university of groningen

faculty of science and engineering

Transfer Learning for Motor Imagery Classification in Low-Cost Brain-computer Interface Systems

Lars Cordes

university of groningen / faculty of science and engineering

University of Groningen

Transfer Learning for Motor Imagery Classification in Low-Cost Brain-computer Interface Systems

Master's Thesis

Under the supervision of
dr. Jelmer Borst (Artificial Intelligence, University of Groningen)
Maruf Dhali, MSc (Artificial Intelligence, University of Groningen)

Lars Cordes (s2965518)

November 22, 2022

All code used in the making of this thesis is available at github.com/lcordes/bci.
Collected experimental data is available at osf.io/fb9pu/.

Contents

Abstract

Brain-computer interfaces (BCIs) enable people with impaired mobility to interact with their environment. BCIs translate brain activity measured through electroencephalography (EEG) into control commands, such as the movement of a prosthesis. Two challenges to the general adoption of this technology are the high cost of medical-grade EEG devices and the high variability of the EEG signal between different users. This study introduces a BCI based on a low-cost EEG device that uses transfer learning between users to reduce the need for individual calibration. This BCI architecture was tested and evaluated on two data sets collected using low-cost EEG as well as a benchmark data set collected using medical-grade EEG. On all data sets, a transfer learning approach employing Euclidean alignment significantly outperformed a baseline system. However, accuracies for the low-cost EEG data sets were generally lower compared to the medical-grade EEG benchmark data set. We conclude that the use of transfer learning in BCI tasks should be encouraged and that the affordability of low-cost EEG devices might not make up for the decrease in data quality.

1 Introduction

Brain-computer Interfaces (BCIs) provide a way to translate the brain's electrical activity into instructions which a computer can act upon. BCIs most often record brain activity through electroencephalography (EEG). Compared to other commonly used neuroimaging techniques such as functional magnetic resonance imaging, EEG affords high temporal resolution which is necessary for real-time user-computer interaction in BCIs. EEG also generally requires less preparation and users are less restrained in their movements during measurements. Over the past decades, EEG-based BCIs have been applied to a broad range of tasks, acting as controllers of physical devices, input devices for virtual environments as well as therapeutic aids for neurorehabilitation (Muller-Putz & Pfurtscheller, 2007; Marshall, Coyle, Wilson, & Callaghan, 2013; Daly & Wolpaw, 2008). In the clinical domain, BCIs enable motor-impaired patients to interact with their environment (Birbaumer et al., 1999).

While the effectiveness of BCIs for patient care has been well-established, the medical-grade EEG devices commonly employed in research bring with them several obstacles to large-scale adoption. For one, medical-grade EEG devices are generally too costly to be employed in individual residential care settings (Duvinage et al., 2013). Additionally, these devices also require the use of salt solution-based gels which are applied between the measurement electrodes and the patient's scalp to improve conductance. As a result, additional time is required before data can be recorded and users need to remove the gel after use. For BCI applications, intermittent-use scenarios would require the gel to be applied before each session, whereas continuous-use scenarios would eventually encounter problems due to the gel drying out. Medical-grade EEG devices are also generally stationary, as the headset, amplifier and processing computer are connected through cables. Especially in control scenarios such as steering a wheelchair, this greatly reduces flexibility of the BCI system. Another difficulty in the design of robust BCI systems is their use across different recording sessions and different users. Due to the high inter-person and inter-session variability of the EEG signal, traditional BCI systems generalize poorly across users and, therefore, require extensive calibration for each new user or session (Ahn & Jun, 2015).

This study introduces a BCI system which addresses several of the aforementioned limitations. First, a low-cost EEG-device is used to record brain activity with dry electrodes, removing the need for conductive gel, while also interfacing wirelessly with the processing computer. Second, a transfer learning paradigm is used to allow for better generalization between users reducing the need for user-specific calibration. Motor imagery (MI), a BCI paradigm aiming to detect different imagined movements in the EEG signal, is chosen for BCI control. Three types of imagined limb movements are translated into three distinct computer commands.

1.1 Low-Cost BCIs

Several commercially available alternatives to medical-grade EEG devices exist, either described as low-cost or consumer-grade EEG instruments in the literature. Systematic reviews generally identify headsets by four companies, namely InteraXon, Neurosky, OpenBCI and Emotiv (Sawangjai, Hompoonsup, Leelaarporn, Kongwudhikunakorn, & Wilaiprasitporn, 2019; Maskeliunas, Damasevicius, Martisius, & Vasiljevas, 2016; LaRocco, Le, & Paeng, 2020). Their product lines differ on several key factors, affecting the type of research paradigm for which they are suitable. The number of electrodes differ per headset, ranging from one electrode (Neurosky's MindWave), to four electrodes (InteraXon's Muse), 14 electrodes (Emotiv's Epoc) up to 16 electrodes (OpenBCI's Ultracortex). For the majority of these headsets, electrode locations are predetermined and cannot be changed by the user, except for the OpenBCI Ultracortex. Here measurement electrodes can be flexibly placed across

35 potential locations (in line with the 10-20 system). The number of electrodes is also generally reflective of the price point, with Neurosky's one-channel headset being the cheapest and OpenBCI's 16-channel headset being the most expensive consumer-grade headset (LaRocco et al., 2020). Most of these headsets employ a dry-electrode setup, except for Emotiv's Epoc. Here, a saline solution is used to soak electrodes, compromising between the improved user comfort of completely dry electrodes and the improved conductance of gel-based electrodes.

Differences in headset characteristics also influence which mental constructs can be studied using a specific device. Headsets featuring fewer electrodes are generally used for research focused on frequency bands such as drowsiness-detection and general concentration. In contrast, headsets with a larger number of electrodes have been used for more complex designs such as event-related potential (ERP) paradigms (Sawangjai et al., 2019). ERPs are generally used to investigate neural activity occurring systematically in response to a presented stimulus (Luck, 2014). A large factor in headset choice is whether the headset affords electrode locations close to the neural sources of the mental construct under investigation. As an example, motor imagery, the BCI paradigm used in this study, is based on the detection of differential activity in the motor cortex. A potential low-cost EEG device consequently requires electrodes located near the centre of the scalp which only the OpenBCI headset affords. As a consequence, the OpenBCI Ultracortex headset was chosen as the low-cost EEG device used for this study.

1.2 BCI Paradigms

A large number of different BCI architectures and paradigms have been introduced in the BCI literature, differing in their design and proposed applications. This section gives an overview of the most commonly used paradigms and provides arguments for why motor imagery is chosen as the paradigm used in this study. BCIs can be differentiated as being invasive or non-invasive, where the former involves the surgical implantation of measurement electrodes and the latter employs external electrodes placed on the scalp (Steyrl, Kobler, Müller-Putz, et al., 2016). While invasive designs offer higher resolution due to electrodes being closer to the neural sources of brain activity, there are obvious disqualifying factors for their widespread consumer adoption. The majority of BCI designs are, therefore, non-invasive.

Concerning the implementation, several dominant paradigms focusing on different EEG patterns have emerged. BCIs based on steady-state visual-evoked potentials (SSVEPs) capitalize on the fact that the brain, when observing visual stimuli flashing at a certain frequency, will produce neural activity at the same frequency in the visual cortex (Zhu, Bieger, Garcia Molina, & Aarts, 2010). By presenting stimuli flashing at different frequencies on a screen, such as letters on a virtual keyboard, one can then decode which stimulus a user is attending by comparing the measured brain activity frequency with the respective stimulus frequencies. While this approach works well for scenarios in which BCI control is purely stimulus-driven, the user's actions are limited to the presented stimuli. As such, the user can only exert control by choosing which stimulus to attend to, rather than producing specific neural signals of their own volition.

