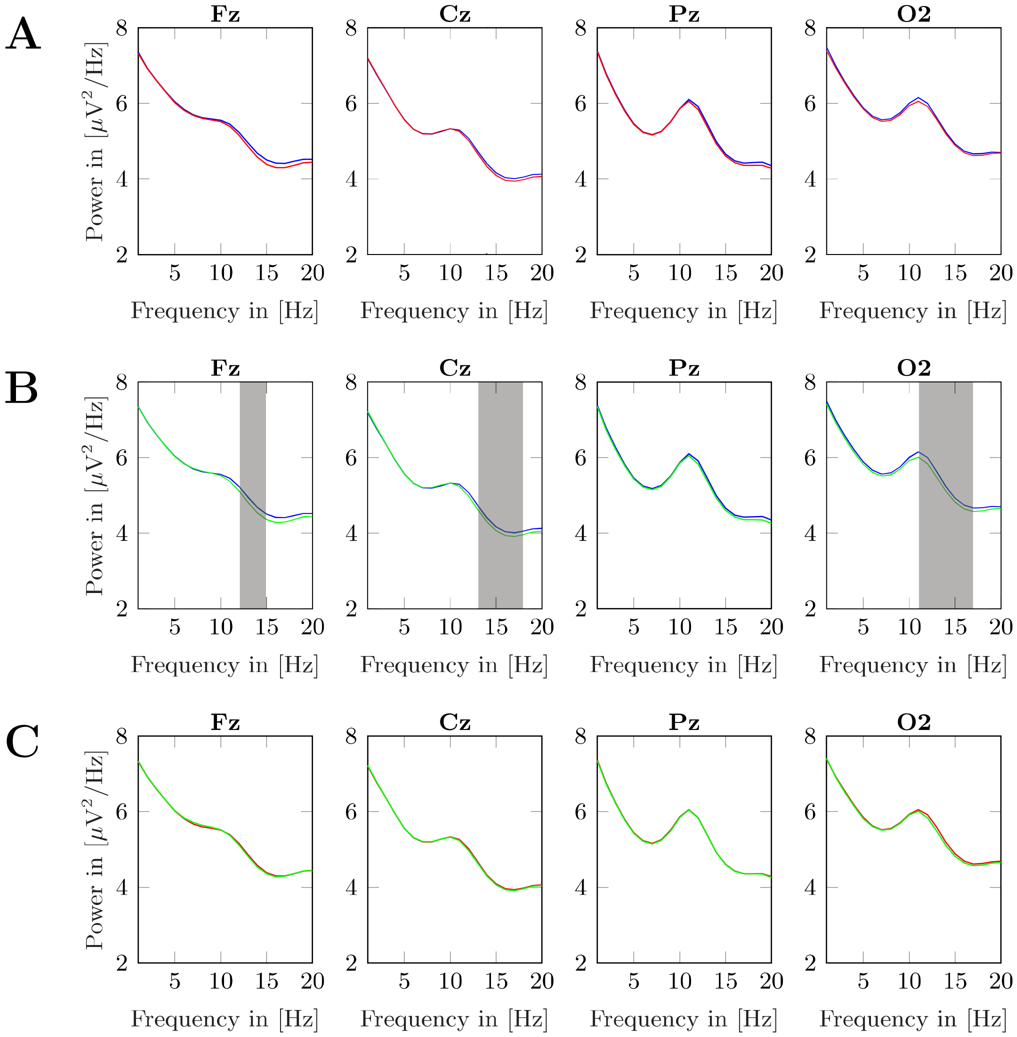


Figure 10.5: Grand average Power spectra for inhibition demands - Displayed are the electrode positions Fz, Cz, Pz and O2 during Single blocks. A pairwise comparison of trials with congruent only, congruent and incongruent flankers can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). **A**: Congonly vs Cong, **B**: Congonly vs Incong, **C**: Cong vs Incong, —: congonly, —: cong, —: incong.

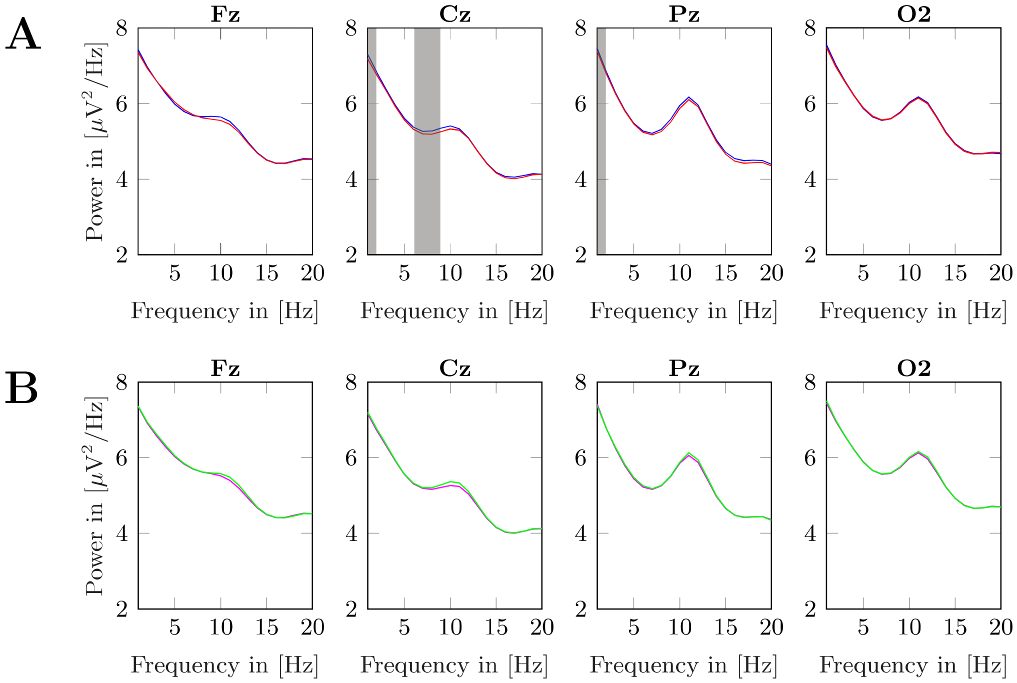


Figure 10.6: Grand average Power spectra for shifting demands - Displayed are the electrode positions Fz, Cz, Pz and O2. A pairwise comparison of trials with different shifting levels during a congruent Flanker can be seen in the three subfigures. The grand average has been calculated over all 21 subjects. Grey areas indicate statistically significant differences between the two conditions ($p<0.05$ Bonferroni corrected, according to number of time points).**A**: Single vs Mixed, **B**: Switch vs Repeat, —: single, —: mixed, —: repeat, —: switch.

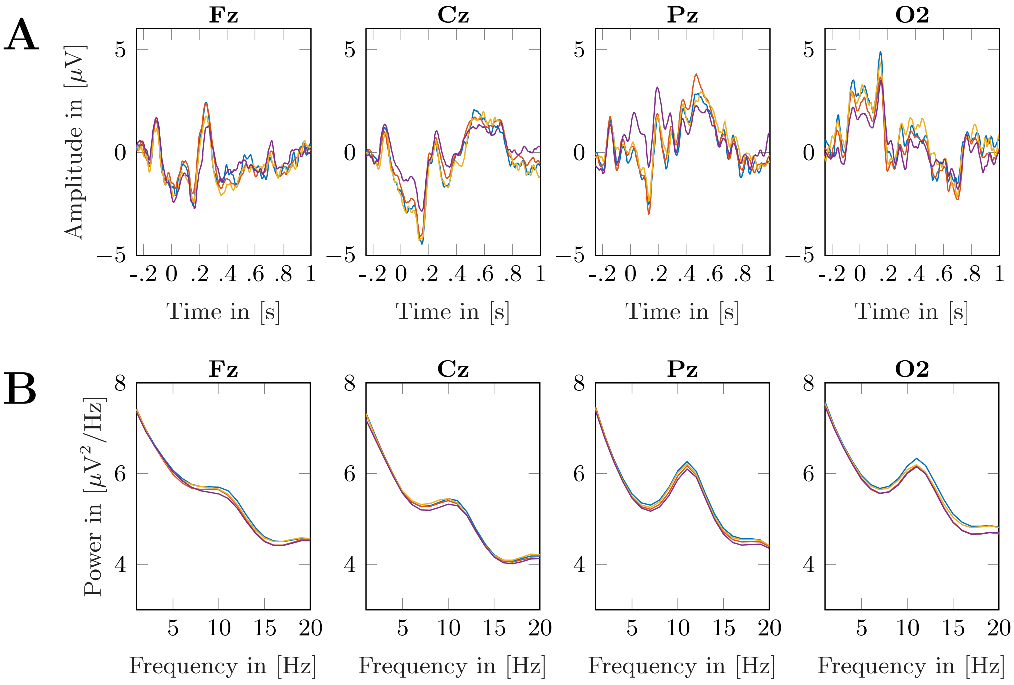


Figure 10.7: Grand average All in One: Displayed are the electrode positions Fz, Cz, Pz and O2. The grand average has been calculated over all 21 subjects. A: ERP, B: Spectra, —: single congonly, —: single cong, —: single incong, —: mixed congonly.

10.3.3 ML-based classification

The Tables 10.4 - 10.6 provide the results for the classification approach in. In Table 10.4 it can be seen, that there are clear differences between the three comparisons. The cong vs. incong comparison has by far the lowest accuracy values, whereas the other two comparisons reach statistical significance with values between 57 - 67 %. Interestingly, the performance is better for the power spectral features than for the ERP features. Table 10.5 shows the results for shifting vs. baseline demands, which was manipulated blockwise. Results above 70% can be achieved for ERP features and up to 68 % for spectral features. A clear difference can be seen for the separation during the presence of an incongruent flanker compared to a congruent or congruent only flanker. As in the previous study, the trialwise manipulation of shifting demands (switch vs. repeat) does not exceed the chance level classification performance (see Table 10.6). Regarding the classification of the two EFs against each other, it can be seen (see Table 10.7) that this is possible with over 70% accuracy. Interestingly this works better for cong trials than for cong only trials concerning ERP features, but inversely for power spectral features.

Table 10.4: Classification inhibition demands - Classification accuracies achieved with ML approach with (Mixed) and without (Single) an additional load factor of shifting. Displayed is the classification accuracy achieved with an SVM and a linear kernel during a 10 fold cross-validation. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Shift	Task	Features	cong only vs. cong	cong only vs. incong	cong vs. incong
Single	OE	ERP (CCA)	59.80 %*	62.46 %*	48.83 %
		Power (1-20)	67.68 %*	57.02 %*	53.81 %
	SG	ERP (CCA)	-	-	53.51 %
		Power (1-20)	-	-	52.36 %
Mixed	OE	ERP (CCA)	52.47 %	62.46 %*	53.01 %
		Power (1-20)	59.39 %*	57.02 %*	51.55 %
	SG	ERP (CCA)	-	-	52.26 %
		Power (1-20)	-	-	53.33 %

Table 10.5: Classification shifting demands (block) - Classification accuracies achieved with ML approach with (Flanker = incong) and without (Flanker = cong) an additional load factor of inhibition. Displayed is the classification accuracy achieved with an SVM and a linear kernel during a 10 fold cross-validation. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	Single vs. Mixed	Single vs. Mixed	Single vs. Mixed
		cong only	cong	incong
OE	ERP (CCA)	71.57 %*	68.61 %*	66.15 %*
	Power (1-20)	67.90 %*	67.34 %*	61.15 %*
SG	ERP (CCA)	-	73.53 %*	66.84 %*
	Power (1-20)	-	63.72 %*	63.59 %*

Table 10.6: Classification shifting demands (trial) - Classification accuracies achieved with ML approach with (Flanker = incong) and without (Flanker = cong) an additional load factor of inhibition. Displayed is the classification accuracy achieved with an SVM and a linear kernel during a 10 fold cross-validation. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution with permutation tests and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	Repeat vs. Switch	Repeat vs. Switch	Repeat vs. Switch
		cong only	cong	incong
Both	ERP (CCA)	52.84 %	53.33 %	50.99 %
	Power (1-20)	51.29 %	51.39 %	48.97 %

Table 10.7: Classification EFs Shifting vs. Inhibition - Classification accuracies achieved with ML approach. Displayed is the classification accuracy achieved with an SVM and a linear kernel during a 10 fold cross-validation. The used time frame contains 1.3 s from stimulus onset (650 samples) from 15 channels. ERP features were additionally filtered with canonical correlation analysis (CCA), whereas power spectral features were calculated with Burgs maximum entropy method from 1-20 Hz. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. Significance level was determined to be at $p < 0.05$.

Task	Features	Inh vs. Switch cong only	Inh vs. Mixed cong only	Inh vs. Switch cong	Inh vs. Mixed cong
Both	ERP (CCA)	67.78 %*	64.36 %*	71.92 %*	73.38 %*
	Power (1-20)	65.12 %*	64.56 %*	58.91 %*	59.56 %*

Neural activation patterns

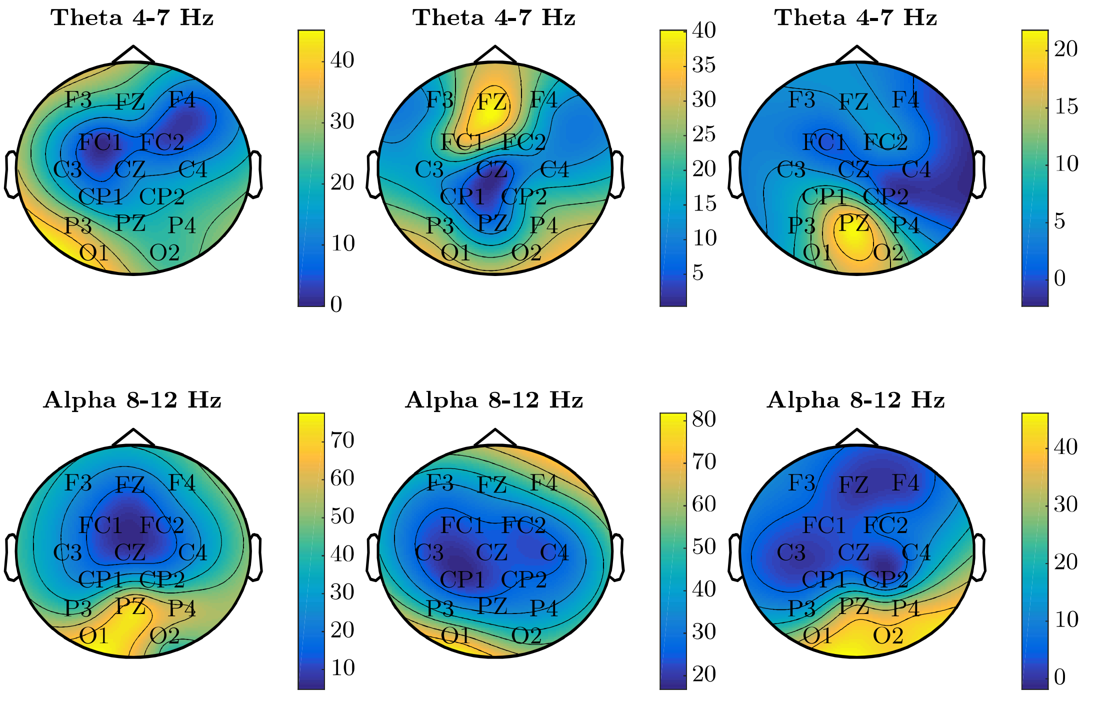
The neural activation patterns that resulted from the classification approach can be seen in the following. When looking at all three Figures (10.8 - 10.10), it needs to be stated, that fewer commonalities for the individual comparisons, but also fewer commonalities concerning the previous studies can be found in the neural activation patterns. The alpha band pattern is similar through the three comparisons but no overlap can be found for the theta band (see Figure 10.8) for the EF inhibition. The neural activation patterns for the shifting demands, block and trialwise, can be seen in Figure 10.9. The two patterns do not seem comparable, but since there were no significantly above chance level classifications for the Switch vs. Repeat classification, the depicted patterns might be due to chance and do not necessarily represent a neural process. The Single vs. Mixed distinction is characterized by parietal theta and alpha activity, which can in parts also be seen in Study 2. For the distinction of the two EFs shifting and inhibition from each other, it can be stated that occipital theta and parietal alpha seem to play an important role as can be seen in Figure 10.10.

10.3.4 Cross-class classification

Again as a last step in the analysis, the results for the cross-class classification are shown. It can be seen in Table 10.8 that most results are around the chance level. An exception is the case in which the classifier was trained on single vs. mixed blocks and therefore on blockwise manipulated shifting demands, and was tested on inhibition trials. Inhibition trials are significantly more often categorized into the single class compared to the mixed class.

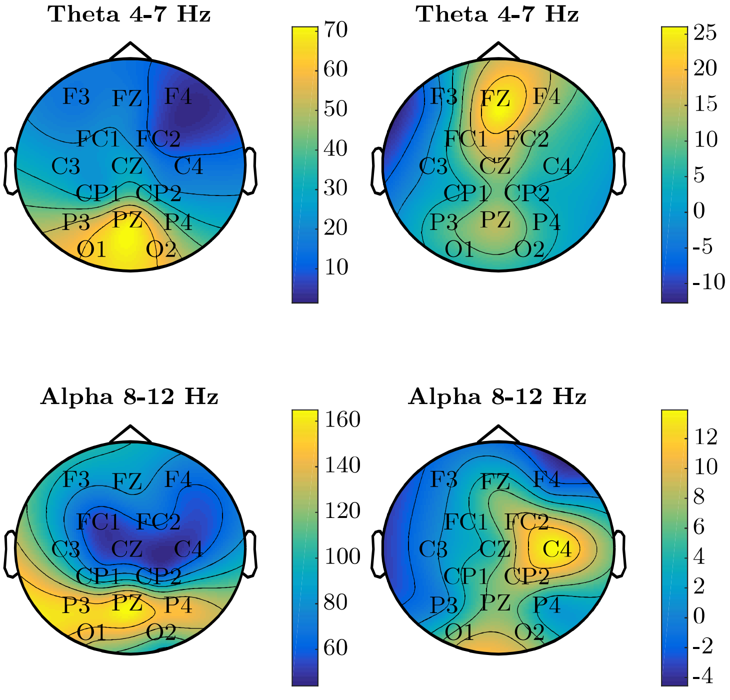
Table 10.8: Cross-class classification: The table provides cross-class classification for ERP features as well as for the power spectra. The classifier was trained on Demand 1 vs. BL and tested on trials belonging to Demand 2 only. Therefore, the here presented accuracies represent the percentage of trials classified as BL and 100 - the here displayed percentage reveals the share of trials classified as Demand 1 respectively. Statistical significance was determined by calculating an empirical null distribution and is indicated by *. The level of significance was determined to be at p < .05

Features	Trainset	Inh vs. BL	Inh vs. BL	Single vs. Mixed	Repeat vs. Switch
	Testset	Mixed	Switch	Inh	Inh
ERP	congonly	50.30 %	49.24 %	35.46 %	46.40 %
Power (1-20 Hz)		53.95 %	50.08 %	51.41 %	47.03 %
ERP	cong	50.28 %	50.72 %	29.01 %*	43.55 %
Power (1-20 Hz)		50.83 %	51.60 %	42.64 %	54.81 %



(a) Congonly vs. Cong (b) Congonly vs. Incong (c) Cong vs. Incong

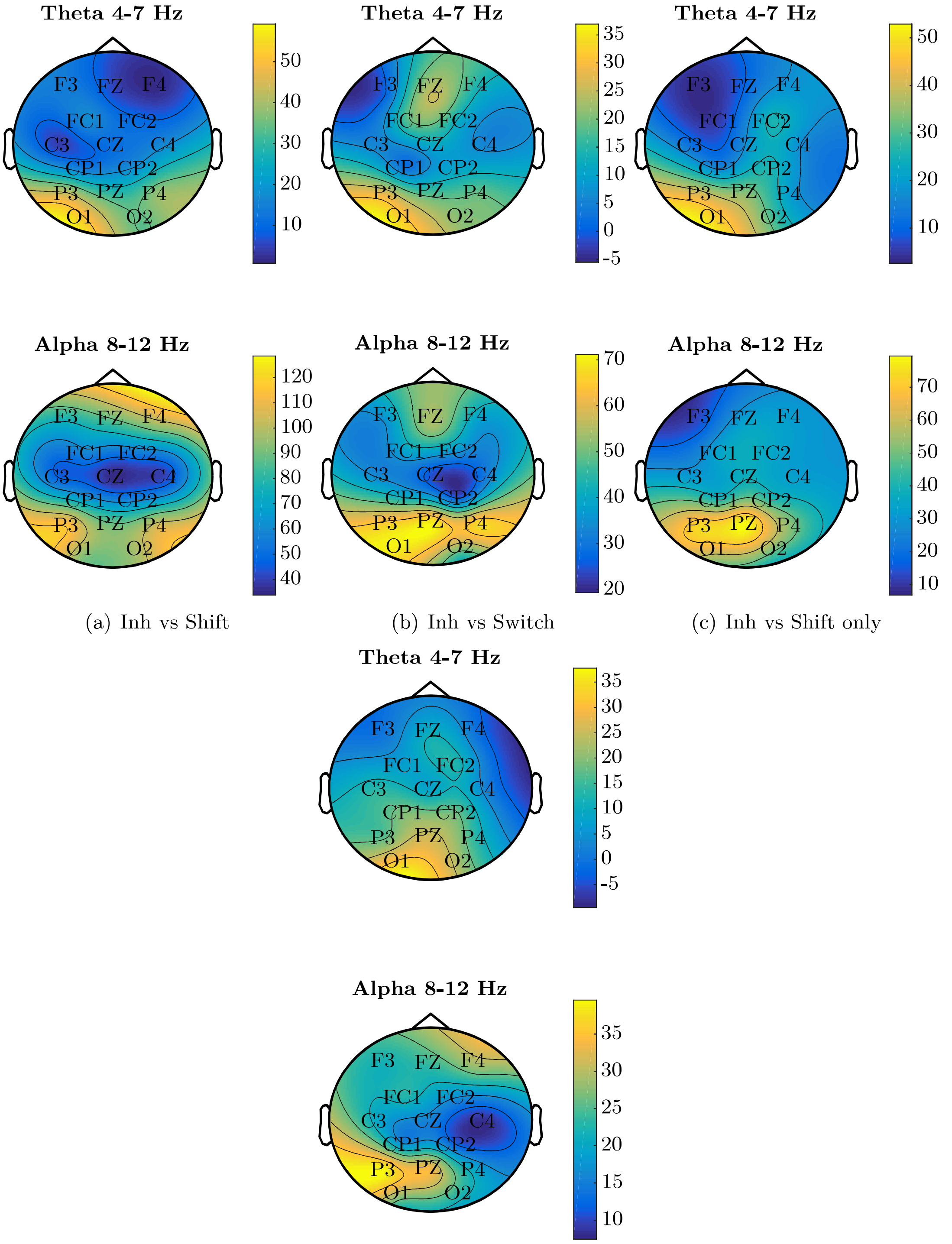
Figure 10.8: Neural activation pattern inhibition demands - Displayed is the color coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A pairwise comparison of conflict conditions congruent only (cong only), congruent (cong) and incongruent (incong) is shown for the shifting level Single. The resulting values are an average over the individual patterns of all 21 subjects. a) Congonly vs. Cong, b) Congonly vs. Incong, c) Cong vs. Incong



(a) Single vs. Mixed

(b) Switch vs. Repeat

Figure 10.9: Neural activation pattern shifting demands - Displayed is the color coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of shifting conditions Single and Mixed, as well as Switch vs Repeat is shown for the congruent flanker condition. The resulting values are an average over the individual patterns of all 21 subjects. a) Single vs. Mixed, b) Switch vs. Repeat



(a) Inh vs Shift

(b) Inh vs Switch

(c) Inh vs Shift only

(d) Inh vs Switch only

Figure 10.10: Neural activation pattern EFs Inhibition vs Shifting Displayed is the color coded activation pattern A, for the frequency bands alpha and theta in a topological distribution. The neural activation pattern has an arbitrary and undefined unit. A comparison of shifting and inhibition demands. The resulting values are an average over the individual patterns of all 21 subjects. a) Mixed vs Incong, b) Switch vs. Incong, c) Mixed vs. Incong only, d) Switch vs. Incong only

10.4 Discussion

10.4.1 Behavioral data

Since mainly the blockwise manipulation of flanker demands is of interest, the discussion will mainly focus on this aspect of the results. The behavioral data showed that congonly trials have a significantly reduced reaction time in mixed blocks compared to single blocks, which is equally present for switch and for repeat trials. This is an indicator that there seems to be a blockwise difference regarding the influence of the flanker items. In Study 3 this was described and hypothesized to be an effect of 'readiness' and therefore an increased level of attention that is required throughout blocks in which flankers varied compared to a constantly congruent flanker. This effect only gets visible in the mixed block, therefore in a condition with a high amount of cognitive load. A similar result was achieved in study 3 when the effect of reduced reaction time only appeared during the 1-back but not the 0-back condition. Therefore, this can be seen as one more indicator for shared attentional resources [110]. Single blocks in which no shifting is required are easier to solve and require less cognitive load than single blocks leaving enough resources available to not cause a measurable difference. In the mixed block, however, more resources are required, which is why the impact of constantly varying flanker items does produce a measurable effect, as attentional capacities seem to reach a limit. Not finding a measurable influence on accuracy could mean that the limit of attention is not overstepped, but still within bounds leaving the error rate unaffected. Overall, all behavioral effects of Study 2 could be replicated.

10.4.2 Neurophysiological data

Regarding the neurophysiological, the same observations as in Study 2 could be made, regarding the Single vs. Mixed and the Switch vs. Repeat comparisons. Likewise for the cong vs. incong comparisons of the flanker. Therefore, potential reasons and hypothesis will not be repeated here. For the newly introduced manipulation of inhibition demands, almost no differences have been found, except in the power spectra for the congonly vs. incong comparison. Overall, again this leads to the assumption that either potential differences between the experimental conditions cannot be assessed with EEG or that existing difference vanish and lose statistical significance due to averaging over 21 subjects.

10.4.3 ML-based classification

The classification performance that has been achieved in this experiment is very comparable to Study 2. The main finding was that only the chance level performance was achieved for inhibition vs. baseline as well as for switch vs. repeat classifications, but a good performance for single vs. mixed classifications. Concerning the newly introduced blockwise manipulation of inhibition demands, it can be stated that the influence is rather small compared to the blockwise manipulation of shifting demands. Since this result was already achieved in Study 3, when comparing inhibition to blockwise manipulated updating demands, this was to be expected. Despite the rather small influence, there are measurable differences caused by the blockwise manipulation of inhibition demands. Since the classification performance drops notably in the distinction of inhibition and shifting when congonly trials are used compared to when cong trials are used, could indicate that the additional load factor simplifies the distinction. More attentional focus is needed with varying flankers, increasing the overall cognitive load during the shifting task and thereby increasing the difference between the two. Cross-class classification again reveals that in most cases there are no significant amounts of joint feature sets between the two EFs shifting and inhibition. Interestingly though, inhibition trials (incongruent from single blocks) were categorized significantly more often into the single block than into the mixed block. This is theoretically not surprising, as this is the only case in which the train and test set are from the same 'block'. In practice, it leads to the question of why this result did not already come up in Study 2. It could be investigated that the overall mean is influenced by a small number of subjects with high accuracies or it the general tendency of all subjects is above the chance level. If the result is only based on individual subjects, this difference between the two studies can easily be explained.

