Brain-Computer Interfaces using Machine Learning: Reducing calibration time in Motor Imagery

Nina Proesmans

Supervisors: Prof. dr. ir. Joni Dambre, Dr. ir. Pieter van Mierlo
Counsellor: Ir. Thibault Verhoeven

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Biomedical Engineering

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik Van de Walle
Faculty of Engineering and Architecture
Academic year 2015-2016

UNIVERSITEIT GENT

Brain-Computer Interfaces using Machine Learning: Reducing calibration time in Motor Imagery

Nina Proesmans

Supervisors: Prof. dr. ir. Joni Dambre, Dr. ir. Pieter van Mierlo
Counsellor: Ir. Thibault Verhoeven

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Biomedical Engineering

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik Van de Walle
Faculty of Engineering and Architecture
Academic year 2015-2016

UNIVERSITEIT
GENT

Preface

With this preface I would like to thank everyone that helped me completing this master thesis as the finalisation of my engineering studies.
First, I would like to thank my supervisor Ir. Thibault Verhoeven, as he always made time for questions and gave proper, enthousiastic guidance, giving me extra motivation to work on this master thesis. I would also like to thank my promotor prof. dr. ir. Joni Dambre for teaching me in the big and interesting world of Machine Learning and giving me the opportunity to apply these concepts on Brain-Computer Interfaces.
Last, but definitely not least, I want to thank my parents and family for the endless support and encouragement they gave me, making these past six years a truly great experience.

Permission for usage

Nina Proesmans
Ghent, June 1st, 2016

Brain-Computer Interfaces using Machine Learning: Reducing calibration time in Motor Imagery

Nina PROESMANS

Supervisors:
Prof. dr. ir. Joni DAMBRE
Dr. ir. Pieter VAN MIERLO

Counsellor:
Ir. Thibault VERHOEVEN

Master's dissertation submitted in order to obtain the academic degree of
MASTER OF SCIENCE IN BIOMEDICAL ENGINEERING
Faculty of Engineering and Architecture
Ghent University
Academic year 2015–2016

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik VAN DE WALLE

Abstract

Brain-Computer Interfaces (BCI's) are new ways for human beings to interact with a computer, by using only the brain. BCI's can be very useful for people who have lost the ability to control their limbs, as BCI's can give these people the opportunity to, for example, steer a wheelchair, using Motor Imagery. Motor Imagery is the process where the patient imagines a movement, resulting in a signal originating from the brain and measurable through EEG.

The biggest challenge for BCI's is that not everyone has the same brain. Using Machine Learning, for every new session, the BCI has to learn from the user's brain, but this learning takes time. The time that the BCI needs to adapt to the user's brain in order to correctly classify their thoughts, is known as the calibration time. Up until now, this calibration could take up to 20 - 30 minutes, which is an exhausting and tiring amount of time that the patient has to wait until the system is fully functional.

To solve this problem, the goal of this thesis was to reduce this calibration time as much as possible. In the first part of this work, a first attempt is done by finding the optimal amount of features needed for reasonable functioning of the BCI, using all calibration data available. Averaged over five subjects, the amount of correctly classified thoughts only reached 67± 15%.

To increase the performance of the BCI while reducing the calibration time, Transfer Learning was used. In Transfer Learning, information extracted from previously recorded subjects is used as good as possible to reduce the amount of calibration needed for classification of thoughts coming from a new target subject. Existing techniques were compared and a new technique was developed, resulting in the need for only 24 seconds of calibration data, classifying 86±8% of the thoughts correctly.

Keywords

Brain-Computer Interfaces, Machine Learning, Calibration time, Motor Imagery, Transfer Learning

Brain-Computer Interfaces using Machine Learning: Reducing calibration time in Motor Imagery

Nina Proesmans

Supervisor(s): Joni Dambre, Pieter van Mierlo, Thibault Verhoeven

Abstract—**The goal of this article is to find a method that is able to reduce the calibration time needed for Motor Imagery classification, without a loss of performance. For this purpose, Machine Learning is applied to the subject of Brain-Computer Interfaces. To illustrate the level of difficulty to find a good Machine Learning model that performs well for every subject, a person-specific BCI is optimised with data available from seven subjects. For the purpose of calibration time reduction, several Transfer Learning techniques are exploited and a new technique is proposed. The new Transfer Learning technique reduces calibration time while performing even better.**

Keywords—**Brain-Computer Interfaces, Machine Learning, calibration time, Motor Imagery, Transfer Learning**

I. Introduction

BY using Brain-Computer Interfaces, people who have lost the ability to control their limbs are given the opportunity to, for example, steer a wheelchair by using Motor Imagery. Motor Imagery is the process where the patient imagines a movement, resulting in a signal measurable from the brain, which is similar to the brain signals when actually planning and performing the movement [1]. In this article, imaginary left and right hand movement will be the topic of interest.

To learn from the user's brain, Machine Learning is applied to Brain-Computer Interfaces. General concepts of different techniques needed to build a Machine Learning algorithm are explained in Section II, a first BCI is built in Section III. Due to interperson and intersession differences, the BCI needs to adapt to the user's brain for every new session, to be able to correctly classify their thoughts. The time that the BCI needs for this adaption is known as the calibration time and up until now, this calibration could take up to 20 - 30 minutes. With the intention of using these BCI's with, for example, ALS patients, this is an exhausting and tiring amount of time that the patient has to wait until the system is fully functional. To overcome this problem, several existing Transfer Learning techniques are explained in Section IV. Using these techniques, previously recorded data can be reused or adapted to improve prediction of tasks performed by new subjects. Different Transfer Learning techniques are proposed in Section VI with their corresponding simulations in Section V. A new Transfer Learning technique is designed in Section VI with its results given in Section VII.

II. Machine Learning

Machine Learning is used to make data-driven predictions, based on properties of example inputs, known as the **training data** or **training set**. If an underlying model exists, containing the properties of the data, a Machine Learning algorithm can construct a model based on the available training data as close as possible to the underlying model. If the ML algorithm succeeded, it should be able to correctly predict class labels of new input samples, known as the **test set**. Applying Machine Learning to the subject of Brain-Computer Interfaces, a set-up as in Figure 1 is used.

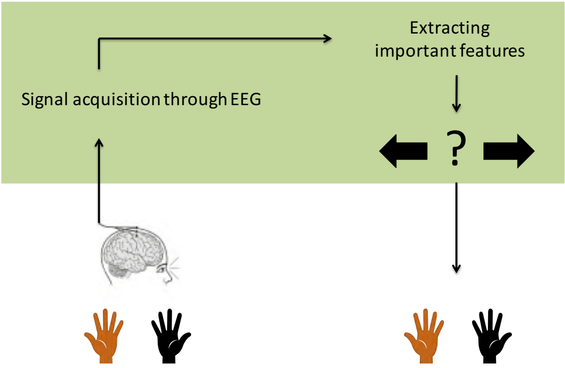


Fig. 1. Brain Computer Interface: Overview

Starting with different trials of imaginary movement of the left and right hand, the signals produced by the user are recorded using EEG. Before giving this data to an ML algorithm, the trials are pre-processed by filtering the data within a specific range. This range will determine which brain wave categories are included for further experiments. The main brain waves measured by EEG include delta waves (<4 Hz), theta waves (4-7 Hz), alpha waves (8-13 Hz) and beta waves (14-30 Hz) [2].

After pre-processing, the most useful features are extracted using **Common Spatial Patterns** [3]. By using Common Spatial Patterns, the original EEG-channels will be linearly transformed, making it easier to discriminate between two conditions. Feature selection is performed by selecting an amount of new CSP-channels, also called filter-pairs, and taking their log variance. These features are given to a classification model, **Linear Discriminant Analysis** [4], that will decide whether the original thought of each trial was a left or right hand movement. After classification this signal can be used to, for example, steer a wheelchair.
In this work, test accuracy will be the measure of classifier performance, calculated as the amount of correctly classified trials divided by the total amount of trials.

III. A first BCI

By applying the techniques as explained in Section II, insight is gained in the working principles of a Brain-Computer Interface. In the attempt of constructing a general BCI with optimal performance, hyperparameters are optimised. For a first BCI, the hyperparameters will be the amount of CSP-filter pairs used and the amount of splits in the frequencyband. By splitting the frequency band in equal parts, every part will have their own specific CSP-filters, increasing the amount of detail the ML algorithm can capture. By increasing the amount of CSP-filter pairs, more information will be available, but being less discriminative. For this experiment, data from 7 subjects from BCI Competition IV [5], is used. The same setup is as in Section II is used, filtering the data from 0 to 40 Hz with a 6^{th} order Butterworth filter, as this frequency range includes the alpha and beta band.

To determine the best set of hyperparameters, a cross-validation scheme is used. Per subject, the first 80% of the data is defined as the training set (here: 160 trials), the last 20% is used as the test set (here: 40 trials). To prevent data leakage while optimising the hyperparameters, the training set is split in 10 equal folds, with 9 folds serving as training set and the 10^{th} fold as the validation fold. The amount of splits in frequency band ranges from 0 to 9, the amount of CSP-filter pairs from 1 to 9. The best person-specific hyperparameters are obtained by using the combination that gives the highest validation accuracy. The corresponding test and train accuracies are given in Figure 2.

This figure shows that the average test accuracy is only 67±15%. The big difference in test accuracy between subjects illustrates the problem that, even with features optimised per subject and using all calibration data available, there is no guarantee that the classifier will perform well. The high train accuracies of 98±3% indicate the possible occurance of overfitting, even when using regularisation. The overall conclusion that can be drawn, is that the level of difficulty to construct a general BCI for every subject is high and other measures have to be taken, especially with the goal of reducing the calibration time without a decrease in performance.

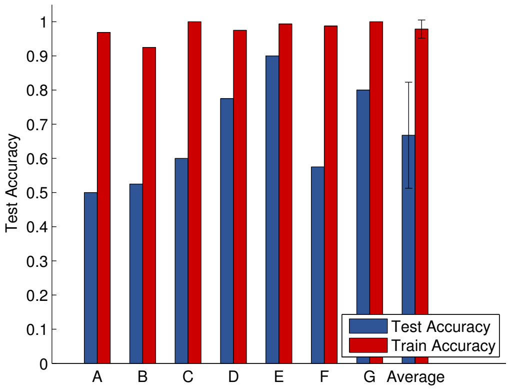


Fig. 2. The train and test accuracies reached when using person-specific optimal features.

IV. Transfer Learning

Using Transfer Learning, the reduction of training data due to a lower calibration time can be compensated by using training data from previously recorded subjects. By using more data, the classification performance can increase, however, due to the difference in the statistical distribution of data recorded from previous subjects, this data has to be transferred to the new subject in a way that it is used as efficient as possible.

For this purpose, several Transfer Learning techniques were investigated. Two naive Transfer Learning techniques, called **Majority Voting** and **Averaging Probabilities**, who respectively, classify according to a majority of votes of different classifiers, or average the probabilities of the predictions of different classifiers. As these techniques don't use data from the target subject, they are not suited for proper Transfer Learning. A third technique, called **Covariance Shrinkage** (CS), regularizes the CSP and LDA algorithms based on data from a selected optimal subset of source subjects [6].

The most interesting Transfer Learning technique for calibration time reduction is **Data Space Adaption** (DSA) [7]. The goal of DSA is to reduce the dissimilarities between the target subject and the k^{th} source subject, by adapting the target subject's data in such a way that their distribution difference is minimised. This first step of the DSA algorithm is called the *subject-to-subject adaptation*. Arvaneh et. al [7] assume that the difference between the source subject k and the target subject's data can be observed in the first two moments of the EEG-data and construct a transformation matrix accordingly.

For each source subject k the optimal linear transformation matrix M_k (see Formula 1)is built based on the covariance matrices Σ_1 and Σ_2 for each condition of the target subject and $\tilde{\Sigma}_{k,1}$ and $\tilde{\Sigma}_{k,2}$ for each condition of the source subject k. The covariance matrices are calculated using the sample covariances. † stands for taking the pseudo-inverse of the matrix.

$$M_k = \sqrt{2\left(\tilde{\Sigma}_{k,1}^{-1}\Sigma_1 + \tilde{\Sigma}_{k,2}^{-1}\Sigma_2\right)}^{\dagger} \qquad (1)$$

Using this transformation matrix, the target subject's data V is transformed to minimise the distribution difference with the k^{th} source subject according to

$$V_k^{transformed} = M_k V \qquad (2)$$

The second step of the algorithm is the *selection of the best calibration model*. After the transformation of the target subject's data, the distribution difference ought to be minimised, but one source subject may be more similar to the target subject's data than another one. Therefere, the most similar source subject has to be found.

This is done by first adapting the target subject's data according to the transformation matrix M_k as in Formula 1 and classifying the adapted data using the model trained on the corresponding source subject. The source subject that results in the highest validation accuracy is selected as the best calibration session.

If more than one source subject would result in the same classification accuracies, a selection is done based on the smallest KL-divergence [7] between the target subject's transformed data and a source subject.

The Data Space Adaption algorithm has been proven to substantially reduce calibration time, without the need for a large database of previously recorded sessions. Another advantage is that it can easily be implemented in online applications, as the calculation of the transformation matrix and the adaption of the new target's data can be done in less than a second[7].

V. Simulations

In this work, for comparison of the four Transfer Learning techniques, only 5 out of 7 subjects from the competition set are used, as these 5 subjects performed the same imaginary movement tasks being left and right hand movement. Averaging test accuracies of these 5 subjects, calculated on the last 40 trials of each target subject, the test accuracy is plotted against the amount of calibration trials used from the source subject in Figure 3, ranging from 10 to 160 training trials, using different Transfer Learning techniques. For this set-up, the data was filtered from 8 - 35 Hz, using a Butterworth filter of the 6^{th} order, as this range includes the most important brain waves categories for Motor Imagery classification[8]. 3 CSP-filter pairs are extracted as features and given to a regularised Linear Discriminant Analysis model for classification. No error bars are shown for clarity of the graph, as the standard deviation varies from 7 to 20%, with no relation to the amount of calibration trials.

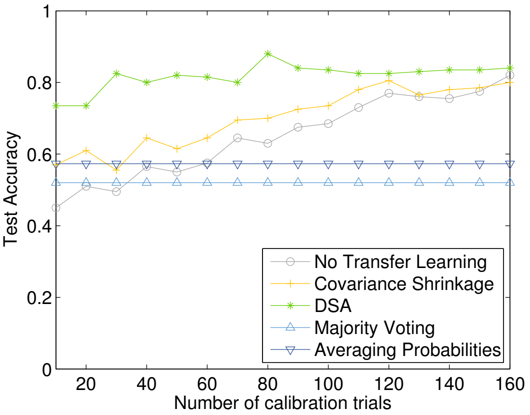


Fig. 3. Average

Figure 3 shows that, on average, Data Space Adaption will always result in the highest test accuracy, independent of the amount of calibration trials used from the target subject. As a baseline, the grey line indicates the test accuracy when no Transfer Learning is used. The naive Transfer Learning methods, obviously, result in the same test accuracies for every amount of calibration trials, as it does not use information from the target subject and is quickly outperformed by the standard approach without Transfer Learning. Covariance Shrinkage does, on average, give higher test accuracies than when using no Transfer Learning, but when looking subject-specific, in 3 out of 5 cases, the test accuracies were the same as without the usage of a Transfer Learning technique. As DSA can, on average, reach test accuracies of 74±14% using only 10 calibration trials, further investigation is done on how to improve this technique to be able to reach even higher test accuracies with a minimal amount of calibration data.

VI. Design of a new Transfer Learning technique

As explained in IV, the DSA algorithm consists of a step where the best calibration model is chosen. When leaving out this selection and looking at the test accuracies reached with consecutively every source subject serving as the calibration model, it was clear that the selection process did not always pick the best source subject for classification of the target subject's data. For some target subjects, this erroneous assignment of best calibration model results in sudden decreases in test accuracy if the addition of new calibration data has large influences on the covariance matrices needed in Formula 1.

These shortcomings in mind, a new technique is developed exploring three different paths:

DSA/CS - Accumulate source subject data
In this approach, data of a subset of source subjects is accumulated. The subset is chosen according to the Subject-Selection Algorithm (as used by Lotte and Guan [6]). Based on the accumulated data, the transformation matrix M (Formula 1) is calculated, the classifier is trained on the accumulated data and tested with the target subject's data.

DSA - Averaging probabilities (AP)
The same set-up as for DSA is used, except, instead of selecting the best calibration source subject, multiple calibration models are used, and for each trial, the class probabilities are averaged.

DSA - Maximum probability (MP)
The same approach as in **DSA - Averaging probabilities** is used, but instead of averaging the class probabilities, the highest probability produced by a classifier determines the class label.

Using these three methods, new experiments are performed using the same parameters as in Section V. As our goal is to reduce calibration time, the aim of the new method should be to give high test accuracies with a low amount of calibration data, hence Figure 4 only shows test accuracies higher than 0.5, for an amount of calibration data varying from 2 to 40 trials, averaged over 5 subjects.

From Figure 4, it is clear that, on average, DSA - MP gives the highest test accuracies for every amount of calibration data. DSA - AP is the second best method. DSA/CS doesn't even always give better results than the standard DSA approach, therefore, this technique will be left out for further research. When looking subject-specific, these conclusions are a little different, as DSA - MP does not always results in the highest test accuracies, but in 82.5% of the experiments, either DSA, DSA - AP or DSA - MP give the highest test accuracy. To guarantee that the highest test accuracy is reached, a new method is constructed that should result in an upper boundary of these three methods.

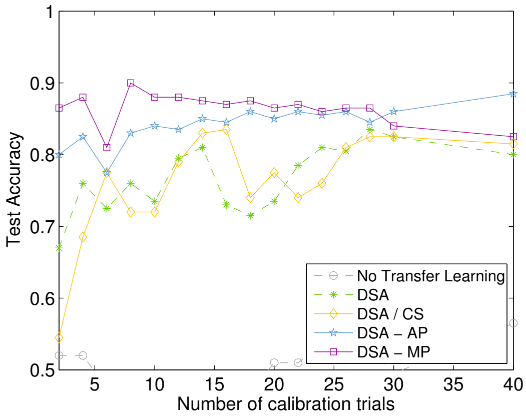


Fig. 4. Average

Upon further investigation of DSA - AP and DSA - MP on why they misclassify certain trials, it became clear that these methods are complementary. If the classifiers produce probabilities that indicate that the classifier is indecisive (probabilities close to 50%), DSA - MP should be used, to find the most certain classifier. If, on the other hand, all classifiers produce reliable probabilities, it is better to use DSA - AP, as a single classifier with a slightly higher probability (in the case of DSA - MP) could shift the class label in the other direction, even if the majority of classifiers would predict otherwise.

These findings are used to construct an algortihm for a new Transfer Learning technique. The algorithm will, dependent on whether a classifier is biased or if it's validation accuracy is low, remove the respective source subject from the further decision making process. To determine whether a classifier is biased, a method was constructed that can predict if the predictions of the corresponding classifier are consistenly the same (if 90% of the class labels are equal). In that case, the corresponding source subject is removed. If the validation accuracy, as the validation accuracy for standard DSA in Section IV, is lower than 70%, the source subject is also removed from the further decision making process.

VII. Results

The results of the final method are plotted in Figure 5. On average, the final method outperforms every other technique. When looking subject-specific and only at the experiments using 40 calibration trials or less, for subject B (see Figure 6), the final method doesn't always leads to the highest test accuracies, but at least it doesn't drop towards test accuracies of only

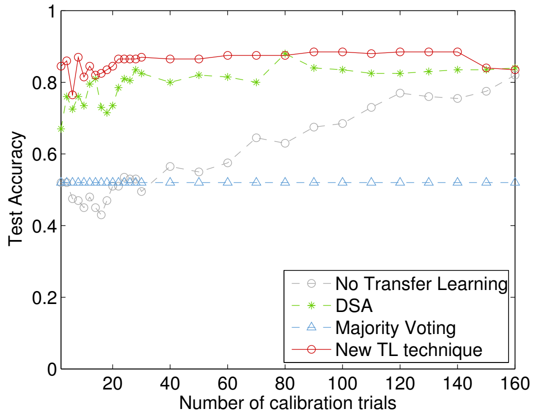


Fig. 5. Average

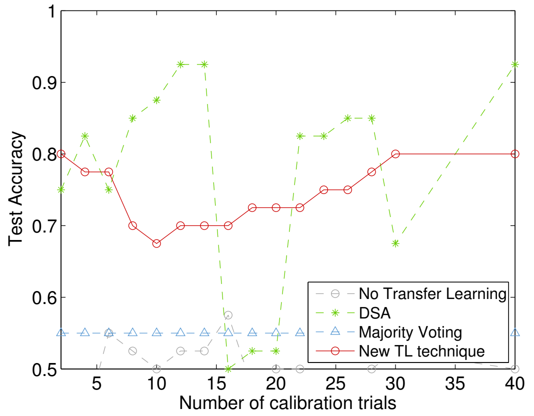


Fig. 6. Subject B

50% like DSA. For subject C (see Figure 7), the final method always performs best. For other subjects, not illustrated here, similar conclusions can be drawn.

Based on the findings for every subject, with a single exception for 6 calibration trials when testing for subject E, the test accuracy of the final Transfer Learning technique never drops below 67%. The most important gain in performance is, that when only having 2 calibration trials available, the test accuracy is minimally 70%. With respect to, when having 4 calibration trials available, only reaching 52±6% on average when not applying Transfer Learning, this is an improvement of at least 15% in the worst case scenario. On average, when using the final method and 4 calibration trials, the test accuracy is 86±8%. This clearly manifests that, by applying Transfer Learning and reducing the amount of calibration, there was absolutely no reduction in performance.

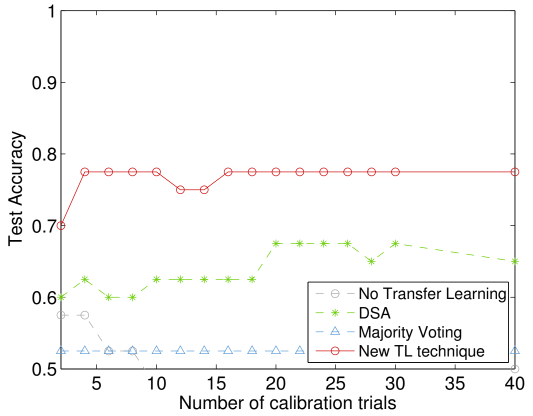


Fig. 7. Subject C

VIII. Conclusion and Future Work

In order to construct a Machine Learning model for Brain-Computer Interfaces that uses as little calibration data as possible, without a reduction of system performance, a first step was taken by constructing a person-specific BCI. Eventhough sufficient training data was available, for some subjects, a testing of 70% wasn't even reached [9]. The constructed BCI was very user-dependent and needed a lot of calibration data. By investigating some Transfer Learning techniques, as in Section IV, improvements were observed in the context of increasing test accuracies with less calibration time. With room for improvement being present, a new method was developed giving promising results. The strength of this final method is its robustness in comparison to DSA. Where the performance of DSA can suddenly decrease when new calibration data becomes available, the new method is less sensitive to alterations in calibration data. With regards to an application of a BCI to steer a wheelchair, this advantage of the final method is an important aspect in sending the wheelchair towards the right direction, even if the user was confused or distracted for a short period of time. The final method might not always be stated as the best method for every amount of calibration data, but by further alterations in the selection criterion, the results can be promising.