Another commonly employed BCI paradigm is the P300 speller (Farwell & Donchin, 1988). This approach makes use of the well-studied P300 event-related potential (ERP) which appears approximately 300 ms after the onset of a stimulus (Sutton, Braren, Zubin, & John, 1965). The P300 has been historically studied using the oddball paradigm in which participants observe a sequence of common recurring stimuli (Squires, Squires, & Hillyard, 1975). At times a low-probability, "oddball" stimulus is interjected into the stimulus sequence which elicits the P300 ERP. This observable reaction to the onset of rare low-probability stimuli is employed in the P300 speller paradigm. Users

are presented with a matrix of letters and told to attend to the letter they mean to communicate. Rows and columns are then repeatedly highlighted through a flash, which elicits a differentiable P300 if the currently highlighted row or column contains the attended symbol (Farwell & Donchin, 1988). In this way, motor-impaired users are able to spell out intentions. However, user actions are again limited to options presented as stimuli in the environment, as is the case for the SSVEP paradigm.

An alternative to externally guided approaches is the motor imagery paradigm (MI) which uses imagined movements as its control patterns. Just as activity in the motor cortex is observable for different body movements, imagined movements also evoke neural activity (Miller et al., 2010). Moreover, the motor cortex features a spatial map of different extremities, where the left side of the body is represented in the right hemisphere and vice versa (Penfield & Boldrey, 1937). Thus, different measurement electrodes distributed over the motor cortex can then observe differential activity when, for example, one thinks about moving their left hand compared to their right hand. Specifically, event-related desynchronization can be observed in brain areas related to movement planning involving particular extremities in the mu frequency band (8 to 13 Hz) as well as a subrange of the beta frequency band (18 to 25 Hz) (Neuper & Pfurtscheller, 2001). Capitalizing on these EEG patterns, MI-based BCIs aim to detect different types of motor imagery and then translate the result to different instructions for a computer.

In contrast to SSVEP and P300 paradigms, MI-BCIs operate without the presence of stimuli in the environment. This study aims to increase the applicability of BCI systems and thereby facilitate their large-scale adoption. Thus, we found an MI approach to afford the most flexibility across different application scenarios. An additional advantage of MI-BCIs is the existence of several benchmark data sets in the literature which allow for direct comparison and evaluation of new algorithms (Sajda, Gerson, Muller, Blankertz, & Parra, 2003; Blankertz et al., 2004, 2006).

Several MI-BCIs using an OpenBCI Ultracortex headset as an EEG recording device have been introduced. However, differences in the type of imagery investigated and problems with low sample size make direct performance comparisons difficult. Sterk (2022), for example, employed the same set of motor imagery examples used in this study for two participants. Saragih, Basyiri, and Raihan (2022) and Peterson, Galván, Hernández, Saavedra, and Spies (2022) both investigated binary classification of the dominant hand and a rest condition for one participant and ten participants, respectively, and Shen et al. (2022) contrasted left-hand with right-hand imagery in eight participants. Results for these studies will be contrasted with the current study in more detail in Section 4.

1.3 Feature Extraction and Classification

Key to the feasibility of BCIs are reliable methods of detecting EEG patterns related to the target construct of interest. Since its inception, the MI literature has overwhelmingly employed machine learning solutions for this task (Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007; Lotte et al., 2018). As such, labelled examples of motor imagery are used to train a classifier which can predict the specific type (i.e. label) of new motor imagery examples. Moreover, the use of machine learning within the BCI field is not limited to classification and also extends to other parts of the BCI pipeline. Several supervised feature extraction techniques, for example, employ machine learning to learn data transformations which improve class separation (Krusienski, McFarland, Principe, & Wolpaw, 2012). Due to the low signal-to-noise ratio of the EEG signal, well-performing feature extraction methods are crucial to enable successful subsequent classification (Subha, Joseph, Acharya U, Lim, et al., 2010).

Generally, feature extraction involves spectral as well as spatial filtering. Whereas spectral filtering relates to specific frequency bands, such as the aforementioned mu- and beta-rhythms, spatial filtering involves different spatial data sources, i.e. EEG channels. A large number of studies employ

the Common Spatial Patterns (CSP) algorithm for spatial feature extraction (Koles, Lazar, & Zhou, 1990; Ramoser, Muller-Gerking, & Pfurtscheller, 2000). This approach involves learning a set of spatial filters in a supervised manner, such that differences in intra-class variance are maximized. Once learned, novel data can then be transformed using these spatial filters.

Lotte et al. (2018) conducted a comprehensive review of different approaches to EEG classification in the literature. Commonly employed machine learning classifiers, such as Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and Random Forest were found to perform well on BCI tasks. The combination of CSP feature extraction and LDA classification especially constitutes one of the most commonly used approaches for BCI pipelines. Recent work has also stressed the importance of regularization for both CSP as well as LDA. One approach to this is covariance shrinkage, where the intra-class covariance matrices are regularized by applying shrinkage to their coefficients (Lotte et al., 2018). Especially when working with limited data, as is often the case for BCI studies, the estimated covariance matrices may be distorted by extreme coefficients (Lotte, 2015; Ledoit & Wolf, 2004). Approaches such as shrinkage attenuate the effect of these extreme coefficients.

Recently a large number of BCI designs have also started using Riemannian geometry-based methods (Congedo, Barachant, & Bhatia, 2017). Here, data transformations are used to transport EEG data from Euclidean space to Riemannian space. Feature extraction and classification are then conducted in this new space. Data in Riemannian space occupies a curved surface, referred to as the manifold. As distances between data points on the manifold vary more smoothly, classification in this space tends to be more resistant to noise in the data (Lotte et al., 2018). Riemannian methods also seem to better account for the non-stationarity of the EEG signal, which might explain its recent gain in popularity (Yger, Berar, & Lotte, 2016). An example of classification in Riemannian space is the Minimum Distance to Riemannian Mean (MDRM) classifier (Barachant, Bonnet, Congedo, & Jutten, 2011). MDRM first computes the Riemannian Mean of the covariance matrices for all examples of a specific class. A new example is then assigned to the class whose Riemannian mean is closest to the covariance matrix of the new example.

Finally, the general increase in deep learning-based methods seen in the machine learning literature is also present in the BCI literature. However, deep learning methods have so far failed to achieve state-of-the-art performance, potentially due to the generally small amount of training examples in BCI data sets (Lotte et al., 2018). Recent studies have also increasingly focused on adaptive classifiers and transfer learning. Adaptive classifiers are not only trained offline but continue their learning process when encountering novel data (Sun & Zhou, 2014). Transfer learning, on the other hand, refers to a set of machine learning techniques focused on transferring knowledge learnt on one task to a related, but different task (Wu, Xu, & Lu, 2020; Jayaram, Alamgir, Altun, Scholkopf, & Grosse-Wentrup, 2016). Both paradigms are uniquely suited to increase the generalization ability of BCIs. For an overview of how the different BCI components introduced in this section form an overall pipeline, refer to Figure 2.

1.4 Transfer Learning

A central challenge for the design of robust BCI systems is their use across different recording sessions and across different users. Due to the high inter-person and inter-session variability of the EEG signal, traditional machine learning-based approaches often do not generalize well between sessions and users (Ahn & Jun, 2015). Therefore, well-performing MI-BCI systems generally require extensive calibration for new users.

Transfer learning provides a well-suited approach to the problem of between-user and between-session generalization. While the probability distributions underlying motor imagery of different

users vary to some degree, the machine learning task is the same. Transfer learning could, therefore, be used to learn user-invariant characteristics within the EEG signal. Previous research has employed data space transformations to move individual user data to a latent feature space in which user-invariant features can be extracted and used for classification (Arvaneh, Guan, Ang, & Quek, 2013). This type of transfer learning is often applied at the feature extraction step and many approaches use a modified version of the CSP algorithm (Kang & Choi, 2014; Blankertz et al., 2007). Novel data is then similarly transformed to the estimated invariant space and classification is performed on the transformed data (Jayaram et al., 2016). Other approaches focus on the use of regularization and sparsity criteria to force BCI systems to only learn patterns which are present across the majority of users (Lotte & Guan, 2010). Furthermore, ensembles have also proven useful for between-user generalization. In this case, one classifier is trained per user in the training set. These are combined into an ensemble which is employed for the classification of novel user data, using, for instance, a simple majority vote (Fazli et al., 2009).