10.4.4 Neural activation patterns

Despite the similarities between Study 2 and 4 regarding all so far discussed results, the neural activation patterns do not match well. Concerning the comparisons in which only the chance level performance values have been achieved, this is not necessarily surprising. In those cases, no distinguishable features have been found, and therefore, no meaningful patterns. Regarding differences between Study 3 and 4 in terms of neural activation patterns, it can be hypothesized that the difference between numbers and letters as stimuli is one of the reasons for this. Processes regarding number processing are known to be strongly active independent of the task at hand because it is such a deeply rooted mechanism. Repeating both Studies, with numbers and letters would be necessary to evaluate this hypothesis. Further, it can be assumed that the differences between shifting and inhibition are more variate between subjects, making it harder to find a repeating pattern. The classification results clearly speak in favor of unique variance of the shifting and inhibition. The difficulties of finding a stable pattern that characterizes the two would explain why Miyake and Friedman claim that no unique variance between the two EFs exists.

10.5 Conclusion

Overall, it can be concluded that the results are very similar but not identical to study 2 and 3. Again it could be shown that a blockwise manipulation of inhibition demands has an influence on behavior and measurable signals, but the difference is not as severe as for the manipulation of shifting demands. Further evidence for the unity and diversity of the EFs inhibition and shifting could be collected with the help of machine learning approach. The machine learning played a crucial role in the determination and quantification of the influence of the blockwise manipulated inhibition demands, which could not be made visible with conventional group-level statistics. Overall, many more studies need to be performed to find patterns that fully characterize each EF on a comprehensive level.

Chapter 11

Discussion

What can be learned from the results of all four studies based on characterizing executive functioning in EEG with the help of machine learning ?

In this chapter, the results from all four presented studies will be discussed and concluded in an overall manner. The discussion will be divided into four categories, summarizing the results concerning Miyake's model of EFs (unity and diversity), potential effects that arise through experimental design (block effects), practical implications of the findings and lastly the advantages that are achieved by adding an ML approach to conventional group-level analysis on psycho-physiological data.

11.1 Unity and diversity

The central objective of the four performed studies was to reveal if Miyake's model of EFs can be corroborated with EEG data. The question was investigated by introducing a particular task design that allows to pairwise compare EFs within one experiment and the same subject. Since the question is mainly about the characterization of individual variance, a task design that puts the individual in the foreground and allows comparisons within the individual is indispensable. Statistics teach us that it is of great importance in which order dimensions and their values are averaged to be able to make statements about the variance and distribution of the variables to be examined. Since unity and diversity are difficult to quantify, it is difficult to make a statement about the extent to which the variance within an individual is an issue. Besides, there is the well-known problem of latent variables, which can hardly be controlled and recorded, especially between different individuals. So, if the relationship of EFs within a subject is estimated first, the variance of latent variables can be actively minimized, since it should eliminate itself. In general, it is more likely to find constants in the relationship than in the expression of the individual EFs themselves, especially in rather small sample sizes.

The chosen task design, combined with the single-subject analysis via machine learning enables to pursue this endeavor and provided promising results. Findings could be replicated within the set of performed studies with different groups of subjects and although the repeated measurements were no exact replications. Overall it can be stated that especially the calculation of neural activation patterns and the cross-class classification are tools that provided the most value regarding the question of unity and diversity if EFs. On the one hand, they complete the classical group-level statistics and add information to the analysis that statistics only cannot provide. Patterns can be extracted, and a statement about potential effect sizes of the patterns can be made by looking at the classification accuracy at the same time. On the other hand, they also validate the usage of an approach like this because the neural activation patterns make the machine learning aspect transparent and comprehensible. It can easily be assessed if the used patterns 'make sense' considering the literature of previous research in the field. In the performed studies they did because all relevant features are known to strongly correlate with WML.

The most substantial evidence for the diversity of EFs is provided by the cross-class classifications. The classification accuracies indicate that no significant joint overlap of features is present because the values are not significantly above chance level. Also, individual activation patterns could be found for all EFs which could in part be replicated over all four studies.

Concerning unity, it could be shown that the amount of working memory load differs but is commonly distributed over the three functions. They all share the typical ERS/ERD components that are known to be characteristic for WML. The amount of load is in these studies must be seen as relative to the respective task, as they differ in their subjective difficulty level and cannot be regarded as equal. The amount of load can, therefore, be seen as a common EF property.

For future research, it is indispensable to perform more studies in a similar design, but with different tasks that induce load on the individual EFs. Only by this, it can be validated if the found neural activation patterns can be linked to the EFs directly or just to task-specific properties. Also, it is necessary to further characterize the relationship of the EFs to each other, to validate if there is (no) unique variance for inhibition, as has been questioned by Friedman et al. [52, 53]. To this end, the results of this thesis would indicate individual differences, but this needs to be confirmed.

11.2 Block effects

During the analysis of the studies, it has been found that there is a major difference between a blockwise compared to a trialwise manipulation of EF demands. In blocked conditions, usually, trials with two different levels of difficulty are compared to each other, whereas in a trialwise comparisons, trials are compared with similar trials that are only minimally different from each other. Therefore, the assessment of the two manipulation types block vs. trial measures constant vs. spontaneous effects. It is rather intuitive that constant effects are easier to capture than spontaneous effects because they usually include less variation. For some tasks and effects, this is a well-known fact, but for others, this aspect has so far been neglected and in some cases not even been considered at all. For the Eriksen-Flanker task, for example, no preexisting study compares the influence of blockwise variation of flanker congruency, only various variants of within block manipulations. Usually, the focus of these studies is different from the one set within the scope of this thesis, but nevertheless, knowing about blockwise effects within a Flanker task can help to understand processes and mental tactics that are used by subjects. Interestingly two different forms of block effects were found. One type, that has been detected in Study 2 for the EF shifting, shows indicators of differences between block and trial wise manipulation within in all measured variables. Behavioral data, as well as neurophysiological data in the classical group-level statistics, show significant differences. The second form of block effect was found for the EF inhibition in Study 3 and 4, in which no signs of differences are found under low levels of cognitive load, neither in behavioral nor in neurophysiological measures, but significant effects under high levels of cognitive load in both measures while using group-level statistics. Especially the latter form leads to the hypothesis, that a limit of attentional capacity was reached that in consequence leads to longer RT and (in parts) in reduced accuracies on task. Many studies and theories emphasize shared attentional resources of EFs and define this aspect as a common underlying mechanism [110], [116], [42]. Finding that an additional load factor influences the performance and leads to poorer results concerning RT and accuracy can be seen as an indicator for limited resources. As long as there is enough capacity the task can be fulfilled without loss, overstepping the capacity, however, leads to the neglect of at least one of the task to ensure an overall stable and goal-directed performance.

For future research, this effect should also be tested for updating demands. Within the current design, this was not possible, as the n-back task does not allow a trial wise manipulation of updating demands. Another task that induces updating demands needs to be used instead. Overall, the detection of the effect leads to the conclusion that an even more careful experimental design is required. It seems to be of great importance to find the individual variance that can be linked to an EF itself and not only to general WML effects. It seems likely that mental strategies vary between a block and trialwise manipulated conditions and therefore also the way of mental processing. This is a crucial aspect that needs to be kept in mind when mental processes are the main focus of the investigation.

11.3 Practical implications

Apart from the gain in theoretical knowledge of psychological processes, it can be argued if there are other practical implications of the here presented results. To be able to detect and identify specific types of load and not only workload in general would be a desirable goal. Cognitive load theory (CLT) [117] describes the type and amount of cognitive load that should be imposed on a learner to achieve optimal learning results. Assessing the mental state a learner is in, based on the amount of WML, can be very useful to keep a learner in a comfortable range to achieve optimal learning success. Being able to identify which specific load a learner is under could be a useful extension to educational applications to find an ideal way of presenting information to avoid an overload. Unfortunately, the usage of 'simple' SVM classification approaches did not provide indicators for a possible use case. Classification accuracies do not reach values that are drastically above chance level, which is necessary to provide any usability and improvement for human subjects [118]. Nevertheless, it would be interesting to follow up on this thought by using more sophisticated and powerful machine learning approaches that aim to separate the EFs. To date, some approaches make even neural networks more transparent to understand what the algorithm learns such as the relevance backpropagation [119]. Exploring neural networks with the framework of this studies would, therefore, a potential gain that should be pursued.

11.4 The advantage of the ML approach

In summary and conclusion, it can be stated that the usage of ML on a single-subject level adds valuable information to classical group-level statistics. There are several aspects, which in part have already been named in the previous sections, but will be recalled here in brevity to complete and round off the chapter. The classification approach adds the single-subject level to the analysis, which can be an interesting factor in understanding cognitive processes. It helps to understand the pronunciation of individual effects within a sample. On the one hand, because it can easily be deduced if every subject has performed on an equal level or if the effect has strong fluctuations between subjects. On the other hand, classification accuracy can also be seen as a kind of effect size for the ensemble of information that has been used in the classification. Both types of information would not be accessible with standard group-level statistics. Calculating effect sizes can only be done on a group-level and only for a small a prior defined excerpt of the data. Using the coefficient of determination is a possibility to assess the strength of an effect for a broader amount of information, but is rarely, if at all, done on a single-subject level. Considering subjects individually is generally only done to determine outliers that affect the overall data negatively. Therefore, measures exist that determine whether a subject individually performs according to the expectations, but are not used to a greater extent.

Within the four studies, the properties of a single-subject helped a lot to understand and quantify the strength of the load that has been imposed on the EFs and to rank the importance and characteristics of the effects. This was of particular interest concerning the trialwise manipulation of shifting demands. It could be shown that despite significant effects on the behavioral data level, no effects on neurophysiological level could be found, neither on the single subject nor on the group level. The exact opposite was the case for the trialwise manipulation of inhibition demands. Under low cognitive load, effects were invisible on the behavioral level but could be made visible on the single subject level for neurophysiological data.

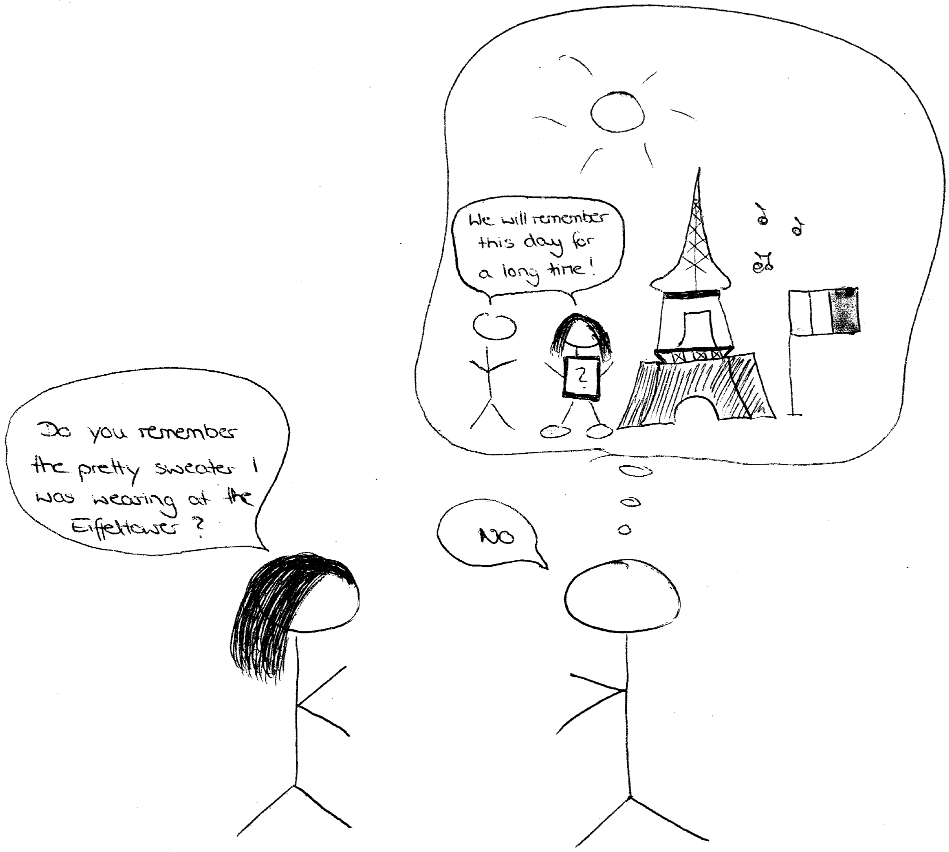
The use of SVMs enabled to calculate neural activation patterns from the parameters of the machine learning algorithm that were crucial for the decision making. One the one hand, this was a great advantage because the SVM is able to take the full high-density data into account in the process of distinguishing the two EFs from each other. Group-level statistics rarely do that due to limitations of multiple comparisons. On the other hand, the machine learning approach enabled to visualize patterns that are characteristic and therefore, discriminative for each EF. Within the studies, the group-level analysis of ERP and power spectra only provided indicators for the unity of the EFs. Differences between the individual experimental conditions were found that correlated well with WML in general, but no more specific information could be drawn out of the analysis. The calculated neural activation patterns showed more precisely where differences between the two EFs can be found and how they are characteristic for the individual EF.

A last important advantage of the machine learning approach was the usage of cross-class classification. It provides information about the overlap of patterns or features between the classes of interest. If a classifier that is trained on a different set of features, performs above chance level, it can be assumed that the tested features have joint properties that the classifier is able to pick up. In the case of the four studies, cross-class classification was not significantly above chance level, with minor exceptions, indicating that each EF has its own variance. If one of the EFs does not have own variance, it would be more highly unlikely to achieve the here presented results.

Overall it can be stated that the explanatory approach that is pursued by standard group-level statistics can be extended and well complemented by the machine learning approach. More insights and more levels of information are added to the analysis that round off the picture of the acquired data. The new insights are able to explain more details and aspects that are of importance to understand the cognitive processes at hand.

Part III

Episodic memory processes


We will remember this day for a long time!
Do you remember the pretty sweater I was wearing at the Eiffeltower?
No

Chapter 12

Introduction into episodic memory

The second case study will focus on episodic memory constructs. To complete the overview of Chapter 5, a more detailed description of episodic memory is given, with a major focus on memory encoding, decisions based on episodic memory content and in terms of confidence and content.
This part will deal with one recognition memory task paradigm, which is inspired by a study performed by Fukuda and colleagues [120]. For conciseness, only the relevant information regarding the task and the involved cognitive processes will be given.

Episodic memory belongs to the explicit (declarative) part of long-term memory. Aside from episodic memory, which contains events and facts that have occurred in personal life, there is semantic memory, which refers to a factual knowing of things and general world knowledge. Episodic memory is therefore responsible for storing experiences of specific situations as well as visio-spatial and timing information of events. In contrast, semantic memory is responsible for the storage of, for example, all the knowledge learned in school including historical dates and biographical information, Latin names of plants and animals or laws of physics.
To store and retrieve information at a later point in time to and from episodic memory, various memory processes must be activated. In the following, three processes will be defined and presented to get an overview of certain aspects of episodic memory. The three processes are the central anchor points of this case study and form the basis of the research questions to be answered. Since all the processes can also be summarized under the term perceptual decision making, a few more sentences will be used to depict the concept.

According to Sternberg and colleagues [71], perceptual decision making can be broken down into three stages: sensory encoding, decision formation, and motor execution. Sensory encoding and decision formation are described in different theoretical frameworks. In a detailed review, Gold and Shadlen made efforts to identify and dissociate the two processes [121]. Two main theoretical groundworks are the basis of this: signal detection theory [122] and sequential sampling framework [123]. Signal detection theory deals with the inability to discriminate between the real sensitivity of subjects and their (potential) response biases caused by conditions of uncertainty. The concept of sensitivity describes the objective difficulty of the task, whereas the bias describes the effect of the consequences a decision could have. Missing or detecting a stimulus according to its sensitivity level can be quantified by reaction time and in terms of EEG correlates by P300 latency [124]. The sequential sampling theory, on the other hand, states that the performance of a subject in an experimental task depends on two main factors: the quality of the stimulus information and the quantity of information required before making a response.

12.1 Memory encoding: Quality of memorization

The process by which information enters long-term memory or episodic memory can be described as memory encoding. According to Baddeley [35], this is controlled and gated by the so-called episodic buffer, which acts as a link between short- and long-term memory. The success of memory encoding is defined by the ability to recall the information later, whereas failure would be characterized by forgetting the information. The memorization of information can be captured in EEG data and characterized by several correlates. On the one hand, a larger sustained positivity at frontal electrodes has been found to be present during the successful encoding of information [125], [126], [127]. On the other hand, it has been found that alpha-band activity is more suppressed during the encoding of items that are later remembered than for those that are later missed [128], [129], [130]. Apart from increased synchrony in the alpha-band, it has also been reported that beta-band power synchrony increases [129], [128], [131], as well as an increased frontal theta coinciding with enhanced gamma power in posterior cortex regions for later remembered (success) versus later forgotten items (failure) [132].

In particular, one further study is of importance, which will be the model for the performed studies within this part. Fukuda and Woodman performed an old/new recognition test in which a series of pictures was shown that should be remembered [120]. In a test phase, memory was tested with known and new pictures, which should be distinguished one by one. The authors claimed that a prediction of the success of memory encoding is possible in real-time. However, this was only done in a post-hoc group-level statistics approach. One study that tried to realize a machine learning based prediction of memory encoding was performed by Noh and colleagues [133]. They used three individual classifiers based on pre-encoding (power), and the actual encoding process (ERP and power). A combination of the three achieved performance values around 59.65 %.

12.2 Memory retrieval: Familiarity and recognition

The process of accessing information stored in memory can be described as memory retrieval. Retrieving information from memory can, in many cases, be described as recognizing something as already known. Judging if something is familiar is often not difficult and can be made intuitively with high precision. Retrieving more detailed information about a situation or specific item, however, is often more difficult. According to dual-process theory, the quality and quantity of recalled information are divided into two different processes that can be distinguished on a neuronal level, namely familiarity and recollection [134, 135]. Familiarity is a form of retrieval in which no context or detail information is available. An item, a word, or a face seem familiar, but qualitative information cannot be recalled [136, 137]. Recollection, in contrast to that, is a more effortful process describing the retrieval of information, including specific details of the item or the context from which it is known [134]. Both processes can be combined in a phenomenon called old/new effect, describing differences between old and familiar items and new so far, not familiar items. The early old/new effect at frontal areas 300-500ms after stimulus onset (FN400) is presumed to reflect familiarity [138, 139]. There is also a parietal and late old/new effect, which is a positive going event-related potential (ERP) observed between 500 and 800 ms, which associates with the process of recollection [140]. Animal studies (mainly rodents and primates) have been performed to get a better understanding of processes that are going on the visual cortex that is responsible for the initial processing of visual information. These kinds of studies often apply invasive measurement techniques, for example, single-cell derivations. As a general result of these studies, it has been found that the anterior inferior temporal cortex is important for visual discrimination and plays a big role in recognition memory [141]. Several types of neurons have been identified that fire during item recognition, depending on how new or familiar the item is. For example, there are individual neurons that fire when something is new (novelty neurons) [142], but there are also recency neurons that fire when an item has recently been seen [143]. It has been found that the information about novelty or recency holds within the level of neurons up to 24h [143] before it vanishes. Due to this memory system on a neuron level, it is possible, with single-cell derivations, to distinguish between new and familiar items within 100 ms after the presentation by the neuronal signals.

12.3 Decision making: Confidence about memory

Certainty in decision making is an important prerequisite in everyday life, helping to make informed and reasonable decisions in difficult circumstances. Decision confidence plays a role in facilitating adaptive regulation of behavior and supports decision making in complex situations. It affects how subsequent actions are planned or how something can be learned from mistakes that have been made. Decision confidence is also crucial for planning actions in a complex environment, especially when subsequent decisions depend on each other or the final outcome of a situation [144, 145]. Unfortunately, it is not straightforward to extract decision confidence from behavioral or neurophysiological data as the concept is deeply intertwined with other concepts. One example is evidence or situation evaluation, which is essential to judge a current state correctly. An accumulation of evidence and constant reevaluation of the available facts requires a broad chain of thoughts which interacts with decision confidence [146]. Another example is the strong correlate of reaction time with decision confidence as well as the error rate, which also varies greatly with the confidence of the current and previous decisions. It was found that certainty is inversely correlated with reaction time and directly correlated with accuracy and motion strength [147]. Therefore, the level of confidence can easily be confounded with sensory evidence or the planning and execution of motoric actions. Also, it could be shown that previous choices and the respective feedback influence future decisions [148]. More general, Gherman and colleagues [149] state that establishing certain confidence for a decision relies on the same mechanism as the choice formation itself. Kiani and colleagues [150] found that neurons in the lateral intraparietal cortex (LIP) represent evidence accumulation in monkeys. Since it has been established that the outcome of previous choices influences the current ones, it seems appropriate to suggest to add a fourth stage to Sternberg and colleagues concept of perceptual decision making, namely outcome and feedback evaluation. General effects concerning the neural correlates following positive or negative feedback that can be found in almost all settings are error-related potentials and feedback-related negativity (FRN). Both belong to the class of event-related potentials (ERPs). The error-related negativity (ERN), for example, was observed in 1990 by Falkenstein et al. [17], time-locked to the presentation of an erroneous event peaking at 80-150 ms. The potential appears strongest at frontal and central electrode sites, has its origin in the anterior cingulate cortex (ACC) [151], and it seems to be linked to error processing [152] and reward prediction [153]. The error-related negativity is often followed by error-related positivity peaking 250-500 ms after stimulus onset, which is generated in the posterior cingulate cortex (PCC). This positive component is associated with conscious error perception [154]. There are event-related potentials that are specifically associated with feedback perception. Especially the feedback-related negativity (FN) is a phenomenon often reported as a negative deflection 145-300 ms after unexpected feedback [155]. It is located frontocentrally and seems to be equal to the N200 component. Interestingly the FN only appears when the feedback is presented immediately after a decision or reaction. The time frame, which still counts as immediate, is at least one second long, according to Weinberg and colleagues [156]. When too much time passes, the FN is no longer visible, but the P3 component remains unaltered, even if the delay is up to six seconds long. Concerning decision confidence, literature states that error-related EEG signals vary in a graded way with the level of confidence [157], and also that, it was found that error positivity (Pe) varies in amplitude with subjective confidence. Both facts show that decision confidence and error detection are closely related processes [158].

12.4 Idea and hypothesis

The literature regarding episodic memory and its related processes is manifold and difficult to keep track of. Not only does episodic memory comprise the components of information encoding, storage, and retrieval, but also other processes that can be associated with the latter. Examples are the level of detail that can be remembered or the confidence with which certain details can be recalled. Many studies from the field focus on one of the processes in isolation within a very specific setting. Examples include but are not limited to the ability to recall familiar and unfamiliar faces, free recall and cued recall of information, confidence during gambling/situations of uncertainty, or the unfolding of confidence with growing amounts of information. To create a systematic overview that comprises of all core processes in a common and simple setting, a set of four studies was created within the scope of this thesis. All four studies follow the same design of an old/new recognition test, in which a set of pictures is studied during a training session (see Figure 12.1). In a test session, familiar pictures from the learning session are shown interleaved with new and unfamiliar pictures. The task is to decide for each picture in the test session if the picture is familiar or not. The processes under investigation within all four conducted studies will be:

- depth of memory encoding (remembered or not remembered)
- familiarity of the presented stimulus (known or unknown)
- levels of decision certainty (100 and 75 %)

The results for each of the processes will be dealt with in a separate chapter starting with memory encoding in Chapter 14, stimulus familiarity in Chapter 15, and lastly decision certainty in Chapter 16. The four studies will be the data basis for all three investigations, which is why the task design will be described only once in the upcoming chapter (Chap. 13). The chosen design of the task allows investigating all processes throughout different phases of the task, from the encoding to the outcome evaluation. In total, four different phases are defined, starting with the **encoding** of the stimuli during the learning session of the study. As a second phase the **presentation** during the recognition test is considered, followed by the **decision** phase which covers the decision making of the subject via a button press and at last, the **feedback** phase in which the subject processes the categorical feedback stating if the given answer was correct or wrong. The aim is to acquire findings that cover the basic mechanisms of the three stated processes without additional restrains.

For each of the three processes, one phase is of particular interest. For the process of memory encoding, this is, naturally, the encoding phase in which the subject tries to actively remember the just presented stimulus during the training phase of the experiment. For the process of stimulus familiarity, this is the presentation phase. In this phase, the subject actively tries to recall the stimuli from the training of the experiment. Lastly, in the process of decision confidence, the decision formation but also the decision evaluation are of the most interest, making the decision and the feedback phase the target of the investigation. For each of the just stated combinations of cognitive processes and phases of the experiment, literature-based hypotheses exist about the anticipated effects. Therefore, it is legit to perform classical hypothesis-driven group-level statistics. For all other combinations, however, no concrete hypotheses exist. Nevertheless, it could be of great interest to investigate whether correlations between the cognitive process in question and other phases of the experiment exist. Overall, this can be seen as following up on all processes that can be linked to the decision of the subject under the frame or label of one specific research question, which could be highly beneficial for the general understanding of human behavior. However, as stated in the problem statement, such an exploratory and multiple testing of undirected hypotheses is highly problematic with methods from the classical group-level statistic. For this purpose, the power of machine learning is exploited to perform exploratory data analysis. The properties of ML, including the generalizability and the avoidance of overfitting, enable to do a more broad analysis of the data without running into the trap of losing significance due to multiple comparisons. For this reason, ML will be applied in this part as a tool for exploratory data analysis, to allow to make more of the available data.