With regards to the goal of reducing the calibration time as much as possible, this requirement is fullfilled, as the amount of trials is reduced to the minimal amount of trials possible, still reaching average test accuracies of 85±10%, needing only 24 second for calibration.

In the construction of a new method, based on comparisons between old and new techniques, it might not be overlooked that an optimal method was constructed based on data from only 5 subjects. To work around this restriction, the steps taken in the process to construct a final method, were not based on the averages of the performance of these 5 subjects, but on target-specific results. As these results still might depend on the specific data used in the experiments, it might be useful to go over the same steps and reasonings, but with a larger or a different dataset. By expanding the dataset with different categories of imaginary movement, like foot movement or eye blinking, the generalisation properties of the Transfer Learning techniques could also be studied.

REFERENCES

[1] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, 1995.

[2] E Marieb and K Hoehn, *Human Anatomy & Physiology*, 2006.

[3] Herbert Ramoser, Johannes Müller-Gerking, and Gert Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[4] M Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., 2006.

[5] B Blankertz, "BCI Competition IV - Dataset 1," 2008.

[6] Fabien Lotte and Cuntai Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, no. 2, pp. 614–617, 2010.

[7] Mahnaz Arvaneh, Ian Robertson, and Tomas E Ward, "Subject-to-Subject Adaptation to Reduce Calibration Time in Motor Imagery-based Brain-Computer Interface," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 6501–6504.

[8] G Pfurtscheller and C Neuper, "Motor imagery and direct brain- computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[9] Andrea Kübler, Nicola Neumann, Barbara Wilhelm, Thilo Hinterberger, and Niels Birbaumer, "Predictability of Brain-Computer Communication," *Journal of Phychophysiology*, vol. 18, no. 2-3, pp. 121–129, 2004.

Contents

List of Figures

List of Tables

Acronyms

ALS Amyotrophic lateral sclerosis.

AUC Area Under Curve.

BCI Brain-Computer Interface.

CAR Common Average Referencing.

CSP Common Spatial Patterns.

DSA Data Space Adaption.

ECoG Electrocorticography.

ERD Event related desynchronisation.

ERP Event-Related Potential.

ERS Event related synchronisation.

fMRI Functional Magnetic Resonance Imaging.

HCI Human Computer Interface.

LDA Linear Discriminant Analysis.

LOOV Leave-One-Out-Validation.

MEG Magnetoencephalography.

ROC Receiver Operating Characteristic.

SVM Support Vector Machine.

Chapter 1

Introduction

1.1 Brain-Computer Interfaces

A Human Computer Interface (HCI) using a keyboard or mouse as an interface to communicate between human and computer is very common. Unfortunately, people unable to generate the necessary muscular movements cannot use these standard HCI's.

With growing recognition of the needs and potential of people with disabilities and new understanding of the brain function, Brain-Computer Interfaces (BCI's) needed to be developed. BCI's only use the brain as a way of communicating between the human brain and an external device, giving people a way to communicate or to control technology, without the need for motor control [1].

By doing so, BCI's can be a way to improve or recover the mobility of patients with severe motor disorders, e.g. amyotrophic lateral sclerosis (ALS) [2], brainstem stroke, cerebral palsy or spinal cord injury. A wheelchair can be controlled with Motor Imagery [3], a P300-speller allows word spelling but can also be used to control a house environment; opening doors, turning on lights [4], etc. The future may even hold options to bypass damaged sections of the spinal cord, allowing actual movement of the paralysed limbs with only the thought of movement [5].

However, the application of a BCI reaches further than only for injured people. Applications can be found in the gaming area or in surgery, as a surgeon may need more than muscles to control movements. And even while focussing on applying BCI's, new knowledge is gained about the functionality of the brain.

1.2 From brain to computer

To measure brain activity, the methods can be divided in three main categories: invasive, partially-invasive and non-invasive measurements.

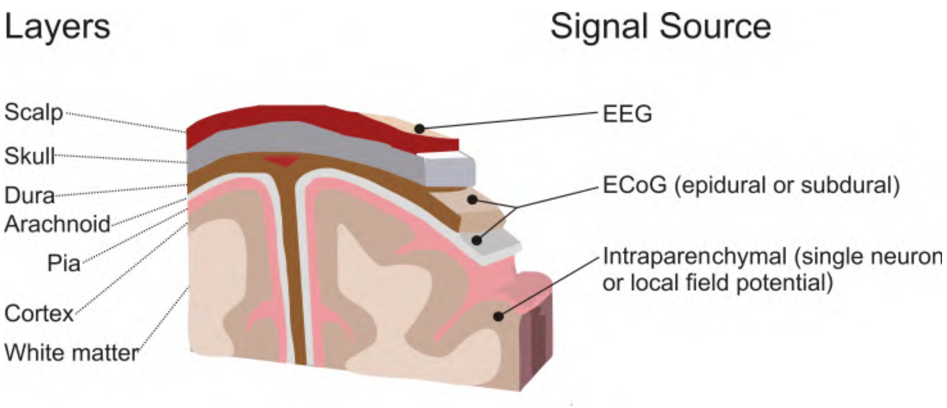

Layers
Signal Source
Scalp
Skull
Dura
Arachnoid
Pia
Cortex
White matter
EEG
ECoG (epidural or subdural)
Intraparenchymal (single neuron
or local field potential)

Figure 1.1: EEG as an example of non-invasive measurements, ECoG for partially-invasive measurements and Local Field Potentials as invasive measurements. The invasiveness depending on interaction with the respective layers of the brain (Source: `http://www.schalklab.org/research/brain-computer-interfacing`.

1.2.1 Invasive measurements

The most invasive way to record brain signals is by implanting electrode arrays into the patient's cortical tissue, recording extracellular potentials from nearby neurons. The recordings have high spatial resolution, but require tens or hundreds of small electrodes being implanted in the brain. These are prone to failure on biocompatibility level if brain tissue reacts with the implants and therefore not suitable for long-time performance stability [6].

1.2.2 Partially-invasive measurements

Electrocorticography (ECoG) is a less invasive technique that does require surgery, but electrodes are implanted subdurally on the surface of the brain, without the need for cortical penetration. The signals acquired by ECoG have a very high signal-to-noise ratio, are less susceptible to artifacts than EEG and have a high spatial and temporal resolution ($<$1 cm and $<$1 ms respectively) [7]. Therefore they are useful to reveal functional connectivity in the brain and resolve finer task-related spatial-temporal dynamics, giving new insights in our understanding of large scale cortical processes, which can improve communication and control.

The clinical risk is lowered and gives better long-term stability as the surgery is less invasive, but has the downside of having to record data within clinical settings, making it hard to obtain lots of data [8].

1.2.3 Non-invasive measurements

In this work, non-invasive BCI's will be used, measuring signals from outside of the skull. The big advantage includes not having to perform surgery, but has the disadvantage of signals being deformed and deflected by the bone tissue of the skull, creating noise and making it harder for a computer to interpret [1].

EEG (electroencephalograpy)

When the billions of neurons in our brain communicate, they do this by generating and propagating action potentials. These electrical changes induce dendritic current, creating an electric field that can be measured by the electrodes. In electroencephalography the oscillations of potential differences are measured outside the scalp using electrodes, representing the synchronous activity of these neurons [9].

When using wet electrodes, a conductive gel is applied for better transduction of charge between scalp and electrode. In comparison to dry electrodes, which don't use a gel, wet electrodes are less pleasant for the user due to the sticky products and the long application process (about 30 minutes), but they are relatively cheap and disposable making them more readily available in clinical settings [10].

To be able to better compare recordings for different persons, the electrodes are placed on the head at fixed locations according to the international 10-20 system, based on standard landmarks of the skull (see Figure 1.2). These marks are labeled according to the different areas with Fp, F, C, P, T and O representing the fronto polar, frontal, central, parietal, temporal and occipital areas, respectively [11].

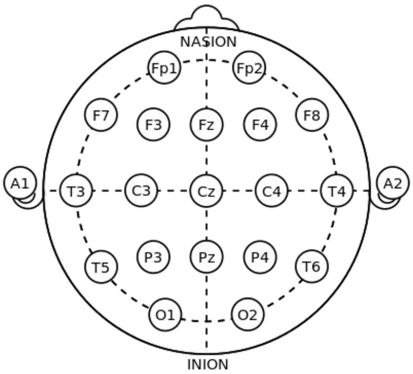


Figure 1.2: Position of the electrodes according to the international 10-20 system (Source: `http://www.nrsign.com/eeg-10-20-system/`).

The patterns recorded by the electrodes are called brain waves. These brain waves are as unique as our fingerprints, but change with age, sensory stimuli, brain disease and the chemical state of the body. The brain waves shown in an electroencephalogram fall into four general categories, based upon their frequency content [12] (see Figure 1.3).

- Delta waves (<4 Hz): high amplitude waves seen during deep sleep.
- Theta waves (4-7 Hz): more common in children and sometimes seen in adults when concentrating.
- Alpha waves (8-13 Hz): relatively regular, rhythmic, low-amplitude waves when in a relaxed state.
- Beta waves (14-30 Hz): less regular than alpha waves and occur when mentally alert, focussing on a problem or visual stimulus.

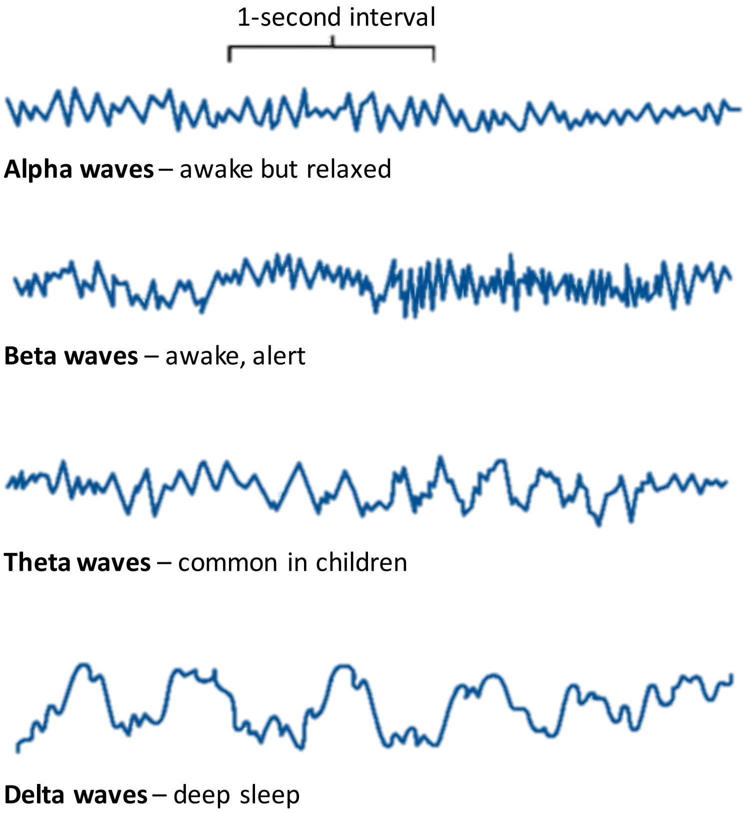


Figure 1.3: 4 main categories of brain waves [12].

Because of the distance between the electrodes and the origin in the brain, the measured signal is the result of the activity of thousands of neurons, making it hard to distinguish exactly where the activity came from, resulting in poor spatial resolution. The measurement of signals not originating from the brain, called artifacts, is another drawback. These may arise due to power line noise (50 or 60 Hz), or due to biological reasons, such as eye blinking, limb movement, chewing or heartbeats. But as the temporal resolution is high, being in the millisecond range, it is the best modality for real-time applications and will be used for this master thesis.

MEG (magnetoencephalography)

MEG works according to the same principle as EEG, where electric fields arise due to the induced dendrite currents from the synchronous firing of the neurons, but measures magnetic fields instead of electric fields. Whereas EEG uses electrodes placed on the scalp, MEG uses sensor coils that do not touch the patient's head [13].

Regardless of the higher spatiotemporal resolution that MEG achieves, the popularity of this method is rather low due to some technical issues. As shielding from the relatively stronger earth's magnetic fields is necessary, the measurements must be done in a shielding room, making it unattractive for daily, mobile use [14].

fMRI (functional magnetic resonance imaging)

In contrast to EEG and MEG, fMRI measures the blood oxygen level-dependent (BOLD) signal with a high spatial resolution, covering the whole brain. BOLD imaging uses hemoglobin as a contrast agent, whereas oxygenated Hg and deoxygenated Hg have different magnetic properties. [15] As this is no direct measurement of neuronal activity, there is a physiological delay of 3 to 6 seconds before signal changes are observed. Data processing introduces a second delay of 1.3 seconds [16], but with increased availabilty of high-field MRI scanners and fast data acquisition scanners, this delay might be further reduced in the future [17]. However, the delay due to the hemodynamic response will remain constant, even with faster measuring and calculation techniques, making fMRI not suitable for fast real-time applications [18].

1.3 The brain signal as information carrier

Depending on the task we want the BCI to fullfill, different types of brain signals can be used. It is important that the signal can easily be identified and is easy to control by the user. The two main groups that are most frequently used in EEG recordings can be distinguished as evoked signals and spontaneous signals, used for the P300-speller and in Motor Imagery respectively.

1.3.1 P300-speller

The P300-speller is used for spelling words or sentences by flashing rows and columns on a screen [19]. The user is asked to react to the target stimulus, which induces an amplitude increase in the measured signal called an Event-Related Potential (ERP) . The P300-wave is a type of ERP waveform, shown in Figure 1.4, occuring as a positive deflection after a latency period of approximately 300 ms [20].

ERP's fall under the category of evoked signals, as they occur by sensory, tactile or cognitive stimuli. Independent on the type of stimulus, the P300-wave is measured best at the level of the parietal lobe.

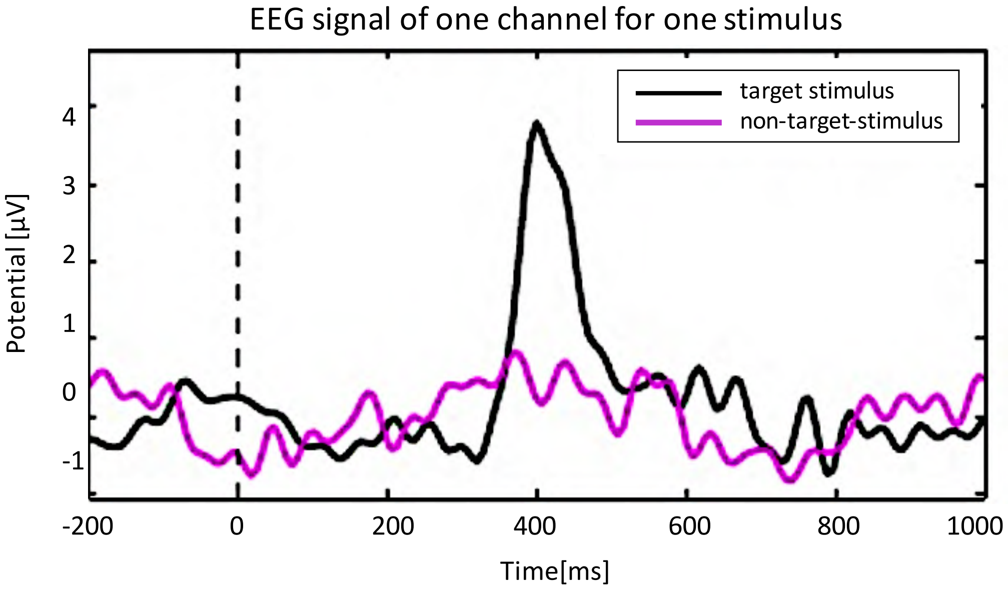


Figure 1.4: The P300 wave starts to peak 300 ms after the target stimulus (Source: `http://www.extremetech.com/extreme/134682-hackers-backdoor-the-human-brain-successfully-extract-sensitive-data`).

1.3.2 Motor Imagery

In this master thesis, imaginary movements will have to cause the change in EEG signals. These signals will be classified as spontaneous signals, as they are not evoked by an external stimulus. As it is broadly accepted that mental imagination of movements creates similar brain signals in the same regions as when preparing or performing the movement, [21] Motor Imagery can be seen as mental rehearsal of a movement, without the respective motor output.

When preparing and planning the movement, this leads to amplitude suppresion, called event-related desynchronisation (ERD) , followed by an amplitude enhancement, called event-related synchronisation (ERS). In the alpha band (mu band), the desynchronisation starts 2.5 seconds before movement-onset, peaks after movement-onset and recovers back to baseline within a few seconds. In the beta band, the desynchronisation is only short-lasting, immediatly followed by synchronisation reaching a maximum in the first second after the movement. In the gamma band, synchronisation reaches a maximum right before movement-onset, but these gamma oscillations are rarely found in a human EEG [22]. In Figure 1.5 the time course for ERD and ERS can be seen for the three different frequency bands, the vertical line indicating the offset of movement.

The most prominent EEG changes will be localised over the corresponding primary sensorimotor cortex contralateral to the movement as indicated in Figure 1.6 [20]. E.g. When executing/imaging left-hand or right-hand movement, the ERD and ERS will be seen over the contralateral hand area, showing different time courses over the alpha and beta bands. When performing classification in Motor Imagery, it comes down to finding the place and the frequency band where the movement is expressed best.

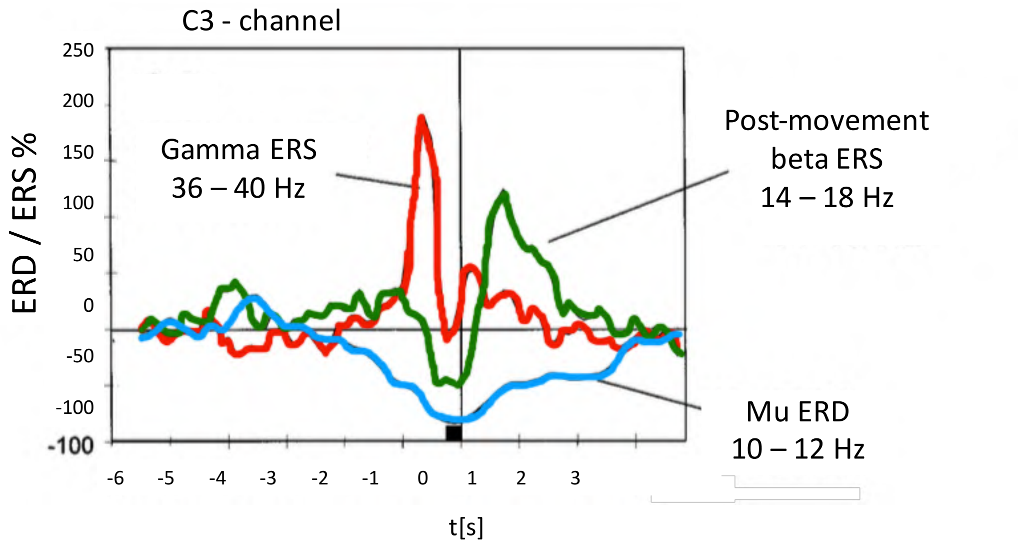


Figure 1.5: ERD and ERS in the alpha, beta and gamma band measured by the C3 electrode during right finger lifting (Source: `http://www.bbci.de/supplementary/conditionalERD/`).

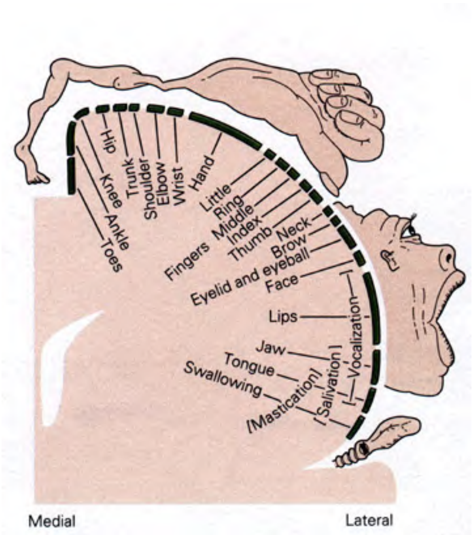


Figure 1.6: Sensorimotor cortex showing the origin of different movements (Source: `http://keck.ucsf.edu/~sabes/SensorimotorCortex/M1.htm`).

1.4 Difficulties

Synchronicity Synchronous BCI's work in a cue-paced mode, meaning that the time intervals in which communication is possible, is paced by the BCI. The EEG-signal can be analysed in predefined time windows, but this severely limits the autonomy of the user, allowing only one thought per time window. Asynchronous BCI's on the other hand, allow the user to communicate whenever they want, making it self-paced and much more flexible. This freedom of communication leans more towards reality, but as it expects continuous analysing, classification will become a more difficult task [23].

Inter-subject variability As not everyone has the exact same brain, or has the same capability to steer their thoughts, BCI performance depends strongly on the user. Algorithms [24] are being developed where the BCI can automatically identify its current user and adapt the classification parameters to maximise BCI performance, making it easier to initialise the BCI, without the need for manual setup.

Intersession differences Between different sessions, variations may also occur, due to fatigue, medication, sickness, hormones, but also due to a slightly different placing of the cap. Current development of data space adaption techniques [25] to minimise the intersession differences, may be a solution.

1.5 Goals

Using this data, it will become clear that due to interperson and intersession differences, it is nearly impossible to find a good Machine Learning model that performs well for every subject. Therefore calibration has to be done at the beginning of every new session. Unfortunately, this calibration is tiring and demotivating for patients, especially for our target audience, being for example ALS patients.

As there is always a trade-off between calibration time and performance of the system, my goal is to reduce this calibration time as much as possible without losing performance, by applying Transfer Learning techniques [26]. Using these techniques, previously recorded data can be reused or adapted to improve prediction of tasks performed by new subjects, hopefully resulting in a reduction of calibration time and ultimately in an unsupervised Motor Imagery BCI.

1.6 Overview

In Chapter 2, general concepts of different techniques needed to build a Machine Learning algorithm are explained. Using these concepts, the framework of Transfer Learning is described, introducing methods as found in literature.
With this knowledge, in Chapter 3, a first person-specific BCI is built and optimised. Second, simulations are performed using a BCI with Transfer Learning, constructed based on the

Transfer Learning principles as explained in Chapter 2.7. These simulations will show the need for a new Transfer Learning method that reduces the calibration time as much as possible, without a decrease in performance. Therefore, a new Transfer Learning technique is proposed in Chapter 4. The results obtained with the new technique are illustrated in Chapter 5.

Chapter 2

Methods

2.1 Machine Learning

According to Arthur Samuel in 1959, Machine Learning is “The field of study that gives computers the ability to learn without being explicitly programmed”.

Machine Learning is indeed used when static instructions, like handcrafted rules and heuristics, are unfeasible to design and program. A Machine Learning algorithm is used to make data-driven predictions based on properties of example inputs, known as the **training data** or **training set**. Take a look at for example face recognition. It is nearly impossible to make rules for the recognition of a face, as there will be an overload of rules and exceptions to these rules. Therefore, if an underlying model exists, containing the properties of the data, Machine Learning is the preferred technique that will try to construct a model based on the available training data, as close as possible to the underlying model.

The general framework for Machine Learning is stated as follows (see Figure 2.1) [27]:
A training set contains N labeled samples (x_1, x_2,... x_N), making it a supervised learning problem, as the class labels for all samples are known and are kept in a target vector (t_1, t_2,... t_N). The training samples x together with their target vector t are used to tune the parameters of the model, constructing an output function $y(x)$. If the hypothesis for the model performs well, $y(x)$ should be able to predict the correct class labels of new input samples x, known as the **test set**.