Concurrent with the increased use of Riemannian approaches in BCI classification, several approaches of Riemannian transfer learning have been proposed. Zanini, Congedo, Jutten, Said, and Berthoumieu (2017) introduced Riemannian Alignment as a transfer learning approach in conjunction with the Minimum Distance to Riemannian Mean classifier (RA-MDRM). RA essentially aims to normalize the covariance matrices of data originating from different users, such that they can then be pooled and jointly used for the training of an MDRM classifier. This process involves computing an alignment matrix on a set of resting trials, i.e. examples of users not engaging in a particular experimental task, and subsequently using this alignment matrix to transform experimental trials. The result of this procedure is that the average covariance matrix of trials for a given user is identical across all users, allowing for their combined use in feature extraction and classification. However, as this procedure is conducted in Riemannian space, subsequent feature extraction and classification approaches also need to be Riemannian in nature.

He and Wu (2019) extended the idea of RA to Euclidean space and introduced Euclidean alignment (EA). Similarly to RA, EA involves the estimation of an alignment matrix which can be used to transform user data, normalizing the respective covariance matrices in the process. While EA also features faster computation compared to RA, its main advantage is that data alignment is performed in Euclidean space. As such, well established feature extraction methods such as CSP and a wide range of machine learning classifiers can be used with EA, whereas RA is limited to Riemannian approaches. He and Wu (2019) also showed that a BCI pipeline using EA outperforms or achieves equal performance to one using RA on a commonly used BCI benchmark data set. Additionally, they also showed that estimation of the alignment matrix on regular experimental trials leads to equivalent performance compared to estimation on resting trials as originally proposed by Zanini et al. (2017). In this study, EA is chosen as our transfer learning paradigm due to its simplicity and overall ability to be used with other well established feature extraction and classification methods.

1.5 Current Study

A large number of applications for BCI systems have emerged, but many BCI architectures are ill-suited for large-scale adoption from both a hardware and a software viewpoint. This study explores several avenues for improving the applicability of MI-BCIs to non-clinical settings.

First, we investigate whether a low-cost, dry-electrode EEG device can provide adequate data quality for motor imagery classification. This is assessed by comparing the performance of our BCI architecture on three data sets, namely a training and an evaluation data set collected using an OpenBCI EEG device as well as a historical benchmark data set collected using medical-grade EEG. We

also expand the more common binary classification task to a three-class task, extending the number of possible inputs available to the user.

Second, we investigate whether a transfer learning-based BCI using only minimal user-specific data can achieve performance similar to a BCI trained on a larger set of user-specific data. Specifically, we employ a Euclidean alignment approach for transfer learning which we compare to a baseline transfer learning system as well as a BCI system using only user-specific data. Both transfer learning systems are trained on a training data set and subsequently tested on a separate evaluation data set. Here we also evaluate the BCI architectures online, that is motor imagery predictions were generated in real-time during the collection of the evaluation data set. As use cases for BCIs are generally online, but evaluation of transfer learning paradigms for BCIs is largely conducted offline on established benchmark data sets, this also gives insight into the applied efficacy of transfer learning-based BCIs.

2 Methods

2.1 Experimental Setup

Two separate data collections took place. During the first, training data was gathered for the proposed BCI system, whereas the second aimed at evaluating the trained system. In both cases, participants were students of the University of Groningen. All participants gave written informed consent and received monetary compensation.

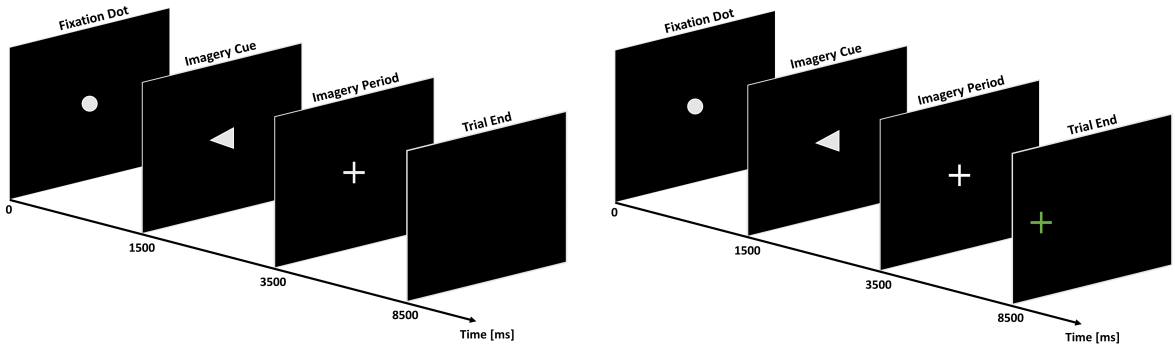
2.1.1 Training Data Collection

Twenty participants took part in the first data collection aimed at generating training data. During data collection, EEG was recorded while participants engaged in motor imagery in response to instructions presented on a computer. A single session lasted from 45 to 60 minutes. Approximately 10 to 15 minutes were spent to physically adjust the EEG headset to the participant's head, as well as to test whether all electrodes were recording with sufficient data quality. The actual experiment lasted approximately 30 minutes. Participant ages ranged from 19 to 28 (Mdn = 22, SD = 3.195). Nine participants reported their gender as female, ten as male and one as other.

The presentation of stimuli during the experiment was implemented in Python using the PyGame module (Van Rossum & Drake Jr, 1995; Shinners, 2011). During the experiment, participants were repeatedly given a cue indicating which of three possible types of motor imagery they should engage in. These three classes of motor imagery consisted of imagined movement of the left hand, right hand or feet. Participants received verbal and written instruction to perform motor imagery by imagining either rhythmic squeezing of the left hand, right hand or curling of the feet. Each of the three motor imagery classes was cued 45 times, resulting in a total of 135 trials. The trial sequence was shuffled across the entire experiment to prevent potential order effects and then split into three blocks. After the first two blocks, participants were asked to take a short break and continue at their own pace. Additionally, participants engaged in six practice trials before the main trial sequence (two examples of each class), after which they could ask the experimenter for clarification. Practice trials were discarded for any of the subsequent analyses.

A single experimental trial lasted for 11.5 seconds. Figure 1a provides an overview of the elements making up a trial. All stimuli were presented as a white shape at the centre of a black background screen. First, participants were shown a fixation dot which lasted for 1.5 seconds. Simultaneously with the onset of the fixation dot, a single-tone audio cue lasting for 200 ms was used to additionally indicate the start of a new trial. Next, a triangle-shaped arrow cue shown for two

seconds instructed participants as to which type of motor imagery they should subsequently engage in (where pointing left, right and down corresponded to left hand, right hand and feet, respectively). Following this, a fixation cross was shown for five seconds, indicating that participants should keep their gaze centred on the cross while engaging in the previously cued motor imagery. From this five-second interval, the motor imagery period of interest was later extracted as discussed in Section 2.3. Lastly, a three-second period without on-screen stimuli, paired with a second distinct 200 ms audio cue communicated the end of the trial to participants.



(a) Training setup.

(b) Evaluation setup.

Figure 1: Experimental setup for the training and evaluation data collections. During the evaluation session participants received feedback as to how their motor imagery was classified through a coloured fixation cross at the end of a trial.