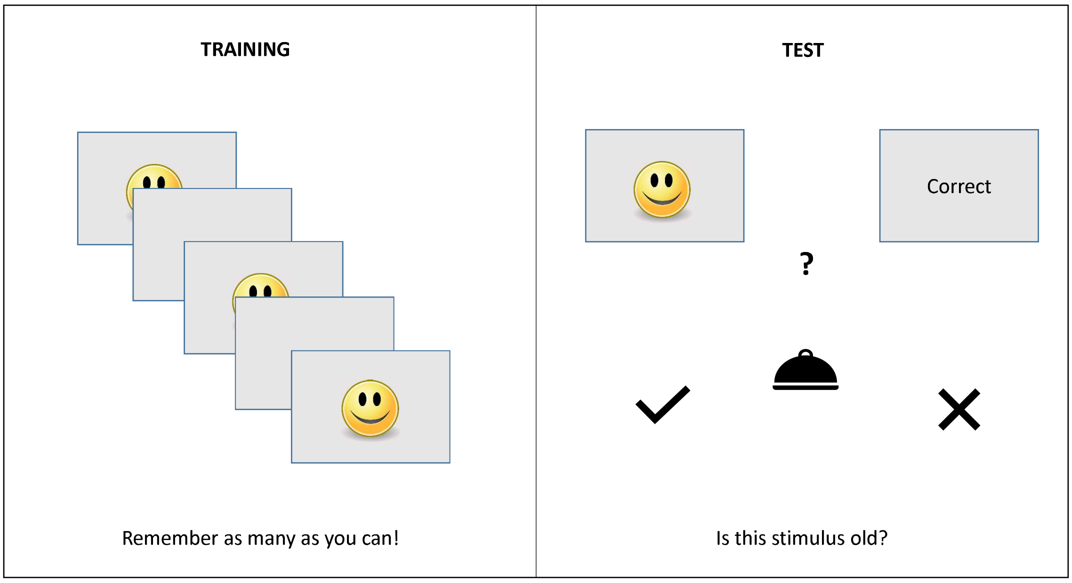


Figure 12.1: Exemplary depiction of a recognition test design - In a training session pictures are presented in a serial manner and the subject is asked to remember as many as possible. In a recognition test, the subject is again presented pictures in a serial manner, but this time the subject is asked to decide and answer if the picture is familiar from the training or new. All four studies that will be conducted in this part, follow this design.

Chapter 13

Studies I - IV: Task design and data analysis

In this chapter, a series of four studies will be presented that all evolve about a simple old/new recognition paradigm in which the subject is asked to learn items during a serial presentation, which shall be recognized in a testing phase. Starting as a reproduction of Fukuda and Woodman's study from 2015 [120], the row of studies resulted in a hypothesis-generating research approach that covers central questions in episodic memory. In total, four studies have been performed covering processes such as information retrieval, decision making under different levels of confidence as well as feedback and performance evaluation within the same experimental setting. Due to the similarity of the studies, the design, technical setup, and methodology will be presented once in the beginning. Differences between the studies will subsequently be highlighted, and the results are sorted and accumulated by a question to show the findings in a focused manner. The results of the following studies have in part been published in [159] and [160], but will be presented in detail in the following chapters.

13.1 Task design

The task design is based on a study originally performed by Woodman and Fukuda [120]. The task was modified and implemented with minor differences concerning the choices of possible levels of decision confidence. In general, the experiment is an old/new recognition test and therefore, divided into a study phase in which the subjects were asked to memorize as many pictures as possible and in a test phase in which a mixture of new and already studied pictures were presented and the shown pictures needed to be identified as known or unknown. In the study phase, a series of 500 pictures were presented. A schematic sketch of the course of the experiment can be seen in Figure 13.1. All pictures are presented in a block design to enable breaks in between the study phase. One block consisted of 50 pictures, after which a break can be made for as long as needed. The continuation of the experiment is controlled manually by button press by the subject. Each picture presentation can be seen as a separate trial. As stimuli, the same picture dataset as in the Fukuda and Woodmans study was used [161]. The dataset contained pictures displaying daily life objects on a white background without many details. In the test phase, a series of pictures are presented to the subject, again in a

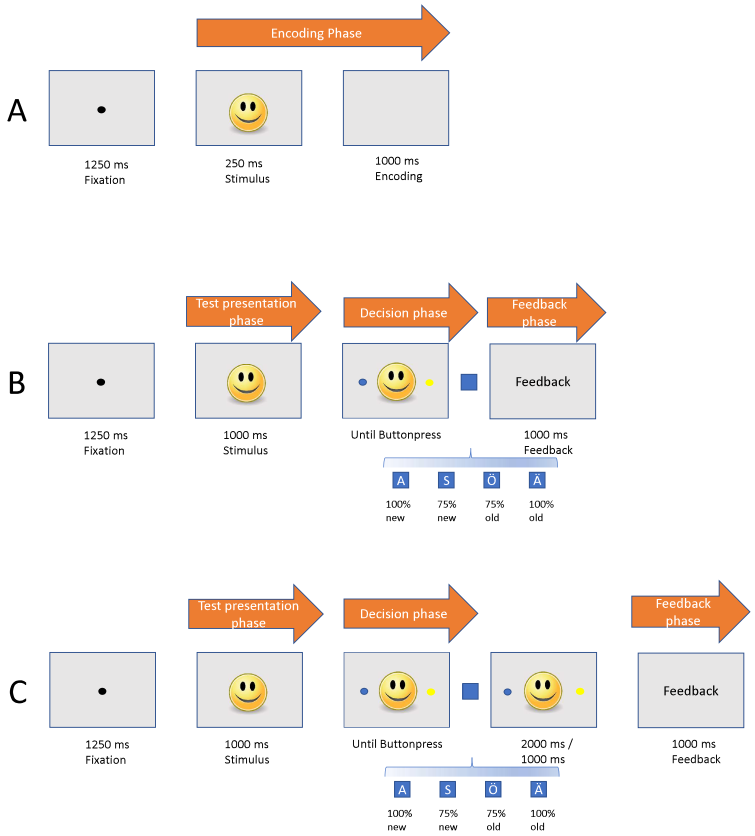


Figure 13.1: Experimental setup - The four different phases of the experiment (Encoding, Presentation, Feedback, and Decision) are indicated by the orange arrows over the appropriate time frames, to indicate which time frames have been used in the analysis. The different rows, indicated by the capitalized letters at the front of each row, represent the following stages. A: Training phase (study I-IV), B: Test phase of study I, C: Test phase of study II, III, and IV. The stimulus is depicted by a smiley and the button press by a blue rectangle. The timing and duration of each frame and phase is noted below each frame.

block design with a break after 50 pictures each. The subject is asked to decide after each picture presentation if the picture is new or already familiar. To answer the question, one out of four options must be chosen: 100 % new, 75 % new, 100 % familiar, 75 % familiar. The percentage represents the decision confidence with which the answer is given. To choose one of the four possible answers the keys A, S, Ö or Ä on a standard German keyboard were used. A stands for (100 %) and S for (75 %) on the left side, Ö (75 %) and Ä (100 %) for the same values on the right side. Which side to choose (right or left) is indicated by circles that appear next to the picture as soon as an answer is required (1000 ms after stimulus presentation). The circles are blue and yellow, blue represents familiar, yellow represents new. The sides on which the blue or yellow circle appears switches after each block to exclude confounds due to the handedness of a subject. The switch is indicated before the new block starts, and within a block, no changes are made to avoid confusion.

After the choice was made by the subject, feedback is presented for 1 s indicating if the choice, independent of the certainty, was right or wrong. After the feedback presentation, the next picture is presented, and the subject has to decide again about the familiarity of the picture. The same set of pictures was chosen for each subject, whereas the order of presentation and group affiliation (new or studied) was randomized.

13.2 Differences between studies

The four studies differ in their design only in the test phase of the experiment. In particular, three criteria have been altered, which are enlisted in the following:

- Total number of pictures in the test phase
- Ratio of old and new pictures in the test phase
- Time between button press and feedback presentation

Table 13.1 gives an overview of the studies concerning the three named criteria. In addition, the recording time of the studies is added in the table, since this also differed between the studies due to the enlisted changes in the criteria. The changes in design evolved dynamically after the assessment of the results of the individual studies. Only a brief reasoning is given here to not anticipate the interpretation and discussion of the results. Nevertheless, for the sake of clarity, the main reasons for the changes are outlined briefly. A more detailed reasoning will be given later. Firstly, the number of presented pictures in the test phase altered between 750 and 500, to be able to compare a balanced as well as an unbalanced number of old vs. new stimuli. The ratio of old and new pictures and the total number of presented pictures go hand in hand and can, therefore, be seen as one factor instead of two. To avoid findings that are solely based on effects by the ratio of stimuli, balanced and unbalanced ratios were tested to evaluate the resulting differences and their cause. Secondly, the time between button press and feedback presentation was varied between 0 and 2 s. The reason for this was that the original design (time lag of 0 s) did not offer the possibility to separate feedback evaluation from decision making. The button press is immediately followed by a feedback presentation by which any correlates related to the decision making might get lost due to new input processing. The choice of a time lag of 2 s brought other difficulties, which will later be stated and evaluated, which

Table 13.1: Overview of design differences - The parameters that vary between the four studies are listed tabularly. The number of pictures refers to the number of presented pictures in the test phase of the experiment. The ratio of pictures presented in the test phase was either balanced (1:1) or unbalanced (1:2), whereby in the unbalanced case, twice as many old as new pictures were presented. The time lag between button press and feedback presentation is specified in seconds (s), the total recording time of the experiment in hours (h).

Study	Number of pictures	Ratio of pictures	Time lag	Recording time
		New : Old	Button and Feedback	Total
I	750	1:2	-	1.15h
II	500	1:1	2s	1.2h
III	750	1:2	1s	1.3h
IV	500	1:1	1s	1.1h

after all lead to a final value of 1 s. For a better understanding, Figure 13.1 also visualizes differences regarding in part B and C.

13.3 Participants

For each study, a new group of subjects was recruited. The participation was voluntary for every subject and could be ended at any time. All subjects gave written and informed consent and received a reward of 8 euros per hour or credits relevant for their study for their participation. The study was approved by the local ethics committee of the University of Tübingen and performed in accordance with the declaration of Helsinki.

13.3.1 Study I

In study I, a total of ten subjects participated, five males and five females, that were 22.7 (±3.91) years old on average. From the ten subjects, seven were right-handed, and all had normal or corrected to normal vision. One subject needed to be excluded from the analysis due to technical difficulties during the recording.

13.3.2 Study II

In study II, a total of 11 subjects participated, two males and nine females, that were on average 20.45 (±1.13) years old. All subjects were right-handed and all had normal or corrected to normal vision. Two subjects needed to be excluded from the analysis due to technical difficulties during the recording.

13.3.3 Study III

In study III, a total of 11 subjects participated, one male and ten females, that were on average 20.27 (±2.19) years old. Ten subjects were right-handed, and all had normal or corrected to normal vision.

Two subjects needed to be excluded from the analysis due to technical difficulties during the recording.

13.3.4 Study IV

In study IV, a total of 10 subjects participated, seven males and three females, that were on average 22.6 (±4.93) years old. Eight subjects were right-handed, and all had normal or corrected to normal vision. One subject needed to be excluded from the analysis due to technical difficulties during the recording.

13.4 Technical setup

To perform the experiment, the subjects were seated in front of a computer screen (19 inches) on which the task was presented. The experiment was programmed and presented in Matlab using the Cogent graphics extension. A standard keyboard was used for entering the answers by the subject. For the recording of the electroencephalogram (EEG) data, the software BCI2000 [162] was used sampling the data with a frequency of 512 Hz. A

Brain Products Acticap system and two 16 channel g.tec g.USBamp amplifiers were set up for the EEG recording. The integrated high pass filter was set to 0.1 Hz and the integrated low pass filter to 100 Hz. Additionally, a notch filter between 48-52 Hz was applied to eliminate power line noise. 29 electrodes were used for the recording and placed according to the extended 10-20 system (FPz, AFz, F7, F3, FZ, F4, F8, FC3, FCz, FC4, T7, C3, Cz, C4, T8, CP3, CPz, CP4, P7, P3, PZ, P4, P8, O1, Oz, O2, PO7, POz, PO8) and three additional electrodes were used for electrooculogram (EOG) recordings at the outer canthi of the eyes and one on the forehead between the eyes. The ground and reference electrodes were placed on the right and left mastoid respectively, and impedances were kept below 10 kΩ. To ensure that stimulus timing is accurately saved in the data, we used the parallel port connected to the EEG amplifier.

13.5 Data analysis

Three different processes are of particular interest in the data analysis: **Depth of encoding**, **Stimulus familiarity**, and **Decision confidence**. The three processes will be analyzed separately in the following chapters within the set of the four conducted studies. Irrespective of the process, the data will be analyzed in terms of reaction time and task accuracy, neurophysiological signals, and classification accuracy based on neurophysiological features. Since the methodological approach is the same in all four studies, again, the procedure will be described once to avoid unnecessary repetitions.

The data can be categorized according to the predefined answers a subject was able to give for each trial. The question ”Is this stimulus old or new?” could be answered with one of the following four options:

- Old - 100 % or 75 % sure
- New - 100 % or 75 % sure

Since each of those answers can either be correct or wrong according to the true label of the stimulus (Old or New), a total of eight categories can be distinguished. The task itself allows a division into four phases covering the **encoding** and **presentation** of the stimulus, the **decision** making and the **feedback** evaluation. Despite the vast number of distinctions that can be made, there are two main categories for each of the processes that are of particular interest:

- Depth of encoding: Remember - Forgotten
- Stimulus familiarity: Old - New
- Decision confidence: 75% - 100 %

For a better overview, the data structure of categories is visualized in Figure 13.2.

At this point, a further remark on the labeling and assigning of categories to each trial might be necessary. For the depth of encoding, the categories can only be assigned after the subject has made the decision about the familiarity of a trial in the recognition test. Otherwise, it is not possible to know whether the subject remembers a trial or has

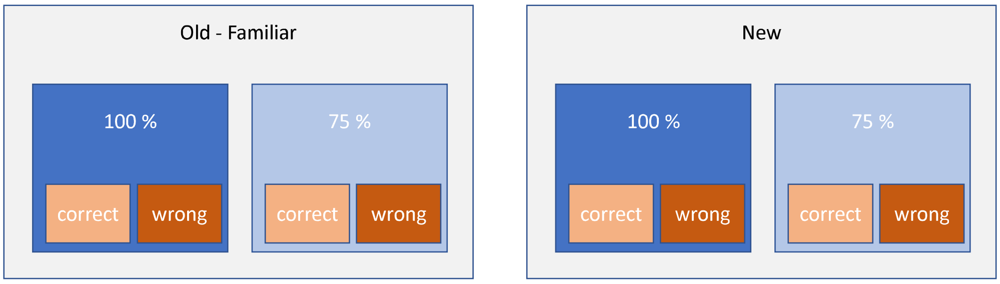


Figure 13.2: Data structure of episodic memory studies - The gray boxes represent the level of familiarity in the data, which can be distinguished into old and new stimuli. The blue boxes represent the level of decision confidence with which the subjects answer for each stimulus and the orange boxes represent the level of encoding which either succeeded or failed. To be precise, only the orange boxes on the right represent the level of encoding, since new stimuli are not encoded in memory yet.

it forgotten. The labeling of trials in the encoding phase, which happens long before the recognition test, is therefore made in retrospective. This also holds for trials in the presentation and decision phase. For the stimulus familiarity, the categories old and new are defined by experimental design and are known for each trial beforehand. The labeling is, therefore, based on objective instead of subjective measures, compared to the depth of encoding. Lastly, for the decision confidence, again, the categories can only be assigned after the subject has made a decision about each trial with either high (100 %) or low (75 %) confidence. Therefore, the labeling is subjective and assigned in retrospective to trials of the encoding, presentation and decision phase, but directly linked to the respective feedback phase.

It is clear that not every phase is representative of each of the three to be investigated processes. In some of the phases, it might not even be possible to directly link potential correlates and effects to one of the processes. But, nevertheless, it is an interesting concept to follow up on correlates that are related, even if only indirectly, to each of the processes through the entire experiment. Therefore, it has been chosen to analyze the data of each process with classical and ML-based approaches concerning the most representative phase. For the process of encoding the encoding phase will be investigated because in this phase the process of stimulus encoding into memory is represented best. For the process of stimulus familiarity, the presentation phase will be investigated. Lastly, for the process of decision confidence, the feedback phase will be investigated.
In addition to this focused analysis on one phase of the experiment, however, all other phases are also investigated with ML in an exploratively driven approach to gain further knowledge regarding the respective process. Therefore, for each of the three processes of interest there is a classical and an exploratory analysis of the data, to get the most complete picture of the data and a deeper understanding of the processes that are investigated.

13.5.1 Behavioral data analysis

The behavioral data will be analyzed with respect to reaction time and task accuracy. A two-sample t-test is applied to find out if the performance and speed of the subjects throughout the task differ concerning the main process. Also, an ANOVA will be performed on a linear regression model to be able to assess statistically significant differences between and across all four studies.

13.5.2 Neurophysiological data analysis

The neurophysiological data will be analyzed on an ERP basis. The data preprocessing and analysis was performed in Matlab 2015b [96]. Firstly, a bandpass filter between 1 and 40 Hz was applied to the recorded EEG signal, and the signal was corrected for EOG artifacts using a regression method proposed by Schoegl [23].
For each of the above described categories, the data was cut into trials of 1 s length, starting with stimulus onset (see Figure 13.3). For the encoding phase, the stimulus onset starts with stimulus presentation, likewise in the presentation phase in the recognition test. In the decision phase, the onset of the phase is chosen to be at 250 ms before button press which marks the decision of the subject/ onset of the phase starts with the presentation of the circles, the point in time at which the subject was allowed to press a button. The feedback phase starts with the onset of the feedback presentation.
Each trial was baseline corrected (-100 ms to 0 ms prestimulus or relative to the corresponding event) and cut into trials of one or 1.25 s length depending on the respective categories. Figure 13.3 shows a schematic overview of how the data is preprocessed from the recording to individual trials. For the ERP analysis, additional filtering was performed to exclude trials exceeding 80 or -80 μV from the analysis. After choosing representative channels for the evaluation, the grand averages over all subjects were calculated for each category individually to reveal if there are differences in the time domain. The statistical significance of the differences in ERPs was established by using a Wilcoxon Ranksum Test [102]. The resulting p-values were Bonferroni corrected according to the number of used observations [103]. To chose adequate channels that are representative for the investigation the coefficient of determination (R^2) is computed to see which variance in the data can be explained by the class label (for a detailed explanation see Section 2.1.4 in Chapter 2).

Since the latency is of significance here to precisely locate the effects, the ERPs are corrected for all displayed studies concerning the raster latencies of monitors [163], which are always present but rarely accounted for. For the specific model that was used for the performed studies, a latency of 28 ms occurs from the trigger until stimulus presentation. The displayed ERPs are, therefore, all corrected according to this latency to have an adequate stimulus onset.

Additional analysis: Memory Encoding

To be able to compare the results achieved within the four studies in this thesis and the original study of Fukuda and Woodman [120], additionally, a time-frequency representation was chosen for the analysis. In a time-frequency representation, the EEG signal is transferred into the frequency domain, as has been done previously in this thesis by

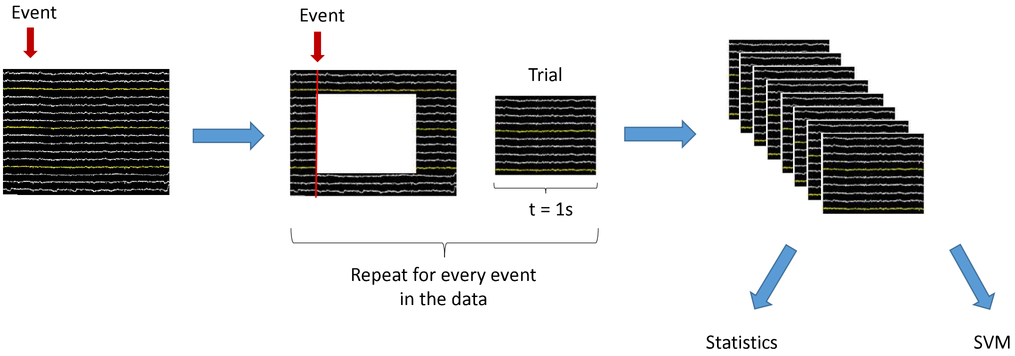


Figure 13.3: Neurophysiological data analysis - Visualization of preparation of data for the neurophysiological analysis. The data of the EEG recording is cut into trials according to events that specify the beginning of a trial. For that, the epoch (time) and channels (locations) of interest are extracted from the full recording and saved in a new data structure to easily access it in the later analysis. The analysis is then either performed with standard group-level statistics or the machine learning approach.

applying Burgs maximum entropy method [20], but in dependency of time. By this, changes over time within the power spectra can be visualized. The standard approach is to use a sliding window, that slides over a larger time frame of interest with a fixed window size (here: 400 ms) and a fixed window overlap (here: 380ms). A power spectrum is then calculated for each window. For points in time that were sampled and calculated several times during this process, a mean value is calculated from all available power spectra. This results in a time series of power values which can then be graphically displayed.

In the EEG data analysis very often graphically represented time series are compared with each other. Instead of simply plotting them next to each other, it is also possible to subtract them mathematically from each other and thus plot only the differences graphically. This method is also known as the difference of means because the mean ERPs of the conditions to be compared are subtracted from each other. For some effects, this is the preferred technique because the differences can be assessed more concise at one glance.

13.5.3 Classification

A SVM with a linear kernel (C = 1) [94], [8] was applied to differentiate between the categories introduced above using the libsvm implementation for Matlab [95], [96]. To ensure stable results 10-fold cross-validation for each comparison was implemented. The datasets (training set as well as the test set) were balanced for each comparison, by removing all spare trials if one of the classes had more trials than the other. For classification, two different types of features were used: ERPs in the time-domain and power spectra of the EEG data. The time-domain features are based on 0-1000 ms epochs, starting at stimulus onset of the 21 channels (AFz,FPz, F3,FZ,F4,FC3,Fz,FC4,C3,CZ,C4,CP3,CPz,CP4,P3,PZ,P4,O1,Oz,O2). All other electrodes were discarded to reduce the influence of noise and artifacts in the data. Due to the sampling rate of 512 Hz, one trial of ERP data is represented by 512 x 21 features. As a way to improve the signal-to-noise ratio of the data, a spatial filtering method based on canonical correlation analysis (CCA) was applied [26]. The filter aims to minimize the variance within a class and to maximize the variance between classes to improve the separability. Classification using features from the frequency domain was conducted on the power spectra between 1-20 Hz calculated on the same time frame (0-1000 ms after stimulus onset) with Burgs maximum entropy method for the same 17 channels (20 x 21 features).

To evaluate the performance of the classification approach, the accuracy was reported, averaged over all subjects. To evaluate the potential influence of the reaction time (RT), the classification was performed on the RT as well. Since a certain level of accuracy can already be reached by chance, depending on the number of classes and used trials per class, the statistical significance of the classification results needs to be established. To achieve that an approach is used that estimates the chance level of classification performance by calculating the binomial cumulative distribution [164]. This approach gives rather generalized and conservative bounds, based on sample size and the number of classes. Classification results exceeding the estimated chance level can, therefore, be seen as statistically significant.

$$P(z) = \sum_{i=z}^{n} \binom{n}{i} \cdot \left(\frac{1}{c}\right)^{i} \cdot \left(\frac{c-1}{c}\right)^{n-1} \tag{13.1}$$

Equation 13.1 describes the binomial cumulative distribution, in which P(z) represents the probability of predicting the correct class at least z times by chance. An appropriate z can be chosen by multiplying the number of samples n with the desired significance level (chosen to be at 0.05). C thereby represents the number of classes. The approach should only be applied when the classes are balanced. Since they are in the classification approach, this is a suitable measure.

13.5.4 Activation patterns

To inspect the features used for the distinction in the classification approach, a method developed by Haufe and colleagues [29] was used that transforms the weights of the SVM classifier into neurophysiological interpretable values, in so-called neural activation patterns. One classifier model is trained for each subject on the data of the respective categories. Applying the method on the model results is one activation value for each feature that was used in the classification. To create a comprehensive picture of the resulting neural activation pattern, the values are shown in heatmaps presenting the calculated activation pattern for each data point in the used time window of 1s for all EEG channels utilized in the classification. This is done for each subject individually, but the median values across subjects will be depicted in a color-coded topological distribution, to visualize the results. By calculating the activation patterns the underlying neurophysiological patterns that are responsible for the distinction can be inspected, which can provide valuable information analyzing each of the three processes of recognition memory (encoding, stimulus familiarity, and decision confidence).

Additional analysis: Stimulus familiarity

For stimulus familiarity, the results suggested that a direct comparison of behavioral and classification performance could be of interest (see Chapter 15). For visualization, a barplot was chosen to easily spot differences between the two measures for each subject and each study. Of particular interest was to evaluate if classification accuracy based on the ERPs in the presentation phase is significantly better than the behavioral accuracy. Hence, a one-sided t-test was performed on the values of the behavioral accuracy and the average classification accuracy of the 10-fold cross-validation of each subject. Subjects for which a p-value of $p < 0.05$ was achieved were marked accordingly in the barplot.

Chapter 14

Memory encoding

In this chapter, the results of all four conducted studies will be presented while focusing on the aspect of the depth of memory encoding. The aim is to find out if the depth of memory encoding is reflected in the encoding phase of the experiment because this is where the process of memory encoding takes place. The two classes or conditions under investigation are items that are later remembered, and items that are later forgotten. The label remembered and forgotten is assigned to the items by the subjects themselves, as only their subjective rating can qualitatively identify an item as remembered or forgotten. Therefore, the labels are assigned to the trials during the encoding phase in retrospective based on the decision of the subject which takes place in the decision phase, and therefore, at a much later point in time in the experiment.

Especially concerning the inspiration of the studies performed in part III, which was the publication of Fukuda and Woodmann [120], it is also of interest to investigate the process of memory encoding in terms of practical applications. The original study claims the feasibility of real-time prediction of memory encoding. In other words, the authors state that items that will later be forgotten can be identified based on their EEG correlates during the encoding phase. Having knowledge about the success of learning during the learning itself and not only in a query test after the learning has been completed would be of high value for the development of learning applications. However, an indispensable prerequisite for this would be the reliable prediction and accurate identification of the depth of memory encoding. Therefore, another question of this chapter will be: Can the depth of encoding be predicted based on physiological signals, and is the accuracy high enough to use it in a practical application? Lastly, the possibility to do exploratory data analysis with the help of machine learning will be explored to evaluate possible benefits that can be gained by this methodology.

Is there a difference between good and bad encoding in neurophysiological data during the encoding of a stimulus?

14.1 Results

As in the previous sections, the analysis will start with classical group-level statistic approaches and continue with the ML-based analysis approach. The conditions good and bad encoding will be represented by remembered and forgotten items, respectively.

14.1.1 Behavioral data

Table 14.1 shows the task accuracy and the RT for remembered and forgotten items of all four studies. It can be seen that except for study I, answers have been given faster for remembered items compared to forgotten items. The ratio of remembered vs. forgotten items, representing the accuracy on task, is similar throughout all four studies. An ANOVA, however, revealed that no significant effects concerning the memory encoding were found within the behavioral data (all p-values above 0.05). The above-stated difference in reaction time is, therefore, not significant.

Table 14.1: Encoding Acc and RT - Represented are the numbers of remembered and forgotten items in percent with respect to the number of all familiar trials from the training phase which were presented in the test phase. The reaction time for each trial is presented in seconds. Both measures are averaged over all trials and subjects of the respective study of the experiment.