The ability of a model to correctly classify new samples that are different from the training set is known as **generalisation**. Generalisation is an important aspect of Machine Learning, as it is unlikely to encounter all possible inputs during training.

A second important aspect in Machine Learning is **pre-processing** the data. Pre-processing, also referred to as **feature extraction**, is used for two reasons:
The first reason is to hopefully be able to solve an easier problem in a new (lower-dimensional) space, for example, using averages of image intensities in subregions of an image as features has been proven to work well in face detection. [28]

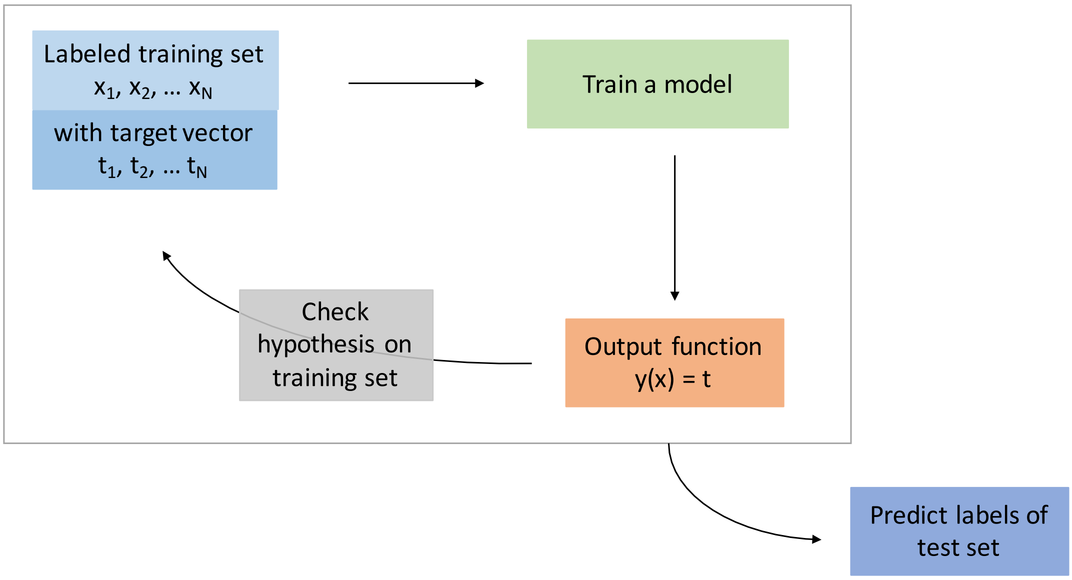


Figure 2.1: The general framework for a Machine Learning algorithm.

The second reason is to speed up computation time. For example, if face-detection has to be applied on video data using only the pixel values as inputs, the input data that is given to a complex ML algorithm is high-dimensional, resulting in long computation times. If instead features are found that are fast to compute without losing discriminative information, this process is better suited for real-time face detection.

In the case of supervised classification, with the importance of regularisation and the process of feature extraction in mind, a Machine Learning algorithm should be able to predict reproducable, reliable class outcomes. Because a Machine Learning algorithm is built on previously learnt relations or trends in the data from the training set and can have the capability to uncover hidden insights in the data, it can find regularities that are used to classify new data.

2.1.1 Pitfalls

Overfitting

The first major challenge in Machine Learning that has to be tackled is overfitting. Stated simply, if the hypothesis of our model is too complex in comparison to the amount of data available, our model will overfit. In terms of classification accuracy, overfitting occurs when with increasing complexity, the training error decreases, but the testing error increases.

In the case of polynomial fitting, it is shown in Figure 2.2 that with increasing order M of the polynomial, the curve is better fitted to the training data, but will perform poorly on new data. This can be interpreted as the curve becoming tuned to random noise that may be present in the training set, instead of being able to fit new data.

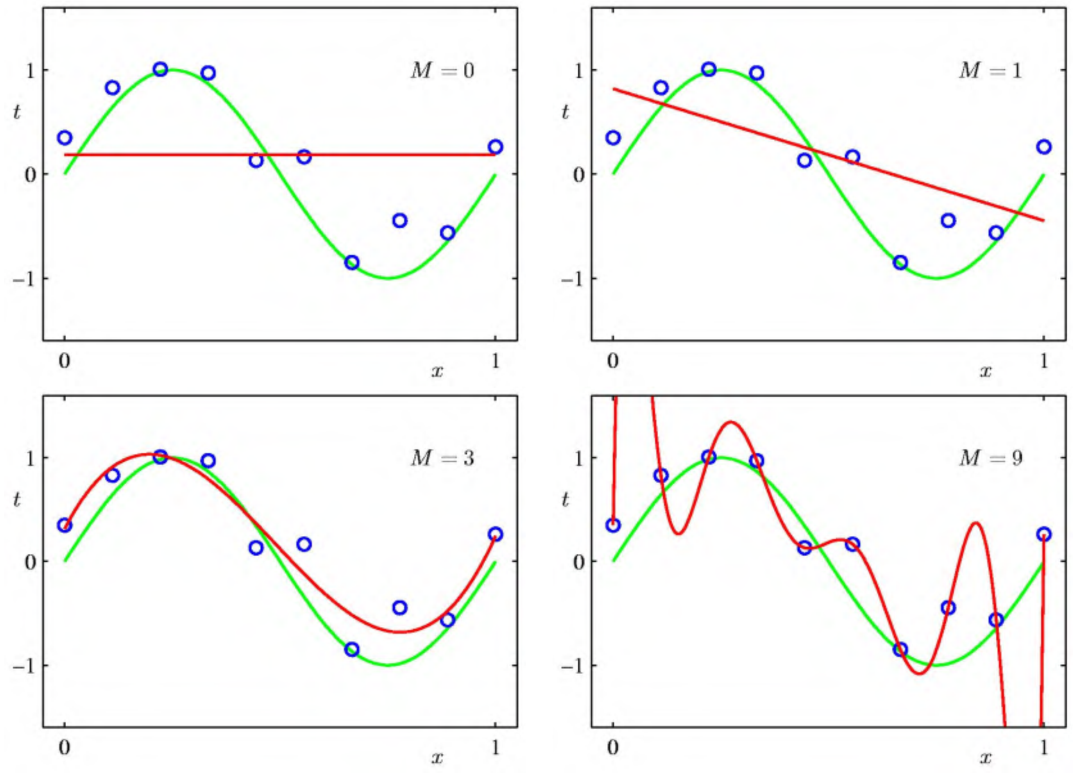


Figure 2.2: Overfitting example: The green line shows the underlying polynomial curve, the red line is the polynomial curve with order M used by the model to predict new data. With increasing order of the polynomial, the training data is fitted better, but if M becomes greater than 3, the model starts to overfit and the test error will increase [29].

A straightforward way to solve the problem of overfitting is by using more data, as this allows the model to be more complex, but with training data often being sparse, other solutions are needed.

Two other good ways to solve overfitting is by reducing the number of features by feature selection and by performing generalisation. In the example of polynomial curve fitting, this would mean reduction of the order of the curve. Using generalisation, large weights are penalised to prevent the coefficients of the model reaching too large values and becoming too complex.

Curse of dimensionality

The curse of dimensionality, as named by R. Bellman [30], is most easily explained using an example of a classification problem.

An example of a classification problem in 2D-space is illustrated in Figure 2.3. The output of the classification problem should be whether it will rain the next day or not. The axis x_1 and

x_2 can be seen as two features, for example being the humidity of the air and the pressure. The data-space is divided in nine equal subcells and if the classification occurs based on the majority of labels present in a subcell, the label of the question mark would be a green cross.

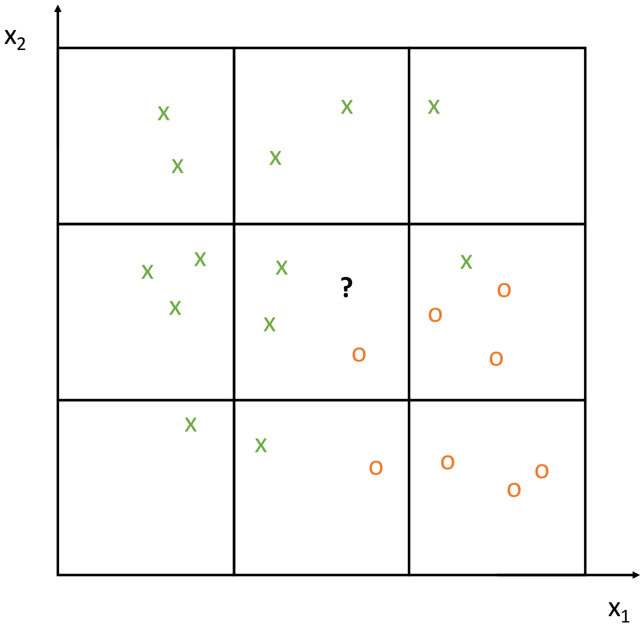


Figure 2.3: Example of a 2D classification problem using majority voting.

In the 2D-space example of Figure 2.3, the data-space is divided in nine equal subcells. When expanding to three-dimensional space, the data would be divided in nine times nine equal parts, showing that the number of cells grows exponentially with increasing dimension of the space (see Figure 2.4). Giving a third feature to the classification problem, means that the training data should also grow to avoid having empty cells. As examples are needed in every subcell to be able to make a decision for new incoming data points, this is known as the curse of dimensionality, as with every dimension that is added, the training data has to grow exponentially.

To anticipate the curse of dimensionality, a rule of thumb for simple classification algorithms such as Fisher's linear discriminant, is that if the number of training samples is at least ten times higher than the number of features, the classification algorithm will still perform well [31]. If not, dimensionality reduction has to be applied to reduce the number of features, by feature extraction or feature selection.

2.1.2 Cross-Validation

If parameters of a model would be learnt from the same data as it would be tested on, the model would always have maximal accuracy. For this reason, when using this model on new unseen data, the performance would be much lower.

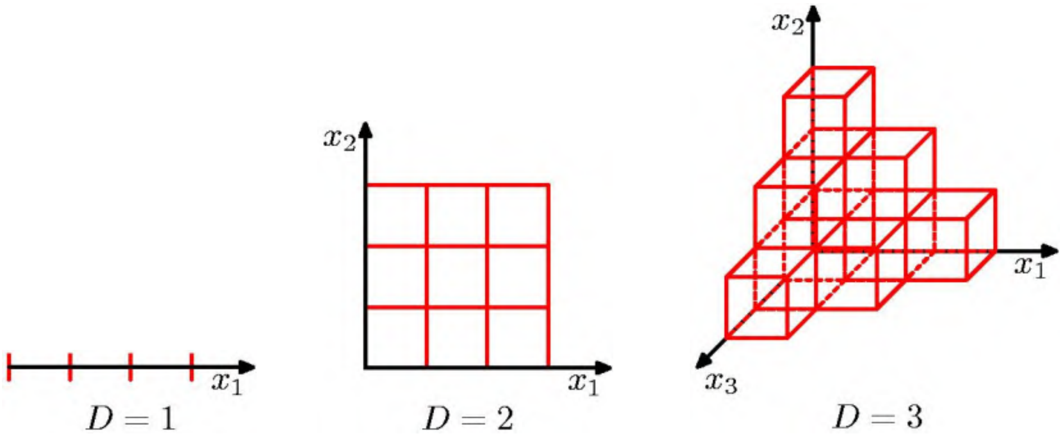


Figure 2.4: Illustration of the curse of dimensionality. With D going from one to three dimensions, the number of cells grows, inducing the need to have more data points to have at least one data point in each cell [29].

This is the case if, like in Figure 2.2 for M = 9, the model is built based on the blue dots, and also tested on these blue dots, resulting in 100% accuracy. This situation is known as overfitting. Therefore it is common practice to split the data in a training and a testing set. The model and its parameters are built based on the training set, holding the test set aside until the model is finished.

When optimising the parameters of the model, such as the amount of CSP-filters used (explained in Chapter 2.5), the risk of overfitting on the test set is still present. Due to the parameters of the model being tuned until optimal performance is reached, information from the test set can leak into the model, referred to as data leakage, and the generalisation properties of the model are jeopardized. Therefore, another part of the dataset, called the **validation set**, is held out. Training will still be performed on the training set, but evaluation is done on the validation set until satisfying results are obtained. The final model is applied on the test set.

As the set is now divided in three parts, a training set, a validation set and a test set, the amount of data available for training is again reduced and parameters can depend on a particular choice of training and validation set. This is the point where Cross-Validation comes forward.

k-fold Cross-Validation splits the training set in k folds, whereas $k - 1$ folds are used for training and the k^{th} fold is the validation fold. This process is repeated k times, resulting in k validation accuracies, which are averaged to give an indication on how well the model performs. Optimising parameters of a model using k-fold Cross Validation is computationally more expensive, but uses all data to make a reliable model. The process of k-fold Cross-Validation using ten folds is illustrated in Figure 2.5.

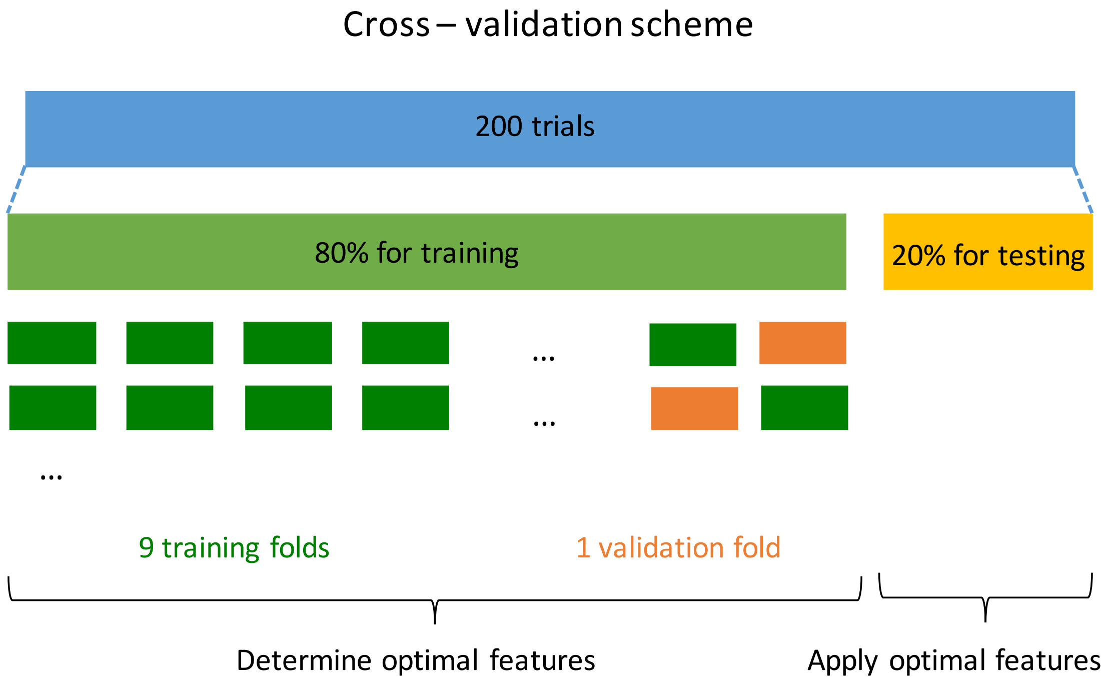


Figure 2.5: A cross-validation scheme showing the process of training a classifier using different subsets of the training data.

An important remark in the partitioning of the dataset, is the ensurance of a balanced dataset. This means that in every set or fold, there should always be as much trials from class one as there are from class two. If the sets are be unbalanced, having mostly trials from one class, it is possible that the model overfits to the class most present and doesn't learn features from the nearly absent class.

2.1.3 Machine Learning for Brain-Computer Interfaces

Applying Machine Learning to the subject of Brain-Computer Interfaces, a set-up as in Figure 2.6 is used.

Starting with different trials of imaginary movement of the left and right hand, the signals produced by the user are recorded using EEG. Before giving this data to an ML algorithm, the trials are **pre-processed** using a filter (see Chapter 2.3) and the most useful **features** are extracted (see Chapter 2.4 and 2.5). After pre-processing, the features are given to a **classification model** (see Chapter 2.6) that will decide whether the original thought of each trial was a left or right hand movement. Using this output from the classifier, it can be applied for e.g. steering a wheelchair.

[1] Sources: `www.wadsworth.org/educate/wolpaw.htm;www.planningdemocracy.org.uk`; Motor Imagery and Direct Brain-Computer Communication, G. Pfurtscheller et al.; Optimizing Spatial Filters for Robust EEG Single-Trial Analysis, B. Blankertz et al.; `www.spinlife.com/images/product/19949.jpg`

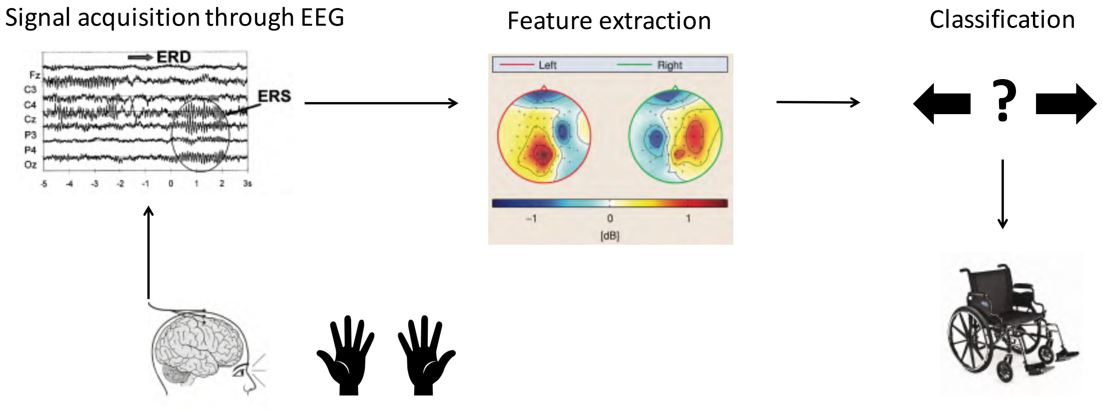


Figure 2.6: Brain-Computer Interface: Overview[1].

2.2 Data

For the research in this thesis, dataset 1 from BCI Competition IV [32] is used. For the construction of this dataset, healthy subjects were asked to perform cued Motor Imagery without feedback. Each subject could select two tasks from three classes being left hand, right hand and foot.

7 subjects were recorded, each instructed to perform a Motor Imagery task while a visual cue of 4 seconds was displayed, repeating this process 200 times. Those 200 iterations or trials were interleaved with 2 seconds of blank screen and 2 seconds of a fixation cross in the centre of the screen. With a sample frequency of 100 Hz, 4 seconds of Motor Imagery give us 400 useful samples, each containing information from 59 EEG channels.

2.3 Signal acquisition and preprocessing

To measure the signals produced by the brain, EEG is used. With the data recorded using 59 electrodes, the electrodes are placed conforming the 10-20 system, adding extra electrodes according to a 10% division [33] to fill in intermediate gaps between the existing fixed locations by the 10-20 system.

As explained in section 1.2.3, brain waves can be divided in four different categories based on their frequency range. Gamma waves are added as the fifth category, covering the range from 25 - 100 Hz as they have proven to play a role in all sensory modalities [34].

To remove unwanted artifacts and extract the most important information from the EEG-measurements, the data is pre-processed. By filtering the data within a range of 0 to 40 Hz, the frequency spectrum includes the alpha and beta band. For this purpose, a Butterworth-filter of the 6th order will bandpass-filter the signal between the desired frequencies. A

Butterworth-filter is commonly used, as it is maximally flat in the passband and rolls off towards zero without ripple.

2.4 Spatial Filtering

The importance for spatial filtering arises due to the poor spatial resolution of EEG measurements. As mentioned before, this poor spatial resolution is the result of a signal caused by the activity of thousands of neurons. A simulation using a volume conductor model of the head, showed that only as little as 5% of the measured signal comes from sources directly under a 1 cm diameter of the respective electrode. 50% comes from within a 3 cm diameter and 95% from within a 6 cm diameter [35]. This confirms that it is hard to distinguish exactly where the activity came from.

For left and right hand movement it is known that the main signal will be above the contralateral corresponding primary sensorimotor cortex, but with possible effects of artifacts or noise, the task still remains difficult [36].

In spatial filtering, signals from multiple electrodes are linearly combined, which makes it easier to locate the source origin, as the increase in signal-to-noise ratio results in being able to extract more discriminative information from the EEG signals.

Three different spatial filtering techniques will be discussed: Common Average Referencing, Laplace Filtering and Common Spatial Patterns. The first two methods are based on channel re-referencing and are explained shortly. The last method makes use of class information and is the method used in this thesis and in many other researches.

2.4.1 Common Average Referencing (CAR)

In Common Average Referencing, the average value of all EEG-channels is subtracted from the electrode of interest [37].

$$V_i^{CAR} = V_i - \frac{\sum_{j=1}^{N} V_j}{N} \tag{2.1}$$

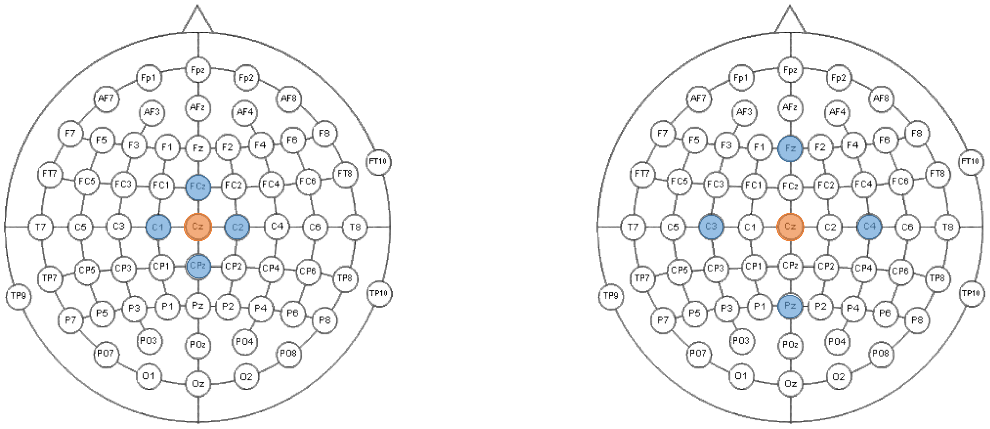
V_i is the potential difference measured at electrode i, with N being the total number of EEG channels.

By taking the average, this method can reduce the impact of signals that are present in a lot of channels and will highlight local signals. On the other hand, if not all channels contain this signal, ghost potentials may arise in the channels that don't.

2.4.2 Laplace Filtering

Using Laplace Filtering, the average of only the neighbouring electrodes is subtracted, instead of the average of all electrodes. In this way, the noise is only reduced in a region of interest.

There are two types of Laplace filters: the small Laplacian and the large Laplacian. The small Laplacian subtracts the average of the nearest four neighbours, whereas the large Laplacian subtracts the average of the four next-nearest neighbours [38]. As illustrated in Figure 2.7a & 2.7b, the orange marked electrode is the one being re-referenced and the blue marked electrodes indicate the respective neighbours.



(a) Small Laplacian.

(b) Large Laplacian.