2.1.2 Evaluation Data Collection

A second data collection consisting of nine participants was used to evaluate and compare the different BCI architectures. The experiment again involved 10 to 15 minutes of headset adjustment followed by 20 to 25 minutes of the actual experiment. Participant age ranged from 18 to 26 (Mdn = 19, SD = 2.47). Five participants reported their gender as female and four as male. In contrast to the training data collection, participants now received feedback on how their motor imagery was classified after each trial. The trial structure was kept as similar as possible to the trial structure described in Section 2.1.1 to focus solely on transfer between participants, rather than also adding transfer between experimental conditions as an additional variable. The fixation dot, imagery cue and imagery period of a trial were identical to the training setup. For the trial end segment, rather than showing a blank screen for three seconds, participants instead were shown a coloured fixation cross for three seconds acting as feedback (see Figure 1b). Specifically, a green fixation cross indicated that their motor imagery had been correctly classified and a red cross indicated incorrect classification. Additionally, the fixation cross moved left, right or down from its previous position at the centre of the screen, depending on whether the preceding motor imagery had been classified as left hand, right hand or feet, respectively.

The overall experiment consisted of twelve calibration trials, as well as 90 experimental trials split into three blocks of 30 trials. The calibration trials were used to estimate an alignment matrix for later use with a Euclidean alignment-based transfer learning classifier. Each of the three experimental blocks consisted of an equal number of examples of the three motor imagery classes in a randomized order. Between blocks, participants were again asked to take a short break.

Blocks differed only by the type of classifier used for online MI prediction. The first block

used a baseline classifier, trained on the combined data of users from the training data set. For the second block, a classifier was similarly trained on the training data set users, however, their data was first transformed using Euclidean alignment. For the third block, no between-users classification took place and its classifier was trained solely on the current user's data from the previous two blocks. Lastly, to prevent training and fatigue effects, the order of the first two blocks was counterbalanced between users.

2.2 BCI Pipeline

Obtaining motor imagery class predictions from raw EEG data involved a multi-step processing pipeline. Figure 2 provides an overview. The available EEG data was first filtered to remove noise and frequency-specific artefacts and then segmented into individual trials of motor imagery. Depending on the specific BCI architecture under investigation, this was followed by Euclidean alignment to normalize the covariance matrices of trials per user, facilitating subsequent transfer learning. Next, the common spatial patterns algorithm was used to extract a set of features for each trial. These features were then fed into a classifier which yielded a single scalar class prediction. Where possible, parameters of pipeline components, such as the filter frequency band or the number of CSP components, were estimated through hyperparameter estimation on the training data set.

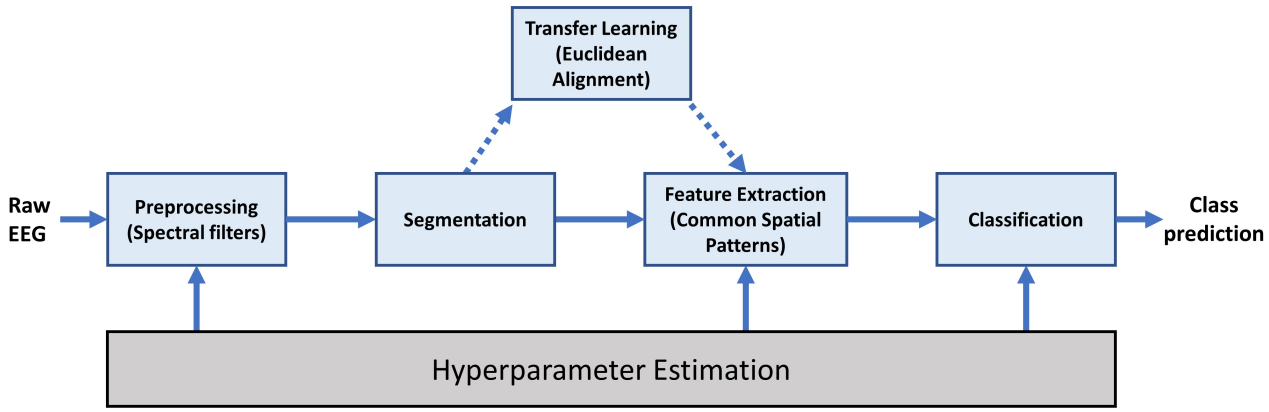


Figure 2: Individual components of the BCI processing pipeline.

2.3 EEG Recording and Preprocessing

For all experiments, an OpenBCI Ultracortex Mark IV headset was used to collect EEG data (OpenBCI, 2022). Measurements were taken using eight dry Ag-AgCl electrodes whose locations were consistent with the default 10-20 system. Seven of these were chosen based on their spatial proximity to the brain region of interest (Cz, C3, C4, CP1, CP2, FC1, FC2) and one (FPz) was chosen to pick up on information related to eye movements and blinks. Figure 3 showcases the headset and Figure 4 gives a visual indication of the chosen electrode locations.

While the Ultracortex headset theoretically allows for the placement of up to 16 electrodes, we decided on a parsimonious approach using eight electrodes, similar to the setup of Yohanandan et al. (2018). First, potential measurement locations are spaced further apart compared to medical-grade EEG devices, preventing the placement of additional electrodes over our region of interest. We also found it difficult in practice to achieve adequate impedance values for all 16 electrodes. As our

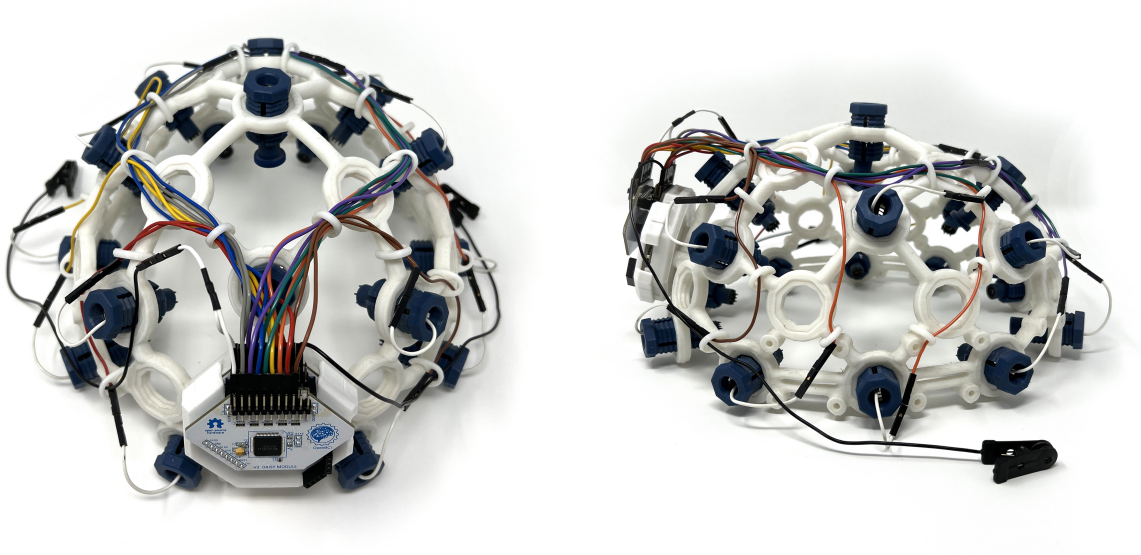


Figure 3: The Ultracortex Mark IV EEG headset. The central amplifier board receives measurements from dry electrodes situated at different positions around the user's head and relays them wirelessly to a processing computer. Earclip electrodes act as ground and reference. Graphics reproduced from the official OpenBCI documentation (OpenBCI, 2022).

feature extraction approach relies on learning spatial filters based on channel covariation, large voltage variations in several channels between recording sessions could reduce overall system performance and increase the difficulty of transfer. As a consequence, we chose to prioritize data quality over additional measurement locations.