Study I	# Pictures [%]	ReactionTime [s]
Remembered	74.40	1.213
Forgotten	25.60	1.119
Study II		
Remembered	72.80	0.953
Forgotten	27.20	1.148
Study III		
Remembered	79.08	0.702
Forgotten	20.92	1.028
Study IV		
Remembered	68.04	0.591
Forgotten	31.96	0.765

14.1.2 Neurophysiological analysis

The neurophysiological analysis of the data regarding the two conditions remembered vs. forgotten items revealed that there seems to be no common pattern within the elicited ERPs between the four studies. Indicators for that can be found in Figure 14.1, 14.2 and 14.3. Figure 14.1 shows the calculated R^2 values in a heatmap. As previously stated, the R^2 values put the variance that can be explained by the two classes in relation to each other. The Figure shows that the values are small and that the distribution of the values over all electrodes positions and points in time varies greatly between the four studies. Figure 14.2 shows the grand-average ERPs of remembered vs. forgotten items. To make the results relatable to the original study of Fukuda and Woodman [120], the ERPs are depicted in three categories at channel Fz, as in the original study. The first two categories

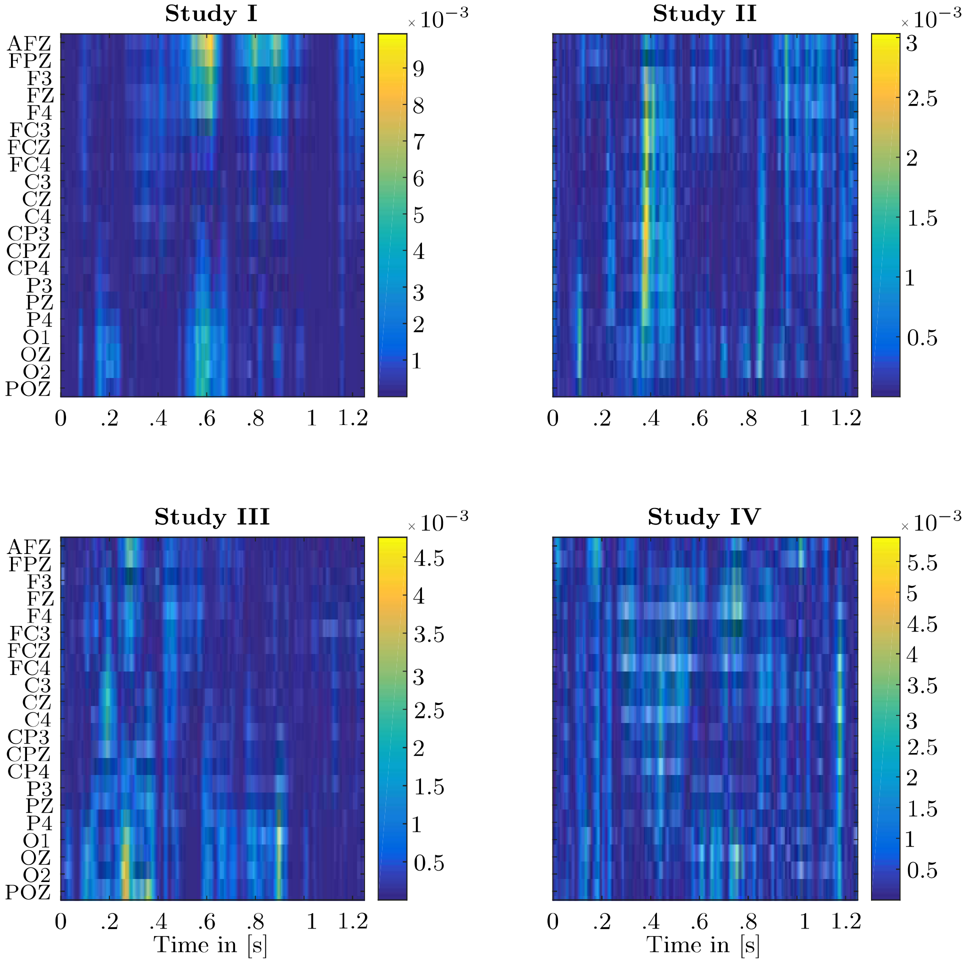


Figure 14.1: R^2 Values for ERPs remembered vs. forgotten in the encoding phase - R^2 values describe the variance in the data that is explained by the class label, in this case trials that were later remembered or forgotten. The subfigures account for the experimental studies I-IV in ascending order. The R^2 values are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

represent remembered items with high or low confidence, and the third represents forgotten (missed) items. Again, it can be seen that the plots vary greatly between the four studies, and only slight commonalities can be found. Especially intriguing is the variation of the categories between the studies. Therefore, no common pattern regarding the relation between remembered vs. forgotten items can be found. Lastly, Figure 14.3 shows a time-frequency representation of signal generated for channel Oz (as in the original study). This plot gives the same impression as the previous two presented before. There is no characteristic of the signals that is equal or steady between the four experimental studies. The relation of remembered (high or low confidence) and forgotten items is different in each study.

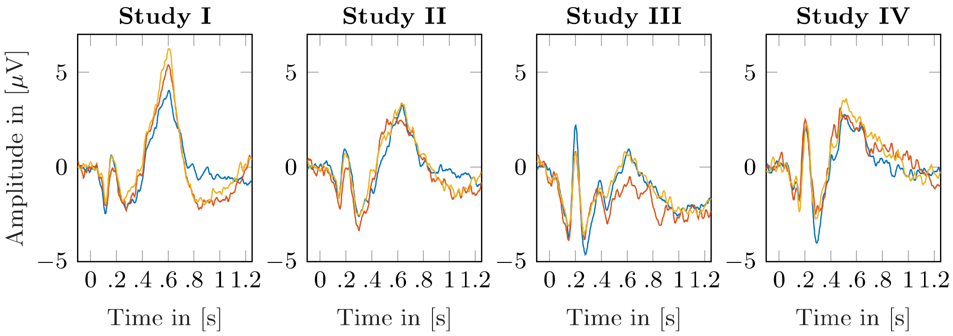


Figure 14.2: Grand average ERPs in encoding phase remember vs. forgotten - Displayed is the electrode positions Fz, calculated for each point in time of the 1250 ms time frame (x-axis) and amplitude of the signal in μV (y-axis). The grand average has been calculated over all 9 subjects). —: high confidence, —: low confidence, —: missed

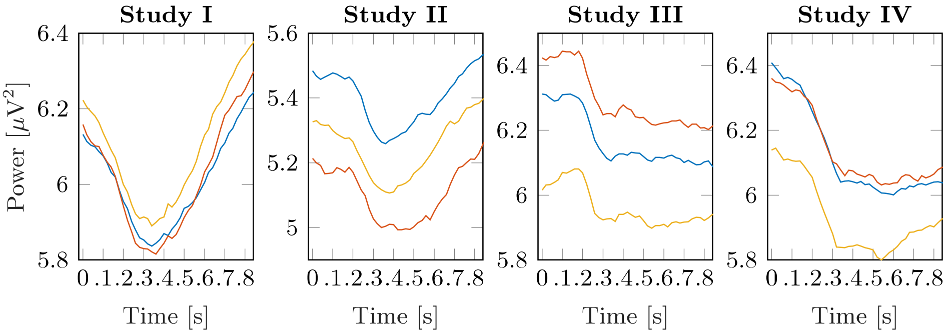


Figure 14.3: Time-frequency representation of encoding phase - Displayed is the electrode position Oz. Frequency values have been calculated in a sliding window over time frames of 40 samples each, along the 1250 ms encoding time frame. —: high confidence, —: low confidence, —: missed

14.1.3 ML-based classification

The ML-based classification approach has been performed to evaluate if a prediction of remembered vs. forgotten items is feasible on a single trial basis. The performance values achieved by the SVM can be seen in Table 14.2. It can be seen that in all four studies, accuracy values around 50 % have been achieved. Therefore, it needs to be stated that no accuracy values significantly above chance level can be found for the prediction of remembered vs. forgotten items in the encoding phase.

Table 14.2: Classification on ERPs remembered vs. forgotten - Classification on ERPs (CCA filtered) of the encoding phase, 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - remember vs. forgotten. Accuracies marked with * are significantly above chance level (0.05) according to binomial cumulative distribution.)

	Study I	Study II	Study III	Study IV
Encoding	51.45 %	51.39 %	50.40 %	54.08 %

14.1.4 Exploratory analysis approach

Since the task offers a lot more than only investigating the cognitive processes during the encoding phase, all other phases of the experiment (presentation, decision, and feedback) are also investigated regarding correlates of remembered and forgotten items. SVM classification was performed on the CCA filtered ERPs of remembered and forgotten items in the remaining phases, which can be seen in Table 14.3. For the presentation and the feedback phase, it can be seen that again, classification accuracies around 50 % have been achieved which do not reach statistical significance. One exception is the decision phase of study II, which accounts for 68 %. However, the feedback phase reveals that in all four studies, a performance above 70 % is possible. For a more detailed view of the achieved accuracies, the classification results of each individual subject can be seen in Figure 14.4 (including the results from the encoding phase). The Figure shows that, with small exceptions, performance above chance level is possible for every subject in every study concerning the ERPs from the feedback phase. All other phases come off equally badly and are located around the chance level for each subject and study. To understand what the caused the exceptionally high classification results from the feedback phase was, the activation pattern of the SVM has been calculated. A heatmap has been chosen as a form of visualization, to get the information about the importance of each feature at one glance. The results can be seen in Figure 14.5. The plots indicate that there

Table 14.3: Classification on ERPs remembered vs. forgotten - Classification on ERPs (CCA filtered) of the remaining phases, presentation, decision and feedback, 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - remembered vs. forgotten. Accuracies marked with * are significantly above chance level (0.05) according to binomial cumulative distribution.)

	Presentation	Decision	Feedback
Study I	51.80 %	51.40 %	71.62 %*
Study II	52.86 %	68.47 %*	69.90 %*
Study III	54.48 %	51.53 %	71.76 %*
Study IV	54.69 %	57.69 %	71.10 %*

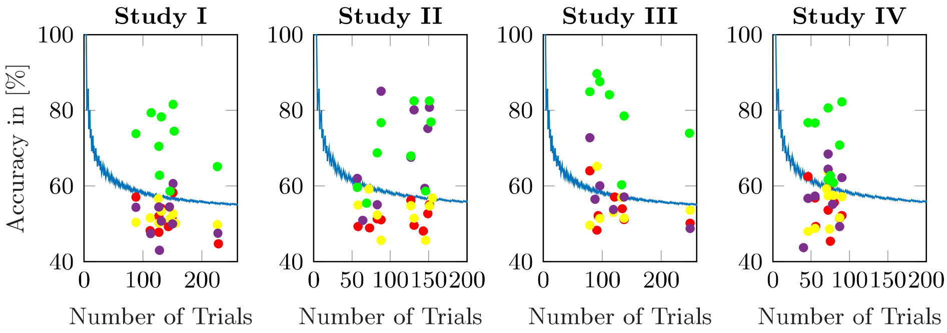


Figure 14.4: Individual classification performance recognized vs. forgotten - The horizontally aligned dots mark the reached classification accuracies for each subject individually, for all three phases of the experiment. Each phase is represented in a different color, and the blue line marks the chance level of each classification based on the binomial cumulative distribution. The y-axis displays the classification accuracy in percent, while the x-axis displays the number of available trials in each classification. — chance threshold, ● encoding, ● presentation, ● decision, ● feedback

are especially two point in time that are of value in the distinction of remembered vs. forgotten items. Around 450-500 ms and 800 ms.

As a representative channel, due to the hypothesis that this could be an error related potential (ErrP), electrodeposition FCz has been chosen to display the differences between the ERPs of remembered and forgotten items. An explanation for the hypothesis will follow in the discussion of this chapter. Figure 14.6 shows, therefore, the difference of means of the two ERPs from the feedback phase of the experiment. The high amplitudes of the difference of means representation indicate that there are major differences between the two classes at this electrode position. It can be seen that overall, the difference of means looks similar for all four studies and is characterized by a positive peak around 500 ms, followed by a negative peak between 600 and 800 ms.

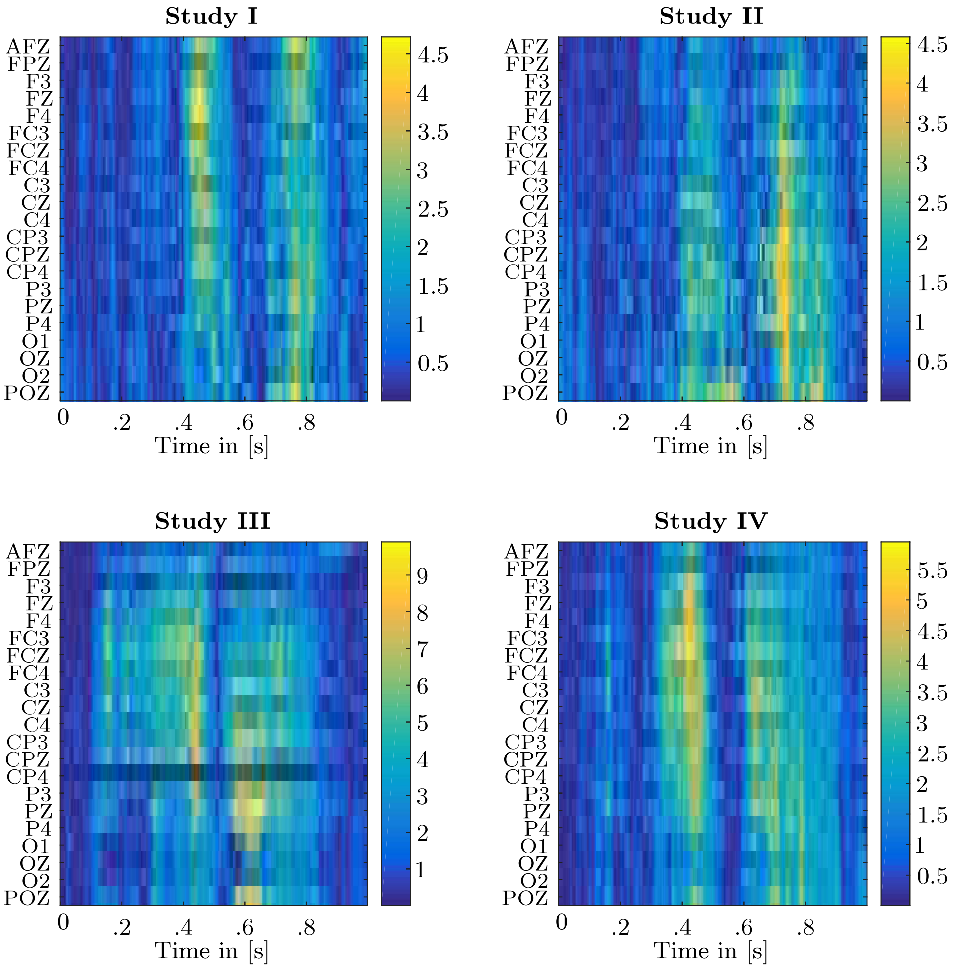


Figure 14.5: Activation pattern for ERPs remembered vs. forgotten in the feedback phase - Activation pattern that has been calculated from the parameters of the SVM. The subfigures account for the experimental studies I-IV in ascending order. The values of the activation pattern are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

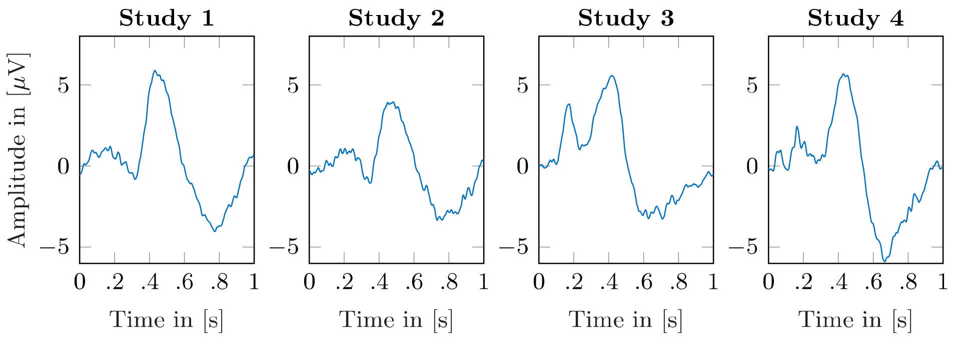


Figure 14.6: Difference of means ERP in the feedback phase remember vs. forgotten - Displayed is difference of means at the electrode position FCz, calculated for each point in time of the 1000 ms time frame (x-axis) and amplitude of the signal in μV (y-axis). The grand average has been calculated over all 9 subjects.

14.2 Discussion

The discussion will focus on a comparison of the results that have been achieved in the four performed studies and the results from the original study performed by Fukuda and Woodman [120]. Additionally, the benefits and new insights that were gained by adding the exploratory ML-based approach to the data analysis.

14.2.1 Behavioral and neurophysiological data

Regarding the question raised at the beginning of this chapter (Is there a difference between good and bad encoding in neurophysiological data during the encoding of a stimulus?) it needs to be said, that neither the neurophysiological nor the behavioral data give rise to an effect concerning the quality of memory encoding. Especially the ERPs show great variance between the studies during the encoding phase. Since there are many reports about effects regarding the encoding strength, it can be argued that the number of subjects within the studies of this thesis is insufficient and not representative enough to obtain an effect. This might be true to some extent, but a variance of this magnitude across studies gives rise to the assumption that the effect is either very small or not stable enough to be easily captured. Since the signal gives the same impression over all evaluated measures, the evidence is multiplying towards a very small effect, that is unfortunately not reflected in the presented data. The differences between the studies concerning the varied parameters, cannot account for these major differences or it seems at least very unlikely. The process of encoding happens at an earlier point in time than the subject could be aware of the ratio of stimuli and be influenced by the time lag between button press and feedback presentation. However, since the labels for the remembered and forgotten stimuli are collected during the test phase and are only assigned retrospective to the trials from the training, it is still possible that the ratio or the time lag does have an influence on the data.

14.2.2 ML-based classification

The ML-based classification also revealed in all four studies that no effect for memory encoding could be found. Classification accuracies are around the chance level for each individual subject, but also on a group average level. Therefore, no differences between the signals of remembered and forgotten items can be found that make them distinguishable.

Remarks to the original study

Since the starting point of this project was a reproduction attempt of the study by Fukuda and Woodman [120], a few words should be said, that address the discrepancies that have been found. The study was chosen because they claimed that a real-time detection of memory encoding was possible in single trial EEG. When taking a closer look, it can be seen that they provided proof of a group effect concerning memory encoding. Interestingly, they explicitly formulated the questions of whether the signal can be used for real-time prediction if the effects provide a sufficient magnitude on a single trial level. Depending on how this term is interpreted, Fukuda and Woodman failed to hold and verify this promise. A general remark needs to be made here regarding the used vocabulary. Especially in BCI research, but also in other areas close to computer science, real-time prediction stands for an online prediction of the signals. This means that immediately after the recording of a trial, this trial is evaluated and its class affiliation is determined. Therefore, this also means that each trial is treated as a standalone signal irrespective of the previous or following trial. For the authors, the vocabulary seems to have a different meaning, because none of the above-stated characteristics fit their performed study. In their study, Fukuda and colleagues arranged all recorded trials in quintiles according to signal strength and magnitude, after the experiment, to show that trials with a strong signal are decent indicators for a good memory encoding. In other fields of research, this would be considered an offline analysis of the results, because the signals are evaluated after the recording has ended. It needs to be assumed that Fukuda and Woodman refer to real-time prediction in a sense, that the computation of the prediction can be done in a short amount of time immediately after the experiment. Independent of this difference in terminology, there is another difference that needs to be mentioned. Fukuda and Woodman used all the data for the evaluation. This can be seen as a group-level analysis of signal strength.In the classification approach of this thesis, the evaluation was done on a single subject level, including validation of the build classification model. By disentangling this methodological and terminological difference, it gets clear why misunderstandings and failures in reproduction can occur. Overall, the effects of the original study cannot be reproduced by classical group-level analysis because the effect seems to be very small and the samples of the four conducted studies in this thesis were too small and not representative enough to find an effect. The effects regarding the prediction of the quality of memory encoding can also not be found, which can be explained by major differences in methodology. A reproduction of the study from Fukuda and Woodman was therefore, not possible.

14.2.3 Exploratory analysis approach

The additionally performed exploratory analysis approach revealed that remembered and forgotten items can be distinguished in the feedback phase with up to 70 % accuracy. Due to this the ERPs in the feedback phase have been investigated more closely. Since the classes remembered and forgotten are in this case identical with the categories correct and wrong, the underlying effect that might be the cause of this can also be attributed to the correctness of the answer. With this assumption in mind, it is not surprising that the difference of means of the ERPs from the feedback phase can be identified as an error-related potential (ErrP) [165]. Regarding the previous statement of an insufficient number of subjects within the studies, it needs to be mentioned that for stable effects such as the ErrP, even small amounts of subjects are sufficient to make the effect visible and to consider it reliable.

The good and significantly above chance level classification results during the feedback phase can be traced back to ErrPs, based on the investigation of the activation patterns from the SVM and the subsequent classical group-level ERP analysis. Despite the chosen class labels - remembered and forgotten -, this effect has no direct relation to the quality of memory encoding. Nevertheless, the two classes correspond perfectly to correct and wrong answer which differ greatly during feedback perception and processing in relation to ones own actions. It has been shown that ErrPs are stable enough to use them reliably during a single-trial classification approach, even in online scenarios [166, 167, 168]. Therefore, it is valid that also in this scenario good classification results can be achieved on a single-subject and single-trial approach.

14.3 Conclusion

Regarding the strength of memory encoding, it needs to be stated, that within the four performed studies no indicators for an effect of memory encoding could be found. Neither in the classical group-level analysis, nor in the ML-based classification approach. Since there are effects concerning the quality of memory encoding that have repeatedly been reported in the literature, it needs to be assumed that on the one hand, the sample sizes have been too small to find an effect within the here collected data. On the other hand, it can also be assumed that the magnitude of the effect is very small on a single trial basis since the classification approach equally failed in all studies and subjects. This finding should be verified with bigger sample sizes though to make a more founded statement. Overall, the analysis showed that there is a discrepancy of vocabulary between different scientific disciplines which needs to be overcome to ensure better reproducibility and cooperation. Regarding the benefit that could be achieved by additionally performing ML-based data analysis, it can be stated that legit exploratory data analysis can be performed. The combination of classification accuracy and activation pattern revealed an effect in the feedback phase of the task, which could later be identified as ErrPs with classical analysis approaches.

Chapter 15

Stimulus familiarity

In the following, the results of all four studies will be presented with a focus on the neurophysiological characteristics of stimulus familiarity. The aim is to find out if there are differences between the processing of old and new stimuli during the time in which the stimulus is presented to the subject. This point in time refers to the presentation phase of the experiment during the recognition test. The label old and new for the individual stimuli is, therefore, an objective label and is defined by experimental design. Since the effect is commonly known as the old/new effect, the two categories will be named and referred to accordingly from here on for the sake of simplicity. In addition to the classical and ML analysis of the categories old and new, again the possibility to do exploratory data analysis with the help of ML will be explored to evaluate possible benefits that can be gained by this methodology. The results of this study have, in part, been published in [160] but will be presented here in full detail.

Is there a difference between the processing of old and new stimuli that can be found in the neurophysiological data?

15.1 Results

As in the previous sections, the analysis will start with classical group-level statistic approaches and continue with the ML-based analysis approach. The conditions old and new will be based on the stimuli that have been presented in the training (old) and stimuli that have only been presented in the recognition test (new). In contrast to the other chapters dealing with memory encoding and decision confidence, there will be no evaluation of the encoding phase concerning the process of stimulus familiarity. The simple reason for this is that the encoding phase took place in the training of the experiment. New items have only been shown in the recognition test in a later part of the experiment. Therefore, no comparison between new and old items can be made.

15.1.1 Behavioral data

The results regarding differences between old and new stimuli in terms of behavioral data can be seen in Table 15.1. It shows the task accuracy, and the RT averaged over all subjects for each study individually. Regarding task accuracy, it can be seen that similar proportions of accurate answers have been given for both categories in all four studies. Regarding reaction time, however, it can be seen, that in study IV subjects were exceptionally fast with a tendency of answering faster for old than for new pictures, as well as for correct compared to wrong answers. The same tendencies can also be found in study III. Interestingly, study I and II do not agree with these tendencies. When looking closer at study I and II to estimate the precise differences, it can be seen that study I stands out because correct and wrong answers have been given approximately equally fast for old stimuli and study II differs because the response to old and new stimuli was given equally fast. The ANOVA confirmed that there are significant differences between studies concerning the overall reaction time (p = .0029). Therefore, it is not surprising that no overall effect for stimulus familiarity was found. However, a significant interaction between the level of confidence and stimulus familiarity was observed (p = 3.712e-06) concerning task accuracy.

Table 15.1: Stimulus Familiarity: Behavioral data - Numbers of correct and wrongly answered trials split according to known (old) and unknown (new) pictures in percentage with respect to the full amount of pictures that have been presented in the recognition test. The RT is presented in seconds. Both measures RT and Acc are averaged values over all subjects.