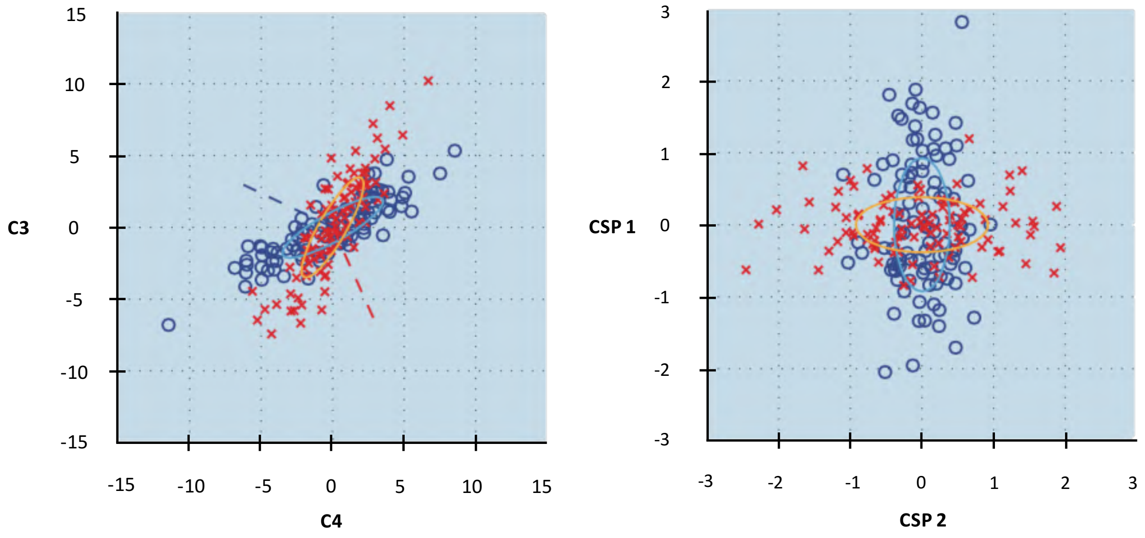
Figure 2.7: Laplacian filtering example (Source: `www.fieldtriptoolbox.org`).

2.4.3 Common Spatial Patterns (CSP)

As explained before, ERD and ERS are the signals that indicate Motor Imagery activity. By observing these simultaneously attenuated and enhanced EEG rhythms, classification of different types of brain states can be done. To this extent, Common Spatial Patterns will be used, as it is a technique that was already of common use in statistical pattern recognition and has proven its efficiency in finding spatial structures of ERD and ERS in BCI settings over the last 15 years [39] [40].

The problem tackled by CSP is illustrated in Figure 2.8. Figure 2.8a shows an example of some original EEG data of two different trials over time, as measured from channel C3 and C4. These are the two channels that measure signals originating from the main location for left and right hand movement. As the origin of the signal is determined as the channel that shows the highest variance over time, the assignment to the axis that shows the highest variance should be straightforward as these trials are measured by the two electrodes that are the main source of two different Motor Imagery signals. Nevertheless, it is not clear whether the variance for the trial as indicated by the red crosses shows highest variance on the C3 or C4 axis. The same problem arises for the trial as plotted in blue circles.

CSP will solve this problem by linearly transforming the data from the original EEG-channels into new channels, resulting in the transformed trials as in Figure 2.8b. Using the new CSP-channels, CSP1 and CSP2, makes it easier to discriminate between two conditions.
The general framework of CSP is built on maximising the variance of one condition while minimising it for the other. In Figure 2.8b it is indeed indisputable that the red crosses show highest variance on the CSP2 axis and the blue circles on the CSP1 axis. The linear transformation of the original EEG data is also known as CSP-filtering the data. The manner in which the linear transformation is constructed and executed, is explained in the next section.



(a) Before CSP-filtering

(b) After CSP-filtering

Figure 2.8: (a) With the original EEG-data of two different trials it is hard to discriminate in which direction the variance is highest. (b) After CSP-filtering the data, it is much more clear that the red crosses show the highest variance on the axis of CSP2 and the blue dots on axis CSP1 [36].

Calculation of the CSP-filters

A CSP-filter is calculated based on the potential differences V_j, with a dimensionality of N_{ch} x T_j, with N_{ch} the number of EEG channels and T_j the number of samples for that trial. Each trial V_j is labeled, implicating that the use of CSP-filters is a supervised technique.

To CSP-filter the signal, a projection matrix is calculated to transform the original EEG data (see Figure 2.9).

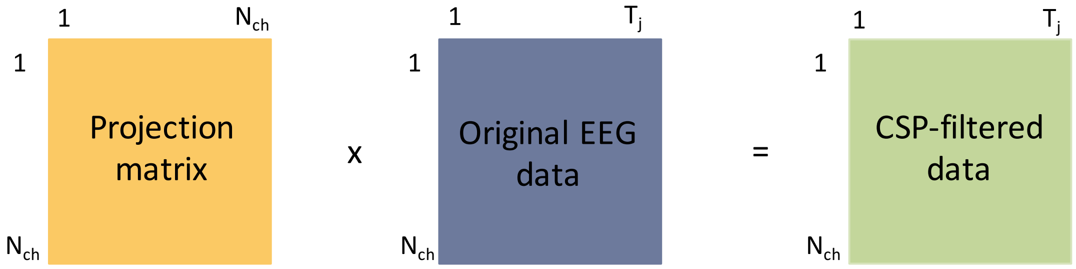


Figure 2.9: The transformation of the original EEG data with CSP-filters is a linear transformation. N_{ch} is the number of EEG channels and T_j the number of samples for that trial.

The projection matrix W will have as much filters as there are channels and the columns of the matrix will carry the weights to make linear combinations of the original EEG channels, thereby deciding which EEG-channels carry the most information.

The first half of the projection matrix will maximise the variance for class one and minimise it for class two, while the second half of the projection matrix will maximise the variance for class two and minimise it for class one.

Under the assumption that the signal is band-pass filtered, the projection matrix is constructed as follows:

Starting with the potential differences V_j from trial j, the covariance matrices are calculated for both classes with C_1 holding the left hand trials and C_2 holding the right hand trials:

$$\Sigma_1 = \sum_{j \in C^1} \frac{V_j V_j^T}{trace(V_j V_j^T)} \tag{2.2}$$

$$\Sigma_2 = \sum_{j \in C^2} \frac{V_j V_j^T}{trace(V_j V_j^T)} \tag{2.3}$$

The overall covariance matrix is composed as $\Sigma = \Sigma_1 + \Sigma_2$.

This covariance matrix is diagonalised and the eigenvalues and eigenvectors can be found in E and M respectively.

$$M^T \Sigma M = E \tag{2.4}$$

To make sure that $U \Sigma U^T = I$, the whitening transformation [40] is performed as follows:

$$U = P^{-\frac{1}{2}} M^T \tag{2.5}$$

Next, R_1 is calculated and diagonalised, with D and Z containing the eigenvalues and eigenvectors respectively.

$$R_1 = U \Sigma_1 U^T \tag{2.6}$$

$$Z^T R_1 Z = D \tag{2.7}$$

It is important that the eigenvalues on the diagonal of D are sorted in ascending order, ≥ 0 and ≤ 1.
With Z sorted according to D, the filters W can be calculated as

$$W = Z^T U \tag{2.8}$$

The original EEG-data can be transformed to $V^{CSP} = WV^{Original}$, whereas each row of V^{CSP} can be seen as a new CSP-channel. Due to the sorting of D, the first filter-pair, as shown in Figure 2.10, contains the most discriminative information. The second filter-pair will contain less discriminative information. When looking at the variances of each channel, the first row of a CSP-filtered signal from class one will have the highest variance and the last row the lowest, whereas if the same CSP-filter would be applied on a class two signal, the first row of this CSP-filtered signal would show a low variance and the last row the highest variance. This results in the biggest difference in variance being between the first and the last channel of the CSP-filtered signal. By extracting more filter pairs or CSP-channels, more information is available, but with increasing amount of channels, this information becomes less discriminative as illustrated in Figure 2.10.
If the projection matrix W, as mentioned before, was able to properly maximise the variance for one class, while minimising it for the other, it will be much easier to correctly classify a trial.

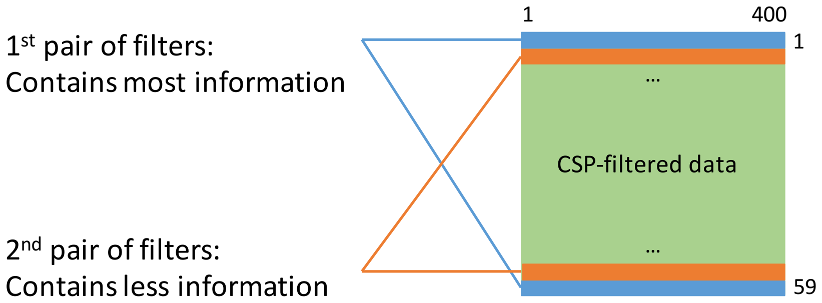


Figure 2.10: The first filter-pair extracts the most discriminative information, the second filter-pair extracts less discriminative information. 400 is the number of samples in a trial, 59 is the amount of EEG-channels used for measurements.

Limitations

The first limitation of using CSP-filters for feature extraction, is that the classification problem is limited to a two-class problem. There are extensions to multi-class CSP approaches [41], but this is beyond the scope of this thesis.

A second limitation of CSP-filters, is that CSP-filters are not based on neurophysiological information, but that it is just an algorithm that maximises and minimises variances between two classes without any background information. As estimating covariance matrices with sample covariance is known to be a nonrobust estimator, it will be sensitive to artifacts and noise, which will have big influences on the CSP-filters. Especially when the training set is relatively small, CSP is well-known to overfit.

2.5 Feature Extraction

If all EEG-data points would be used, the dimensionality of the data would be too high to give to a classifier, so the most relevant features are extracted. As CSP is designed to discriminate between conditions by optimizing the variances, the log variances of the CSP-filtered signal can be used as features f_i, i being the respective CSP-channel.

$$f_i = log(VAR(V_i^{CSP})) \tag{2.9}$$

Based on the amount of features needed, an amount of CSP-channels, also called filter pairs, is used. When selecting a pair of filters, the outermost channels are chosen, as those filters correspond to the highest and lowest eigenvalues by construction, and thus contain most information.

2.6 Classification

The last important part of a Brain-Computer Interface is the correct prediction of the intended movement, using the extracted feature vectors. This prediction depends strongly on the type of classifier, but the number of features and the amount of training data available also plays a significant role.

Two commonly used classifiers in Motor Imagery classification are Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) [42].

The base of both techniques relies on the ability to discriminate between two conditions based on features vectors f_i. As these feature vectors are multi-dimensional, the problem would become simpler if the dimensionality of these feature vectors is reduced.

The goal of LDA is to create a hyperplane that separates both classes with the class of the feature vectors depending on the side of the vector regarding to the hyperplane. SVM's use the same principle to discriminate between classes, but the hyperplane is selected based on a maximisation of margins. Regardless of the fact that SVM's have good generalisation properties and are insensitive to overtraining and the curse-of-dimensionality, LDA is more popular, as it is a simpler technique with lower computation time. On the other hand, LDA does not perform well on complex nonlinear data, but the advantages of SVM's come at the cost of speed of execution, which is not desirable bearing in mind the goal of an online system. Therefore only LDA is discussed.

2.6.1 Linear Discriminant Analysis (LDA)

As mentioned, the goal of this technique is to find a projection vector w that maximally separates the feature vectors f_i by projecting the multi-dimensional feature vectors onto one dimension, without inducing loss of information [29].
The simplest way to separate the two classes is by projecting onto one dimension based on maximal separation of the projected class means.

With C_1 holding the left hand trials and C_2 holding the right hand trials, the means of the feature vectors are calculated based on class. N_s^1 being the total amount of left hand trials, N_s^2 being the total amount of right hand trials.

$$m_1 = \frac{1}{N_s^1} \sum_{i \in C^1} f_i \tag{2.10}$$

$$m_2 = \frac{1}{N_s^2} \sum_{i \in C^2} f_i \tag{2.11}$$

When only using the means as a separation measure, the maximisation problem becomes:

$$m_1^* - m_2^* = w^T(m_1 - m_2) \tag{2.12}$$

with m_1^* and m_2^* the class means of the projected data. A Langrange multiplier is used to perform constrained maximisation, where $w = m_1 - m_2$ is the solution to the maximisation problem.

As seen in the left part of Figure 2.11 this approach results in too much overlap, which is due to the class distributions having strongly nondiagonal covariances.

Fisher's solution to this problem is to maximise a function that, while maintaining a large separation between the class means, also makes the variance small within each class, resulting in a minimal class overlap.

With the within-class covariance matrix S_w being

$$S_w = \frac{1}{N-1} \left(\sum_{i \in C^1} (f_i - m_1)(f_i - m_1)^T + \sum_{i \in C^2} (f_i - m_2)(f_i - m_2)^T \right) \tag{2.13}$$

and the between-class covariance matrix S_B:

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \tag{2.14}$$

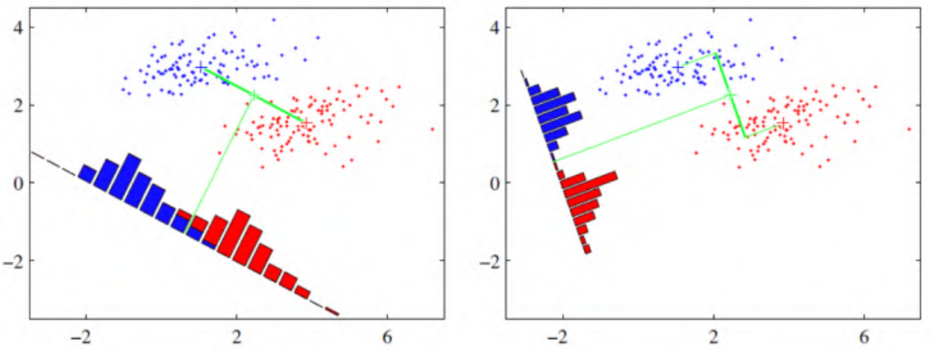


Figure 2.11: Left: The construction of a hyperplane based on maximising the difference between the class means results in too much overlap. Right: A better separation of the two conditions as a result of also minimising the within-class variance while maintaining a large difference between class means [29].

the Fisher criterion can be written as:

$$J(w) = \frac{w^T S_B w}{w^T S_w w} \tag{2.15}$$

By differentiating this formula with respect to w, the Fisher criterion is maximised when

$$(w^T S_B w) S_w w = (w^T S_w w) S_B w \tag{2.16}$$

As we do not care about the magnitude of w but only the direction, the scalar factors can be dropped. By multiplying both sides with S_w^{-1} we obtain:

$$w^* = S_w^{-1}(m_1 - m_2) \tag{2.17}$$

This result is also known as Fisher's linear discriminant, but despite the name containing the word discriminant, it is a more a direction for projection.

With this projection vector, the transformed means of the classes can be calculated as

$$m_1^* = w^T m_1 \tag{2.18}$$

$$m_2^* = w^T m_2 \tag{2.19}$$

and the transformed feature vectors become

$$f^* = w^T f \tag{2.20}$$

As seen in the right part of Figure 2.11, the feature vectors are now optimally separated.

To assign a label to each trial, the euclidian distance is calculated between the transformed feature vectors f^* and the transformed means m_1^* and m_2^*. Based on the smallest euclidian distance to either means, the trial is assigned a left or right class label.

Regularisation

Linear Discriminant Analysis is known to overfit in high-dimensional feature spaces when only a few data points are given. If overfitting occurs, the within-class covariance matrix S_w can become singular, as the large eigenvalues are estimated too large and the small eigenvalues too small, making the matrix impossible to invert or the estimation of w imprecise.

To solve this problem a technique called shrinkage will be used. This technique compensates for this systematic bias of S_w, that could cause a decrease in classification performance. To counterbalance this error, S_w is replaced by

$$\tilde{S}_w = (1 - \lambda S_w) + \lambda \nu I \tag{2.21}$$

with ν the average eigenvalue trace of S_w.

This results in a shrinked covariance matrix $\tilde{S}_w$ having the same eigenvectors, but with shrinked eigenvalues towards the average ν. This is illustrated in Figure 2.12: the left part shows data points drawn from a Gaussian distribution, with their true covariance matrix in orange and the unregularised estimated covariance matrix in cyan. When applying shrinkage, an approximation of the true covariance matrix is made as a linear interpolation between the empirical covariance matrix in cyan, where $\lambda = 0$ and the LDA is unregularised, and a spherical covariance matrix in black, where $\lambda = 1$.

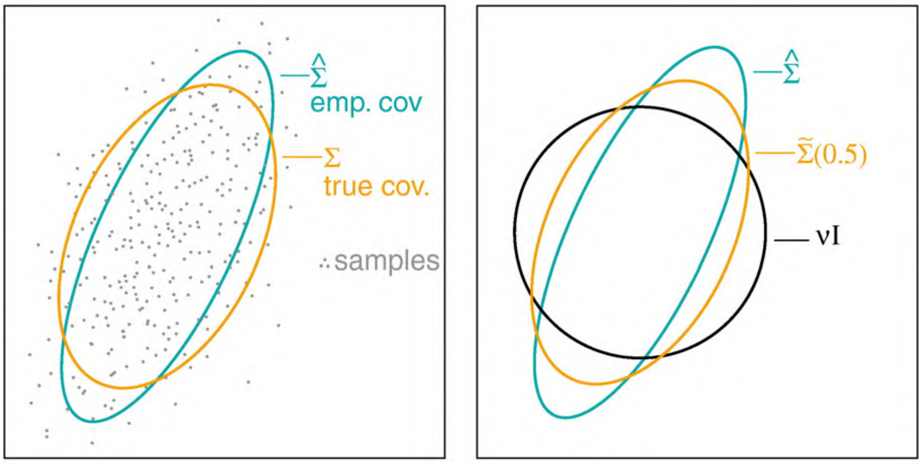


Figure 2.12: Left: data points from a Gaussian distribution in grey, the true covariance matrix in orange and the unregulzarized estimated covariance matrix in cyan. Right: The same unregularised estimated covariance matrix in cyan, a spherical covariance matrix in black and a linear interpolation between these two as the shrinked estimate of the covariance matrix in orange[43].

For determination of the optimal shrinkage parameter λ, an analytic method found by Ledoit and Wolf is used [44]. This formula is designed to minimise the Frobenius norm between the shrunk covariance matrix and the unknown true covariance matrix and has as a consequence that stronger shrinkage is applied when the sample-to-sample variance of entries from the empirical covariance matrix is large.

Using the Ledoit-Wolf formula doesn't always lead to the best classification results, especially if the number of trials available is low, resulting in possibly bad estimations of the covariance matrices used in S_w. But as this technique is easy to implement and computationally cheaper, it is the preferred technique for regularisation. In the formula for the computation of the optimal parameter, N is the number of trials, Σ indicates taking the sum and $T = \nu I$ with I the identity matrix.

$$\lambda = \frac{N-1}{N} \frac{\sum \left(\sqrt{S_w}(N-1) - \sqrt{S_w \frac{N-1}{N}} \right)}{\sum \sqrt{S_w - T}} \tag{2.22}$$

2.6.2 Classifier performance

For evaluation of the classifier, different measures can be used, all derived from the confusion or error matrix[45]. A confusion matrix gives a more detailed view on correct and incorrect classification of trials. In the case of a binary classification, the confusion matrix holds the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) as illustrated in Table 2.1.

Table 2.1: Confusion matrix of a binary classification problem.

		Prediction outcome	
		+1	-1
Actual Value	+1	TP	FN
	-1	FP	TN

As no measure can be defined as the best measure for every purpose and different measures can even be contradictive as they don't represent accuracy in the same way, it is critical to choose an appropriate accuracy measure.

Here, two accuracy measures are presented:

Accuracy The simplest performance measure is the test accuracy, which is calculated as the number of correctly classified samples against the total amount of samples, or according to Formula 2.23 as calculated with the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.23}$$

Area Under Curve A very popular performance measure is the Area Under Curve. The AUC is calculated as the area under the Receiver Operating Characteristic (ROC) curve, which is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) (see Formula 2.24 and 2.25) [46]. A higher AUC indicates higher performance of the classifier. An AUC of 0.5 corresponds to random assignment, an AUC of 1 indicates perfect classification.

$$TPR = \frac{TP}{TP + FN} \tag{2.24}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.25}$$

If the dataset is unbalanced or certain outcomes have more weight (like in cancer recurrence), accuracy is not the preferred performance measure. For example, classifying 90 samples labeled with ‘negative for cancer’ and 10 labeled with ‘positive for cancer’, the accuracy can be 90% if all samples are labeled with ‘negative for cancer’, without capturing the essence of the data, and in this case, missing the recurrence of cancer for 10 patients. For unbalanced datasets, AUC is a more powerful measure as, in this case, it would only be 0.5, implying random assignment of the labels.

As, in this thesis, a balanced dataset is used with equal importance for left and right hand assignment, accuracy will be the preferred classifier performance measure.

2.7 Transfer Learning

As the goal of this thesis is to reduce calibration time without a decrease in performance, the reduction of training data due to a lower calibration time can be compensated by using training data from previously recorded subjects. Using more data will result in an increase in classification performance, however, due to the difference in the statistical distribution of data recorded from previous subjects, this data has to be transferred to the new subject in a way that it is used as efficient as possible.

First, two naive Transfer Learning techniques will be discussed that combine information from different classifiers, without adaption of data. Next, a Transfer Learning technique is explained that regularises the CSP and LDA algorithms based on data from previously recorded subjects, referred to as 'Covariance Shrinkage' throughout this thesis. The last discussed technique uses a Data Space Adaption (DSA) algorithm that linearly transforms the target subject's data based on previously recorded subjects to minimise the distribution differences between the target subject and the previously recorded subjects, also known as source subjects.

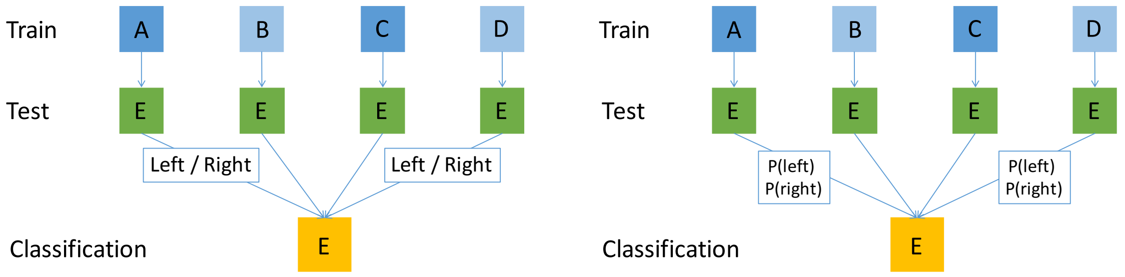
2.7.1 Naive Transfer Learning

Majority Voting

The simplest technique to combine the outcomes of different classifiers is by majority voting. Per trial, each classifier predicts an outcome and the outcome receiving most votes wins. This technique is simple, but overlooks other information given by the classifiers.

Averaging probabilities

By averaging the probabilities of the predictions of the classifiers, more information is used and the assignment of the label is based on the class with the highest summed probability.



(a) Majority Voting. **(b)** Averaging Probabilities.

Figure 2.13: Naive Transfer Learning algorithms.