In cases where a channel exceeded an absolute voltage threshold of 100 millivolts, we considered the channel flatlined and, therefore, unusable. Generally, flatlined channels observed during headset adjustment could still be calibrated to bring measurements into an acceptable range, but in some cases even repeated adjustment did not improve data quality. In other cases, channels flatlined during the experiment. To account for this, we considered a channel as flatlined for a single trial if more than 10% of its data values exceeded the threshold of 100 millivolts and flatlined for the entire experiment if more than a third of the experimental trials were considered flatlined. These channels were then removed from analyses for the specific user. In practice, CP1 was considered flatlined for four participants in the training data set and FC1 flatlined for one participant. In the evaluation data set, CP1 flatlined for one participant, FC1 for one participant and Cz for three participants.

In addition to measurement electrodes, two additional electrodes were placed on the earlobes to act as ground and reference. The EEG signal was digitized at a sampling rate of 125 Hz. While this is somewhat lower than the sampling rates of medical-grade EEG devices, our frequency bands of interest were all below 30 Hz. Since the Nyquist-Shannon theorem states that one's sampling rate should be at least twice as large as the highest frequency of interest, 125 Hz should be adequate for our use case (Shannon, 1949).

In regards to preprocessing, a bandpass and a notch filter were applied to the EEG signal. For the bandpass filter, different frequency bands were explored as part of hyperparameter estimation (see Section 2.7). A notch filter at 50 Hz was employed to account for power-line noise during recording. Lastly, trials were segmented into epochs starting from 500 ms after the onset of the imagery period fixation cross. This offset was used to remove any visual processing activity related to the fixation cross. The length of the subsequent imagery period was varied as a hyperparameter as well.

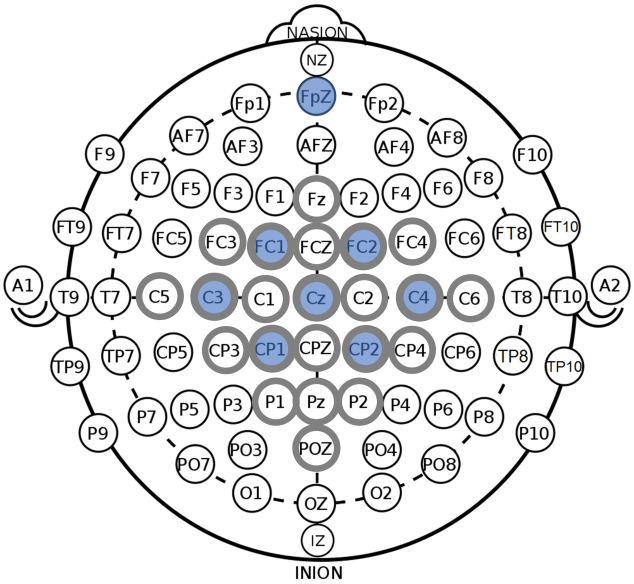


Figure 4: Electrode locations used for the OpenBCI training and evaluation data sets (blue) and the medical-grade EEG benchmark data set (grey).

As the evaluation session involved using the BCI pipeline for online classification, several preprocessing steps had to be adjusted. For online filtering, we employed a bandpass filter length of ten seconds. After the imagery period of a trial had ended, the preceding ten seconds of recorded data were filtered and the segment corresponding to the imagery period was extracted from the filtered data. An additional challenge for online transfer learning involved dealing with flatlined channels. In cases where channels had flatlined for the current user but were present with sufficient data quality in the training data set, performance suffered significantly. To account for this, flatlined channels were determined at the beginning of the experimental session and subsequently not used for online classification. The transfer classifiers were then also retrained on the training data without using the currently flatlined channels.

2.4 Feature Extraction

After preprocessing, spatial filters were extracted from the data using the Common Spatial Patterns (CSP) algorithm as implemented in the Python module MNE (Gramfort et al., 2013). This implementation follows the original two-class CSP algorithm as described by Koles et al. (1990) and the subsequent extension to multiclass tasks by Grosse-Wentrup and Buss (2008). Similar to Principal Component Analysis (PCA), CSP aims to learn a set of components which characterize the variation within a set of data. Whereas components in PCA correspond to the axes of maximal variation in the data overall, CSP is a supervised technique and maximizes differences in variation between classes (i.e. types of motor imagery). As a first step, one computes the class-specific covariance matrices Σ_i using only the trials of class i for all c classes. Class-specific trials are concatenated along the time point axis, creating a matrix X_i where rows correspond to channels and columns to time points. The class covariance matrices are then given by

$$\Sigma_i = X_i X_i^T N_s^{-1} \tag{1}$$

with N_s being the total number of time points in X_i. Having obtained the class covariance matrices, the next step is to jointly diagonalize them, yielding a characteristic set of eigenvalues and eigenvectors.

In a two-class scenario with classes a and b, joint diagonalization is equivalent to the eigenvalue decomposition of $\Sigma_a \Sigma_b^{-1}$, i.e. the ratio of the two class covariance matrices. For more than two classes, joint diagonalization can be performed using the approximate joint diagonalization (AJD) algorithm (Pham, 2001). Similar to the two-class scenario, this process yields a single set of eigenvectors and eigenvalues and maximizes differences between the set of covariance matrices which were jointly diagonalized. The ordered eigenvectors are then stored as the learnt spatial filters S.

For a single trial, one can then use the spatial filters to extract the features F using

$$F = \log \frac{1}{N_s} \sum_{s=1}^{N_s} (SX_s)^{\circ 2}. \tag{2}$$

Here we first compute the dot product of the spatial filters S and a single time point X_s of the given trial. The result is squared element-wise. We then average across the transformed time points and take the natural logarithm, yielding a one-dimensional matrix with its size being the number of spatial filters. The resulting features represent the average power per filter for the given trial. While the dimensionality of the data is greatly reduced by averaging across the time point dimension, motor imagery as a paradigm does not expect there to be temporal information within a trial. Differences are thought to be present between channels and classes and power within a trial is not thought to vary in a way which would yield additional information. Lastly, the feature vectors estimated for each trial are concatenated into a two-dimensional matrix with dimensions corresponding to trials and features. This matrix is used as input for a subsequent classifier.

For feature extraction in our final BCI architecture, we also applied a shrinkage transformation to the computed class covariance matrices. Shrinkage essentially regularizes the covariance matrices and attenuates more extreme coefficients. Ledoit and Wolf (2004) showed that shrinkage can improve performance, especially in scenarios were the number of examples is small. The shrinkage extent was estimated empirically using the Ledoit-Wolf lemma (Ledoit & Wolf, 2003).

2.5 Classification

Several different classification architectures, including Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and the Random Forest were explored as candidate classifiers for the final BCI architecture (Fisher, 1936; Cortes & Vapnik, 1995; Breiman, 2001). All classifiers were implemented using the Python module Scikit-learn (Pedregosa et al., 2011).

As LDA was eventually chosen as the final classification approach for our pipeline we introduce it in more detail here. LDA shares several similarities with CSP and both essentially learn a data rotation which maximizes differences between classes. While CSP maximizes differences in variation between the classes, LDA maximizes differences between class means. Specifically, LDA aims to maximize the scatter between classes while also minimizing the scatter within classes (Fisher, 1936). This process yields a set of linear discriminants (similar to CSP's spatial filters) which can be used to transform novel data and draw decision boundaries for classification in the transformed space. For learning linear discriminants we follow the approach described by Hart, Stork, and Duda (2000) who solve the learning process through eigenvalue decomposition. We first compute the within covariance matrix

$$\Sigma_W = \sum_{i=1}^{c} \Sigma_i \tag{3}$$

as a sum of the within covariance matrices Σ_i for the c different classes. For a given class i, Σ_i can be

computed as

$$\Sigma_i = X_i X_i^T N^{-1} \tag{4}$$

where the rows of X_i represent the extracted features, i.e. CSP components, columns correspond to the trials of class i and N is the number of trials. Similarly, the total covariance matrix Σ_T can be computed as

$$\Sigma_T = X X^T N^{-1} \tag{5}$$

where X consists of all trials. As the within and between covariance matrices sum up to the total covariance matrix, the between covariance matrix S_B can be computed as

$$\Sigma_B = \Sigma_T - \Sigma_W. \tag{6}$$

For a formal derivation of this property refer to Hart et al. (2000). Finally, computing the eigenvalue decomposition of $\Sigma_W^{-1}\Sigma_B$, essentially a ratio of between-class and within-class variance, yields a set of eigenvalues and eigenvectors. The sorted eigenvectors are the final linear discriminants which can be used to transform novel data. For a new trial X with an unknown label y we first transform X using the linear discriminants and compute the probability $P(y = class_i|X)$ for each class i using maximum likelihood estimation. We then assign the class label for which this probability is highest to X. For a more thorough treatment of how these probabilities are computed, we again refer to Hart et al. (2000).