# Pictures [%]	# Pictures [%]		ReactionTime [s]	
Study I	Old	New	Old	New
Correct	74.40	58.00	1.213	0.968
Wrong	25.60	42.00	1.119	1.150
Study II				
Correct	72.80	64.40	0.953	0.959
Wrong	27.20	35.60	1.148	1.185
Study III				
Correct	79.08	79.56	0.702	1.028
Wrong	20.92	20.44	1.028	1.293
Study IV				
Correct	68.04	80.80	0.591	0.686
Wrong	31.96	19.20	0.765	0.724

15.1.2 Neurophysiological analysis

The neurophysiological analysis revealed that there seem to be discriminative areas in the ERPs of the two categories, old and new stimuli. More precise, the signal at the occipital electrodes seems to differ significantly during the presentation in the test phase. Indicators for that can be found in the Figures 15.1 and 15.2. Firstly, the R^2 values have been evaluated, which are depicted in the heatmaps in Figure 15.1. They are highest around 100 ms at the occipital electrode positions (O1, Oz, O2) in all four studies, which is indicated by the yellow color in the plots. When comparing the four studies to each other, it can be seen that the values are smaller in the studies I and II compared to studies III and IV, but the pattern they create is the same for all studies. To get a closer look at what kind of information is provided at the occipital electrode positions, Figure 15.2 shows the ERPs at electrode position Oz. The Figure shows that around 100 ms a shift in time occurs between the two ERPs of old and new stimuli. The ERP elicited by old stimuli occurs a little earlier than the ERP elicited by new stimuli. In all four cases, this difference in latency is statistically significant. Overall, it can be seen that the ERPs of

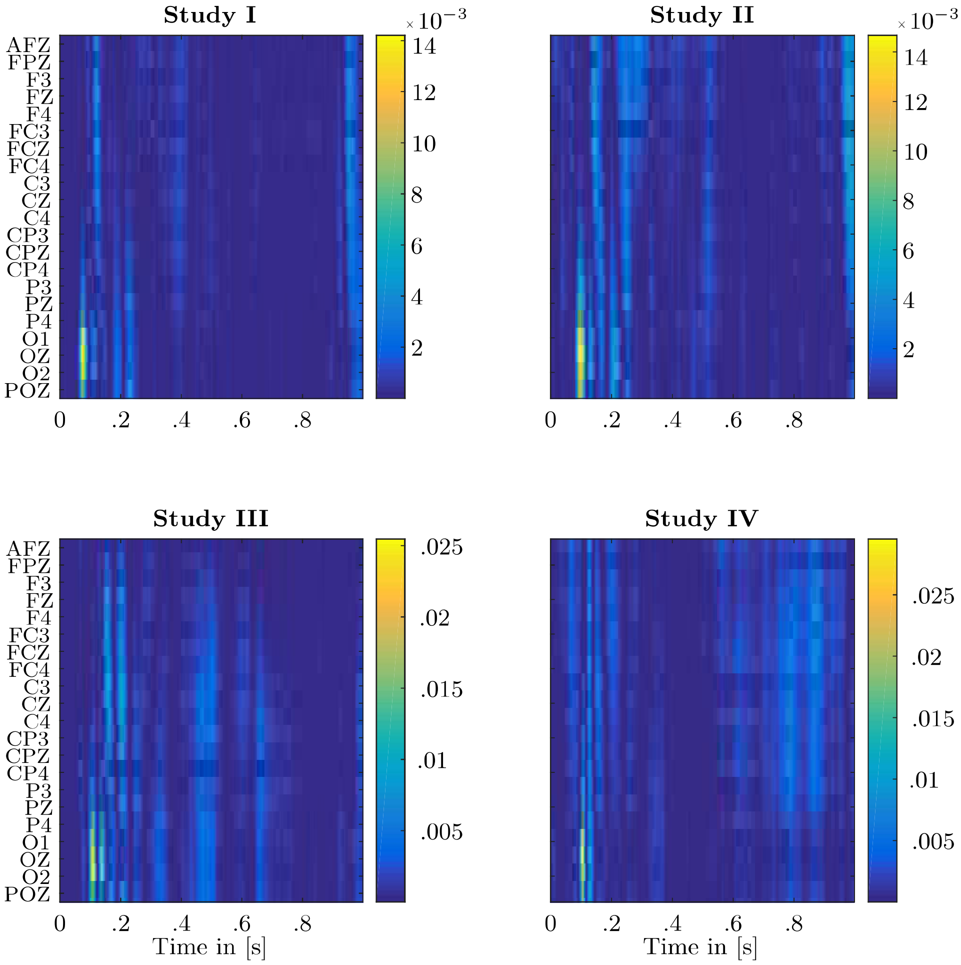


Figure 15.1: R^2 Values for ERPs old vs. new in the presentation phase - R^2 values describe the variance in the data that is explained by the class label, in this case trials that were old or new. The subfigures account for the studies I-IV in ascending order. The R^2 values are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

study I and II differ in some aspects from the other two studies, but in general, it can be stated that the same components can be found in each experimental study.

15.1.3 ML based classification

In the following, the results of the SVM classification will be presented. Table 15.2 shows the performance of the SVM in terms of accuracy for all four studies. A further division of the results was made concerning the correctness of the subjects' answers, to reveal if any differences in performance occur when all trials are used compared to using only the correctly answered trials. At first glance, it can be seen that in study II no statistically significant results can be found concerning the prediction of stimulus familiarity. For this reason, study II will be omitted for now but will be dealt with later during the discussion. When excluding study II, two interesting findings can be made. The first shows that significant results of the SVM prediction can be found for the presentation phase of all

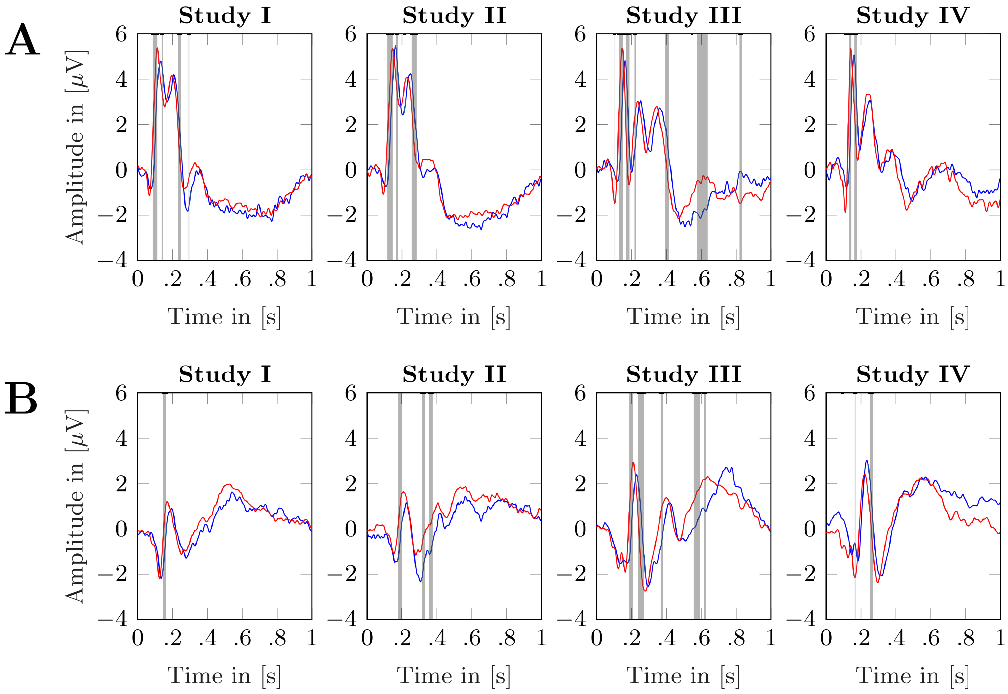


Figure 15.2: Grand average ERPs in presentation and decision phase Old vs New - Displayed is the electrode positions Oz for the presentation phase, calculated for each point in time of the 1000 ms time frame (x-axis) and Amplitude of the signal in μV (y-axis). The grand average has been calculated over all 9 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). ▬: new, ▬: old

studies. Performance values between 63 and 75 % can be reached. Due to the similarity of the average performance values of subjects and SVM classification, it was decided to compare the two measures on a single-subject level. Figure 15.3 shows the performance values one by one in a barplot, for each study individually. It can be seen that in 12 cases, the SVM classification outperforms the human ability to decide correctly. For 7 out of these 12 cases, the difference in performance is statistically significant. Especially interesting are those subjects which performed below chance level. For all of them, it was found that classification performance outperforms the behavioral performance significantly.

To quickly evaluate if the subjective label reveals any interesting correlations, the classification analysis was also performed on the same set of trials but grouped and labeled according to the subjects' answers. The results of that can also be seen in Table 15.2. The numbers reveal that the accuracy values do not significantly exceed chance level when all trials are taken into account. Since the correct answers only are identical to the objective correct answers, no additional classification was performed for this subset of trials.

Table 15.2: Classification on ERPs (CCA filtered) of the presentation phase - 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - known vs. unknown. Accuracies marked with * are significantly above chance level (0.05) according to binomial cumulative distribution.)

Presentation		Study I	Study II	Study III	Study IV
Objective	All	65.33 %*	52.55 %	74.78 %*	72.52 %*
	corr.	63.57 %*	52.32 %	74.93 %*	72.12 %*
Subjective	All	53.41 %	53.39 %	55.54 %	53.19 %

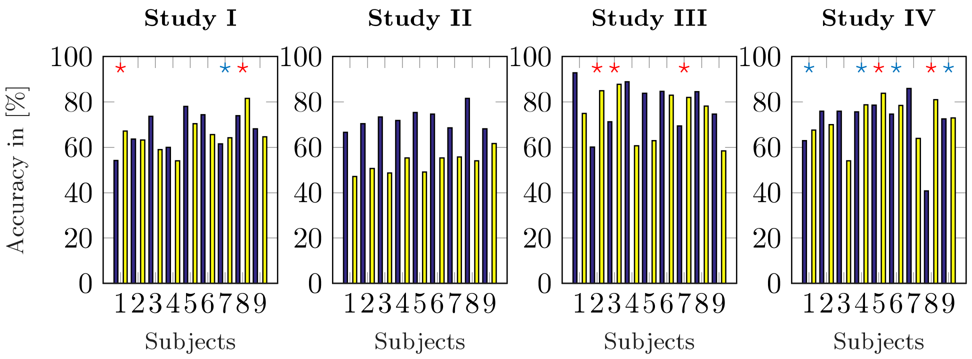


Figure 15.3: Accuracy subject vs machine - Comparison of classification vs behavioral performance Old/New. The subfigures represent the studies I-IV, respectively. The yellow bars display the accuracy of the classifier, the blue bars the behavioral accuracy. Cases in which the classification performance exceeds the behavioral performance are marked with a pentagram. The y-axis represents the accuracy in percent, and the x-axis displays the subjects in ascending order. The blue pentagrams mark the cases in which the machine is better than the subject, and the red pentagrams mark the cases in which the machine is significantly better than the subject.
🟦 Subject, 🟨 SVM

15.1.4 Exploratory analysis approach

Since the task offers a lot more than only investigating the cognitive processes during the presentation phase, all other phases of the experiment (decision and feedback) are also investigated regarding correlates of old and new items. In addition, the reaction time was used as a single feature for the classification, to see if this singular value can also be predictive of the stimulus familiarity. Table 15.3 shows the respective results. It can be seen that similar performance values can be found for the feedback phase, as in the presentation phase. This is mainly valid for study I and III, but not for study IV. Regarding the overall performance, there was no major difference between all and correct answers only.

For the reaction time, it can be seen that no significant results can be reached, making the RT a non-predictive feature for stimulus familiarity. Further, Figure 15.4 shows the classification performance of each individual subject in relation to the significance threshold. Except for the decision phase, it can be seen that the performance of the majority of subjects reaches statistical significance. To get an insight into the cause of the good classification performance in the feedback phase, the activation patterns based on the parameters of the SVM have been calculated. Figure 15.5 shows the patterns for all four studies. It can be seen that there seems to be an area of interest around 350-450 ms across the many, but mostly the central channels. Therefore, it has been decided to take a closer look at the ERPs at channel Cz, which can be seen in Figure 15.6. It can be seen, that study I and II again differ from study III and IV. Study III and IV show two positive peaks within the first 400 ms after stimulus onset, whereas study I and II only show one broad positive deflection in this time frame. Interestingly, there is no major difference between the studies II and IV compared to the studies I and III even though the classification accuracies, indicated that there is a difference regarding this issue.

Nevertheless, statistically significant differences can be found in all four studies, regarding the categories old and new.

Table 15.3: Classification on ERPs (CCA filtered) of the remaining phases - The remaining phases (decision and feedback) and the reaction time of the subjects, 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - known vs. unknown. Accuracies marked with * are significantly above chance level (0.05) according to binomial cumulative distribution.)

		Decision	Feedback	Reactiontime
Study I	All	47.31 %	66.60 %*	51.51 %
	corr.	52.38 %	64.98 %*	51.33 %
Study II	All	54.80 %	52.44 %	50.41 %
	corr.	56.12 %	54.97 %	51.11 %
Study III	All	53.31 %	71.16 %*	53.73 %
	corr.	58.07 %*	71.45 %*	55.94 %
Study IV	All	50.75 %	53.36 %	50.82 %
	corr.	57.09 %*	57.51 %*	53.72 %

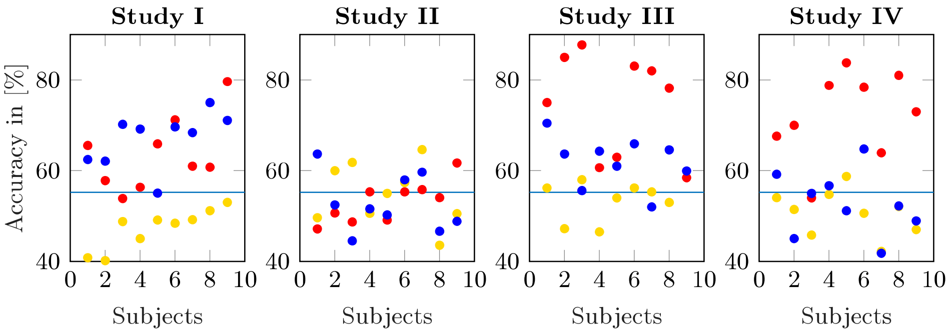


Figure 15.4: Individual classification performance old vs. new - Individual classification performance old/new: The horizontally aligned dots mark the reached classification accuracies for each subject individually, for all the three phases, presentation, decision and feedback of the experiment. Each phase is represented in a different color, and the blue horizontal line marks the chance level of each classification based on the binomial cumulative distribution, indicating when the chance level is exceeded significantly (p<0.05). The y-axis displays the classification accuracy in percent, while the x-axis displays the subjects in ascending order. — chance threshold, ● presentation, ● decision, ● feedback,

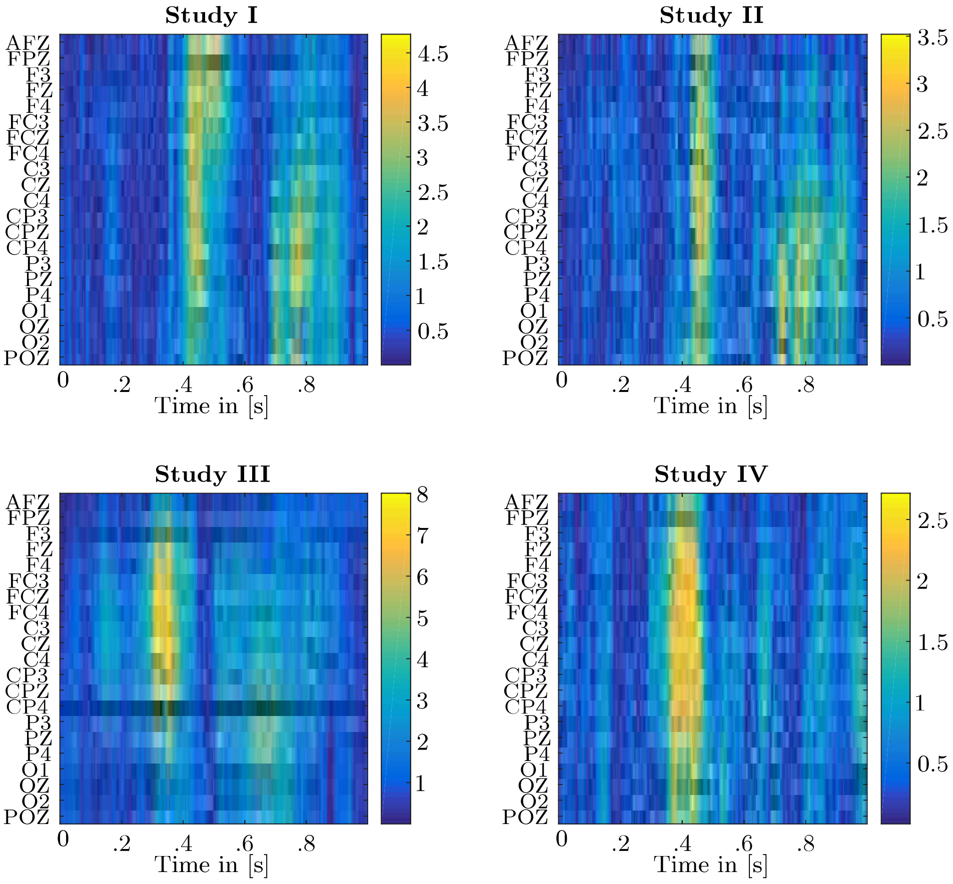


Figure 15.5: Activation pattern for ERPs old vs. new in the feedback phase - Activation pattern that has been calculated from the parameters of the SVM. The subfigures account for the experimental studies I-IV in ascending order. The values of the activation pattern are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

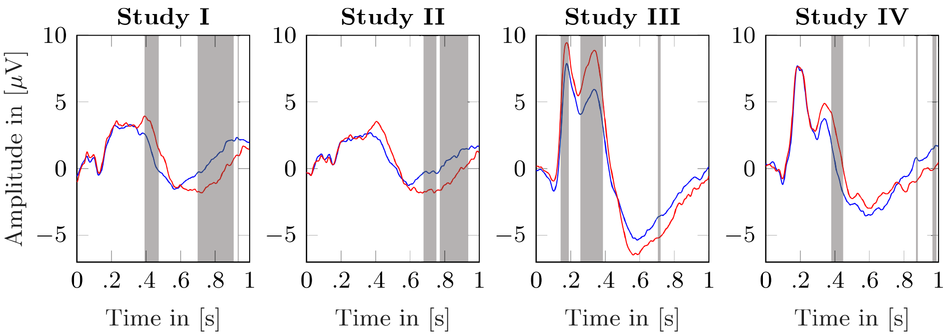


Figure 15.6: Grand average ERPs in feedback phase old vs new - Displayed is the electrode positions Cz for the decision phase, calculated for each point in time of the 1000 ms time frame (x-axis) and Amplitude of the signal in μV (y-axis). The grand average has been calculated over all 9 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). ━━: new, ━━: old

15.2 Discussion

15.2.1 Behavioral data

In the behavioral data, mixed results can be found with respect to stimulus familiarity. Overall, subjects were able to solve this task with more than 70 % accuracy on average, for 500 as well as 750 tested items. Except for study I, the RT shows a tendency for quicker reactions towards old compared to new stimuli. However, no significant effect regarding the reaction time was found. Even the classification approach did not reveal classification above the chance level on the RT. It can be assumed that this is due to the manifold of variables that needed to be considered in the analysis (decision confidence, correctness, familiarity) and also to the studies I and II which seem to have produced results that neither support study III and IV nor the literature in some of the investigated aspects. Study I and II differ from III and IV regarding the time lag between feedback presentation and button press in each trial of the experiment. In study I there was no time lag at all, whereas in study II, the time lag accounted for two seconds, compared to study III and IV in which the time lag accounted for one second. It can be assumed that two seconds might have been too long to remain focused on the task in between trials, whereas no time lag at all might have been too short to switch the attention immediately to the next trial after feedback presentation. Since similar results regarding the behavioral data of study I and II have been found in Chapter 16 for the process of decision confidence, it seems likely that the time lag is responsible for the irregularities in the behavioral data. The stimulus ratio which has also been altered between the four studies can be ruled out as a reason for those changes as both, balanced and unbalanced ratios, occur in the studies III and IV, which are otherwise stable in their results.

15.2.2 Neurophysiological data

Regarding the neurophysiological data, it was found that differences in the EEG signal between old and new stimuli already occur at 100 ms after stimulus presentation. Old stimuli seem to be processed a little faster, leading to a shift in latency of the ERPs between old and new stimuli that lasts until 450 ms after stimulus onset. The mentioned characteristics in the ERPs for old and new items do not necessarily suit the literature concerning the existing old/new effect from recognition memory. Since the strongest and most discriminative shift can be located at the occipital electrodes, the processing can very likely be located in the visual cortex. Further, since the effect occurs as fast as 100 ms after stimulus onset, this can be seen as an indicator for early visual processing, likely in V1, which is a particular area in the visual cortex. None of the existing studies reported a temporal shift for old vs. new items in the context of familiarity or the old/new effect. However, the findings are very similar to the effects that have been measured in V1 via invasive electrodes in rodents and primates [142, 143]. Compared to other studies, the presentation and therefore, the study time is relatively short (250 ms) for each stimulus. Eventually, this short presentation triggers slightly different mechanisms for storage and information retrieval, which could be subconscious only. Many studies in recognition memory rely on words as stimuli, for which a presentation of 250 ms would very likely not be sufficient to process the stimulus fully. These two points can be seen as the major reasons why the results are not comparable to the effects described in the literature. Regarding the N400 which is reported to be an indicator of familiarity [138, 139], it can be said that there are no significant differences in amplitude between old and new stimuli in our data. But it needs to be noted that, effects caused by familiarity seem in general to be hard to compare because the epochs used for investigation are not always explicitly stated, therefore, the studies might not use the same baseline. Also, in many cases, subjects are prompted to answer if the stimulus is old or new with a signaling word or queue that replaces the stimulus. In the here presented studies I-IV, the stimulus remains at the center of the screen while the addition of two colored dots flanking the image left and right prompt the subjects to answer. This form of prompting could also be a meaningful difference, depending on the exact epoch that was under investigation.

15.2.3 ML-based classification

The classification accuracies showed that old and new stimuli could be separated during the presentation phase. They reached more than 70 % throughout all studies and in individual cases even exceeded the performance of the behavioral accuracy. This is an interesting finding that supports the fact that the found differences in the signals belong to a stable effect. Since subjects and the machine learning based classification performed equally well on average, a direct comparison of the performance values was made on a single-subject level. It surprisingly revealed that human memory could be outperformed by the SVM classification in individual cases. This fact allows the development of two hypotheses. On the one hand, it could mean that the initial visual processing of a stimulus is independent of memory retrieval. This would explain both cases, the ones in which humans perform better than the machines and vice versa. More precisely, it would mean that both measures, behavioral and SVM performance would thereby be independent. This hypothesis would be backed up by the finding of Fahy and colleagues [143] who stated that in the stage of early visual processing, different neurons fire for familiar and non-familiar items within 100 ms after stimulus presentation. Further, it was found that V1 has memory on its own on this level of neurons that lasts up to 24 h. However, another hypothesis would also be likely, stating that there are one or more processes in the chain from visual perception to memory retrieval that can simply fail. In that case, the two measures would be related to each other. Due to the bad spatial resolution of EEG, the effect cannot be located exactly making it hard to follow up on either of the two hypotheses. Other recording techniques are needed to shed more light onto this question. Nevertheless, it seems likely that the human visual cortex has its own memory system on a neuron level. About the connection between visual processing and memory retrieval, however, no further assumptions can be made.

15.2.4 Exploratory analysis approach

The additionally performed exploratory analysis approach revealed that a separation of old and new stimuli is also possible in the feedback phase in study I and III. Study I and III have in common that the ratio of old and new stimuli that are presented in the recognition test are unbalanced compared to study II and IV in which the ratio is balanced. It can be hypothesized that this ratio has an influence on feedback perception and processing, which makes the two classes separable in the feedback phase. Even though the ratio is not communicated to the subjects, they could subconsciously be aware of it and adapt their behavior to achieve better performance. Overall, it seems very likely that the ratio of old and new stimuli in the recognition test influences the decision making and also feedback processing. A subconscious knowledge about an imbalance can alter the expectation for correct and wrong answers. Identifying this as a factor with an impact on the results can be equated with the identification of a latent variable. Pointing out that the ratio influences the data could only be achieved with the ML approach because the standard ERP analysis did not give rise to any significant differences. This is an interesting finding since many studies in memory psychology perform studies based on a remember/know procedure. Most of the studies use unbalanced sets of pictures with ratios of 2:1 or 3:1 (known to unknown pictures). Depending on the process that is under investigation the influence of the ratio might not be of relevance. But nevertheless, it is important to keep it in mind to not report confounded results.

Due to the high classification results during the feedback phase, and the respective activation patterns, the ERPs of this phase have also been investigated at electrode position Cz. Interestingly, there was no clear indicator in the group-level analysis that could explain the differences between study I and III compared to II and IV. Therefore, it needs to be assumed that the difference can be found on a multi-electrode or on a single-subject level.

15.3 Conclusion

Stimulus familiarity is a property that can be detected and distinguished with classical group-level analysis and with ML based classification. Differences between old and new stimuli occur as early as 100 ms after stimulus presentation, which are visible especially in the time-locked ERPs in occipital regions. Old stimuli are processed faster, which is reflected in a shift of latency of the ERPs. Due to the timing and location of the effect it can be assumed that it can be associated with early visual processing, possibly in V1. Since ML based classification and human behavior performed equally well on average, a direct comparison was made on single-subject level. It could be shown that in some cases the machine outperforms human memory performance. This lead to the hypothesis that the initial visual processing of a stimulus is independent of memory retrieval. Another hypothesis could also be that there are one or more processes in the chain from visual perception to memory retrieval that can fail, which would explain the here presented findings. Lastly, it could be shown that machine learning can be a useful tool for the identification of latent variables in a tightly knit network of experiments.

Chapter 16

Decision confidence

In this chapter, the results of all four studies will be presented concerning the aspect of decision confidence. It will be evaluated if the confidence with which the subject decides if the stimulus is old or new, is reflected in physiological signals throughout the task. Two levels of confidence have been assessed in the task, between which the subjects were able to choose for each decision. A distinction between 100 % and 75 % confidence, which can also be seen as high and low confidence, will be made in the following data analysis. The label is, therefore, based on a subjective decision of the subject. Since the level of confidence is likely to have the greatest effects in the decision phase and the feedback phase, one of the two phases should be chosen for the investigation. For brevity, it has been decided only to evaluate the feedback phase with the classical group-level approach. Nevertheless, an evaluation of all phases was done by an ML-based classification approach to explore the full data set.The results of this study have in part been published in [159], but will be presented here in full detail.

Is there a difference between high and low levels of confidence in the neurophysiological data?

16.1 Results

As in the previous sections, the analysis will start with classical group-level statistic approaches and continue with the ML-based analysis approach. The conditions high and low confidence are labeled according to the subjects' decision and will be investigated mainly in the feedback phase of the experiment.

16.1.1 Behavioral data

Table 16.1 shows the behavioral data that was collected throughout the experiment sorted according to the two levels of decision confidence. Presented are the averaged values over all subjects for each study individually.
It can be seen, that except for study 2, more correct answers have been given with 100 % than with 75 % confidence. In contrast to that, more wrong answers have been given with 75 % than with 100 % consistently in all four studies. Overall, more answers were given correctly than wrong. Concerning reaction time, it can be seen that except for study I, answers have been given faster when the level of confidence was high (100 %), compared

to when the level was low (75 %). An ANOVA revealed that this effect on reaction time is statistically significant ($p < 5.06e^{-5}$).

Table 16.1: Confidence Acc and RT - Represented are the numbers of answers given with 100 % or 75 % confidence in percent with respect to the number of all trials from the recognition test. The reaction time for each trial is presented in seconds. Both measures are averaged over all trials and subjects of the respective study.

	# Pictures [%]		ReactionTime [s]	
Study I	100 %	75 %	100 %	75 %
Correct	44.27	19.33	1.161	1.107
Wrong	16.00	20.4	1.123	1.137
Study II				
Correct	33.80	32.80	0.840	1.126
Wrong	15.20	19.20	1.006	1.264
Study III				
Correct	44.00	35.24	0.586	1.093
Wrong	4.23	16.53	0.775	1.202
Study IV				
Correct	42.40	32.02	0.536	0.783
Wrong	8.40	17.18	0.611	0.816

16.1.2 Neurophysiological analysis

The results for the neurophysiological analysis of the feedback phase regarding differences between high and low levels of confidence can be found in the Figures 16.1 and 16.2. Figure 16.1 displays which electrode positions and points in time provide information that helps to dissociate the two levels of decision confidence using R^2 values. The R^2 values do not show a consistent pattern over the four studies. It can be assumed that points of interest are located around 500 and 800 ms broadly distributed over the head. Nevertheless, it has been chosen to have a closer look at electrode position Cz, which can be seen in Figure 16.2. Cz is a central position that is discriminative for many cognitive processes, and at least in Study 3 and 4 the R^2 values suggest certain importance of the channel. Subfigure A shows the ERPs after feedback stating the answer was correct, whereas subfigure B shows the ERPs after feedback indicating the given response was wrong. Interestingly, well-pronounced differences between the two levels of confidence can be found during the perception of affirmative feedback, but almost no differences can be seen during opposing feedback. This holds for all four studies.