2.7.2 Covariance shrinkage

The hypothesis used by Lotte and Guan [47] is that it is possible to find common information in the EEG-data measured from different subjects, despite large inter-subject variabilities. However, it is still important to find the most relevant subjects, as not all subjects may be significant to use. With the correct set of other subjects, their information is integrated in the algorithm as a regularisation term in the estimation of the covariance matrices in CSP and the estimation of the means in LDA:

$$\tilde{\Sigma_t} = (1 - \lambda)\Sigma_t + \lambda \frac{1}{S_t(\Omega)} \sum_{i \in S_t(\Omega)} \Sigma_i \tag{2.26}$$

$$\tilde{\mu_t} = (1 - \lambda)\mu_t + \lambda \frac{1}{S_t(\Omega)} \sum_{i \in S_t(\Omega)} \mu_i \tag{2.27}$$

Σ_t, μ_t and $\tilde{\Sigma_t}$, $\tilde{\mu_t}$ represent the unregularised and regularised version of the covariance matrix and the means. Ω contains the whole set of previously recorded subjects, $S(\Omega)$ holds the selection of subjects used for target t and λ is the optimal regularisation parameter.

With this technique it may be possible to set up a robust BCI with little calibration data from the target subject, by selecting an optimal set of targets and a regularisation parameter.

Subject Selection Algorithm

Instead of using all other previously recorded data, an optimal subset will be chosen based on the algorithm described by Lotte and Guan [47], illustrated in Figure 2.14.

Using subject A as the target subject and B, C, D, E,... as the subjects from which an optimal subset has to be chosen, the algorithm is based on the following guidelines:

- Training of the classifier is always done on the subset (unless it is empty) plus one of the subjects from the remaining set. Testing is done on the training data available from the target subject.
- During the subject selection algorithm, subjects are added to the subset and removed from the remaining set as illustrated in Figure 2.14

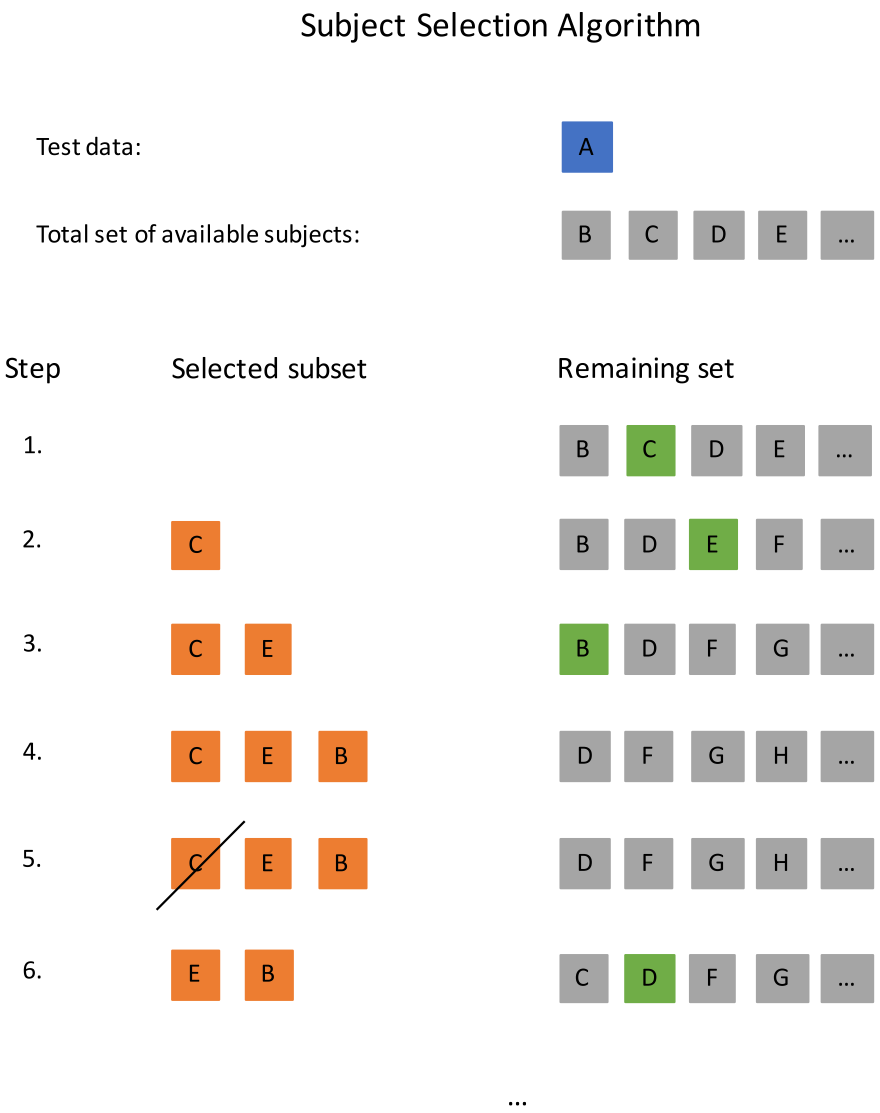


Figure 2.14: The subject selection algorithm. An orange box indicates the addition of a subject to the subset while removed from the remaining set. A green box indicates that this subject is the best choice from the pool of remaining subjects for that specific case.

As described in Figure 2.14, the algorithm is performed as follows:

1. Browsing through the remaining set, the best subject is the one that, when trained on, results in the highest training accuracy on target subject A. In the example of Figure 2.14, subject C performs best.

2. Subject C is added to the subset and removed from the remaining set. To search for a subject that works best in combination with C, for each other subject the covariance matrix is combined with that of C by averaging. Training accuracies are determined in the same way.

3. Subject E is added to the subset and removed from the remaining set. The same steps are performed as in 2.

4. Subject B is added to the subset and removed from the remaining set.

5. If there are three or more subjects in the subset, the algorithm will reinvestigate the option if training accuracies would be higher if one of these subjects would be left out. In this case, subject C can be left out and will be readded to the remaining set.

6. With C removed, the algorithm continues to find a new subject to add to the subset.

7. ...

The algorithm is continued until an optimal subset is found that maximises the training accuracy for our target subject A.

Regularisation parameter

For the determination of the optimal regularisation parameter λ the calculation of three accuracies is needed.

1. **Target Accuracy**: The test accuracy when Leave-One-Out-Validation (LOOV) is performed on only the target subject

2. **Selected Accuracy**: The test accuracy for training with the complete range of other source subjects and testing on the target subject

3. **Random Accuracy**: The test accuracy of the target subject determined by a classifier performing at chance level

With these accuracies, λ can be seen as the amount of data from other subjects that should be included for training of the classifier.

- When tested on only the target (= Target Accuracy), the model performs better than when including all other subjects (= Selected Accuracy), $\lambda = 0$.
- If, when using a classifier performing at chance level (= Random accuracy), the model performs better than by only using the data of the target, the other subjects are used to train the model, or $\lambda = 1$, as the target data is not appropriate to train a satisfactory BCI.
- Otherwise, if Selected Accuracy > Target Accuracy > Random Accuracy:

  $$\lambda = \frac{SelectedAccuracy - TargetAccuracy}{100 - RandomAccuracy} \tag{2.28}$$

  In this case, the higher the Selected Accuracy in comparison to the Target Accuracy, the higher λ should be as more confidence should be given to the EEG signals from the other subjects.

2.7.3 Data Space Adaption (DSA)

Arvaneh et. al [48] propose an new algorithm that reduces calibration time by using a subject-to-subject adaption algorithm, without the need for a large pool of previously recorded sessions, here also called historic sessions or subjects.

The algorithm consists of two main steps: first, subject-to-subject adaption is performed, second, the best calibration model is selected.

Subject-to-subject adaption

The goal of the transformation is to reduce the dissimilarities between the k^{th} source subject and the target subject's data, by adapting the target subject's data in such a way that their distribution difference is minimised. Arvaneh et. al [48] assume that the differences between the source subjects k and the target subject's data can be observed in the first two moments of the EEG-data and construct a transformation matrix based on calculations with the means and covariance of the EEG-data.

This results in an optimal linear transformation matrix M_k for each source subject k, using the covariance matrices Σ_1 and Σ_2 for each condition of the target subject and $\tilde{\Sigma}_{k,1}$ and $\tilde{\Sigma}_{k,2}$ for each condition of the source subject k as calculated in Formula 2.2 and 2.3, using the sample covariances. † stands for taking the pseudo-inverse of the matrix.

$$M_k = \sqrt{2\left(\tilde{\Sigma}_{k,1}^{-1}\Sigma_1 + \tilde{\Sigma}_{k,2}^{-1}\Sigma_2\right)^{\dagger}} \tag{2.29}$$

Using this transformation matrix, the target subject's data is transformed to minimise the distribution difference with the k^{th} source subject according to

$$V_k^{transformed} = M_k\, V \tag{2.30}$$

Selection of the best calibration model

After the transformation of the target subject's data, the distribution difference ought to be minimised, inferring that there still can be a difference. As one source subject may be more similar to the target subject's data than another, the aim of the second step of the algorithm is to find the most similar source subject.

This is done by first adapting the target subject's data according to the transformation matrix M_k as in Formula 2.29 and classifying the adapted data using the model trained on the corresponding source subject. The source subject that results in the highest classification accuracy is selected as the best calibration source subject.

If more than one source subject would result in the same classification accuracies, a selection is done based on the smallest KL-divergence between the target subject's transformed data and a source subject. The best source subject based on smallest KL-divergence is calculated as follows:

$$k^* = argmin_{k\in\phi} \sum_{j=1}^{2} \frac{1}{2}(trace(\Sigma_{k,j}^{-1} M_k^T \Sigma_j M_k) - ln(\frac{det(M_k^T \Sigma_j M_k)}{det(\tilde{\Sigma}_{k,j})}) - d) \tag{2.31}$$

with d the dimensionality of the data.

The Data Space Adaption algorithm has been proven to substantially reduce calibration time, without the need for a large database of previously recorded sessions. Another advantage is that it can easily be implemented in online applications, as the calculation of the transformation matrix and the adaption of the new target subject's data can be done in less than a second [48].

Chapter 3

Simulation Results

3.1 A first BCI

To gain insight in the working principles of a Brain Computer Interface, the previously explained techniques will be applied to the data as recorded from 7 subjects in dataset 1, BCI Competition IV [32].

For every subject in this dataset, the following structure is used to build a machine learning model that predicts the class of the imaginary movement, as recorded with EEG.

1. **Preprocessing**: The EEG-data is bandpass filtered from 0 to 40 Hz, as this frequency range includes the alpha and beta band. In this set-up, a Butterworth-filter of the 6th order is used.

2. **Feature Extraction**: To extract the most important properties from the EEG-signal, the training data is split in different frequency bands and filtered using CSP. By splitting the data in different frequency bands, every band will have their own specific CSP-filters, increasing the amount of detail.

3. **Classification Model**: Using the extracted features, a regularised LDA classifier is trained. The shrinkage parameter is calculated according to Ledoit-Wolf's formula (see Formula 2.6.1).

4. **Classify new data**: To classify new data, based on the model constructed with the training data, CSP-filtering is applied to the test set as calculated in **Feature Extraction** and the LDA as trained in **Classification Model** will predict the class labels.

3.1.1 Hyperparameters

Making an attempt to construct a robust BCI with optimal performance, hyperparameters are optimised. In the set-up of our machine learning model, the amount of splits in the frequency band and the amount of CSP filter pairs are the two hyperparameters that will influence the performance of the BCI.

The first hyperparameter is the amount of splits in frequency band. The frequency range is splitted in equal parts, with no correlation between the scientific brain-wave distribution as in Figure 3.1.

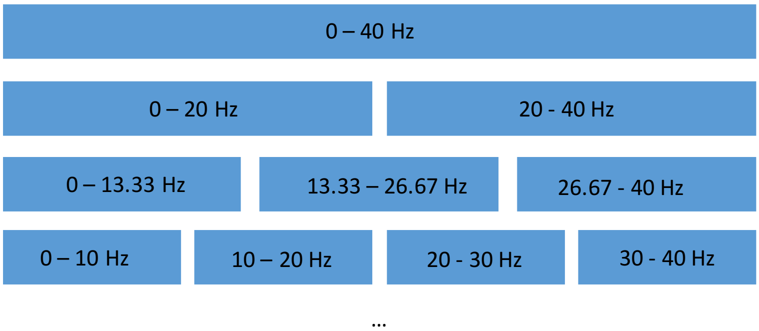


Figure 3.1: The splits in the frequency band are made by equally dividing the ranges from 0 to 40 Hz.

The second hyperparameter is the amount of filter-pairs extracted from the CSP-filtered data as explained in Chapter 2.4.3. One should keep in mind that using too much features in comparison to the amount of training data, could induce overfitting.

3.1.2 Optimisation process

To determine the best set of hyperparameters, the cross-validation scheme as in 2.1.2 is used. Per subject, the first 80% of the data is defined as the training set, the last 20% is used as the test set. To prevent data leakage while optimising the hyperparameters, the training set is split in 10 equal folds, with 9 folds serving as training set and the 10^{th} fold as the validation fold.

For every subject the process as explained in Chapter 3.1 is applied. In the second step the amount of splits in frequency band ranges from 0 to 9 and the amount of CSP filter-pairs from 1 to 9.
For every combination of splits and filter-pairs, due to the use of Cross-Validation, 10 validation accuracies are acquired, averaged and stocked in a 10 x 9 matrix. The person-specific optimal amount of splits and filter-pairs is determined as the combination that leads to the highest validation accuracy.

This optimisation process is repeated for every subject, resulting in a person-specific optimal set of hyperparameters. By averaging the validation matrices obtained over each subject, a set of averaged optimal hyperparameters can be used for the construction of a general BCI, which should be applicable on any new subject. The features that resulted in the highest validation accuracies are documented in Table 3.1. Their corresponding test and train accuracies are illustrated in Figure 3.2. For details, see Appendix A.

	# splits in frequency band	# CSP-filter pairs
Subject A	6	2
Subject B	3	1
Subject C	7	3
Subject D	5	1
Subject E	8	1
Subject F	8	3
Subject G	6	7
Average Subject	8	3

Table 3.1: Person-specific and general optimal hyperparameters.

Figure 3.2 shows that the average test accuracy is only 67±15%. The standard deviaton having high values means that the difference in test accuracies between subjects is rather high, in this case ranging from 50 to 80%. This illustrates the problem that, even with features optimised per subject and using all the calibration data available (160 training trials per subject), there is no guarantee that the classifier will perform well. Subject D, E and G have test accuracies higher than 75%, but the others don't reach the threshold of 70 % of proper BCI performance [49]. Some explanations for these bad results could be the patient being tired or distracted during recording, or that they were just not able to steer the BCI properly.

As the average train accuracy is 98±3%, even with the usage of regularisation, the main explanation for the weak performance of the BCI is the data overfitting to the training set. With 160 training trials available, it is expected that the system would have good generalisation properties if the amount of features is in proportion to the amount of training data available (preferably less than 16 trials in this case, as explained in Chapter 2.1.1). In Table 3.1, with the amount of features calculated as (number of splits in frequency band) x 2 x (number of CSP-filterpairs), only subject B, D and E require a small amount of features, but not necessarily reaching high test accuracies. This reveals the fact that using a low amount of features doesn't guarantee the prevention of overfitting. Using a too low amount of features could also induce underfitting, if the BCI wasn't able to capture the essence of the data.

To see if it is possible to create a robust BCI, the performance of the BCI is compared to a BCI using the averaged best features (8 splits in frequency band and 3 CSP-filterpairs).

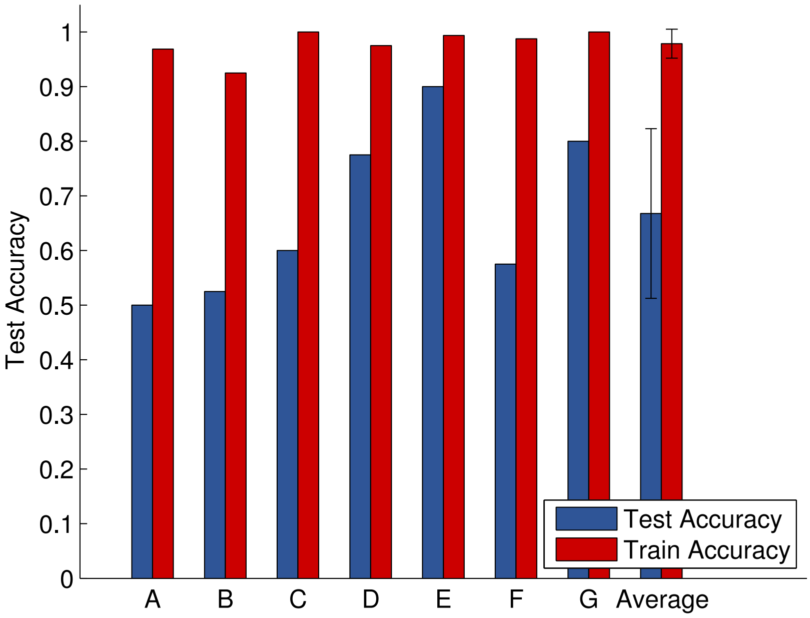


Figure 3.2: The train and test accuracies reached when using the person-specific optimal features.

A second comparison is done with a BCI using only two splits in the frequency band and 3 CSP-filterpairs (see Figure 3.3). These splits are based on the alpha band (8-13 Hz) and beta band (13-30 Hz), as these two frequency bands contain the most relevant information for Motor Imagery applications. The delta (0-4 Hz) and gamma band (4-8 Hz) are left out as literature has proven better performance when restricting the frequency range from 8 - 30 Hz [40] [39].

Figure 3.3 shows that the performance on the test set is not guaranteed to be highest when the personal best features are used. This is due to the possible occurence of overfitting, as the amount of features were optimised for the training data, and not for the testing data. This can result in other combinations of features being better for the unseen test data.
Using only the alpha and beta band, the test accuracy is higher for every subject (except for subject A) than when using the personal best features. This may be due to the proper amount of features (= 2 x 2 bands x 3 CSP-filterpairs), avoiding overfitting and by only using frequency bands that are medically relevant.
On average, the last method performs best, but with the standard deviations being high, it is not possible to draw conclusions on which hyperparameters perform best. The only conclusion that can be drawn is the high level of difficulty to construct a proper BCI model for every subject.

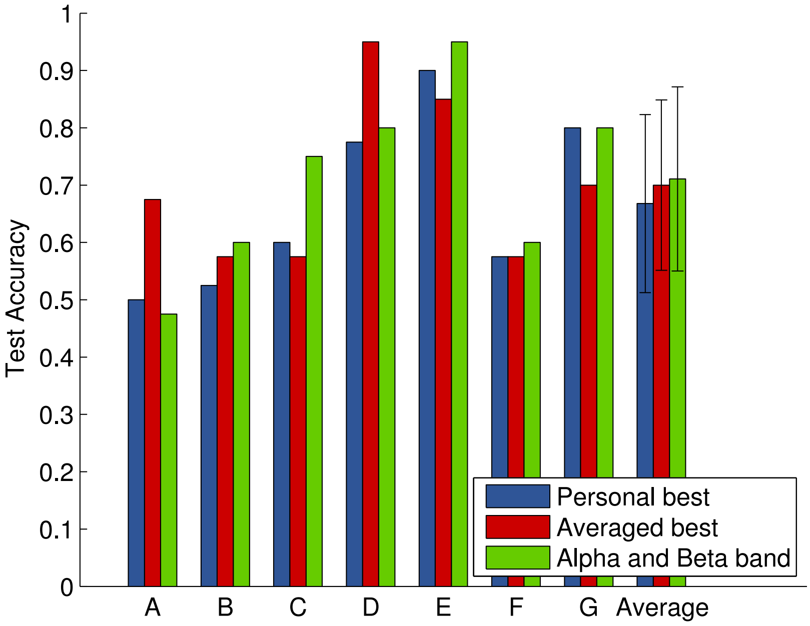


Figure 3.3: Comparison of the test accuracies when using person-specific optimised features, an averaged optimal amount of features and when only using data from the alpha and beta band.

3.2 Transfer Learning

With the problem of constructing highly performing BCI's, illustrated in Figure 3.2 and Figure 3.3, some Transfer Learning techniques will be tested to see if it is possible to reach higher test accuracies, with less training data from the new subject.

In Transfer Learning, the data transferred between subjects doesn't necessarily has to be derived from the same task (e.g., left hand movement, right foot movement,...), but it has to be similar.
In this work, only data from subjects that did perform the same tasks is used, as this gives a first indication of the possibilities that Transfer Learning techniques hold. Therefore, from here on, subject A and subject F will be left out as the imaginary movement of those subjects were left hand and foot movement and the other majority of subjects chose left and right hand movement.
Later on, this can be expanded to transferring data originating from different tasks.

For clarity of the graphs, in the upcoming figures, no error bars are shown when averaging results over subjects, as the standard deviations can vary from 7 to 20%, with no relation to the amount of calibration trials. By only comparing 5 subjects, the average can give an indication of the performance of the methods, but for accurate comparisons, the investigation has to be done subject-specific.

First, four Transfer Learning techniques from literature will be evaluated on our dataset. These standard techniques are compared and explored further to build a new Transfer Learning technique that should reduce the calibration time as much as possible.

3.2.1 State of the art techniques

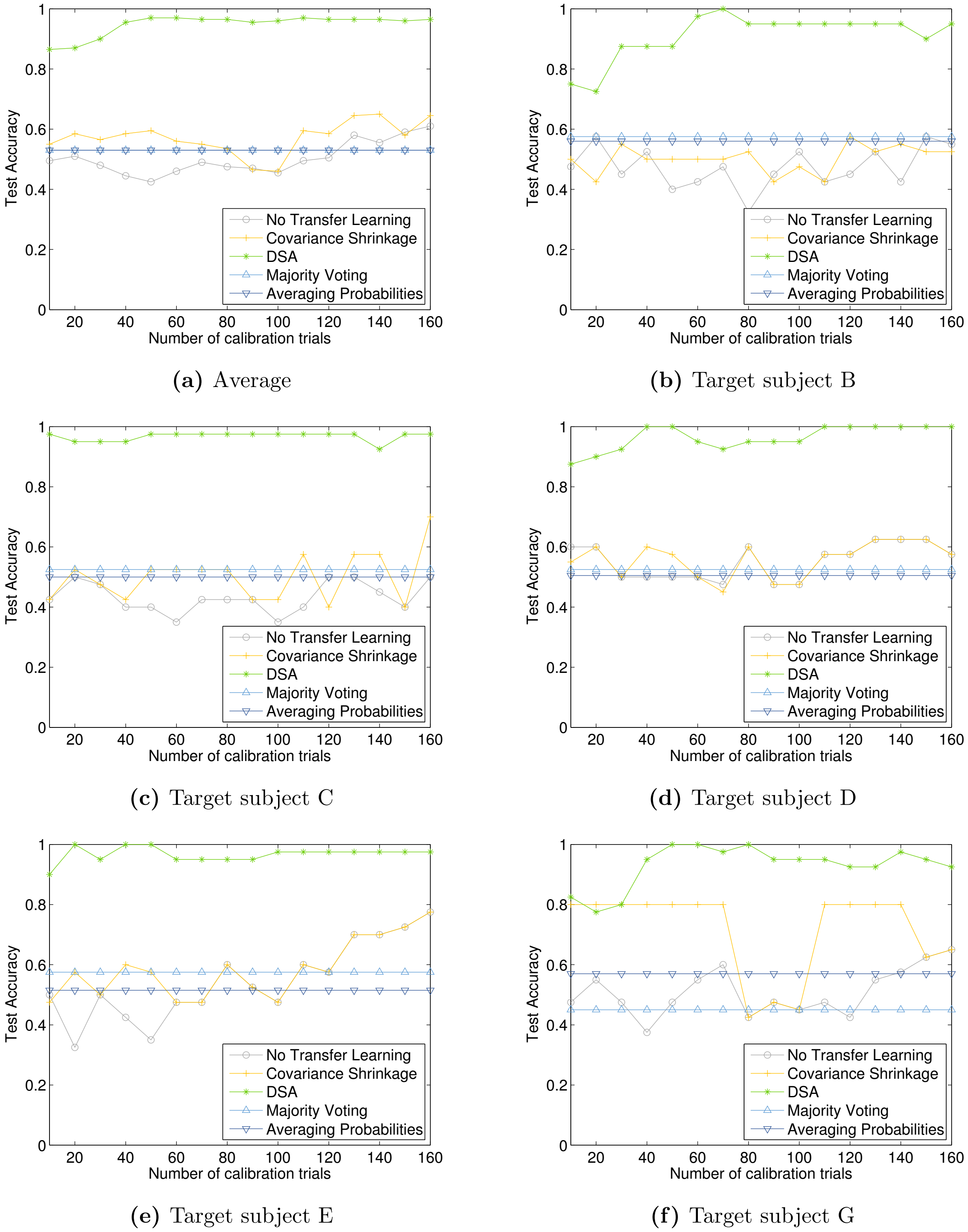
In the first comparison the four following techniques are used: Majority Voting, Averaging Probabilities, Covariance Shrinkage and Data Space Adaption as explained in Chapter 2.7. To see the influence of the amount of calibration trials on the performance of the classifier, several experiments are done with the training set size ranging from 10 to 160 trials, in steps of 10 trials. The test set will contain the last 40 trials.