In regards to evaluation metrics, we chose accuracy scores to capture classifier performance. While other metrics such as precision, recall or the F1-score often are superior evaluation metrics compared to accuracy, we found accuracy sufficient for the given use case. First, accuracy fails to give an accurate depiction of performance if the data is unbalanced, i.e. not all classes have the same number of examples. In our case, the collected data set is balanced by design. Second, precision and recall can better assess performance for use cases where different types of misclassifications carry different weight. In medical applications, false negatives are often far more detrimental than false positives. For the current system, however, all misclassifications have the same impact. Therefore, we decided that accuracy is appropriate as a metric in the current context and preferred due to its more intuitive interpretation.

2.6 Transfer Learning

Due to the high variability of the EEG signal across different users, the BCI literature considers the use of a BCI system for between-user classification transfer learning (Jayaram et al., 2016). In this study, we compared two types of transfer learning approaches in their efficacy. First, a baseline system simply pooled together users from the training data collection and trained feature extraction and classification models on their combined data. A second transfer learning system used Euclidean Aligment (EA) as introduced in He and Wu (2019) before pooling user data together. The general idea of EA is analogous to normalization in that the average covariance matrix for each user after alignment is identical to those of other users. Alignment is performed within a user, meaning that it does not involve information from other users. It is also unsupervised in nature and precedes feature extraction and classification, as it is essentially a data transformation.

For a single user, one first computes the reference matrix

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T N_s^{-1} \tag{7}$$

which represents the average covariance matrix across the users n experimental trials. Here X_i is a single experimental trial where the rows correspond to channels and the columns to time points, with

N_s being the number of time points per trial. The trial covariance matrix $X_i X_i^T N_s^{-1}$ then represents the variation between different channels, essentially encapsulating how channels do or do not co-vary for different types of motor imagery, which CSP also capitalizes upon when learning its spatial filters. Normalizing covariance matrices between users should, therefore, improve performance when pooling user data together and employing their combined data for feature extraction and classification. From the reference matrix $\bar{R}$ one subsequently computes the alignment matrix

$$A = \bar{R}^{-1/2}. \tag{8}$$

Alignment of a trial only involves a simple matrix multiplication of

$$\tilde{X}_i = AX_i \tag{9}$$

where $\tilde{X}_i$ is the aligned trial. Mathematically this alignment has the effect that the average covariance matrix of all aligned trials for a given user will be the identity matrix and consequently identical across users. Performing EA, therefore, requires at least some user-specific trials to estimate the alignment matrix A. For offline transfer learning scenarios, A can be estimated from the available experimental trials, however, online classification using EA requires that one first estimates A on a set of calibration trials.

2.7 Hyperparameter Estimation

Since the proposed BCI system was conceptualized as an end-to-end pipeline, hyperparameters were not limited to the parameters of the classification model. Instead, preprocessing parameters (e.g. length of the imagery window) as well as feature extraction parameters (e.g. number of CSP components) were also experimentally investigated.

Leave-one-out cross-validation (LOOCV) was used to estimate the optimal set of hyperparameters within users (Hastie, Tibshirani, Friedman, & Friedman, 2009). Given a single user's n experimental trials, $n-1$ trials were used to jointly train a feature extraction and classification model which were then evaluated on the left-out trial. This process was repeated n times such that all trials were used as a left-out trial exactly once. For each user, a LOOCV accuracy score was computed in this way, after which the user scores were averaged to yield a single overall accuracy score for the particular set of hyperparameters tested. Table 1 gives an overview of all hyperparameters and their explored values. As several classifier architectures were explored, some hyperparameters are exclusive to a particular classifier.

2.8 Benchmark Data Set

In addition to training and validating the pipeline introduced in this paper on our collected data sets, we also investigated performance on a publicly available benchmark data set. As this data set was collected using a medical-grade EEG device, one can then disentangle whether differences in performance are due to the data recording hardware, or the chosen BCI architecture. The benchmark data set was first introduced in Naeem, Brunner, Leeb, Graimann, and Pfurtscheller (2006), in which the authors compared performance of a CSP-based feature extraction method to one based on independent component analysis. The data set was later made publicly available as part of the fourth BCI competition, an initiative to compare the performance of different BCI approaches on open-access data sets (Tangermann et al., 2012).

The benchmark data set consists of recordings from nine participants who engaged in examples of motor imagery across two sessions. Only eight were analyzed in Naeem et al. (2006) without

Hyperparameter	Values
General Parameters	
Number of CSP Components	2, 4, 8
CSP Regularization	None, Ledoit-Wolf
Imagery Window (in s)	2, 3, 4
Frequency Band (in Hz)	8-13, 10-12, 18-25, 8-25
LDA	
Regularization	None, Ledoit-Wolf
SVM	
Regularization (C)	0.25, 0.5, 0.75
Kernel	Linear, Poly, Rbf, Sigmoid
Random Forest	
Number of Trees	10, 100, 500
Maximal Tree Depth	10, 20, 100

Table 1: General and classifier-specific hyperparameters with their explored values.

the authors stating a specific exclusion criterion. As the current study was interested in user-to-user transfer rather than within-user session-to-session transfer, only recordings from the first session were used. Participants were said to be naive, i.e. had not used MI-BCIs before. Types of motor imagery consisted of imagined movement of the left hand, right hand, feet and tongue, constituting a four-class problem. The experiment consisted of 288 trials of motor imagery, made up of 72 trials per type of imagery. A single trial lasted eight seconds. During a trial, participants were shown a fixation cross for the first six seconds. From 2 to 3.25 seconds after the trial start, an arrow cue was shown in addition to the fixation cross, indicating the type of motor imagery to perform. The imagery period was then defined as lasting from 2.5 seconds to 5.5 seconds after trial start. Following this, a two-second pause without a fixation cross marked the end of the trial.

Data was recorded using 22 Ag/AgCl electrodes placed in line with the 10-20 system. Figure 4 indicates which locations were used. Additional electrodes at the earlobes were used as ground and reference. In total, seven out of the eight electrode locations used in our BCI design were also used in the benchmark setup. The data was digitized at a sampling rate of 250 Hz and preprocessed using a 0.5-100 Hz bandpass filter as well as a 50 Hz notch filter.

3 Results

Several analyses were conducted to investigate the performance of our BCI architecture on three different data sets, namely the training, evaluation and benchmark data sets. We first estimated the optimal set of hyperparameters for our BCI architecture on the training data set. These parameters were also used for the evaluation data set and, where applicable, the benchmark data set. For all data sets, we investigated both performance within users, i.e. the BCI was trained and tested on the same

user, as well as transfer performance, where the BCI was trained on one set of users and tested on another user. While the act of combining user data for generalization to new users already constitutes transfer learning in the BCI literature, transformation approaches such as Euclidean alignment are necessary for transfer learning to perform well. Here, we compared a baseline transfer system to one employing Euclidean alignment. We also visually explored differences between user performance as well as the effect of Euclidean alignment on feature extraction and classification. For the training and benchmark data set, we additionally investigated what effect fewer experimental trials, fewer channels or fewer motor imagery classes would have on performance. Lastly, for the evaluation data set, we contrasted the performance of the BCI architecture estimated online as described in Section 2.1.2 with the performance estimated in an offline manner. BCI performance is generally presented as an average across users, for individual user performance see Figure 16 in the Appendix.