16.1.3 ML based classification

The performance values of the SVM classification approach for the distinction of low and high confidence in the feedback phase can be seen in Table 16.2. Again, it has been differentiated between all, and correct answers only. It can be seen that study II and IV have lower accuracy values than the other two studies. They all reach statistical significance, but for study II and IV only values between 55 and 63 % can be reached,

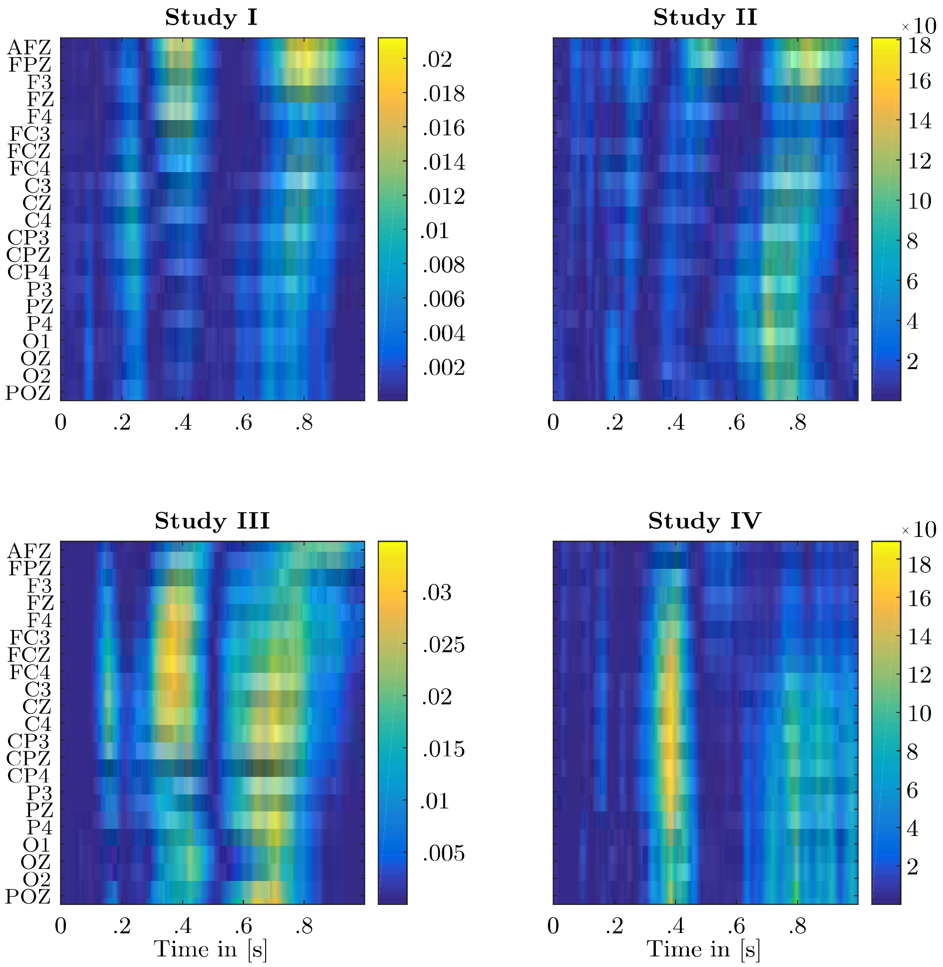


Figure 16.1: R^2 **Values for ERPs 100 vs 75 % in the feedback phase -** R^2 values describe the variance in the data that is explained by the class label, in this case trials that were answered with 100 % or 75 % confidence. The subfigures show the results of the studies I-IV in ascending order. The R^2 values are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

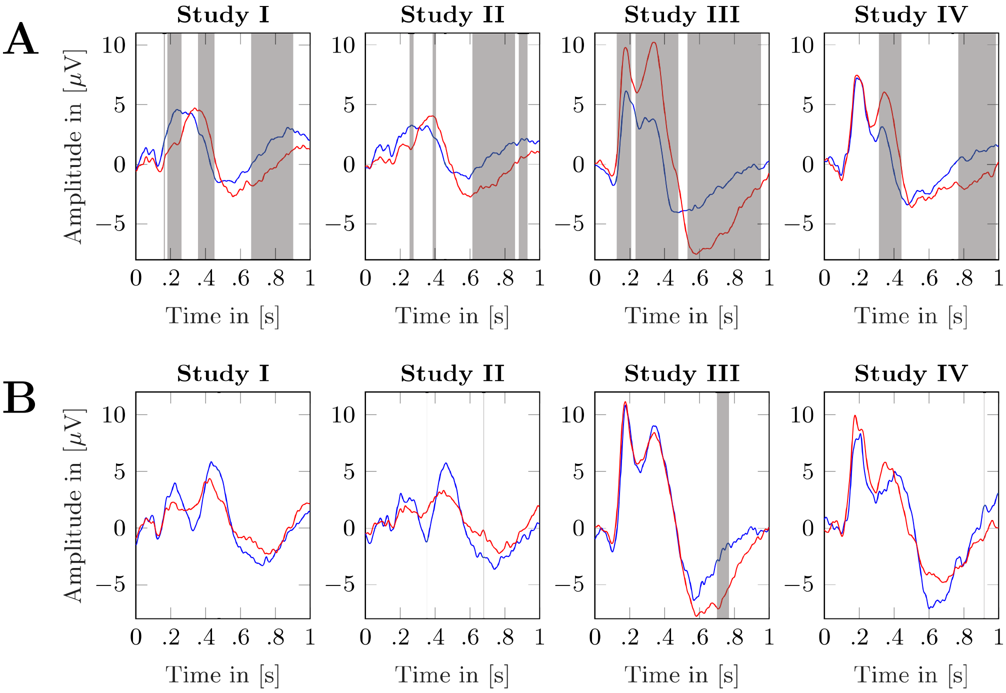


Figure 16.2: Grand average ERPs in Feedback phase 100 vs 75 % confidence - Displayed is the electrode position Cz, calculated for each point in time of the 1000 ms time frame (x-axis) and Amplitude of the signal in μV (y-axis). The grand average has been calculated over all 9 subjects. Grey areas indicate statistically significant differences between the two conditions (p<0.05 Bonferroni corrected, according to number of time points). A: Correct answers B: Wrong answers —: 100 %, —: 75 %

whereas for study I and III values around 70 % are achieved. No major differences between all and correct answers only have been found regarding the performance values.

Table 16.2: Classification ERPs 75/100% - Classification on ERPs (CCA filtered) of the feedback phase of the subjects, 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - 100 % vs 75 % decision confidence. Accuracies marked with * are significantly above chance level ($p < 0.05$) according to binomial cumulative distribution.)

		Study I	Study II	Study III	Study IV
Feedback	All	69.63 % *	55.74 % *	69.83 % *	63.20 % *
	corr.	67.64 % *	59.55 % *	70.13 % *	60.77 % *

16.1.4 Exploratory analysis approach

Since the task offers a lot more than only investigating the cognitive processes during the feedback phase, all other phases of the experiment (encoding, presentation, and decision) are also investigated regarding correlates of high and low confidence answers. In addition, the reaction time was used as a single feature for the classification to see if this singular value can also be predictive of decision confidence. The results can be seen in Table 16.3. Except for the encoding phase, a distinction of the two levels of confidence have been possible with accuracies above chance level in all phases and studies of the experiment. The reaction was partly predictive for the level of decision confidence. Especially study II and III must be highlighted here since values close to 60 % or even up to 70 % have been reached. Regarding the disassociation between correct and wrong answers, there is no nameable difference between the subsets in all cases. For more detailed evaluation of the classification results, the classification performance of each subject individually can be seen in Figure 16.3. For the sake of completeness, Figure 16.5 and 16.4 show the activation patterns for the decision and the presentation phase, respectively. However, no commonalities or clear pattern can be found across the four studies for each of the phases. Therefore, no further investigation was performed.

Table 16.3: Exploratory classification ERPs 75/100% - Classification on ERPs (CCA filtered) of the remaining phases (encoding, presentation and decision) and the reaction time of the subjects, 21 channels, 1s, SVM linear kernel, 10-fold cross-validation - 100 % vs 75 % decision confidence. Accuracies marked with * are significantly above chance level ($p < 0.05$) according to binomial cumulative distribution.)

		Encoding	Presentation	Decision	Reactiontime
Study I	All	51.10 %	57.97 % *	56.84 % *	53.16 %
	corr.	46.08 %	58.48 % *	54.60 % *	54.94 %
Study II	All	52.17 %	56.24 % *	62.31 % *	58.38 % *
	corr.	54.92 %	56.16 % *	63.37 % *	59.52 % *
Study III	All	50.30 %	59.04 % *	58.91 % *	66.37 % *
	corr.	50.33 %	60.63 % *	58.94 % *	67.39 % *
Study IV	All	48.13 %	56.34 % *	54.71 %	57.96 % *
	corr.	- %	59.76 % *	56.65 % *	55.33 %

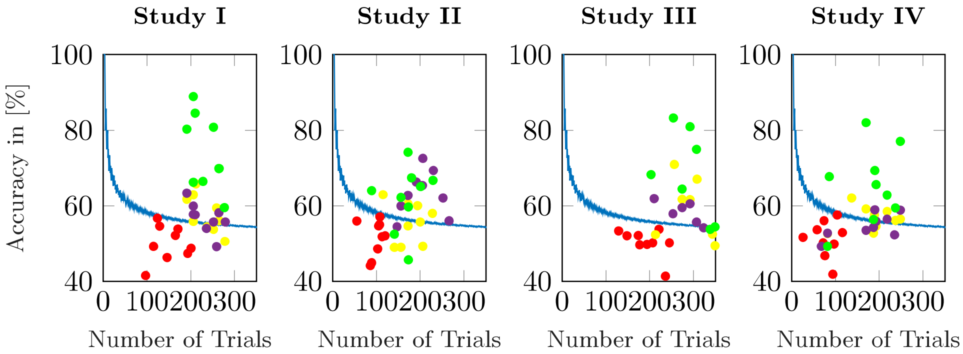


Figure 16.3: Individual classification performance 75/100% - The horizontally aligned dots mark the reached classification accuracies for each subject individually, for all four phases of the experiment. Each phase is represented in a different color, and the blue line marks the chance level of each classification based on the binomial cumulative distribution. The y-axis displays the classification accuracy in percent, while the x-axis displays the number of available trials in each classification. — chance threshold, 🔴 encoding, 🟡 presentation, 🟣 decision, 🟢 feedback

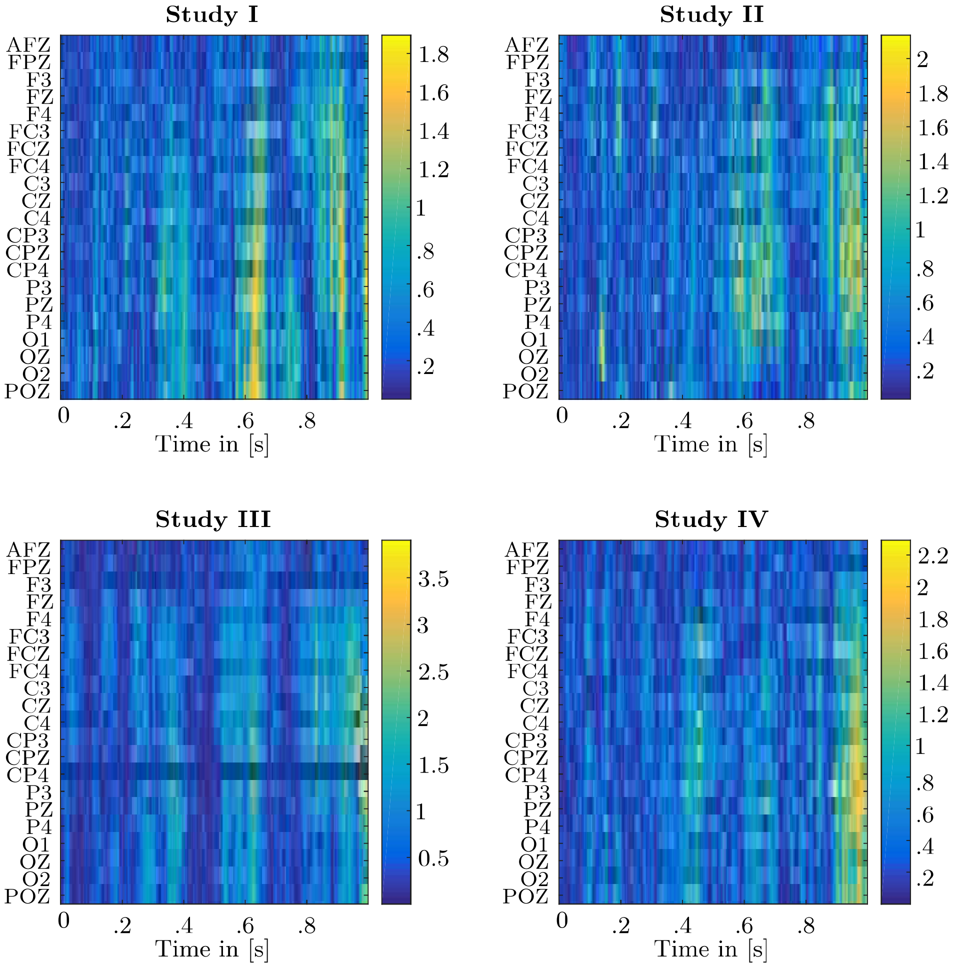


Figure 16.4: Activation pattern for ERPs 100 vs 75 % confidence in the presentation phase - Activation pattern that has been calculated from the parameters of the SVM. The subfigures account for the experimental studies I-IV in ascending order. The values of the activation pattern are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

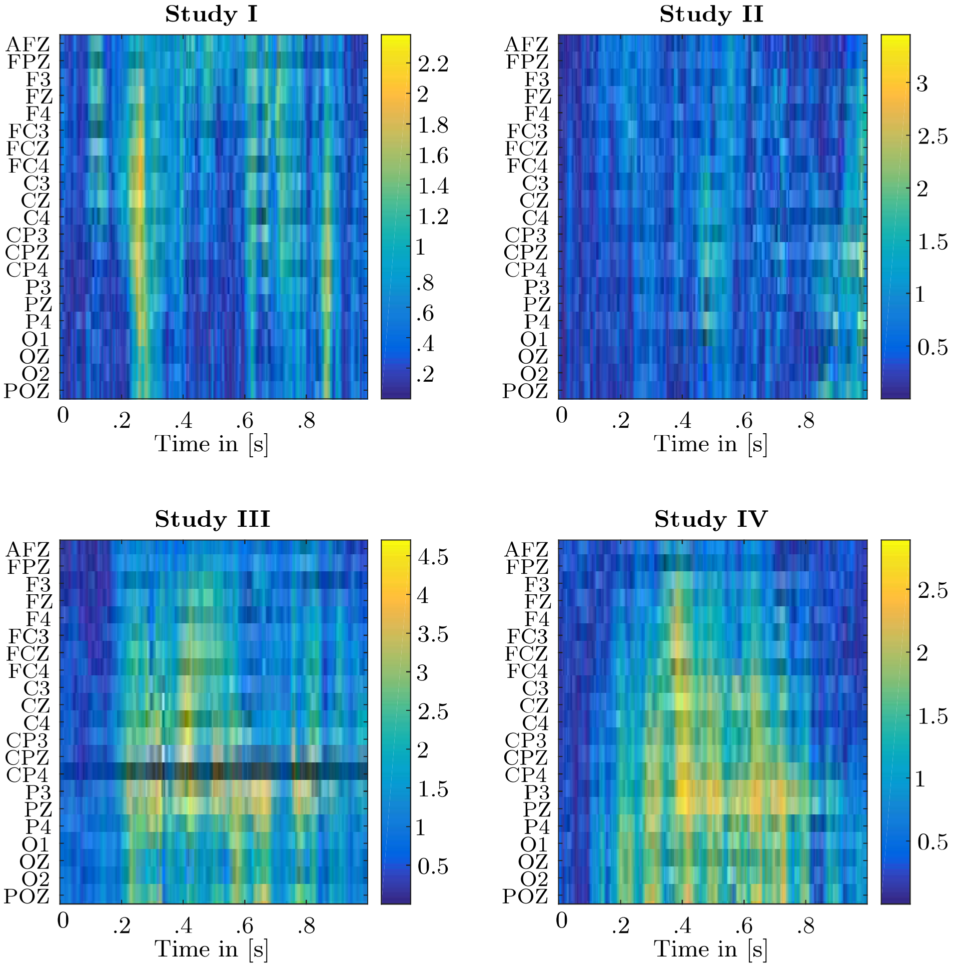


Figure 16.5: Activation pattern for ERPs 100 vs 75 % confidence in the decision phase - Activation pattern that has been calculated from the parameters of the SVM. The subfigures account for the experimental studies I-IV in ascending order. The values of the activation pattern are color coded in a heaptmap, calculated for each point in time of the 1000 ms time frame (x-axis) and for each electrode position (y-axis).

16.2 Discussion

16.2.1 Behavioral data

In the behavioral data, mixed results can be found concerning the dissociation of two levels of decision confidence. The reaction times of study II-IV reflect what can also be found in the literature. The subjects reacted much faster when they were highly confident about their answer, compared to when they were less confident, as well as slower when the answer was wrong than in cases in which the answer was correct [147]. Study I of the experiment does not reflect that. The main reason for that might be the delay that was introduced after study I between feedback and decision of the subject. Answering the trials without mandatory breaks, only between blocks, could lead to a loss of focus. Therefore, the subjects needed more time to refocus on a new stimulus and hence also take more time to answer the trial in study I. Interestingly, the classification results based on the reaction time, mirror this finding. In [169] the hypothesis was formulated that also the ratio of known and unknown stimuli could be the cause of this since in study II, the number of known and unknown pictures were equal and in study I it was a ratio of 1 to 2. Study III and IV, however, mirror the distribution of known and unknown stimuli from the previous studies and in both the expected differences in reaction time could be found.

16.2.2 Neurophysiological data

In the neurophysiological data, it has been found that clear indicators for the different levels of confidence can only be found in reactions to correct feedback for all studies of the experiment. This result is in line with literature [170], [171], [172] and therefore, not surprising. Lack of significance between confidence levels in the wrong answers could also be due to the available number of trials, which were a lot less than for correct answers, consistently for all subjects. In general, it remains unclear if the neural responses are that well distinguishable because the subjects were forced to assess their level of confidence with every answer, or if the level of confidence would be reflected as well if there was no need to quantify it after each trial. This fact is hard to revise because the subjective level of confidence needs to be collected somehow to be able to label and categorize the data. Still, since the self-assessment of the current progress in learning is an important marker for deciding when a specific content has been learned sufficiently well, it is an interesting finding.

16.2.3 ML based classification

The most pronounced differences between the EEG signals of the two levels of decision confidence were found in the feedback phase. The differences lead to classification accuracies of up to 70 %. In study II classification accuracy is inferior but still significantly above chance level. The bad results of study II could be caused by the introduced delay of 2 s between button press and feedback presentation. It is possible and rather likely that the link between own action and the corresponding feedback is weakened by the long pause, and thereby alters the processing and the reaction to the feedback.

16.2.4 Exploratory analysis approach

The exploratory data analysis revealed that classification accuracies above chance level have also been found in the presentation and in the decision phase. For the presentation phase, it can be assumed that the strength of recognition or familiarity influences the neural correlates, which in turns automatically translates into decision confidence, with which the upcoming decision will be made. The same might still be true for the decision phase. Here, however, also the motor preparation and execution for the decision could already be included. It is likely that the accumulation of processes leads to a difference concerning the categories regarding decision confidence, even if they do not directly represent the confidence of the subject. Since the activation patterns did reveal a common pattern across the four studies for either of the two phases, it can be assumed that no identifiable process is taking place during those phases. Further, this can be interpreted as an indicator for the validity of the hypothesis about accumulated processes.

16.3 Conclusion

It could successfully be shown that trials labeled according to subjective decision confidence, can be separated with statistical significance in all investigated phases of a simple recognition task. The differentiability of high and low confidence levels could be shown by classical group-level ERP analysis, as well as with a machine learning classification approach. It is possible to distinguish two levels of decision confidence, with up to 70 % classification accuracy, based on the ERPs of the subjects elicited by categorical feedback to the given answer. The main effect resulting in this difference is based on the reaction to positive feedback and not on negative feedback. While trying to disentangle feedback processing from decision formation, it has been found that after introducing a delay of 2 seconds between entering the decision and receiving the corresponding feedback, the performance drops immensely. This could either be due to not being able to link the made decision to the corresponding feedback anymore or to the disentanglement of the two phases, revealing that the effect is based on an accumulation of the processes of both phases.

Chapter 17

Discussion

What can be learned from the results of all four studies based on characterizing episodic memory processes in EEG with the help of machine learning ?

In this chapter, the results from all four presented studies will be discussed and concluded in an overall manner. The discussion will be divided into five categories, summarizing the results with respect to possible latent mental processes that are present throughout the task in addition to the investigated ones, the disentanglement of phases in the analysis of the task, potential effects that arise through experimental design (ratio of known and unknown stimuli), practical implications that could be of use, and lastly the advantages that are achieved by adding an ML approach to conventional group-level analysis on psycho-physiological data. The primary objective of the four performed studies was to reveal core episodic memory effects without additional restrains of other processes.

17.1 Latent mental processes

Regarding the posed research questions and the used task design, a few remarks need to be made before the concluding discussion. Although all phases allow some form of prediction about all three processes, memory encoding, stimulus familiarity, and decision confidence through classification accuracy and statistical analysis of the ERPs, the found neural correlates might not be directly related to the process in question. Nevertheless, each process can be used as an umbrella for all phases in the experiment, since the subject labels each trial with a category for each process, and every trial passes all phases of the experiment. To evaluate possible explanations besides the respective process, decision confidence, memory encoding, and stimulus familiarity, all phases are quickly scanned for possible effects and processes that could be present.

First, there is the phase where the stimulus was encoded in memory. As stimulus encoding is the only process happening at this time point and no decision is involved, there cannot be direct correlates of confidence or familiarity in this phase. However, attention to the stimulus or the strength of encoding will influence all processes at a later point, and thereby correlates of these processes can be found in the EEG.

As the second phase of interest, the stimulus presentation in the recognition test was investigated. The recognition test started with the instruction to decide about the familiarity of a picture and to state the respective level of confidence. Therefore, in the test presentation phase, mainly information retrieval takes place. The decision phase was chosen as an intermediate step to capture the actual process of decision making by selecting a window that starts with a trigger that indicates the subject that an answer should now be given. Therefore, in this case, it might be legit to speak of decision confidence but also movement preparation and execution.

As a last time window of interest, the feedback phase has been investigated to evaluate if the level of confidence of the decision is also reflected in feedback perception and evaluation. It is possible to speak about affirmation or disappointment, which naturally varies with the level of confidence with which the corresponding decision was made. All of these aspects should be considered, and only careful statements about the actual mental process that is present should be made.

17.2 Disentanglement of phases

Interestingly the experimental design in its details played a major role in the findings that have been made within this part of the thesis. Of particular interest here is the time lag between picture presentation and decision making, as well as the time lag between decision and feedback presentation. First, let's talk about the time lag between picture presentation and decision making. Within all four studies, subjects' needed to wait for one second after the stimulus presentation in the test phase, until they were allowed to answer the task. This was already implemented in the original study of Fukuda and Woodman [120] and therefore, not newly introduced within this thesis. Nevertheless, it is a property that almost no other study implements. In most studies, subjects are instructed to respond as fast as possible as soon as they made a decision after stimulus presentation. In the studies of this thesis, subjects were also instructed to react as fast as possible but only after the one second delay. Reactions before that were not recorded and had to be repeated in the correct time frame. The hypothesis is that this delay causes a disentanglement of visual processing and decision making. Therefore, this could be the explanation of why the effect of faster stimulus processing of familiar stimuli was not reported on humans before.

Concerning the time lag between button press and feedback presentation, it can be stated that the amount of time that has passed also played a crucial role in the results. In study I there was no delay at all, which resulted in overall uniformly distributed answers of the subjects concerning reaction time. The known effects, which are faster reactions for high confidence answers and faster reactions for correct answers, could not be found in the data. Since these effects were present in the remaining three studies, it can be assumed that the lack of a delay between answer and feedback presentation is the cause for it. Possibly there was not enough time to refocus on the processing of the feedback and on the upcoming trial that the behavior diverged from the known patterns. In study II a delay of two seconds was introduced, which changed the outcome of behavioral and neurophysiological data significantly. Concerning reaction time, the anticipated effects could be found, but on a neurophysiological level, the effect vanished completely. This became especially clear when looking at the classification performances which dropped below chance level. Since the effect reappeared in study III and IV, it can be assumed that a delay of two seconds was simply too long and the link between the made decision and the corresponding feedback could not be made anymore. A shortening of the delay to one second seemed to be an appropriate choice and a good compromise, that resulted in stable effects and effects that are to be expected from the literature.

17.3 The ratio of known and unknown stimuli

Choosing different amounts of stimuli to be presented is also a choice of experimental design, which had interesting effects on the data. The model study of Fukuda and Woodman proposed to show 500 stimuli in the training phase and 750 in the test phase from which 250 were new, and all 500 old pictures from the training phase were repeatedly shown. To avoid potential confounds, throughout the development of the studies, it has been decided to test both, a balanced and an unbalanced ratio of known and unknown stimuli in the test phase of the experiment. Therefore, study I and III had an unbalanced ratio as proposed by Fukuda and Woodman, and study II and IV had a balanced ratio of 250 old and 250 new pictures in the test phase. Overall, behavioral data did not give rise to the assumption that any difference between the experimental parts exists concerning the ratio of stimuli. The neurophysiological data on group-level did not show clear indicators that the unbalanced ratio might be responsible for confounds or effects in the data. Classification accuracies of the machine learning approach, however, show that the performance of the classifier differs greatly for the individual studies when separating old from new stimuli. Differences occur in the presentation phase and in the feedback phase, and they are characterized by higher accuracy values for study I and III compared to II and IV.

Differences can also be seen when the two levels of confidence are separated. The two levels of confidence can be separated with much higher accuracies, when the ratio of stimuli is unbalanced, compared to when they are balanced. This could either mean, that subjects have a stronger and more stable pronouncement of confidence within the task because they are aware of the unbalanced ratio and therefore of the increased chance of making the correct choice. But it could also mean that other effects that are based on confounds due to the unbalanced ratio are classified here, which should further be investigated. Depending on the posed research question, this finding is essential and needs to be considered. Most studies based on recognition tests use unbalanced ratios of stimuli for testing.

17.4 Practical implications

For each of the three aspects that have been under investigation, considerations can be made concerning the practical implications and future use of the achieved results. Overall, a firm understanding of the field of episodic memory can lead to new insights regarding memory encoding and retrieval, which in turns would help to find better ways to memorize information. Educational applications based on the assessment of the current cognitive state, have so far shown that it is possible to assess the amount of load a subject is under and to adapt the difficulty of arithmetic tasks to keep the subject within a comfortable range of load [173], [174]. According to cognitive load theory (CLT) [175], the key to successful learning is to avoid cognitive over- or underload and to keep the learner appropriately challenged.