Comparison between methods: 3 - 40 Hz

In Figure 3.4, results from five methods are compared: using no Transfer Learning, using two naive Transfer Learning techniques, being Majority Voting and Averaging Probabilities and using two Transfer Learning techniques from literature, being Covariance Shrinkage and Data Space Adaption. Having data available from five source subjects, different experiments are executed with a variable amount of calibration trials, ranging from 10 to 160 trials from every source subject. Testing is done on the last 40 trials of the target subject. During pre-processing (as in Chapter 3.1), the data is filtered from 3 - 40 Hz. The delta waves are left out, as these occur during relaxation and sleep, not during Motor Imagery. According to literature [47][48], no splitting in the frequency band is done. Concerning CSP-filtering, the three most discriminative pairs of filters are used as features and given to a regularised LDA classifier.

From Figure 3.4a, it is clear that DSA, on average, gives the best results, as the test accuracy is highest for every amount of calibration data. Covariance Shrinkage, on average, doesn't always score better than when using no transfer learning. In some cases (see Figure 3.4f) it does give higher performance, even close to the results obtained by DSA, but for the other subjects, Covariance Shrinkage performs only a little better or the same in comparison to when not applying Transfer Learning.
The sudden decrease in performance of Covariance Shrinkage for subject G when using 80, 90 and 100 calibration trials, may be due to a shift in the calibration data, causing the selection of the best subset or the regularisation parameter (as in Formula 2.28) to change.

The Naive Transfer Learning techniques, being Majority Voting and Averaging Probabilities, on average, give the same results being 53±5% and 3% respectively. When not much calibration data is available (<50 trials), for subject B, C and E, they score better than when applying no Transfer Learning, but as the performance stays relatively low, independent of the amount of calibration data, there is no future in these naive Transfer Learning techniques.



(a) Average

(b) Target subject B

(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 3.4: Comparison of different Transfer Learning techniques, filtering from 3 - 40 Hz.

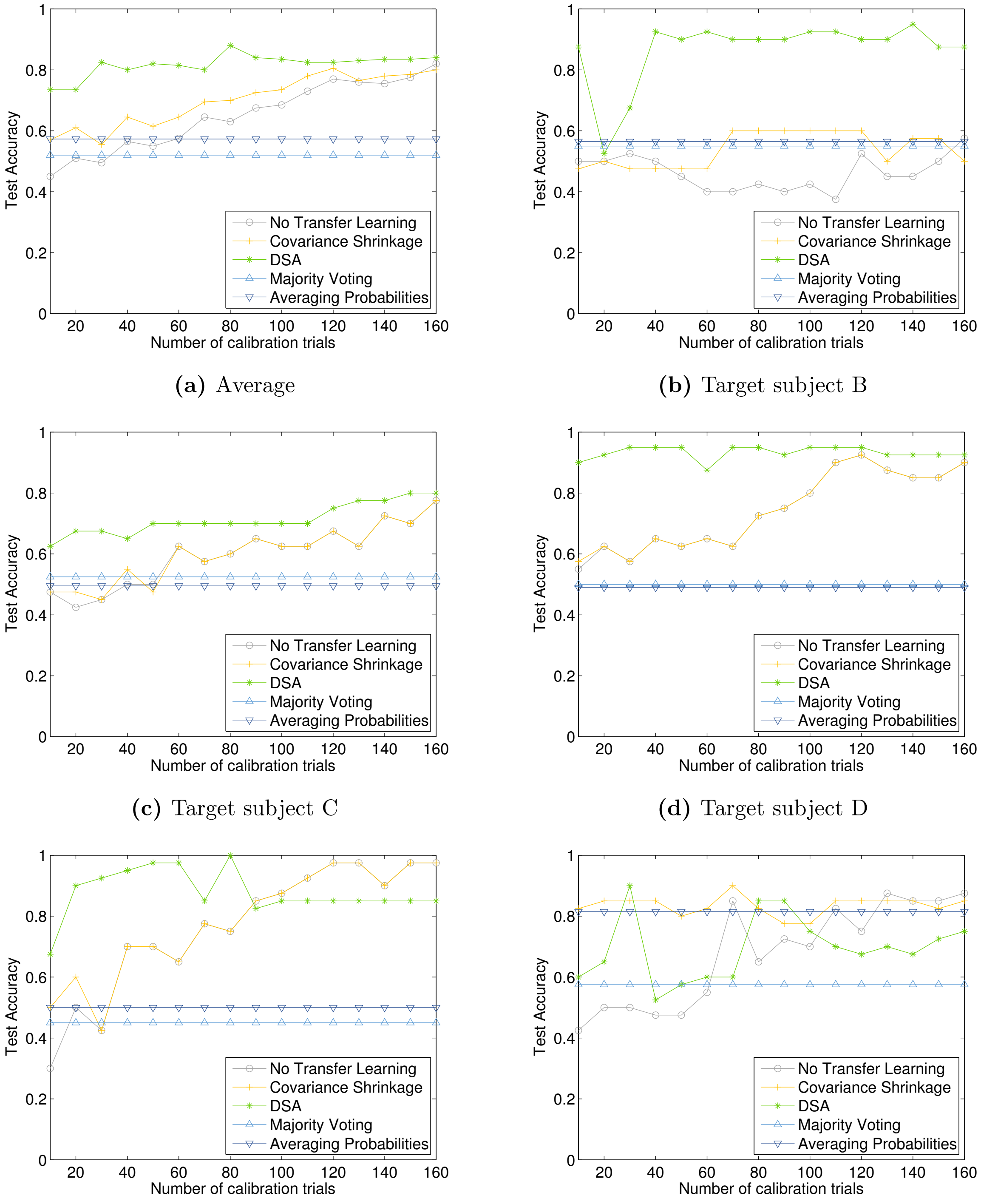
What is odd in these graphs is that the test accuracy for subject B, C and D, when not applying Transfer Learning, does not increase with more calibration trials becoming available. It might be that the range of 3 - 40 Hz is not optimal for Motor Imagery classification, so in the next comparison a different frequency band is used.

Comparison between methods: 8 - 35 Hz

In Figure 3.5, the same comparison as in 3.2.1 is done, with the only difference the data being filtered from 8 - 35 Hz, as this was the frequency range used in the papers describing the method of Covariance Shrinkage and DSA. By only using the frequencies from 8 - 35 Hz, alpha waves, beta waves and the beginning of gamma waves are measured. These include the most import wave categories for Motor Imagery classification. [22]

On average, when comparing Figure 3.4a and Figure 3.5a, it seems like the performance of DSA is lower when filtering from 8 - 35 Hz, but when looking subject-specific, the difference in performance between this frequency range and when filtering from 3 - 40 Hz is not significant when using less than 20 calibration trials ($p>0.2$).
What does decrease, is the robustness of DSA. For subject B and G, the DSA line shows different peaks between certain amounts of calibration data. This is due to the altering in covariance matrices when new training data becomes available, resulting in a different choice of the best source subject. If hereafter, more calibration trials are added, the algorithm stabilises back into their previous better working form and the right source subject is rechosen.
The performance of Covariance Shrinkage, on average, does increase significant when narrowing down the frequency spectrum. On the other hand, when looking at subject B, C and E, the Covariance Shrinkage line stays the same as when no Transfer Learning is used. Only for subject B and G it can give higher test accuracies.
The performance of the BCI when not using Transfer Learning does look more logic in this frequency range. It starts off with a low test accuracy and increases with more training data becoming available.

The naive Transfer Learning techniques still show low performance, for the same reason as in the other frequency range, as they don't use properties of the calibration trials.
The high performance of DSA for the 3 - 40 Hz frequency band, may be due to the particular choice of subjects in our dataset, but might not work as good in a more general system. As the frequency range of 8 - 35 Hz is more related to Motor Imagery classification and gives more logical results when not applying Transfer Learning, this range will be used for further experiments.



(a) Average

(b) Target subject B

(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 3.5: Comparison of different Transfer Learning techniques, filtering from 8 - 35 Hz.

Chapter 4

Design of a new Transfer Learning method

In this chapter, a new Transfer Learning method is proposed to overcome the problem of interperson and intersession differences, resulting in long calibration times. The new method should reduce the calibration time, while maintaining a high level of performance.

By studying the results acquired in Chapter 3, a new Transfer Learning method is built by combining or adapting the Transfer Learning methods from literature, in order to better classify data from a new unseen subject.
To design this method, further investigation is done on the working principles of Covariance Shrinkage (Chapter 2.7.2) and Data Space Adaption (Chapter 2.7.3).

4.1 Working principles of Covariance Shrinkage and Data Space Adaption

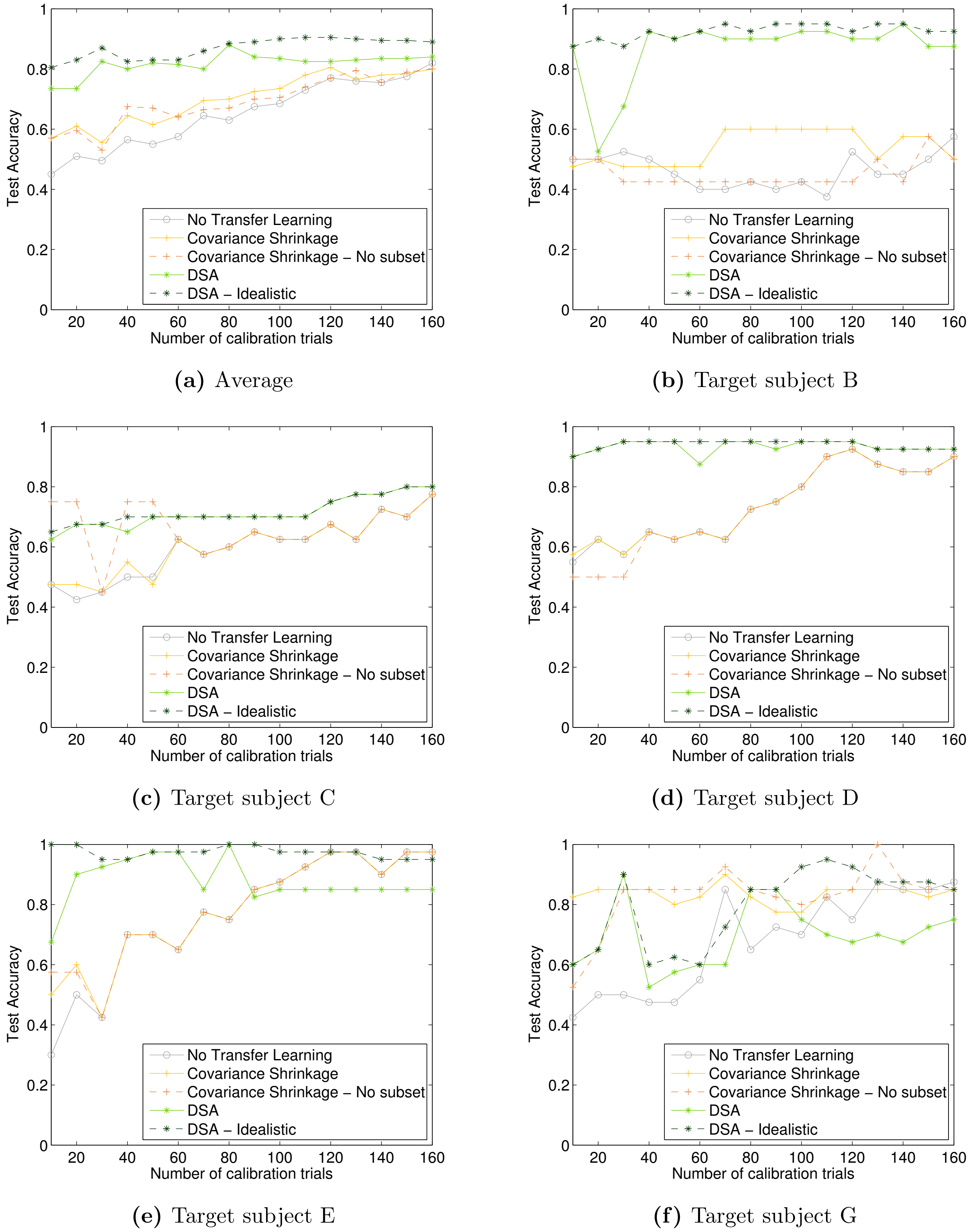
In **Covariance Shrinkage** a Subject Selection Algorithm (see Section 2.7.2) is used, meaning that not necessarily all the data of the previously recorded source subjects is used to construct the covariance matrix in regularizing CSP and LDA. To see the effects of the Subject Selection Algorithm, results were produced by using all source subjects, instead of selecting a subset of source subjects. In the following graphs, this method is called 'Covariance Shrinkage - No subset'.

In **Data Space Adaption**, after having performed the Subject-to-Subject adaptation, there is a selection of the best calibration model (see Section 2.7.3). To see what the effect of this selection is, this selection is left out. Instead of choosing the best calibration model, all models can now predict labels of new incoming trials and afterwards, the best calibration model is chosen based on the highest test accuracy. This technique is not allowed in Machine Learning algorithms, as it uses posterior information being the test accuracy, but here it is only used for illustration purposes, to check whether the selection method of DSA works optimal. This method is called 'DSA - Idealistic' in the following graphs.

In Figure 4.1 five methods are plotted: No Transfer Learning as a baseline, Covariance Shrinkage and DSA as previously plotted and the adapted versions of Covariance Shrinkage and DSA.

Figure 4.1a shows that, on average, the idealistic version of DSA scores better, for every amount of calibration trials. Looking at Figure 4.1b the problem with the sudden decrease in performance when going from 10 to 20 trials, is solved by using the idealistic version. This means that, by the addition of new calibration data, another source subject was chosen for transfer of data, but not the one that resulted in the highest test accuracy. As, for subject B, E and G, the idealistic DSA scores better or equal to the standard DSA, it is clear that there is still room for improvement in the DSA algorithm, including removal of the stability problem for subject B, E and G, as sudden drops in test accuracy are unwanted results.

When comparing the performance of Covariance Shrinkage with and without a selection of subsets, from the average test accuracy in Figure 4.1a no conclusions can be drawn. When looking at every subject individually, choosing all source subjects either works worse (as in subject B), a little better (as in subject C until 60 trials), but for subject C (from 60 trials), D (from 40 trials) and E (from 30 trials), the choice of subset did not alter the results, as the regularisation parameter (see Formula 2.28) stayed at zero, independent of the subset of source subjects. A detailed comparison of the chosen subset and the corresponding regularisation parameter with and without a selection of subsets can be found in Appendix B.



(a) Average

(b) Target subject B

(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 4.1: Illustration of the working principles of Covariance Shrinkage and Data Space Adaption.

4.2 Adaption of existing methods

Based on the comparisons done between different Transfer Learning techniques in Chapter 3.2.1 and 4.1, it is clear that none of these methods result in a classifier that always performs best, for every subject. Therefore, it is interesting to investigate new possibilities that could increase the performance of these techniques with as little calibration data as possible.

4.2.1 DSA/CS - Accumulate source subject data

The first idea is a combination of the two promising Transfer Learning methods, Data Space Adaption and Covariance Shrinkage.

- With the Subject-Selection Algorithm from Covariance Shrinkage, define the best subset.
- Place the trials of these subjects after each other, as if the data would come from one target.
- Transform the calibration data of the target subject according to the transformation matrix M (as calculated in Formula 2.29), using the covariance matrices of the accumulated source subject data.
- Train the classifier with the accumulated source subject data and test it on the transformed new trials.

4.2.2 DSA - Averaging probabilities (AP)

For the second idea a variant of Data Space Adaption is used.

- As in DSA, the calibration trials are transformed with a transformation matrix M_k, based on the covariance matrices of each source subject k.
- Instead of selecting one best calibration model, multiple calibration models are selected, by picking the ones that give the highest validation accuracies.
- As different classifiers are constructed based on different source subjects, for each trial, the classifiers will output probabilities that are **averaged**.

By averaging the probabilities over the remaining classifiers, the drawback from the original DSA algorithm, where new calibration data could induce a different choice of best source subject, could be reduced, by averaging influences of multiple classifiers.

The choice for averaging the probabilities instead of using Majority Voting, is based on the fact that by averaging probabilities, the effect of insecure classifiers is reduced. If three insecure classifiers would predict a left-hand trial, and a fourth confident classifier would predict a right-hand trial, Majority Voting would result in a left-hand trial, while Averaging Probabilities would result in a right-hand trial.

It was visible during the experiments, that almost always every source subject was chosen based on validation accuracy, therefore this pre-selection of subjects is left out, resulting in a method that averages the outputs of all classifiers.

4.2.3 DSA - Maximum probability (MP)

The third idea works according to the same principle as *DSA - Averaging probabilities*, but instead of averaging the probabilities of the different classifiers, the trial will be classified according to the class to which the **highest probability** was given by a classifier. As for DSA - AP, there will be no pre-selection of subjects based on validation accuracy.

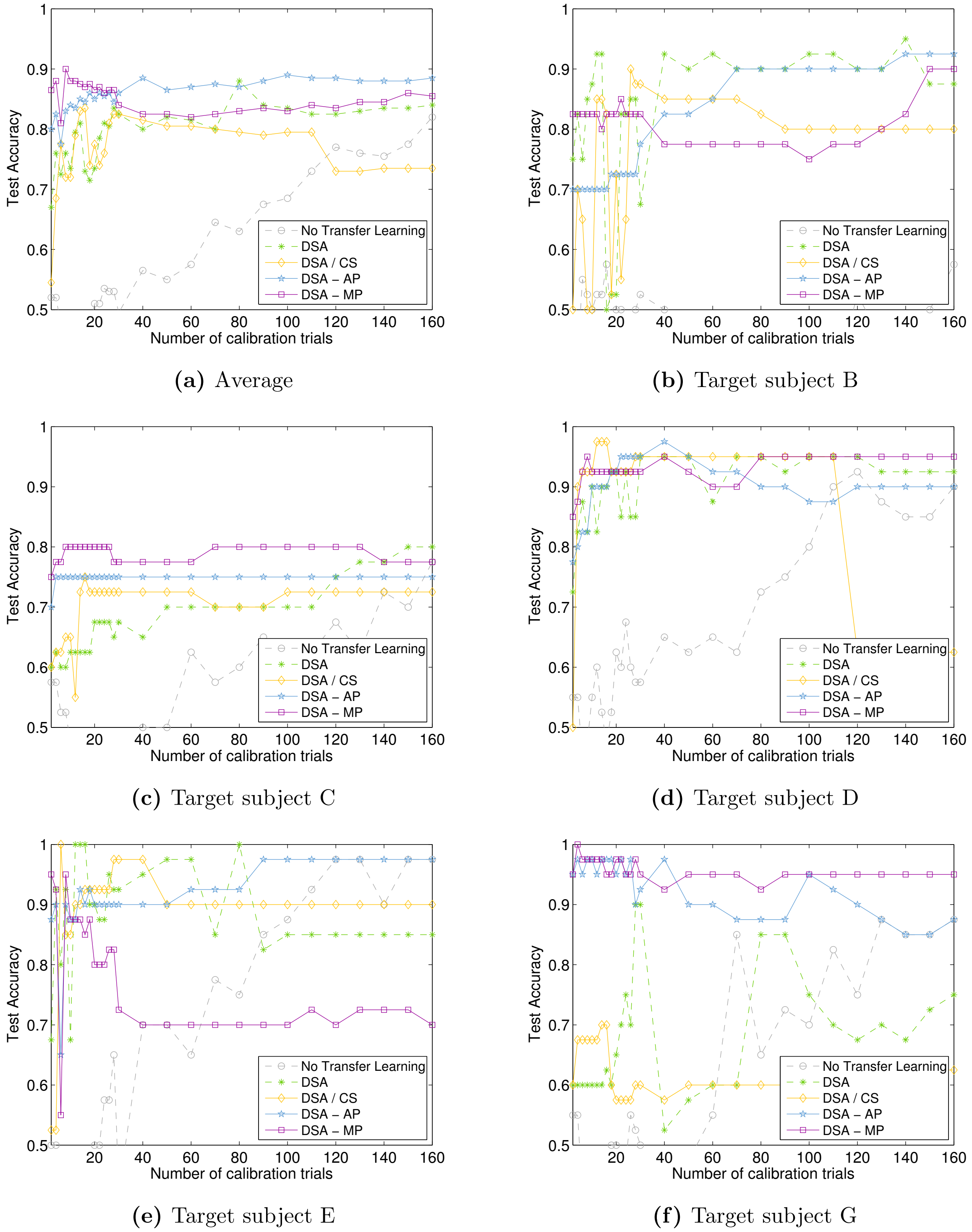
4.2.4 Comparison

In Figure 4.2 and Figure 4.3these ideas are compared to the existing methods for every source subject. The new methods are plotted with solid lines, the old techniques from literature are plotted in dashed lines. As the focus of the new Transfer Learning technique lies on using as little calibration data as possible, these experiments are performed with a training set size varying from 2 to 30 calibration trials, in steps of 2 trials, after which experiments are performed in steps of 10 trials in training set size. To be able to see the small differences between test accuracies, only the range from 50% to 100% is plotted.

From Figure 4.2 it is clear that no new method can be identified as the best performing method for every target, neither based on the average performance, neither on individual performance of the methods.

As our goal is to reduce calibration time, the aim of the new method should be to give high test accuracies with a low amount of calibration data, hence the focus of the comparison is on the left part of the graphs in Figure 4.2. A more detailed view on the experiments done using two to 40 calibration trials is shown in Figure 4.3.

For subject B and subject D, DSA - AP and DSA - MP don't give the best results, but at least the test accuracies remain stable with few calibration data, in comparison to the performance of DSA/CS and DSA.
For subject C and subject G, DSA - AP and DSA - MP do give the best overall results, so in these cases, the first idea is discarded.
For subject E, the best method depends on the number of calibration trials, with test accuracies for all methods fluctuating when few calibration data is given. This may be due to the first consecutive trials of subject E being indecisive or unclear to the BCI.