3.1 Training Data Set

3.1.1 Hyperparameter Estimation

A number of classifier-independent hyperparameters were systematically varied to arrive at a final architecture for use with the evaluation and benchmark data sets. Table 2 presents performance in terms of average accuracy across users for different sets of parameters using the best-performing classifier approach, LDA. To reduce the number of total results, parameter permutations are shown for the best-performing frequency band (10-12) only. Results for the other frequency bands are presented only for the best-performing set of the other general parameters.

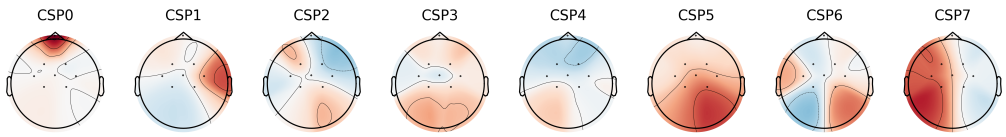
Additionally, classifier-specific hyperparameters were investigated for three different classifier architectures (LDA, SVM and Random Forest) and can be seen in Table 5 in the Appendix. Due to a large number of possible parameter permutations, different classifier results are again only reported for the best set of general parameters. Overall, an architecture using a 10-12 Hz bandpass filter, regularized CSP with Ledoit-Wolf shrinkage, eight CSP components for feature extraction, an imagery window of four seconds and an LDA classifier without shrinkage was found to give the highest accuracy of 0.473. As is often the case for motor imagery paradigms, performance varied widely between users, with accuracies ranging from 0.27 to 0.84. Standard deviations across users ranged from 0.123 to 0.232 and generally did not seem to vary systematically with any of the hyperparameters. For a more detailed description of how the different hyperparameters affected performance, refer to Section A in the Appendix.

3.1.2 Within-User Classification

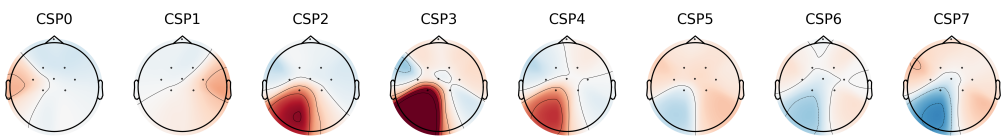
Having estimated the hyperparameters for our final BCI architecture, we visually explored the efficacy of its feature extraction and classification components. Figure 5 provides insight into the CSP-based feature extraction by visualizing the neural sources underlying the learnt spatial filters. Note that values in the figure are in CSP space and do not represent actual EEG voltages, which is why we chose not to include a measurement unit colour bar. As can be seen, the learnt CSP components varied widely between users. For the best-performing user, the first two components seemed to correspond to hemispheric activity relating to the left and right-hand MI classes. In contrast, the first component for the worst-performing user was focused on FPz, potentially due to a large number of eye blinks in their data.

Bandpass	CSP-Regularization	N-CSP	IW	Accuracy	SD
10-12	Ledoit-Wolf	2	2	0.390	0.169
10-12	Ledoit-Wolf	2	3	0.367	0.175
10-12	Ledoit-Wolf	2	4	0.408	0.171
10-12	Ledoit-Wolf	4	2	0.440	0.147
10-12	Ledoit-Wolf	4	3	0.431	0.163
10-12	Ledoit-Wolf	4	4	0.447	0.148
10-12	Ledoit-Wolf	8	2	0.447	0.127
10-12	Ledoit-Wolf	8	3	0.452	0.159
10-12	Ledoit-Wolf	8	4	0.473	0.156
10-12	None	2	2	0.390	0.166
10-12	None	2	3	0.370	0.175
10-12	None	2	4	0.409	0.169
10-12	None	4	2	0.438	0.145
10-12	None	4	3	0.427	0.166
10-12	None	4	4	0.452	0.145
10-12	None	8	2	0.444	0.123
10-12	None	8	3	0.452	0.158
10-12	None	8	4	0.469	0.153
8-13	Ledoit-Wolf	8	4	0.447	0.146
18-25	Ledoit-Wolf	8	4	0.383	0.130
8-25	Ledoit-Wolf	8	4	0.445	0.14

Table 2: Average within-user classification accuracy on the training data set using LDA for classification. Each row corresponds to a specific set of parameters, where the best-performing parameter set is highlighted with a grey background. N-CSP refers to the number of CSP components used during feature extraction and IW refers to the length of the imagery window.



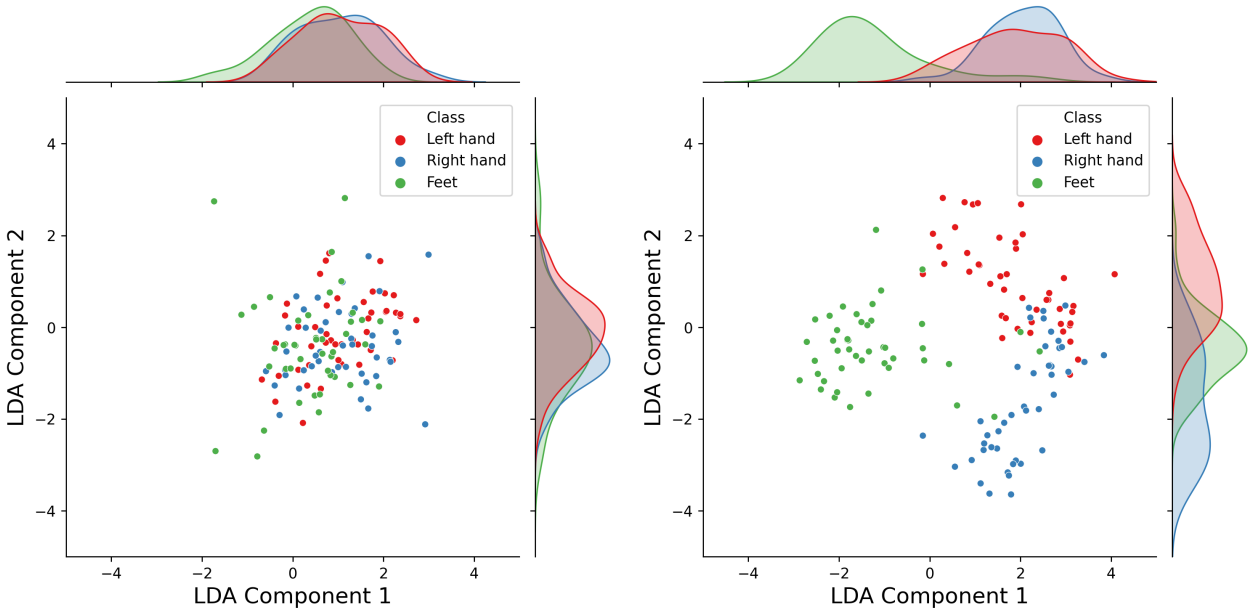
(a) User with the lowest accuracy.



(b) User with the highest accuracy.

Figure 5: Neural patterns underlying the learnt CSP filters for the users with the lowest and the highest classification accuracy in the training data set.

Differences between users and their motor imagery performance also became apparent when visualizing the learnt LDA components as seen in Figure 6. For the user with lowest accuracy, the three classes were still mostly intermixed, whereas the combined CSP-LDA approach seemed to almost perfectly separate the classes for the user with the highest accuracy in the training data set. As indicated by the class density curves, the first LDA component differentiated feet MI from the two hand classes and the second component separated left and right-hand MI.



(a) User with the lowest accuracy.