17.4.1 Memory encoding

Regarding the aspect of memory encoding, it, unfortunately, turned out that a classification of the effect cannot be managed on a single trial level. Predicting the success or failure of memory encoding would have been a great achievement and would be a good example to base a neuro tutor on. It would enable to repeat the content that has been identified as bad or not encoded into memory at all, instead of repeating everything. The paper published by Fukuda and Woodman [120] lead to believe that a neuro tutor based on the strength of memory encoding could be built. Unfortunately, this is not true, which is an interesting and important finding. Without the use of machine learning approaches, this would not have been possible.

17.4.2 Stimulus familiarity

Concerning the aspect of stimulus familiarity, it can be said, that the practical implications that have been found are especially of interest on a theoretical level. The analysis generated hypotheses about the connection between early visual processing and memory retrieval. Testing these hypotheses will be needed to get the inside about the specific connection finally, but having a lead that promises valuable new insights is equally important. Since the classification accuracies were above 70 %, a usage in BCI applications, however, would also be possible. The threshold of 70 % is often seen as the limit that needs to be exceeded to ensure a sufficient level of usability. Building applications that serve as lie detectors or as memory aid would be conceivable.

17.4.3 Decision confidence

Being able to extract the confidence with which knowledge can be retrieved from memory is also something that has practical implications. It can easily be imagined how having knowledge about the level of the decision during decision making in an education based application scenario might be interesting. It would allow to extract and identify content that is not entirely secured in memory. This specific content could be recapitulated until the subject reaches higher confidence during answering the question related to the content. This would be a useful extension to error adaptive learning systems [176].

17.5 The advantage of using the ML approach

In summary and conclusion, it can be stated, that the usage of ML on a single-subject level once again added valuable information to classical group-level statistics. There are several aspects, which in part have already been named in the previous sections, but will be recalled here in brevity to complete and round off the chapter.

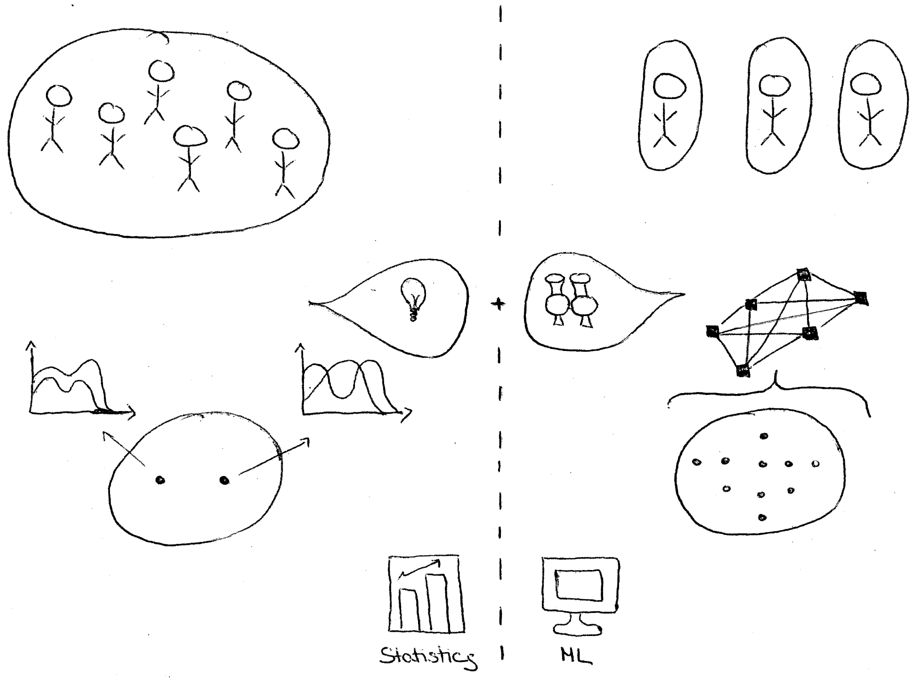
It could be shown that classification accuracy is a very useful indicator of effect size, that is able to quantify the influence of factors and variables. To some extent, classification performance gives an indicator concerning the variance of the data and therefore, the stability of the effect on a single-trial level. This was especially of interest during the investigation of stimulus familiarity. Without this measure of effect size, the influence of the ratio of old and new stimuli might not have been detected. In combination with the analysis concerning decision confidence, it could be seen, that there, the ratio also played a role and is not negligible. Within this analysis, this can be interpreted as a form of latent variable identification. The influence of the ratio cannot directly be measured but is responsible for different outcomes between the experiments. In this example, it became clear that it is possible to show foreign influences that can be observed and identified more closely.

In addition to effect size and latent variable identification, it could also be shown that classification analysis, in combination with classical group-level analysis, can put brain activity in relation to behavioral performance and capabilities. In particular, this became clear by comparing classification performance based on brain activity to behavioral performance. Finding that SVM based classification categorize old and new stimuli better than subjects can, reveals potentially new insights of memory processes. Two hypotheses can be formulated for this issue. Either information gets lost during the processing pathway from the visual cortex through other parts of the brain or that the process of memory retrieval and the early visual processing of a stimulus effect are completely independent processes. Again this is an insight that could not have been made without the additional machine learning component.

Overall it can be stated that in this part of the thesis, ML served as a great tool to generate new hypothesis about the data generating processes. Using classification accuracy is a measure for effect size on complex data can have many advantages that could not be provided by standard group-level analyses. Regarding the collection of all four studies as one dataset, on which three research questions were asked could be seen as an example of a well-controlled exploratory data analysis. It is ensured that overfitting is prevented by using cross-validation mechanisms and despite the exploratory character, the number of performed tests is limited to concrete research questions and controlled by corrections for multiple comparisons. The possibilities that are opened up by this and the general gain of ML are manifold and therefore highly valuable for future research.

Part IV

Benefits of using machine learning in experimental psychology



Chapter 18

Summary

In two parts, representing two case studies, this thesis demonstrated how machine learning could be used to beneficially complement the standard analysis of EEG data in experimental psychology. Before discussing the challenges and advantages involved in using machine learning in experimental psychology, the results of this thesis will be summarized.

Part I - Introduction
gave an introduction to statistics and machine learning by underlining the similarities and differences between the two fields. As an example of how machine learning is already used for experimentally generated neuro-physiological data, BCIs were introduced. The field of BCI research shows successfully what can be achieved by using ML methodology on brain data. In a problem statement, the issues that arise from group-levels statistics, which are current state-of-the-art in experimental psychology were explained to reveal why the introduction of a new methodology would be a good idea. Lastly, an introduction into the field of human memory structures and memory psychology was given, as this was the field of application for this thesis. Two out of the three presented memory structures served as a stage for a proof of concept of the proposed use of the ML methodology within the two case studies.

Part II - Working memory and Executive Functions (EFs)
represented the first of two case studies, on which the benefits of using ML in experimental psychology were shown. The case study dealt with the characterization of components in human working memory based on EEG data. More precisely, the study aimed to identify EFs and tried to capture their commonalities and differences, as these properties are prominently debated in current research. To achieve this a series of four studies have been performed.
In **Chapter 7** the first out of four studies was introduced that focused on the two EFs updating and inhibition. It was found that a general distinction of the two EFs is possible, which can, on the one hand, be attributed to different amounts of WML that they induce. On the other hand, distinct neural activation patterns have been found for each EF by investigating the parameters calculated by the machine learning approach. ML was also able to show that the neural correlates of the two EFs are not interchangeable and do not significantly overlap, emphasizing the diversity of the two EFs.
In **Chapter 8** the second study was introduced, that focused on the two EFs shifting and inhibition. Again it was found that a general distinction is possible, which relies even

stronger on the different amounts of WML that the two EFs induce. Nevertheless, ML was able to provide distinct neural activation patterns, as well as proof for no significant overlap of properties of inhibition and shifting.

Chapter 9 presented the third study of the series. It was a replication of Study 1 with the addition of one more experimental condition. It could be shown that the results of Study 1 can be reproduced in great parts, which can be seen as a strong indicator that neural activation patterns of updating and inhibition are stable and might generalize over a bigger population. The newly introduced condition revealed that not only two but three levels of inhibitory control could be distinguished, depending on whether it is a block or trial wise change of conditions.

Chapter 10 is the last study of part II and can be seen as a replication of Study 2, again with the addition of one more experimental condition. It could also be shown that three levels of inhibitory control exist, which get visible under additional cognitive load. This finding speaks in favor of common attentional resources of the EFs. The patterns for shifting and inhibition could not be replicated entirely, indicating that inter-subject variability could play a bigger role here compared to updating.

Chapter 11 concluded part II by discussing the insights that have been gained within the four studies concerning the unity and diversity of EFs. It also highlighted the advantages that could be achieved by the addition of the ML methodology. Overall, part II showed how explanatory data analysis could be extended and complemented by using ML approaches in combination with classical group-level statistics.

Part III - Episodic memory processes

represented the second of the two case studies on which the benefits of using ML in experimental psychology were be shown. The case study dealt with properties of episodic memory, which can be located in human long-term memory. In total four studies have been performed that try to characterize aspects of memory encoding, memory retrieval and the confidence with which the retrieved information can be expressed.

In **Chapter 13** the design and implementation of all four studies was introduced together with the performed analysis steps for data processing and analysis.

Chapter 14 dealt with the first investigated process within episodic memory, namely memory encoding. The results of all four studies were aggregated concerning this process. The initial assumption that the strength and therefore, also the success of memory encoding can be predicted on the basis of EEG signals had to be refuted. It could be shown that an overall group effect does not necessarily imply that the effect is present in every subject and stable in every trial. Making real-time judgments about the encoding strength was not possible.

Chapter 15 investigated the process of information retrieval during the presentation of information within the four studies. It could be shown that familiar stimuli are processed faster than new stimuli, which is characterized by a shift in latency of about 100 ms in occipital ERPs. ML was able to show that the classification of ERPs can significantly outperform the behavioral accuracy in the recognition test in individual cases. This insight leads to the hypothesis that visual processing is either independent of memory retrieval or part of the processing pathway, which is prone to fail on many occasions. Without the ML approach, this hypothesis could not have been revealed.

Chapter 16 dealt with the aspect of decision confidence while judging if items can be remembered or not. It could be shown that the decision confidence influences feedback processing on a neurophysiological level. The classification accuracy showed that the influence is more distinct in parts in which the ratio of familiar and new stimuli was unbalanced. It was also shown here to what extent the classification quality can and may be interpreted as effect strength.

Chapter 17 concluded part III by discussing the insights that have been gained within the four studies concerning processes of memory encoding. It also highlighted the advantages that could be achieved by the addition of the ML methodology. Overall, part III was an example of how machine learning allows for exploratory and hypothesis generating data analysis within experimental psychology.

Chapter 19

Discussion

How do the insights of part II and III fit together and how do they relate to the posed problem statement?

To conclude this thesis, this chapter represents a final discussion evaluating the insights gained from both performed case studies. The results are discussed mainly concerning the problem statement formulated at the beginning to close the loop and to show the benefits that can be achieved when using machine learning in experimental psychology. To set the expectations right, it needs to be stated that the two methodologies, classical group-level statistics, and single-subject machine learning, both access different levels of information within the data. Hence a direct comparison of the methods is neither fair nor appropriate. Group-level statistics is an essential tool in experimental psychology which can and will not be replaced. Despite the importance, group-level statistics have limitations which were in parts listed in the problem statement. Using machine learning in addition to group-level statistics can attenuate these limitations, which was demonstrated in the two experimental parts of this thesis. The field of application for the approach is not limited to EEG data but can be used for other data types. However, the added value is especially given with dense and high dimensional data, for which EEG is a typical example.

19.1 Patterns

Use high-density data in all its complexity instead of abandoning many dimensions due to statistical technicalities

One of the main issues of classical group-level statistics is the complete omission of potential relations between variables. This omission is a considerable disadvantage in particular for high dimensional data such as EEG recordings because the human brain is known to be a highly functional and interconnected network. The field of machine learning has its origin to a large extent in pattern recognition and the recognition of regularities in large amounts of data. Therefore, it is an excellent candidate to overcome the issue of neglected patterns.

In this thesis, SVMs were chosen to classify data into classes which correspond to a priori determined variables of interest. The SVM integrates all the given data into one classification model to solve the classification problem. In the case of the performed studies electrode sets of 14 to 21 sensors with approximately 500 to 600 values each, were used for the training of the SVMs. The solution of the problem, represented by the resulting accuracy values and neural activation patterns are therefore based on amounts of data that can by no means be captured with classical group-level statistics. The interplay between sensors can be captured, independent of its complexity, as long as it can mathematically be described. Patterns automatically flow into the data model without encountering problems with corrections due to multiple comparisons of single electrodes. However, the accuracy of the classification does not provide any information about the properties of the found patterns. These must be extracted separately by taking a closer look at the generated SVM model. Within this model, each feature received a weight which indicates the importance with regard to the decision line that has been calculated between the two classes. The use of special methods (e.g., Haufe et al. [29]), enables to interpret these values in the sense of neural activity patterns.

Within this thesis, the recognition of patterns has played an important role, especially in the first case study. It could be shown that patterns exist that reflect the relationship between two executive functions, which, to a large extent, also coincide with the existing literature. Although the results could provide important indicators for the unity and diversity of executive functions, which was the central topic of the first case study, they have also shown something much more important. There are ways to make machine learning approaches transparent, making it possible to understand how they work. The fact that the neural activation patterns calculated by the parameters of the SVM are largely consistent with the existing literature makes it clear that the methodology provides meaningful results that can be used for the acquirement of new knowledge.

In this work, only analyses with SVMs were made. There are even more powerful machine learning methods that could be used for this purpose, especially since intensive work is being done to develop methods that can make machine learning methods more transparent. One example are neural networks whose way of working can be made visible by relevance back propagation (cf.,[119]). Overall, it was necessary to start with simple methods to show the value that can be achieved by adding ML to standard analysis techniques while keeping the approach as transparent and comprehensible as possible. In the next step, it would be worthwhile to explore more sophisticated methods to evaluate which additional gains can be achieved to drive psychological research even further forward.

19.2 Single-subject level

Incorporate inter-subject variability instead of eliminating it from the analysis

Group-level statistics reveal if the data of the full group of subjects as a whole supports the posed hypothesis or not. Knowing how each subject behaves individually, in accordance or contrary to the hypothesis is an interesting factor that so far does not get the deserved attention in experimental psychology. Standard group-level statistics do not cover this aspect of the collected data, because the overall effect within the full population is of interest. By assessing confidence intervals it can be estimated how the collected data is distributed, and if adequately marked, it should be possible to reveal if individual subjects diverge from the group or if the group in total is diverse. Correlation analysis can easily be done for individual subjects, but depending on the sample size, and level of interest (feature wise, electrode wise or full time series of the signal) it quickly gets complex and unmanageable to consider every subject individually. So although it is possible to consider the single-subject level, this is rarely or never practiced. In contrast to that, machine learning makes it very easy to build single-subject data models as long as a sufficient number of training data is available for each individual subject (a recommended value are 100 trials per condition). By calculating an accuracy value for each model, it can be estimated how an individual behaves concerning the posed research question. Comparing and evaluating one single value for each subject can easily be managed even in bigger data samples.

In this thesis, the specific performance of an individual subject was especially of importance in the second case study. For the question of stimulus familiarity, the consideration of the single-subject level has led to the generation of new influential hypotheses regarding the memory function of the visual cortex. In these studies, by comparing behavioral and machine learning performance subject by subject, it could be shown that the passively elicited brain signal can be more discriminatory than the knowledge that can be actively retrieved from memory by the subject. This difference between behavior and brain activity indicates gaps in the processing pathway of information or independence of visual processing and knowledge retrieval. This knowledge could only be gained by looking at the single-subject level because the mean values of the two performance measures did not give rise to this assumption. For the process of memory encoding, the single-subject level also has a special significance. Finding that a single-subject and also single-trial classification for the encoding strength is not possible, could indicate that the process is not as stable or even not that discriminative, as the group-level analysis might suggest.

But also for the first case study, the single-subject level was of great importance for the achieved findings. The neural activation patterns were calculated for each subject individually before they were averaged to get an overall estimate. This form of analysis enabled to determine the relationship of the EFs within each subject without paying much attention inter-subject variabilities that can differ from subject to subject. This order of calculation first minimizes the variance within a subject by eliminating interfering elements that are present in both EFs. Through the subsequent averaging across all subjects, the variance between subjects is minimized, but only at the relevant level of patterns and not at the irrelevant level of interfering factors.

In this work only within-subject classifications were performed, because a primary focus was set on the single-subject level. For the sake of completeness, it also needs to be mentioned that a cross-subject classification, i.e., a classification analysis that uses the data of all subjects for the training and testing of a classifier, would also be possible and complete and round off the analyses of all performed studies. Cross-subject analysis can be seen as an equivalent measure for the group-level statistics, but with the inclusion of patterns. With a leave-one-out cross-validation, a classifier could be trained on all subjects except for one, which would then be used for testing and to evaluate the classifier. Again the classification accuracy could be interpreted in terms of effect size for each subject individually, but the information of all subjects could be considered at once. However, a well functioning cross-subject classification is a challenging problem due to the particular issues of EEG recordings. Non-stationarities and big inter-subject variabilities make it hard to find a common feature space in which the collected recordings behave according to the same rules. To tackle this problem, normalization methods have been developed that begin to soften the issue, but there is still a lot of work that needs to be done. To name a few, methods like the Riemann Geometry [177] and the Stationary subspace analysis [178] are examples that have found their way in BCI research to normalize the available data for unified classifiers across subjects. Therefore, a focus of future work could be on cross-subject classification to further deepen the knowledge of cognitive processes on the group-level.

19.3 Latent variables

Detect and identify confounds or other factors of importance

Latent variables are variables that cannot be measured directly but have an influence on the measurable outcome of an experiment. In most of the studies, the influence of latent variables is fully neglected regarding the data analysis. Since the detection and identification of latent variables is a complicated and so far unsolved problem, this can hardly be criticized. Random effects model is a way in current statistics that try to include and capture random effects within experimental data. Other methods like the independent component analysis (ICA) can also be used to investigate the components in the data, that show the most variance and therefore have the most influence on the statistical model. However, they need to be additionally performed and do not necessarily provide an output that helps to understand the data better.

In the second case study, it could be shown that ML can be a useful tool in the identification of latent variables. The ratio of old and new stimuli within the test phase of the experiment did have a significant impact on the neurophysiological signals of subjects, which was not visible in the standard group-level analysis. The single-subject classification approach, however, revealed that there are differences between experimental parts with balanced ratios of old and new stimuli (II and IV) and unbalanced ratios (I and III). In this particular case, it was not only possible to detect a latent variable but also to identify it, due to the structure of the four consecutive experiments. In other cases, this might not be that easy, because of fewer opportunities to compare the influence of individual parameters. Nevertheless, the use of machine learning can help to identify areas of interest that show a high level of classification accuracies, without easily explainable effects. These areas can then be further investigated concerning potential candidates of latent variables that could be responsible for the measured outcome. A big advantage is that the classification performance serves as a measure of effect size and conglomerates a lot of information in one value. Using the computed values as signposts for what and which types of analyses need to follow to better understand the data, seems, therefore, a great and easy to use idea.

19.4 Exploratory and explanatory data analysis

Save data: Allowing exploratory data analysis by using well controllable analysis techniques

Science suffers from high pressure to publish exciting, new, impressive, and influential results. This is not only true for experimental psychology but the whole scientific community. In general, aiming high in your endeavors is essential for succeeding and gaining new insights. However, the pursuit of success always has a drawback. Failures, which in scientific research can be null effects or results that are not in accordance with the literature, are rarely reported as they cannot be used to acquire funding. Since experiments "fail" at regular intervals, due to various reasons that might not be in the control of the experimenter, a lot of research vanishes in cabinet drawers. Using the study results for a different purpose instead of leaving it in a drawer is usually not an option due to the strict limitations of hypothesis-driven designs. In addition, due to probability distributions, the chance of finding effects by pure luck get higher, the more often statistical tests are performed. Performing them anyway would only lead to the publishing of random effects instead of null effects which are also not desired.

Especially for the second case study of the thesis, the possibility to perform exploratory data analysis was key to the achieved findings. Overall, three instead of one research questions were asked for the assembly of studies based on episodic memory processes. They were developed continuously and not based on a-priori defined hypothesis, but on questions that can legitimately be asked when looking at the experimental design. Multiple hypothesis testing is made possible by using different approaches that ensure in their entirety that the findings are not random but tangible. The used methodologies work hand in hand and support each other's findings although their way of working is independent of each other. Especially for the stimulus familiarity aspect of the study, it could be shown that correlation analysis combined with group-level ERP analysis and several classification approaches all supported the same effect.
Regarding the aspect of memory encoding, the exploratory ML approach, in combination with the neural activation patterns revealed error-related potentials during the feedback of the task, which could be validated by classical group-level analysis. Therefore, it could be proven that the effects that can be identified in the ML-based exploratory analysis approach are legit effects that enable greater insight into the available data. The fact that the results are supported by more than one of the performed studies further enhances credibility and reduces the probability of chance findings.

The machine learning methodology has several properties that lift the strict barrier of exploratory data analysis to a certain extent. Apart from the issue of multiple testing, statistical models often tend to overfit. The process of overfitting can also be expressed by learning the data by heart or finding the wrong or no level of abstraction for the data. One of the great strengths of machine learning are mechanisms that counteract overfitting since the generalizability of the model is the priority. Key to finding generalizing models is to simplify the model and also the parameter space. A reduction of the parameter and model space thereby also reduces the chances of finding effects simply by coincidence.

It must be emphasized that rigorous testing without concrete research question is still not possible and should not be encouraged at any time. Explorative data analysis can only be realized through a tightly knit network of methods which combine generalizability and inference of the data. In the case of contradictory results between the methods, caution is required as far as interpretation is concerned. However, if the results work together, this can be interpreted as a strong indication that a well-founded effect exists.

19.5 Reproducibility

Choose generalizability over the goodness of the fit to facilitate reproducibility

The Reproducibility Project [1] from the year 2015 showed impressively that social sciences face a serious problem that needs to be overcome to create credible and well-grounded research. Especially the already stated issue of overfitting, but essentially a combination of all the above-stated issues play a crucial role in the emergence of the crisis. The suggestion to use machine learning methods to eliminate the disadvantages of standard group-level statistics, which are in great parts responsible for the issue, could be demonstrably explained in this thesis as meaningful. In both experimental parts, it was possible to replicate the results of studies and to underpin them with different methodologies. In particular, it was also possible to replicate the results of an external, not self-designed study. Scharinger and colleagues collected the data of study 1 from the working memory part of this thesis. For the design and the survey of the follow-up studies, the expertise of Dr. Scharinger was consulted, but the concrete implementation of the experiment is different from the original study. Nevertheless, the same findings as in the original study of Scharinger and colleagues [57] could be made using the conventional as well as the ML approach. This means that both approaches generate consistent results. However, the ML approach is less error-prone due to the inherent cross-validation and the broader scope, facilitation the group as well as the single-subject level. Furthermore, additional information could be generated regarding the unity and diversity of EFs, which supplements the findings and therefore generate higher confidence in the overall results. Taking this into account, it becomes apparent that the use of ML offers a possibility to increase reproducibility by increasing the confidence in the results.

The second case study also showed that results could be replicated despite new samples for each study. Especially in this part of the thesis, it could be shown that the results achieved by machine learning, can be validated by classical group-level statistics. Therefore, this part strongly legitimizes to use ML in the context of data analysis in experimental psychology because it produces reliable and comprehensible results. The cross-validation and generalizability make ML a strong tool that simplifies data analysis and its validation compared to standard analysis approaches. But especially when using the two approaches in combination with each other the strength and the advantages get undeniably clear, further increasing the credibility of results.

In general, the main focus of this thesis was to make use of this generalizability of the results and the gain in knowledge that can be achieved by the addition of the new methodology. Generalizability is key for reproducibility for which deductions regarding the goodness of the fit, should willingly be accepted. For this reason, many inconspicuous and at first glance, poor results can be found in this thesis, especially concerning the model quality. Many classification accuracy values of the machine learning models lie barely noticeably above the chance level. An improvement of the values would have very likely been possible with parameter tuning or feature engineering. However, even with results that appear poor, a tangible statement can be made, provided that it is ensured that the deviation from the random level is significant. Therefore, the goodness of the fit was neglected for the sake of a reliable prediction, and the credibility and reproducibility of results.

Chapter 20

Conclusion

Group-level statistics and machine learning are two methods with different strengths that can access different levels of information in data. The classical group-based statistics averages the available measures over all subjects and aims to find differences on a group level. The basis of this analysis is to find what is common between all subjects within one condition and to see if this condition is significantly different from a second condition. A priori defined hypotheses that explain the expected effects are confirmed or rejected based on the group-level statistics. The strength of this analysis is that the findings are generalized over big amounts of data that allow inferences about the variance of the effects across populations of subjects.

In contrast to that, the ML approach focused on finding statistically significant differences between the two conditions for each subject individually. Instead of calculating averages or an analysis of variance, this is done by extracting mathematically optimized patterns from the available signals, that make the conditions distinguishable. To avoid over-fitting, cross-validation is implemented, that extracts the patterns from one part of the data and then validates the applicability of the found patterns on a second part of the data, that has explicitly been left out for testing. The strength of this analysis is that it allows applying the gained knowledge to new data points. The approach predicts the condition of each data point individually which implies that the findings are generalized and validated on each subject separately that allow making inferences about the variance of the effects within subjects.

The combination of both approaches, therefore, allows combining an explanatory as well as a predictive approach to create more significant insights into the experimental data. Conclusions can be drawn on a group level, but also on a single subject level. Especially when dealing with EEG data, this is of great importance, as inter-subject variability can be the key to understand complex mental states on the neurophysiological level. In both parts, it could be demonstrated that different results are achieved when using single-subject ML approaches, compared to standard group-level statistics. Recent developments call for the use of both methodologies to make progress in the scientific world [179], [180]. It is therefore suggested and encouraged by the results of this thesis and the scientific community to take both analysis steps into account.

List of Abbreviations

ALS	Amythrophic lateral sclerosis
ANOVA	Analysis of Variance
BCI	Brain-Computer interface
CAR	Common average referencing
CCA	Canonical correlation analysis
CLT	Cognitive load theory
EEG	Electroencephalography
EF	Executive Function
EOG	Electrooculogram
ERD	Event-related desynchronization
ERP	Event-related potential
ErrP	Error potential
ERS	Event-related synchronization
FCE	Flanker congruency effect
FRN	Feedback-related negativity
ICA	Independent component analysis
MEG	Magnet encephalography
ML	Machine learning
MRI	Magnet resonance imaging learning
NIRS	Near infrared spectroscopy
PET	Positron emission tomography
RP	Readiness potential
RT	Reaction time
SVM	Support vector machine
WM	Working memory
WML	Working memory load

List of publications

1. Krumpe, T., Gerjets, P., Rosenstiel, W., & Spüler, M. (2020). Decision confidence: EEG correlates of confidence in different phases of an old/new recognition task. *Brain-Computer Interfaces* Taylor & Francis.