(a) Average

(b) Target subject B

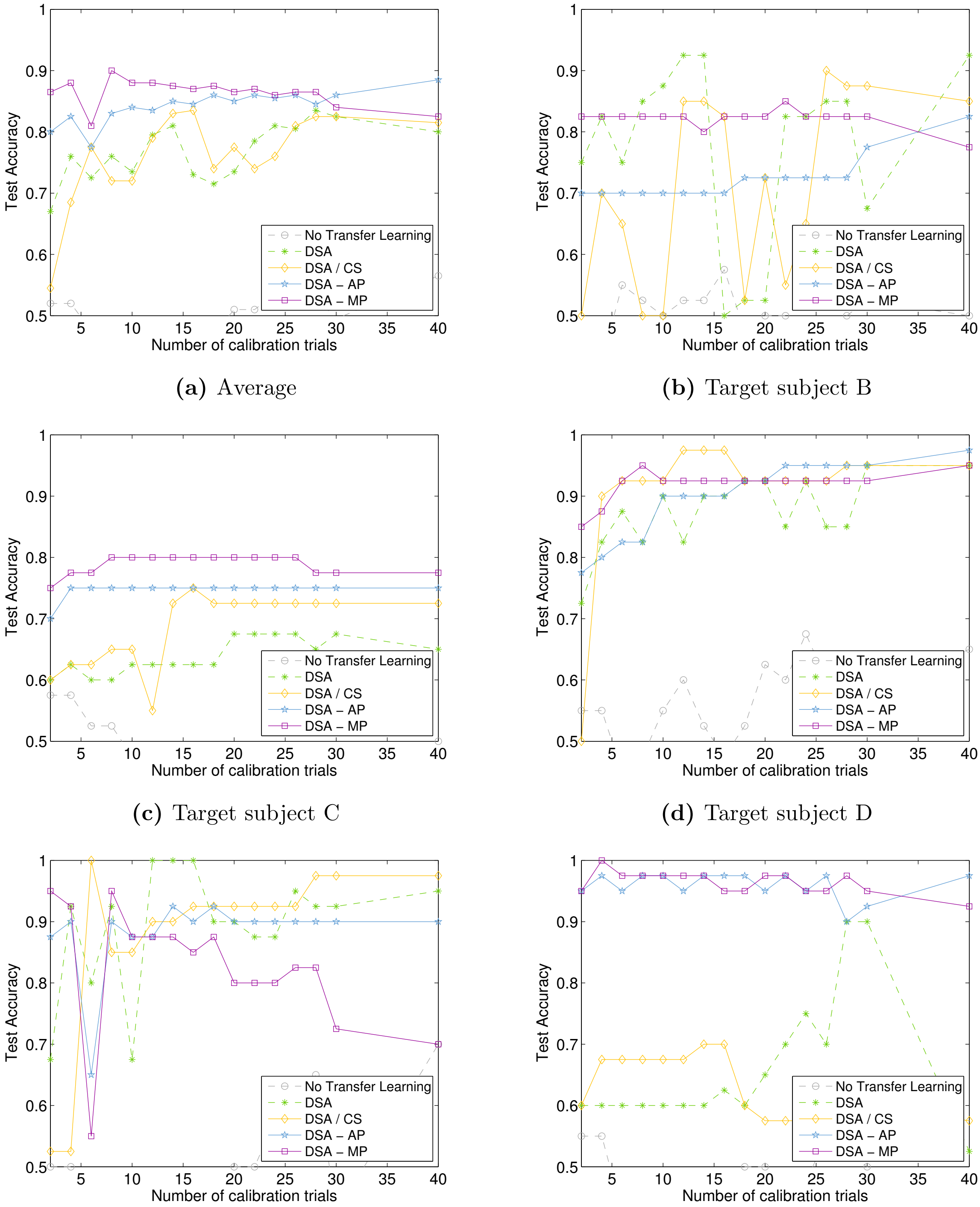
(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 4.2: Comparison of No Transfer Learning, DSA and three new methods.



(a) Average

(b) Target subject B

(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 4.3: Comparison of No Transfer Learning, DSA and three new methods, focusing on the experiments done with 40 calibration trials or less.

Using a dependent t-test, it can be measured whether the test accuracies across 5 subjects differ significantly when using one method or the other. In Figure 4.4, the performance of DSA is compared to DSA-AP and DSA-MP respectively. It is seen that the p-value for both null-hypothesis (H_0 (DSA = DSA-MP) and H_0 (DSA = DSA-AP)), for every amount of calibration trials used, is always bigger than p = 0.05, so the null-hypothesis is not rejected. Only using 2 calibration trials, the null-hypothesis can be rejected with a p-value of respectively 0.02 and 0.03.

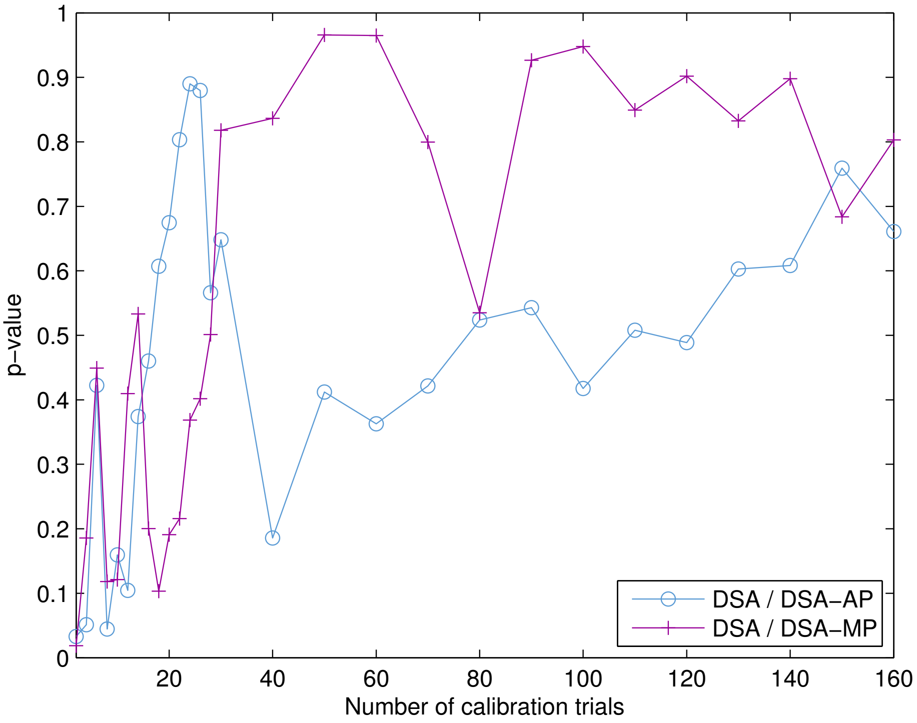


Figure 4.4: P-values for both null-hypothesis H_0 (DSA = DSA-MP) and H_0 (DSA = DSA-AP), showing that the null-hypothesis can not be rejected.

4.3 New Transfer Learning Technique

Based on the comparisons made in Figure 4.2, a new method is constructed. As neither DSA - AP or DSA - MP reached significantly better results than DSA (see Figure 4.4), the new method will combine the approaches of DSA - AP and DSA - MP.

This choice of methods relies on the fact that, depending on the amount of calibration data, one of these three methods will give the highest test accuracy, as was illustrated in Figure 4.2. To guarantee that the highest test accuracy is always reached, a new method is constructed that should result in an upper boundary of DSA, DSA - AP and DSA - MP.

To be able to meet these conditions, further investigation is done on where DSA - AP and DSA - MP make mistakes in classifying trials.

If DSA - AP is used, which averages the probabilities of the classifiers for every class, trials are classified wrong if the decision is made based on a majority of classifiers, which are unsure of their decision, prevailing the classifier that is sure of his decision to dominate. An example can be seen in Table 4.1.

Table 4.1: Example of a possible outcome when predicting the class label for a trial of target subject B, which should be a right-hand trial. When averaging the probabilities, the class label is left, while when using the maximum of the probabilities the class label would be correctly predicted as right.

	P(Left)	P(Right)	Average(Left)	Average(Right)
Source subject C	0.65	0.35	**0.525**	0.475
Source subject D	0.65	0.35	Maximum(Left)	Maximum(Right)
Source subject E	0.6	0.4	0.65	**0.8**
Source subject G	0.2	0.8		

This gives rise to a first criterion that the new method should remove or ignore the outcomes of an unsure classifier.

If DSA - MP is used, which classifies based on the classifier that gives the highest probability, trials are classified wrong if one of the classifiers gives a very high probability, but the other ones were right, with respectively a lower probability, illustrated in Table 4.2. This brings to mind that even though accuracies are high, this doesn't mean that they correctly predict a class label. The new method should thus be able to detect classifiers that give high probabilities, but don't result in high performance.

Based on these findings, it is clear that DSA - AP and DSA - MP are complementary, but a criterion has to be found whether to use the one or the other.

Table 4.2: Example of a possible outcome when predicting the class label for a trial of target subject B, which should be a left-hand trial. When taking the maximum of the probabilities, the decision is a right-hand trial, while the majority of predictions is a left-hand trial.

	P(Left)	P(Right)	Average(Left)	Average(Right)
Source subject C	0.1	0.9	**0.625**	0.375
Source subject D	0.8	0.2	Maximum(Left)	Maximum(Right)
Source subject E	0.8	0.2	0.8	**0.9**
Source subject G	0.8	0.2		

In **Algorithm 1**, the set-up of the new Transfer Learning technique is described. Dependant on whether a classifier is biased or if it's validation accuracy is low, the source subject is removed from the decision making process. These two specifications will be explained later.

The following variables and methods are used in Algorithm 1:

- D_c: training data from the target subject.
- D_t: new test data from the target subject.
- D_k, k $\in \{0, ..., N_s\}$: previously recorded data from the source subjects, with N_s the total amount of source subjects.
- *CalculateM(X,Y)* constructs the transformation matrix M_Y using the estimated covariance matrices of subject X and Y to minimize their distribution differences, according to Formula 2.29.
- *ApplySource(X,Y, M_Y)* will calculate the validation accuracies when transforming the data of subject Y, using M_Y to minimize the distribution differences between subject X and Y.

The threshold of 70% for validation accuracies was chosen by comparing results when varying this parameter from 60 to 90% in steps of 5%.
A classifier is labeled as biased if its predictions are consistently the same. How a classifier is determined to be biased is illustrated with an example in Figure 4.5.
In this figure, 20 calibration trials of subject B are available. To see whether classifier C is a biased classifier, data of other source subjects is used.

In **step 1**, data of the subsequent source subjects are adapted towards the available calibration data, here D', E' and G'. In this way, more data is constructed with the same distribution as target subject B.

In **step 2**, the testing of the BCI is mimiced by transforming D', E' and G' (to minimise distribution differences with subject C) towards D", E" and G", on which the BCI is tested in **step 3**, after training on the 200 trials of source subject C.

Algorithm 1 New Transfer Learning technique

```
Determine the transformation matrices and validation accuracies using the calibration data $D_c$
BiasedClassifier()                          ▷ Will set $Flag_{Biased}$ = True when biased
while $k \leq N_s$ do
    $M_k$ = CalculateM($D_c, D_k$)
    ValAcc = ApplySource($D_c, D_k, M_k$)
    if ValAcc $\leq 0.7$ then
        $Flag_{Validation}(k)$ = True
    end if
end while

Determine which source subjects of $D_k$ should be left out, and whether the probabilities should be averaged (standard) or use the maximum probability of the remaining source subjects (if maxProb = True).
if (# $Flag_{Validation}$ == True) $\geq N_s/2$ then
    if (# $Flag_{Biased}$ == True) <$N_s/2$ then
        Remove the biased source subject from $D_s$
    end if
    maxProb = True
else
    Remove the source subject with $Flag_{Validation}$ = True
    if (# $Flag_{Biased}$ == True) <$N_s/2$ then
        Remove the biased source subject from $D_s$
    else
        maxProb = True
    end if
end if

Determine the test accuracy with $D_t$
while $k \leq N_s$ do
    Probabilities(k) = ApplySource($D_t, D_k, M_k$)
end while
if maxProb == True then
    Use the maximum probability of Probabilities(k)
else
    Average the probabilities of Probabilities(k)
end if
```

The BCI is trained according to the same principles as for DSA, using the same amount of calibration trials from respectively D, E and G, with subject C serving as the source subject. For testing, 40 trials are used. If, for more than half of the temporary classifiers, more than 90% of the trials are labeled the same, the source subject C is flagged as biased.

This threshold of 90% was chosen by comparing results when varying this parameter from 70 to 95% in steps of 5%.

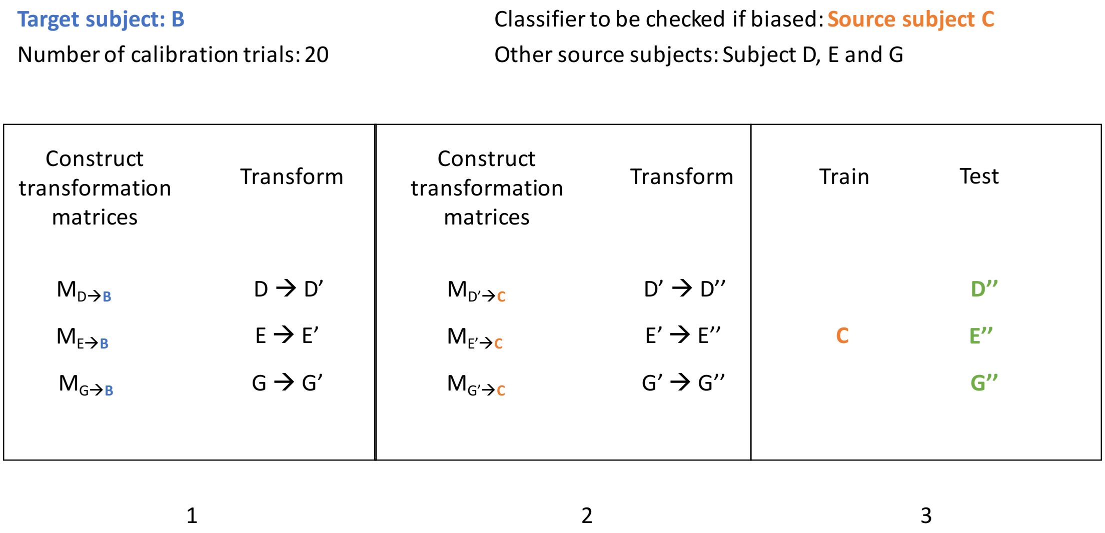


Figure 4.5: From target subject B, 20 calibration trials are available. For the transformation between target subject B and the source subjects in step 1 and the transformation in step 2, the same amount of trials from the source subjects are used.

Chapter 5

Experimental Results

The results of the new Transfer Learning technique are plotted in Figure 5.1. The new method is indicated with a red solid line.

Based on Figure 5.1a, the new method could be labeled as the best performing one, but when looking subject-specific, the results vary.

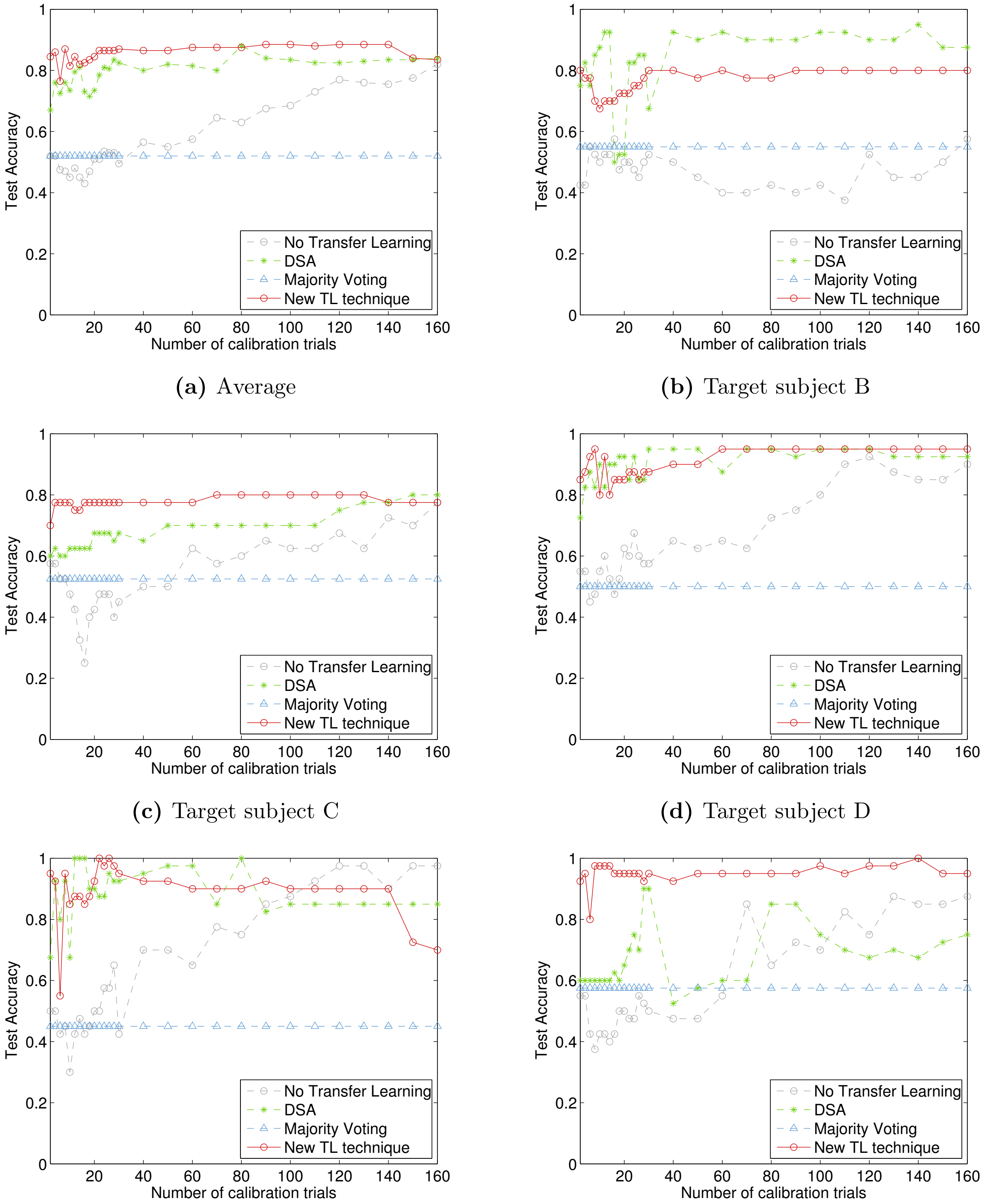
For subject C and subject G, the new Transfer Learning technique scores best for every amount of calibration data. For subject G, there is a slight decrease in performance for 6 calibration trials, but the performance is not as dependant on the amount of calibration trials as in DSA.

For subject B, the new method's accuracy mostly stays below the one of DSA, but at least it doesn't suddenly drop to accuracies below 60%.

For subject D, the results are not optimal with few calibration data. One can see that the accuracies for the new method stay above 80%, but the variation from 2 to 20 calibration trials is high. This may be due to the discrimination based on low validation accuracies or biased classifiers, not being able to point the best source subject/subjects to use, as DSA sometimes results in higher accuracies up until 50 calibration trials.

For subject E, the same conclusion can be made with regards to the performance of DSA, with the difference of the new method being more stable than DSA (except for the peak in the beginning).

In Figure 5.2, the same results are shown, but only the experiments using 40 calibration trials or less, and focusing on the range of test accuracies above 50%. The numeric results can be found in Appendix C.



(a) Average

(b) Target subject B

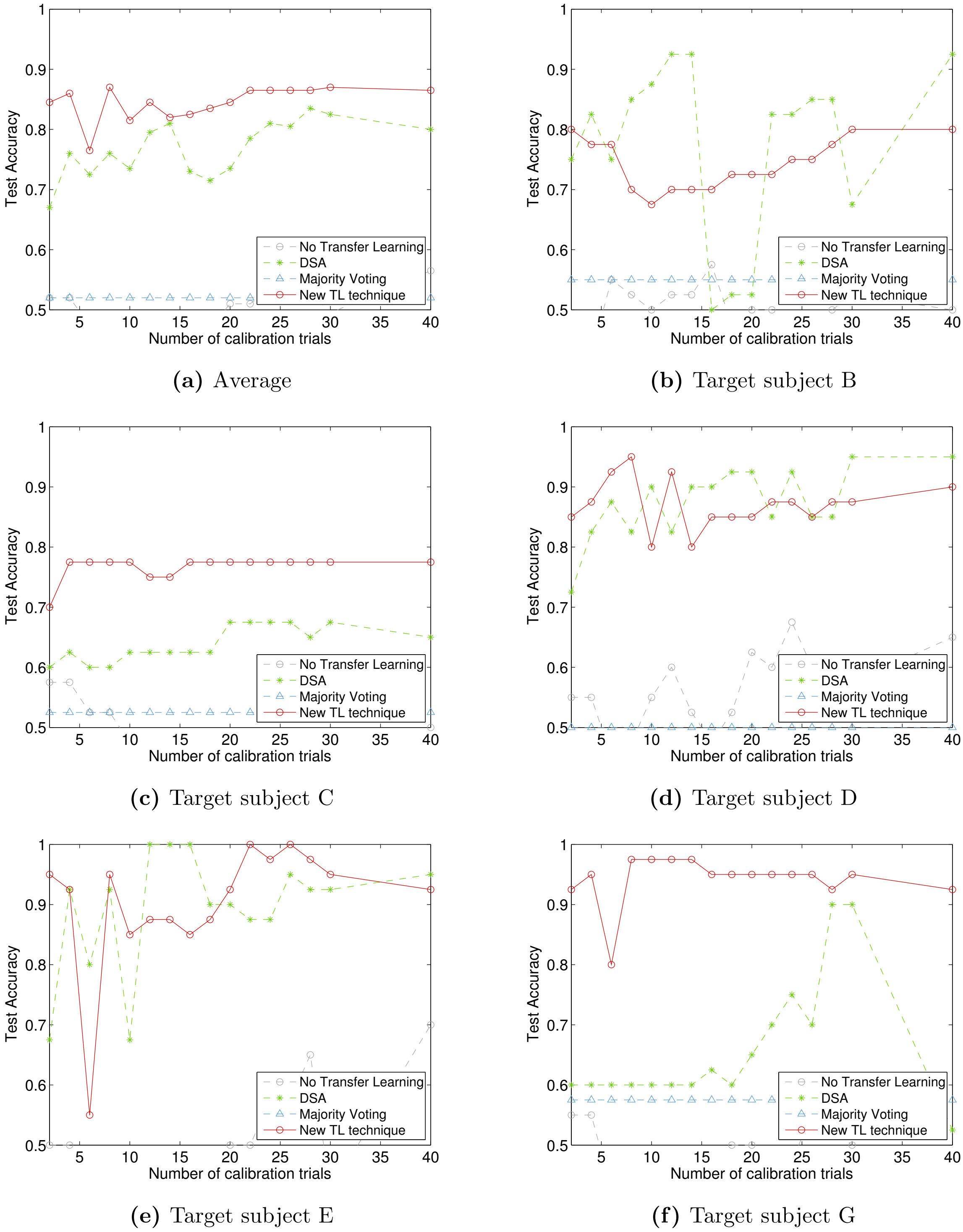
(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 5.1: Comparison of No Transfer Learning, Majority Voting, DSA and the new Transfer Learning technique.