2. Krumpe, T., Scharinger, C., Rosenstiel, W., Gerjets, P., & Spüler, M. (2018). Unity and diversity in working memory load: Evidence for the separability of the executive functions updating and inhibition using machine learning. *Biological psychology*, 139, 163-172.

3. Krumpe, T., Scharinger, C., Rosenstiel, W., Gerjets, P., & Spüler, M. (2018). Using a machine learning approach to complement group level statistics in experimental psychology: A case study to reveal different levels of inhibition in a modified Flanker Task. *bioRxiv*, 502278. (*under review*)

4. Spüler, M., Krumpe, T., Walter, C., Scharinger, C., Rosenstiel, W., & Gerjets, P. (2017). Brain-computer interfaces for educational applications. In *Informational Environments* (pp. 177-201). Springer, Cham.

5. Grissmann, S., Spüler, M., Faller, J., Krumpe, T., Zander, T., Kelava, A., Scharinger, C., & Gerjets, P. (2017). Context sensitivity of EEG-based workload classification under different affective valence. *IEEE Transactions on Affective Computing*

6. Krumpe, T., Rosenstiel, W., & Spüler, M. (2019) Prediction of item familiarity based on ERPs. *2019 7th International Conference on Brain-Computer Interface (BCI). IEEE, 2019.*

7. Krumpe, T., Baumgärtner, K., Rosenstiel, W., Spüler, M. (2017) Non-stationarity and inter-subject variability of EEG characteristics in the context of BCI development *Proceedings of the 7th International BCI Conference Graz 2017.*

Bibliography

[1] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[2] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[3] Björn Rasch, Malte Friese, Wilhelm Hofman, and Ewald Naumann. *Quantitative Methoden 1*. Springe Science & Business Mediar, 2004.

[4] Björn Rasch, Malte Friese, Wilhelm Hofman, and Ewald Naumann. *Quantitative Methoden 2*. Springe Science & Business Mediar, 2004.

[5] Martin Schumacher and Gabriele Schulgen-Kristiansen. *Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung*. Springer DE, 2008.

[6] Ruth Bernstein and Stephen Bernstein. *Schaum's Outline of Elements of Statistics II: Inferential Statistics*. McGraw-Hill Professional, 1999.

[7] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[8] Vladimir N. Vapnik and Alexey J Chervonenkis. Theory of Pattern Recognition. 1974.

[9] Bernhard Schölkopf and A. J Smola. *Learning with kernels*. MIT Press, 2001.

[10] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297, 1999.

[11] Niels Birbaumer. Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.

[12] Hans Berger. Über das Elektroencephalogramm des Menschen. *European Archives of Psychiatry and Clinical Neuroscience*, 87(1):527–570, 1929.

[13] Herbert H. Jasper. The ten-twenty electrode system of the International Federation. *Electroencelography Clinical Neurphysiology*, 10:371–375, 1958.

[14] A. Kok. On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, 38:557–577, 2001.

[15] J. Polich. Updating P300: An integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

[16] M.A. Conroy and J Polich. Normative variation of P3a and P3b from a large sample (N=120). *Journal of Psychophysiology*, 21:22–32, 2007.

[17] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. *Psychophysiological Brain Research*, pages 192–195, 1990.

[18] Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

[19] Lee Osterhout and Phillip J Holcomb. Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806, 1992.

[20] John Parker Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37(2):375–376, 1972.

[21] Edward Niedermeyer. Alpha rhythms as physiological and abnormal phenomena. *International Journal of Psychophysiology*, 26(1-3):31–49, 1997.

[22] Stuart N Baker. Oscillatory interactions between sensorimotor cortex and the periphery. *Current opinion in neurobiology*, 17(6):649–655, 2007.

[23] Alois Schlögl, Claudia Keinrath, Doris Zimmermann, Reinhold Scherer, Robert Leeb, and Gert Pfurtscheller. A fully automated correction method of eog artifacts in eeg recordings. *Clinical neurophysiology*, 118(1):98–104, 2007.

[24] O. Bertrand, F Perrin, and J Pernier. A theoretical justification of the average reference in topographic evoked potential studies. *Electroencelography Clinical Neurophysiology*, 62:462–464, 1985.

[25] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[26] Martin Spüler, Armin Walter, Wolfgang Rosenstiel, and Martin Bogdan. Spatial filtering based on canonical correlation analysis for classification of evoked or event-related potentials in eeg data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(6):1097–1103, 2014.

[27] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

[28] Matthias Schultze-Kraft, Daniel Birman, Marco Rusconi, Carsten Allefeld, Kai Görgen, Sven Dähne, Benjamin Blankertz, and John-Dylan Haynes. The point of no return in vetoing self-initiated movements. *Proceedings of the National Academy of Sciences*, 113(4):1080–1085, 2016.

[29] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[30] Tanja Krumpe, Katrin Baumgärtner, Wolfgang Rosenstiel, and Martin Spüler. Non-stationarity and inter-subject variablility of eeg characteristics in the context of bci developement. In *7th Graz Brain-Computer Interface Conference 2017*, pages 260–265. Graz University of Technology, 2017.

[31] Klaus-Robert Müller, Michael Tangermann, Guido Dornhege, Matthias Krauledat, Gabriel Curio, and Benjamin Blankertz. Machine learning for real-time single-trial eeg-analysis: from brain–computer interfacing to mental state monitoring. *Journal of neuroscience methods*, 167(1):82–90, 2008.

[32] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes1. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.

[33] Alan Baddeley, Graham Hitch, et al. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[34] Alan Baddeley. Exploring the central executive. *The Quarterly Journal of Experimental Psychology: Section A*, 49(1):5–28, 1996.

[35] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

[36] Larry R Squire. Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience*, 4(3):232–243, 1992.

[37] Akira Miyake and Priti Shah. *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge University Press, 1999.

[38] Alan Baddeley, Michael W Eysenck, and Michael C Anderson. *Memory.* Psychology Press, 2009.

[39] Akira Miyake, Naomi P Friedman, Michael J Emerson, Alexander H Witzki, Amy Howerter, and Tor D Wager. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100, 2000.

[40] Pierre Barrouillet, Sophie Bernardin, and Valérie Camos. Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, 133(1):83, 2004.

[41] Nelson Cowan, Emily M Elliott, J Scott Saults, Candice C Morey, Sam Mattox, Anna Hismjatullina, and Andrew RA Conway. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, 51(1):42–100, 2005.

[42] Michael J Kane, David Z Hambrick, Stephen W Tuholski, Oliver Wilhelm, Tabitha W Payne, and Randall W Engle. The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, 133(2):189, 2004.

[43] Klaus Oberauer. Design for a working memory. *Psychology of learning and motivation*, 51:45–100, 2009.

[44] Nash Unsworth and Randall W Engle. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological review*, 114(1):104, 2007.

[45] Nelson Cowan, J Scott Saults, and Christopher L Blume. Central and peripheral components of working memory storage. *Journal of Experimental Psychology: General*, 143(5):1806, 2014.

[46] Akira Miyake and Naomi P Friedman. The nature and organization of individual differences in executive functions four general conclusions. *Current directions in psychological science*, 21(1):8–14, 2012.

[47] Tim Shallice and Paul Burgess. *Supervisory control of action and thought selection.* Clarendon Press/Oxford University Press, 1993.

[48] Paul W Burgess. Theory and methodology in executive function research. *Methodology of frontal and executive function*, pages 81–116, 1997.

[49] Tim Shallice and Paul Burgess. Deficits in strategy application following frontal lobe damage in man. *Brain*, 114(2):727–741, 1991.

[50] Randall W Engle. Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23, 2002.

[51] Tara A Niendam, Angela R Laird, Kimberly L Ray, Y Monica Dean, David C Glahn, and Cameron S Carter. Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, 12(2):241–268, 2012.

[52] Naomi P Friedman, Akira Miyake, Susan E Young, John C DeFries, Robin P Corley, and John K Hewitt. Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, 137(2):201, 2008.

[53] Naomi P Friedman, Akira Miyake, JoAnn L Robinson, and John K Hewitt. Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: a behavioral genetic analysis. *Developmental psychology*, 47(5):1410, 2011.

[54] Timothy A Salthouse, Thomas M Atkinson, and Diane E Berish. Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of experimental psychology: General*, 132(4):566, 2003.

[55] Fabienne Collette, Martial Van der Linden, Steven Laureys, Guy Delfiore, Christian Degueldre, Andre Luxen, and Eric Salmon. Exploring the unity and diversity of the neural substrates of executive functioning. *Human brain mapping*, 25(4):409–423, 2005.

[56] Fabienne Collette, Michaël Hogge, Eric Salmon, and Martial Van der Linden. Exploration of the neural substrates of executive functioning by functional neuroimaging. *Neuroscience*, 139(1):209–221, 2006.

[57] Christian Scharinger, Alexander Soutschek, Torsten Schubert, and Peter Gerjets. When flanker meets the n-back: What eeg and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology*, 52(10):1293–1304, 2015.

[58] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[59] Hans-Lukas Teuber. Unity and diversity of frontal lobe functions. *Acta Neurobiol. Exp*, 32:615–656, 1972.

[60] Charles W Eriksen. The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2:101–118, 1995.

[61] AF Sanders and JM Lamers. The eriksen flanker effect revisited. *Acta Psychologica*, 109(1):41–56, 2002.

[62] Adele Diamond. Executive functions. *Annual review of psychology*, 64:135, 2013.

[63] Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Attention, Perception, & Psychophysics*, 16(1):143–149, 1974.

[64] Carola Lehle and Ronald Hübner. On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review*, 15(4):814–818, 2008.

[65] Gabriele Gratton, Michael GH Coles, and Emanuel Donchin. Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General*, 121(4):480, 1992.

[66] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[67] Michael Inzlicht, Bruce D Bartholow, and Jacob B Hirsh. Emotional foundations of cognitive control. *Trends in cognitive sciences*, 19(3):126–132, 2015.

[68] Henk van Steenbergen, Guido PH Band, and Bernhard Hommel. Reward counteracts conflict adaptation: Evidence for a role of affect in executive control. *Psychological Science*, 20(12):1473–1477, 2009.

[69] Neil Morris and Dylan M Jones. Memory updating in working memory: The role of the central executive. *British journal of psychology*, 81(2):111–121, 1990.

[70] John Jonides, Eric H Schumacher, Edward E Smith, Erick J Lauber, Edward Awh, Satoshi Minoshima, and Robert A Koeppe. Verbal working memory load affects regional brain activation as measured by pet. *Journal of cognitive neuroscience*, 9(4):462–475, 1997.

[71] Saul Sternberg. The discovery of processing stages: Extensions of donders' method. *Acta psychologica*, 30:276–315, 1969.

[72] Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[73] Scott Watter, Gina M Geffen, and Laurie B Geffen. The n-back as a dual-task: P300 morphology under divided attention. *Psychophysiology*, 38(6):998–1003, 2001.

[74] Katie C Ewing and Stephen H Fairclough. The effect of an extrinsic incentive on psychophysiological measures of mental effort and motivational disposition when task demand is varied. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 259–263. Sage Publications Sage CA: Los Angeles, CA, 2010.

[75] Christopher H Chatham, Seth A Herd, Angela M Brant, Thomas E Hazy, Akira Miyake, Randy O'Reilly, and Naomi P Friedman. From an executive network to executive control: a computational model of the n-back task. *Journal of cognitive neuroscience*, 23(11):3598–3619, 2011.

[76] Y Chen, Suvobrata Mitra, and Friederike Schlaghecken. Interference from the irrelevant domain in n-back tasks: an erp study. *Acta Neurologica Taiwanica*, 16(3):125, 2007.

[77] Yung-Nien Chen, Suvobrata Mitra, and Friederike Schlaghecken. Sub-processes of working memory in the n-back task: An investigation using erps. *Clinical Neurophysiology*, 119(7):1546–1559, 2008.

[78] Alan Gevins and Michael E Smith. Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral cortex*, 10(9):829–839, 2000.

[79] JB Krause, John G Taylor, Daniela Schmidt, Hubertus Hautzel, Felix M Mottaghy, and H-W Müller-Gärtner. Imaging and neural modelling in episodic and working memory processes. *Neural Networks*, 13(8):847–859, 2000.

[80] Christina M Krause, Mirka Pesonen, and Heikki Hämäläinen. Brain oscillatory 4–30 hz electroencephalogram responses in adolescents during a visual memory task. *Neuroreport*, 21(11):767–771, 2010.

[81] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

[82] Mirka Pesonen, Heikki Hämäläinen, and Christina M Krause. Brain oscillatory 4–30 hz responses during a visual n-back memory task with varying memory load. *Brain research*, 1138:171–177, 2007.

[83] Stephen Monsell. Control of mental processes. *Unsolved mysteries of the mind: Tutorial essays in cognition*, pages 93–148, 1996.

[84] David A Grant and Esta Berg. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. *Journal of experimental psychology*, 38(4):404, 1948.

[85] Arthur Thomas Jersild. Mental set and shift. *Archives of psychology*, 1927.

[86] Stephen Monsell. Task switching. *Trends in cognitive sciences*, 7(3):134–140, 2003.

[87] Orit Rubin and Nachshon Meiran. On the origins of the task mixing cost in the cuing task-switching paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1477, 2005.

[88] Irving Biederman. Mental set and mental arithmetic. *Memory & Cognition*, 1(3):383–386, 1973.

[89] Padmanabhan Sudevan and David A Taylor. The cuing and priming of cognitive operations. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1):89, 1987.

[90] Tanja Krumpe, Christian Scharinger, Wolfgang Rosenstiel, Peter Gerjets, and Martin Spüler. Unity and diversity in working memory load: Evidence for the separability of the executive functions updating and inhibition using machine learning. *Biological psychology*, 2018.

[91] Martin Spüler, Tanja Krumpe, Carina Walter, Christian Scharinger, Wolfgang Rosenstiel, and Peter Gerjets. Brain-computer interfaces for educational applications. In *Informational Environments*, pages 177–201. Springer, 2017.

[92] Eddy J Davelaar. When the ignored gets bound: sequential effects in the flanker task. *Frontiers in psychology*, 3, 2012.

[93] HH Jasper. The 10/20 international electrode system. *EEG and Clinical Neurophysiology*, 10:371–375, 1958.

[94] John C Platt. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.* MIT Press, 2000.

[95] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[96] MATLAB. *version 7.10.0 (R2015b).* The MathWorks Inc., Natick, Massachusetts, 2015.

[97] Martin Spüler, Wolfgang Rosenstiel, and Martin Bogdan. One class svm and canonical correlation analysis increase performance in a c-vep based brain-computer interface (bci). In *ESANN*, 2012.

[98] Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media, 2013.

[99] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

[100] Katie Ewing and Stephen Fairclough. The impact of working memory load on psychophysiological measures of mental effort and motivational disposition. *Human Factors: A system view of human, technology and organisation. Maastricht: Shaker Publishing*, 2010.

[101] Alan Gevins, Michael E Smith, Linda McEvoy, and Daphne Yu. High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral cortex*, 7(4):374–385, 1997.

[102] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[103] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[104] Patrik Sörqvist and Jerker Rönnberg. Individual differences in distractibility: an update and a model. *PsyCh journal*, 3(1):42–57, 2014.

[105] Stanislas Dehaene, Serge Bossini, and Pascal Giraux. The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122(3):371, 1993.

[106] Hans-Christoph Nuerk, Frank Bauer, Joseph Krummenacher, Dieter Heller, and Klaus Willmes. The power of the mental number line: How the magnitude of unattended numbers affects performance in an eriksen task. *Psychology Science*, 47(1):34–50, 2005.

[107] Robert S Moyer and Thomas K Landauer. Time required for judgements of numerical inequality. *Nature*, 1967.

[108] Hans-Christoph Nuerk, Wiebke Iversen, and Klaus Willmes. Notational modulation of the snarc and the marc (linguistic markedness of response codes) effect. *Quarterly Journal of Experimental Psychology Section A*, 57(5):835–863, 2004.

[109] Tanja Krumpe, Christian Scharinger, Wolfgang Rosenstiel, Peter Gerjets, and Martin Spüler. Using a machine learning approach to complement group level statistics in experimental psychology: A case study to reveal different levels of inhibition in a modified flanker task. *bioRxiv*, page 502278, 2018.

[110] So-Yeon Kim, Min-Shik Kim, and Marvin M Chun. Concurrent working memory load can reduce distraction. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16524–16529, 2005.

[111] Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412, 2010.

[112] Amos Spector and Irving Biederman. Mental set and mental shift revisited. *The American Journal of Psychology*, pages 669–679, 1976.

[113] Robert D Rogers and Stephen Monsell. Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, 124(2):207, 1995.

[114] Ole Jensen and Claudia D Tesche. Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience*, 15(8):1395–1399, 2002.

[115] P Missonnier, M-P Deiber, G Gold, P Millet, M Gex-Fabry Pun, L Fazio-Costa, P Giannakopoulos, and V Ibáñez. Frontal theta event-related synchronization: comparison of directed attention and working memory load effects. *Journal of Neural Transmission*, 113(10):1477–1486, 2006.

[116] David P McCabe, Henry L Roediger III, Mark A McDaniel, David A Balota, and David Z Hambrick. The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology*, 24(2):222, 2010.

[117] J. Sweller, J. van Merrienboer, and F. Pass. Cognitive architecture and instructional design. *Educational Psychology Review*, 10:251–296, 1998.

[118] Andrea Kübler, Nicola Neumann, Jochen Kaiser, Boris Kotchoubey, Thilo Hinterberger, and Niels P Birbaumer. Brain-computer communication: self-regulation of slow cortical potentials for verbal communication. *Archives of physical medicine and rehabilitation*, 82(11):1533–1539, 2001.

[119] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[120] Keisuke Fukuda and Geoffrey F Woodman. Predicting and improving recognition memory using multiple electrophysiological signals in real time. *Psychological science*, page 0956797615578122, 2015.

[121] Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annu. Rev. Neurosci.*, 30:535–574, 2007.

[122] David M Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[123] Philip L Smith and Roger Ratcliff. Psychology and neurobiology of simple decisions. *Trends in neurosciences*, 27(3):161–168, 2004.

[124] Simon P Kelly and Redmond G O'Connell. The neural processes underlying perceptual decision making in humans: recent progress and future directions. *Journal of Physiology-Paris*, 109(1):27–37, 2015.

[125] David Friedman and Charlotte Trott. An event-related potential study of encoding in young and older adults. *Neuropsychologia*, 38(5):542–557, 2000.

[126] Ken A Paller, Marta Kutas, and Andrew R Mayes. Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography and clinical neurophysiology*, 67(4):360–371, 1987.

[127] Ken A Paller, Gregory McCarthy, and Charles C Wood. Erps predictive of subsequent recall and recognition performance. *Biological psychology*, 26(1-3):269–276, 1988.

[128] Simon Hanslmayr, Bernhard Spitzer, and Karl-Heinz Bäuml. Brain oscillations dissociate between semantic and nonsemantic encoding of episodic memories. *Cerebral cortex*, 19(7):1631–1640, 2009.

[129] W Klimesch, H Schimke, M Doppelmayr, B Ripper, J Schwaiger, and G Pfurtscheller. Event-related desynchronization (erd) and the dm effect: does alpha desynchronization during encoding predict later recall performance? *International Journal of Psychophysiology*, 24(1):47–60, 1996.

[130] Ken A Paller and Anthony D Wagner. Observing the transformation of experience into memory. *Trends in cognitive sciences*, 6(2):93–102, 2002.

[131] Uwe Friese, Moritz Köster, Uwe Hassler, Ulla Martens, Nelson Trujillo-Barreto, and Thomas Gruber. Successful memory encoding is associated with increased cross-frequency coupling between frontal theta and posterior gamma oscillations in human scalp-recorded eeg. *Neuroimage*, 66:642–647, 2013.

[132] Daria Osipova, Atsuko Takashima, Robert Oostenveld, Guillén Fernández, Eric Maris, and Ole Jensen. Theta and gamma oscillations predict encoding and retrieval of declarative memory. *Journal of neuroscience*, 26(28):7523–7531, 2006.

[133] Eunho Noh, Grit Herzmann, Tim Curran, and Virginia R de Sa. Using single-trial eeg to predict and analyze subsequent memory. *NeuroImage*, 84:712–723, 2014.

[134] Michael D Rugg and Tim Curran. Event-related potentials and recognition memory. *Trends in cognitive sciences*, 11(6):251–257, 2007.

[135] Michael D Rugg. Memory and consciousness: A selective review of issues and data. *Neuropsychologia*, 33(9):1131–1141, 1995.

[136] Ken A Paller, Joel L Voss, and Stephan G Boehm. Validating neural correlates of familiarity. *Trends in cognitive sciences*, 11(6):243–250, 2007.

[137] Michael D Rugg and Andrew P Yonelinas. Human recognition memory: a cognitive neuroscience perspective. *Trends in cognitive sciences*, 7(7):313–319, 2003.

[138] T Curran. The electrophysiology of incidental and intentionalretrieval: erp old new effects in lexical decision andrecognition memory. *Neuropsychologia*, 37(7):771–785, 1999.

[139] Tim Curran. Brain potentials of recollection and familiarity. *Memory & cognition*, 28(6):923–938, 2000.

[140] Michael D Rugg, Jane E Herron, and Alexa M Morcom. Electrophysiological studies of retrieval processing. *Neuropsychology of memory*, 3:154–165, 2002.

[141] Mortimer Mishkin. A memory system in the monkey. *Phil. Trans. R. Soc. Lond. B*, 298(1089):85–95, 1982.

[142] J-Z Xiang and MW Brown. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37(4-5):657–676, 1998.

[143] FL Fahy, IP Riches, and MW Brown. Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Experimental Brain Research*, 96(3):457–472, 1993.

[144] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

[145] Douglas Vickers. *Decision processes in visual perception.* Academic Press, 2014.

[146] Floris P de Lange, Ole Jensen, and Stanislas Dehaene. Accumulation of evidence during sequential decision making: the importance of top–down factors. *Journal of Neuroscience*, 30(2):731–738, 2010.

[147] Roozbeh Kiani, Leah Corthell, and Michael N Shadlen. Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6):1329–1342, 2014.

[148] Jan Kubanek, N Jeremy Hill, Lawrence H Snyder, and Gerwin Schalk. Cortical alpha activity predicts the confidence in an impending action. *Frontiers in neuroscience*, 9, 2015.

[149] Sabina Gherman and Marios G Philiastides. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *Neuroimage*, 106:134–143, 2015.

[150] Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764, 2009.

[151] C.S Carter. Anterior cingulate cortex, error detection and the online monitoring of performance. *Science*, 280:747–749, 1998.

[152] Olav E Krigolson and Clay B Holroyd. Hierarchical error processing: different errors, different systems. *Brain research*, 1155:70–80, 2007.

[153] Clay B Holroyd, Sander Nieuwenhuis, Nick Yeung, and Jonathan D Cohen. Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, 14(18):2481–2484, 2003.

[154] Jan R Wessel, Claudia Danielmeier, and Markus Ullsperger. Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of cognitive neuroscience*, 23(10):3021–3036, 2011.

[155] Wolfgang HR Miltner, Christoph H Braun, and Michael GH Coles. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a “generic” neural system for error detection. *Journal of cognitive neuroscience*, 9(6):788–798, 1997.

[156] Anna Weinberg, Christian C Luhmann, Jennifer N Bress, and Greg Hajcak. Better late than never? the effect of feedback delay on erp indices of reward processing. *Cognitive, Affective, & Behavioral Neuroscience*, 12(4):671–677, 2012.

[157] Marten K Scheffers and Michael GH Coles. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1):141, 2000.

[158] Annika Boldt and Nick Yeung. Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35(8):3478–3484, 2015.

[159] Tanja Krumpe, Peter Gerjets, Wolfgang Rosenstiel, and Martin Spüler. Decision confidence: Eeg correlates of confidence in different phases of an old/new recognition task. *Brain-Computer Interfaces*, pages 1–16, 2020.

[160] Tanja Krumpe, Wolfgang Rosenstiel, and Martin Spüler. Prediction of item familiarity based on erps. In *2019 7th International Conference on Brain-Computer Interface (BCI)*, pages 90–95. IEEE, 2019.

[161] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[162] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[163] Sebastian Nagel, Werner Dreher, Wolfgang Rosenstiel, and Martin Spüler. The effect of monitor raster latency on veps, erps and brain–computer interface performance. *Journal of neuroscience methods*, 295:45–50, 2018.

[164] Etienne Combrisson and Karim Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250:126–136, 2015.

[165] Michael Falkenstein, Jörg Hoormann, Stefan Christ, and Joachim Hohnsbein. Erp components on reaction errors and their functional significance: a tutorial. *Biological psychology*, 51(2-3):87–107, 2000.

[166] Martin Spüler, Wolfgang Rosenstiel, and Martin Bogdan. Online adaptation of a c-vep brain-computer interface (bci) based on error-related potentials and unsupervised learning. *PloS one*, 7(12):e51077, 2012.

[167] Martin Spüler, Michael Bensch, Sonja Kleih, Wolfgang Rosenstiel, Martin Bogdan, and Andrea Kübler. Online use of error-related potentials in healthy users and

people with severe motor impairment increases performance of a p300-bci. *Clinical Neurophysiology*, 123(7):1328–1337, 2012.

[168] Anna Buttfield, Pierre W Ferrez, and Jd R Millan. Towards a robust bci: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):164–168, 2006.

[169] Tanja Krumpe, Peter Gerjets, Wolfgang Rosenstiel, and Martin Spüler. Decision confidence: Eeg correlates of confidence in different phases of a decision task. *bioRxiv*, page 479204, 2018.

[170] Ben Eppinger, Jutta Kray, Barbara Mock, and Axel Mecklinger. Better or worse than expected? aging, learning, and the ern. *Neuropsychologia*, 46(2):521–539, 2008.

[171] Michael X Cohen, Christian E Elger, and Charan Ranganath. Reward expectation modulates feedback-related negativity and eeg spectra. *Neuroimage*, 35(2):968–978, 2007.

[172] Clay B Holroyd, Kaivon L Pakzad-Vaezi, and Olave E Krigolson. The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, 45(5):688–697, 2008.

[173] Martin Spüler, Carina Walter, Wolfgang Rosenstiel, Peter Gerjets, Korbinian Moeller, and Elise Klein. Eeg-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM*, 48(3):267–278, 2016.

[174] Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, Peter Gerjets, and Martin Spüler. Online eeg-based workload adaptation of an arithmetic learning environment. *Frontiers in Human Neuroscience*, 11:286, 05 2017.

[175] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296, 1998.

[176] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O Zander. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, 8, 2014.

[177] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to bci classification. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 629–636. Springer, 2010.

[178] Paul Von Bünau, Frank C Meinecke, Franz C Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Physical review letters*, 103(21):214101, 2009.

[179] Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017.

[180] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. Points of significance: statistics versus machine learning. *Nature Methods*, pages 1–7, 2018.