(a) Average

(b) Target subject B

(c) Target subject C

(d) Target subject D

(e) Target subject E

(f) Target subject G

Figure 5.2: Comparison of No Transfer Learning, Majority Voting, DSA and the new Transfer Learning technique, focused on the experiments done with 40 calibration trials or less.

Chapter 6

Conclusion and future work

By using Brain-Computer Interfaces, people who have lost the ability to control their limbs are given the opportunity to steer an application using Motor Imagery, but due to interperson and intersession differences, the BCI needs to adapt to the new user's brain to be able to correctly classify their thoughts, for every use of the application. This calibration could take up to 20 - 30 minutes, giving a lot of room for improvement.

The difficulty of constructing a good BCI was proven by the low testing accuracies that were reached when constructing a person-specific BCI, even though sufficient training data was available. As for some targets the testing accuracy didn't even reach 70%, but for some users the BCI could reach 90% performance, the system is clearly user-dependent.

Averaging the optimal settings of the person-specific BCI's resulted in higher accuracies for some targets, but on average, no significant difference was noted (p=0.45). Averaging optimal settings when only using frequency bands common in Motor Imagery classification (the alpha and beta band), could also lead to higher accuracies for some targets, but with these results not scoring significantly worse (p=0.33), it is an indication for further use of this frequency band only.
These first experiments gave rise to two problems: the system being very user-dependent and the need for a lot of calibration data. These findings dictate the fact that advanced techniques will be needed to guarantee a good performance of the BCI, independent of the user performing the Motor Imagery task and preferably with as little calibration data as possible.

In this attempt, several experiments were done using different Transfer Learning techniques. By comparing different Transfer Learning techniques, varying from naive methods as averaging probabilities and majority voting, to methods that transform the test data according to covariance matrices, it became clear that none of these methods could be labeled as the one that would always result in the highest performance. When getting to the bottom of the techniques that performed significantly better than when not using Transfer Learning at all, the need for improvement of DSA arose.

This shortcoming of DSA was the baseline for the construction of a new method. The main problem of the current Transfer Learning techniques is finding the best subject from which information can be transferred, or in some cases, the best subjects. In the case of multiple subjects being a good fit for the job, the outputs of the classifiers could either be averaged or the maximum could be taken. Experimenting with these ideas, showed that neither one of these approaches induces the highest possible performance for every amount of calibration data. Where one method performs bad, the other one will perform well.

These findings led to a new Transfer Learning technique that combined both averaging the outputs of different classifiers and taking the maximum probability produced by a classifier. The amount of classifiers on which Transfer Learning is applied and the preferred method used, is based on a selection criterion.

The strength of this new Transfer Learning technique is its robustness in comparison to DSA. Where the performance of DSA suddenly decreases when new calibration data becomes available, the new method is less sensitive to alterations in calibration data. With regards to an application of a BCI to steer a wheelchair, this advantage of the new method is an important aspect in sending the wheelchair towards the right direction, even if the user was confused or distracted for a short period of time. The new method might not always be stated as the best method for every amount of calibration data, but by further alterations in the selection criterion, the results can be promising.

At the moment, as seen in Figure 5.1, with a single exception for 6 calibration trials when testing for target E, the test accuracy of the new Transfer Learning technique, never drops below 67%. The most important gain in performance is that, when only having 4 calibration trials available, the test accuracy is minimally 77%. With respect to, when having 4 calibration trials available, only reaching 52±6% on average when not applying Transfer Learning. In the worst case scenario, there is even an improvement of 38%. On average, when using the final method and 4 calibration trials, the test accuracy is 86±8%. This clearly manifests that, by applying Transfer Learning and reducing the amount of calibration, there was absolutely no reduction in performance.

With regards to the goal of reducing the calibration time as much as possible, this requirement is fullfilled, as the amount of trials is reduced to the minimal amount of trials possible (one trial of each class), still reaching average test accuracies of 85±10%.

In the construction of a new method, based on comparisons between old and new techniques, it might not be overlooked that an optimal method was constructed based on data from only 5 subjects. To work around this restriction, the steps taken in the process to construct a new Transfer Learning technique were not based on the averages of the performance of these 5 subjects, but on target-specific results. As these results still might depend on the specific data used in the experiments, it might be useful to go over the same steps and reasonings, but with a larger or a different dataset.

Another path that can be explored is by adding more Motor Imagery tasks to the classification problem, to study the generalisation properties of the Transfer Learning method. For the application of steering a wheelchair, this could be the addition of a stop and backwards command. The removal of restricting the BCI to be synchronous would also be a big step in the right direction, as it gives the possibility to the user to communicate at their own pace. But before exploiting these options, further thought should be given on how to eliminate the calibration procedure and making it an online application that learns from the user, while already performing the first movements, preferably in the right direction.

Appendix A

Hyperparameters for a first BCI

To determine the best set of hyperparameters, the validation matrices are averaged over all subjects. This resulted in the validation matrix as in Table A.1. The rows indicate how much splits in the frequency band are used, the columns indicate the amount of CSP-filters used.
These validation accuracies are averages, so the low average validation accuracies are due to some subjects for whom it is hard to construct a proper BCI (like for subject B, who never results in validation accuracies higher than 0.543). On the other hand, subjects like D and G, easily reach validation accuracies of 0.8 or higher.
On average, according to a validation accuracy of 0.693, the best hyperparameters are obtained, as highlighted, when splitting the frequency band in 8 equal splits and using 3 CSP filter pairs.

	1	2	3	4	5	6	7	8	9
0	0.467	0.477	0.472	0.485	0.478	0.489	0.489	0.489	0.487
1	0.515	0.514	0.541	0.563	0.555	0.549	0.537	0.540	0.546
2	0.564	0.587	0.576	0.593	0.581	0.568	0.55	0.542	0.559
3	0.618	0.631	0.634	0.635	0.621	0.615	0.603	0.602	0.615
4	0.604	0.667	0.645	0.645	0.641	0.622	0.646	0.632	0.610
5	0.637	0.644	0.652	0.642	0.620	0.629	0.614	0.610	0.622
6	0.662	0.678	0.683	0.668	0.652	0.647	0.645	0.640	0.636
7	0.654	0.680	0.682	0.6625	0.651	0.65	0.635	0.632	0.630
8	0.657	0.692	**0.693**	0.651	0.649	0.656	0.634	0.622	0.629
9	0.655	0.690	0.685	0.647	0.647	0.657	0.637	0.630	0.622

Table A.1: Validation accuracies averaged over all source subjects. The rows indicate how much splits in the frequencyband are used, the columns indicate the amount of CSP-filters used.

The test and train accuracies reached with person-specific optimal features are seen in Table A.2. The test accuracies obtained when using person-specific optimal features, averaged optimal features and when using only the alpha and beta band are seen in Table A.3.

	Test Accuracy	Train Accuracy
Subject A	0.500	0.969
Subject B	0.525	0.925
Subject C	0.600	1.000
Subject D	0.775	0.975
Subject E	0.900	0.994
Subject F	0.575	0.988
Subject G	0.800	1.000
Average	0.668 ± 0.217	0.979 ± 0.318

Table A.2: Train and test accuracies reached when using the person-specific optimal features.

	Personal best	Averaged best	Alpha and Beta band
Subject A	0.500	0.675	0.475
Subject B	0.525	0.575	0.600
Subject C	0.600	0.575	0.750
Subject D	0.775	0.950	0.800
Subject E	0.900	0.850	0.950
Subject F	0.575	0.575	0.600
Subject G	0.800	0.700	0.800
Average	0.668 ± 0.217	0.700 ± 0.224	0.711 ± 0.230

Table A.3: Comparison of the test accuracies when using person-specific optimised features, an averaged optimal amount of features and when only using data from the alpha and beta band.

Appendix B

Covariance Shrinkage: Subset selection and regularisation parameter

The subset that was chosen by the Subject Selection Algorithm as in Chapter 2.7.2, is illustrated in Table B.1. In the columns, the source subjects chosen for the corresponding target subject in the header are shown. The corresponding regularisation parameter for each experiment is added in Table B.2. In both tables, each row represents a new experiment using a bigger amount of calibration trials.

In this table, the regularisation parameter when using the complete set of source subjects is added for comparison. It is clear that there is no big difference in regularisation parameter when a different subset is used. For subject C, D and E, the regularisation parameter is mostly zero, so an altering in subset can not have an influence on the test accuracy. If the regularisation parameter is non-zero, the subset chosen by the Subject Selection Algorithm is nearly always a set of three out of four subjects, and as visible in Figure 4.1, for target subject B, the addition of a fourth source subject lowers the test accuracy, where for target subject G, the test accuracy slightly increases when adding a fourth source subject.

	Subject B	Subject C	Subject D	Subject E	Subject G
10	C	G	B, C	D, G	C, D, E
20		D, E, G		B, C	B, C, D, E
30	D, E, G				B, C, D, E
40	D, E, G	D, E, G			B, C, D, E
50	D, E, G	D, E, G			C, D, E
60	D, E, G				C, D, E
70	C, D, E				C, D, E
80	C, D, E				C, D, E
90	C, D, E				C, D, E
100	C, D, E				C, D, E
110	C, D, E				C, D, E
120	C, D, E				C, D, E
130	C, D, E				C, D, E
140	C, D, E				C, D, E
150	C, D, E				C, D, E
160	C, D, E				C, D, E

Table B.1: The subset of source subjects as chosen by the Subject Selection Algorithm (see Chapter 2.7.2).

	Subject B		Subject C		Subject D		Subject E		Subject G	
	S	NS	S	NS	S	NS	S	NS	S	NS
10	1.000	0.167	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.778
20	0.000	0.000	1.000	1.000	0.000	1.000	0.143	1.000	1.000	0.769
30	1.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000
40	1.000	1.000	0.087	1.000	0.000	0.000	0.000	0.000	1.000	1.000
50	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.667	0.615
60	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.606
70	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.324	0.308
80	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.529	0.486
90	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.458	0.500
100	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.407	0.407
110	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.321	0.367
120	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.222	0.364
130	0.132	0.134	0.000	0.000	0.000	0.000	0.000	0.000	0.093	0.104
140	0.029	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.061	0.059
150	0.012	0.013	0.000	0.000	0.000	0.000	0.000	0.000	0.048	0.050
160	0.057	0.054	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.025

Table B.2: The regularisation parameter when a subset of source subjects is chosen (S) and when all subjects are chosen as source subjects (NS).

Appendix C

Results of the comparison of No Transfer Learning, Majority Voting, DSA and the new method

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.520	0.670	0.520	0.845
4	0.520	0.760	0.520	0.860
6	0.475	0.725	0.520	0.765
8	0.470	0.760	0.520	0.870
10	0.450	0.735	0.520	0.815
12	0.480	0.795	0.520	0.845
14	0.450	0.810	0.520	0.820
16	0.430	0.730	0.520	0.825
18	0.470	0.715	0.520	0.835
20	0.510	0.735	0.520	0.845
22	0.510	0.785	0.520	0.865
24	0.535	0.810	0.520	0.865
26	0.530	0.805	0.520	0.865
28	0.530	0.835	0.520	0.865
30	0.495	0.825	0.520	0.870
40	0.565	0.800	0.520	0.865

Table C.1: Results of a comparison of different Machine Learning techniques, averaged over 5 target subjects.

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.425	0.750	0.550	0.800
4	0.425	0.825	0.550	0.775
6	0.550	0.750	0.550	0.775
8	0.525	0.850	0.550	0.700
10	0.500	0.875	0.550	0.675
12	0.525	0.925	0.550	0.700
14	0.525	0.925	0.550	0.700
16	0.575	0.500	0.550	0.700
18	0.475	0.525	0.550	0.725
20	0.500	0.525	0.550	0.725
22	0.500	0.825	0.550	0.725
24	0.475	0.825	0.550	0.750
26	0.450	0.850	0.550	0.750
28	0.500	0.850	0.550	0.775
30	0.525	0.675	0.550	0.800
40	0.500	0.925	0.550	0.800

Table C.2: Results of a comparison of different Machine Learning techniques, using subject B as target subject.

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.575	0.600	0.525	0.700
4	0.575	0.625	0.525	0.775
6	0.525	0.600	0.525	0.775
8	0.525	0.600	0.525	0.775
10	0.475	0.625	0.525	0.775
12	0.425	0.625	0.525	0.750
14	0.325	0.625	0.525	0.750
16	0.250	0.625	0.525	0.775
18	0.400	0.625	0.525	0.775
20	0.425	0.675	0.525	0.775
22	0.475	0.675	0.525	0.775
24	0.475	0.675	0.525	0.775
26	0.475	0.675	0.525	0.775
28	0.400	0.650	0.525	0.775
30	0.450	0.675	0.525	0.775
40	0.500	0.650	0.525	0.775

Table C.3: Results of a comparison of different Machine Learning techniques, using subject C as target subject.

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.550	0.725	0.500	0.850
4	0.550	0.825	0.500	0.875
6	0.450	0.875	0.500	0.925
8	0.475	0.825	0.500	0.950
10	0.550	0.900	0.500	0.800
12	0.600	0.825	0.500	0.925
14	0.525	0.900	0.500	0.800
16	0.475	0.900	0.500	0.850
18	0.525	0.925	0.500	0.850
20	0.625	0.925	0.500	0.850
22	0.600	0.850	0.500	0.875
24	0.675	0.925	0.500	0.875
26	0.600	0.850	0.500	0.850
28	0.575	0.850	0.500	0.875
30	0.575	0.950	0.500	0.875
40	0.650	0.950	0.500	0.900

Table C.4: Results of a comparison of different Machine Learning techniques, using subject D as target subject.

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.500	0.675	0.450	0.950
4	0.500	0.925	0.450	0.925
6	0.425	0.800	0.450	0.550
8	0.450	0.925	0.450	0.950
10	0.300	0.675	0.450	0.850
12	0.425	1.000	0.450	0.875
14	0.475	1.000	0.450	0.875
16	0.425	1.000	0.450	0.850
18	0.450	0.900	0.450	0.875
20	0.500	0.900	0.450	0.925
22	0.500	0.875	0.450	1.000
24	0.575	0.875	0.450	0.975
26	0.575	0.950	0.450	1.000
28	0.650	0.925	0.450	0.975
30	0.425	0.925	0.450	0.950
40	0.700	0.950	0.450	0.925

Table C.5: Results of a comparison of different Machine Learning techniques, using subject E as target subject.

	No Transfer Learning	DSA	Majority Voting	New TL technique
2	0.550	0.600	0.575	0.925
4	0.550	0.600	0.575	0.950
6	0.425	0.600	0.575	0.800
8	0.375	0.600	0.575	0.975
10	0.425	0.600	0.575	0.975
12	0.425	0.600	0.575	0.975
14	0.400	0.600	0.575	0.975
16	0.425	0.625	0.575	0.950
18	0.500	0.600	0.575	0.950
20	0.500	0.650	0.575	0.950
22	0.475	0.700	0.575	0.950
24	0.475	0.750	0.575	0.950
26	0.550	0.700	0.575	0.950
28	0.525	0.900	0.575	0.925
30	0.500	0.900	0.575	0.950
40	0.475	0.525	0.575	0.925

Table C.6: Results of a comparison of different Machine Learning techniques, using subject G as target subject.

Bibliography

[1] D. J. McFarland and J. R. Wolpaw, “Brain-Computer Interfaces for Communication and Control,” *Communications of The ACM*, vol. 54, no. 5, pp. 60–66, 2011.

[2] J. D. Mitchell and G. D. Borasio, “Amyotrophic lateral sclerosis,” *Lancet*, vol. 369, no. 9578, pp. 2031–41, 2007.

[3] T. Kaufmann, A. Herweg, and A. Kübler, “Toward brain-computer interface based wheelchair control utilizing tactually-evoked event-related potentials,” *Journal of neuroengineering and rehabilitation*, vol. 11, no. 7, pp. 1–17, 2014.

[4] C. Guger, C. Holzner, and C. Groenegress, “Control of a smart home with a brain-computer interface,” *BrainComputer Interface*, pp. 2–6, 2008.

[5] N. Birbaumer and L. G. Cohen, “Brain-computer interfaces: communication and restoration of movement in paralysis,” *The Journal of Physiology*, vol. 579, no. 3, pp. 621–636, 2007.

[6] V. S. Polikov, P. a. Tresco, and W. M. Reichert, “Response of brain tissue to chronically implanted neural electrodes,” *Journal of Neuroscience Methods*, vol. 148, pp. 1–18, 2005.

[7] N. J. Hill, D. Gupta, P. Brunner, A. Gunduz, M. a. Adamo, A. Ritaccio, and G. Schalk, “Recording Human Electrocorticographic (ECoG) Signals for Neuroscientific Research and Real-time Functional Cortical Mapping,” *Journal of Visualized Experiments*, no. 64, pp. 1–5, 2012.

[8] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. N. Rao, and J. G. Ojemann, “Electrocorticography-based brain computer interface–the Seattle experience.,” *IEEE transactions on neural systems and rehabilitation engineering*, vol. 14, no. 2, pp. 194–198, 2006.

[9] M. Lopez-Gordo, D. Morillo, and F. Valle, “Dry EEG Electrodes,” *Sensors*, vol. 14, no. 7, pp. 12847–12870, 2014.

[10] A. Searle and L. Kirkup, “A direct comparison of wet, dry and insulating bioelectric recording electrodes,” *Physiological Measurement*, vol. 21, no. 2, pp. 271–283, 2000.

[11] H. H. Jasper, “The ten-twenty electrode system of the International Federation,” *Electroencephalography and Clinical Neurophysiology*, vol. 10, no. 2, pp. 371–375, 1958.

[12] E. Marieb and K. Hoehn, *Human Anatomy & Physiology.* 2006.

[13] S. Singh, “Magnetoencephalography: Basic principles,” *Annals of Indian Academy of Neurology*, vol. 17, no. 5, p. 107, 2014.

[14] J. Mellinger, G. Schalk, C. Braun, H. Preissl, W. Rosenstiel, N. Birbaumer, and A. Kübler, “An MEG-based brain-computer interface (BCI).,” *NeuroImage*, vol. 36, no. 3, pp. 581–93, 2007.

[15] O. J. Arthurs and S. Boniface, “How well do we understand the neural origins of the fMRI BOLD signal?,” *Trends Neurosciences*, vol. 25, no. 1, pp. 27–31, 2002.

[16] N. Weiskopf, K. Mathiak, S. W. Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer, “Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI),” *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 966–970, 2004.

[17] M. O. Sokunbi, D. E. J. Linden, I. Habes, S. Johnston, and N. Ihssen, “Real-time fMRI brain-computer interface: development of a ”motivational feedback” subsystem for the regulation of visual cue reactivity,” *Frontiers in behavioral neuroscience*, vol. 8, no. 392, p. 10, 2014.

[18] J. Sulzer, S. Haller, F. Scharnowski, N. Weiskopf, N. Birbaumer, M. Blefari, A. Bruehl, L. Cohen, R. DeCharms, R. Gassert, R. Goebel, U. Herwig, S. LaConte, D. Linden, A. Luft, E. Seifritz, and R. Sitaram, “Real-time fMRI neurofeedback: Progress and challenges,” *NeuroImage*, vol. 76, pp. 386–399, 2013.

[19] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “A comparison of classification techniques for the P300 Speller.,” *Journal of neural engineering*, vol. 3, no. 4, pp. 299–305, 2006.

[20] S. L. Bressler and M. Ding, “Event-Related Potentials,” in *Wiley Encyclopedia of Biomedical Engineering*, pp. 412–415, 2006.

[21] M. Jeannerod, “Mental imagery in the motor context,” *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, 1995.

[22] G. Pfurtscheller and C. Neuper, “Motor imagery and direct brain- computer communication,” *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[23] G. Townsend, B. Graimann, and G. Pfurtscheller, “Continuous EEG classification during motor imagery - Simulation of an asynchronous BCI,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 2, pp. 258–265, 2004.

[24] P. Sovka and M. K. Ilek, “Overcoming Inter-Subject Variability In BCI Using EEG-Based Identification,” *Radioengineering*, vol. 23, no. 1, pp. 266–273, 2014.

[25] M. Arvaneh, C. Guan, K. K. A. Ang, and C. Quek, “EEG Data Space Adaptation to Reduce Intersession Nonstationarity in Brain-Computer Interface,” *Neural Computation*, vol. 25, no. 8, pp. 2146–2171, 2013.

[26] V. Jayaram and M. Grosse-Wentrup, “Transfer Learning in Brain-Computer Interfaces,” pp. 1–20, 2015.

[27] M. Bishop, “Introduction,” in *Pattern Recognition and Machine Learning*, pp. 1–4, Springer-Verlag New York, Inc., 2006.

[28] P. Viola, O. M. Way, and M. J. Jones, “Robust Real-Time Face Detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[29] M. Bishop, *Pattern Recognition and Machine Learning.* Springer-Verlag New York, Inc., 2006.

[30] R. Bellman, *Dynamic Programming.* 1957.

[31] S. J. Raudys and A. K. Jain, “Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[32] B. Blankertz, “BCI Competition IV - Dataset 1,” 2008.

[33] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,” *Clinical Neurophysiology*, vol. 112, pp. 713–719, 2001.

[34] J. R. Hughes, “Gamma, fast, and ultrafast waves of the brain: Their relationships with epilepsy and behavior,” *Epilepsy & Behavior*, vol. 13, pp. 25–31, 2008.

[35] P. L. Nunez, R. Srinivasan, A. F. Westdorp, and R. S. Wijesinghe, “EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales,” *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 499–515, 1997.

[36] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, “Optimizing spatial filters for robust EEG single-trial analysis,” *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[37] D. J. Mcfarland, L. M. Mccane, S. V. David, and J. R. Wolpaw, “Spatial filter selection for EEG-based communication,” *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 386–394, 1997.

[38] G. Bernhard, Z. A. Brendan, and P. Gert, *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction.* 2010.

[39] H. Ramoser, J. Müller-gerking, and G. Pfurtscheller, “Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement,” *IEEE Transaction on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[40] G. Pfurtscheller, H. Flyvbjerg, and J. Mu, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clinical Neurophysiology*, vol. 110, pp. 787–798, 1999.

[41] Z. Y. Chin, K. K. Ang, C. Wang, C. Guan, and H. Zhang, "Multi-Class Filter Bank Common Spatial Pattern for Four-Class Motor Imagery BCI," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 571–574, 2009.

[42] F. Lotte and M. Congedo, "A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces," *Journal of Neural Engineering*, vol. 4, pp. 1–24, 2007.

[43] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-r. Müller, "NeuroImage Single-trial analysis and classification of ERP components A tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[44] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, 2004.

[45] S. V. Stehman, "Selecting and Interpreting Measures of Thematic Classification Accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[46] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627–635, 2013.

[47] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, no. 2, pp. 614–617, 2010.

[48] M. Arvaneh, I. Robertson, and T. E. Ward, "Subject-to-Subject Adaptation to Reduce Calibration Time in Motor Imagery-based Brain-Computer Interface," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6501–6504, 2014.

[49] A. Kübler, N. Neumann, B. Wilhelm, T. Hinterberger, and N. Birbaumer, "Predictability of Brain-Computer Communication," *Journal of Psychophysiology*, vol. 18, no. 2-3, pp. 121–129, 2004.