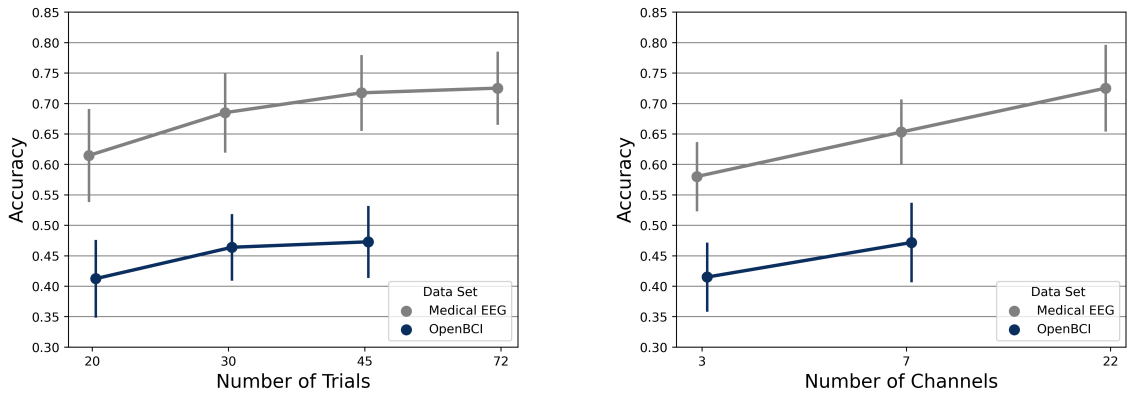
(b) User with the highest accuracy.

Figure 6: Class separability expressed through the learnt LDA components for the users with the lowest and the highest classification accuracy in the training data set. Class densities for the individual LDA components are plotted on their respective axes.

Furthermore, the effect of sparser data on performance was investigated by artificially reducing channels and trials per class, as seen in Figure 7, which provides a visual overview as well as a comparison to the benchmark data set. See Table 6 in the Appendix for numerical results. Differences between the training and benchmark data set will be presented in more detail in Section 3.3.2. In regards to trials per class, we simulated the effect of less training data by using a random subset of trials. In order to minimize the impact of random chance on the resulting accuracies, we picked ten random sets per investigated subset size and averaged the results. Reducing the number of trials per class from 45 to 30 and 20, reduced accuracy from 0.473 to 0.464 and 0.413, respectively. This indicates that 20 class trials are not sufficient and that there is only a marginal advantage of 45 trials compared to 30.

Using only the three most essential channels C3, C4 and Cz, resulted in an accuracy of 0.415 and removing FPz from the original set of eight channels yielded 0.472. Although FPz is not situated over the motor cortex and, therefore, not obvious as a candidate for additional information, its inclusion did not seem to decrease performance, but rather marginally increase it. A subset of only the three most important channels significantly decreased performance, indicating that there is useful additional information in the other five channels. As a last comparison, we also reduced the data set to a binary classification task only using trials of left and right-hand imagery, as is common in many

MI paradigms in the literature. Accuracy was found to be 0.588, which is quite low considering that an accuracy of 0.5 would constitute random chance.



(a) Trial subsets.

(b) Channel subsets.

Figure 7: Classification accuracy for the OpenBCI training data set and the medical EEG benchmark data set on subsets of fewer trials per class and fewer channels. Vertical bars represent standard errors calculated across users. For the three-channel subset, electrodes C3, C4 and Cz were included, whereas the seven-channel subset featured all OpenBCI electrodes except FPz.

3.1.3 Transfer Learning Classification

While the main mode of investigating transfer learning for our BCI architecture was through transfer from the training data set to the evaluation data set, we also investigated transfer within the training data set. For this, transfer learning classifiers were trained on all but one user and subsequently tested on this left-out user. To assess the efficacy of Euclidean alignment (EA) for BCI transfer learning we compared it to a baseline which did not involve a data transformation. Averaged across users, the EA system achieved an overall accuracy of 0.447 compared to 0.396 for the baseline system. Figure 8 compares their performance and those of several classifiers trained within each user. Note that the error bars represent the standard deviation across users and, therefore, represent the inherent differences in performance between users more so than measurement error. While the EA system clearly outperformed the baseline system, it also achieved similar performance to classifiers trained on user-specific data. These classifiers were either trained on half of a user's experimental data and tested on the other half (50-50), trained on two-thirds and tested on one-third (67-33) or trained on all but one trial and tested on the left-out trial (LOOCV). For the first two, 100 data splits were chosen randomly and their results averaged. While training on more trials did increase performance, as indicated by the highest score for the LOOCV classifier, the added gains in accuracy were small. Crucially, the EA classifier achieved the same performance as the within-user 50-50 classifier, without being trained on any user-specific data.

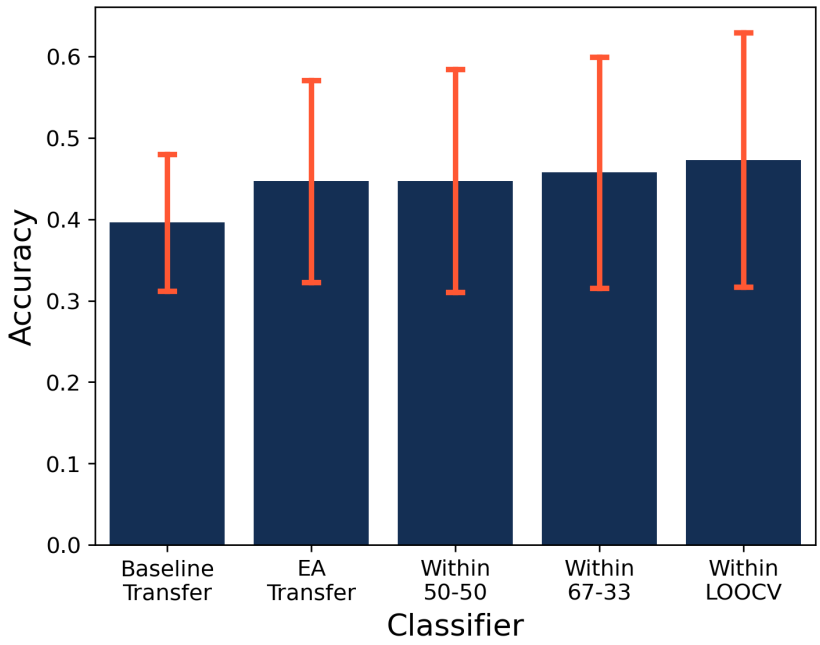
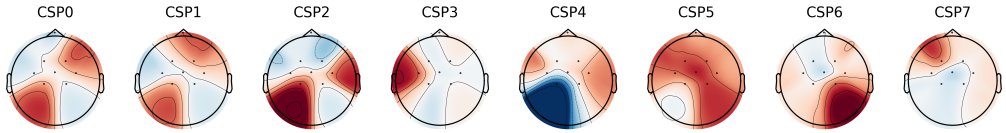
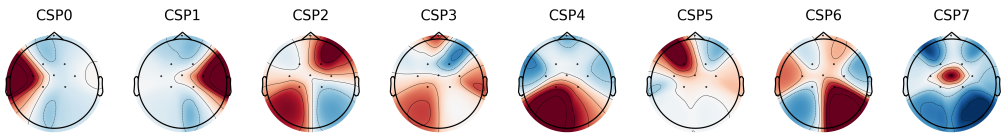


Figure 8: Classification accuracy averaged across users for different classifiers on the training data set. Error bars represent the standard deviation across users.

The effect of EA was also apparent when visualizing CSP and LDA outcomes. As seen in Figure 9, EA significantly changed the estimated CSP components. Whereas components before alignment were less straightforward to interpret, the first two components after alignment very clearly reflected the hemispheric differences inherent to MI.



(a) Without Euclidean alignment.

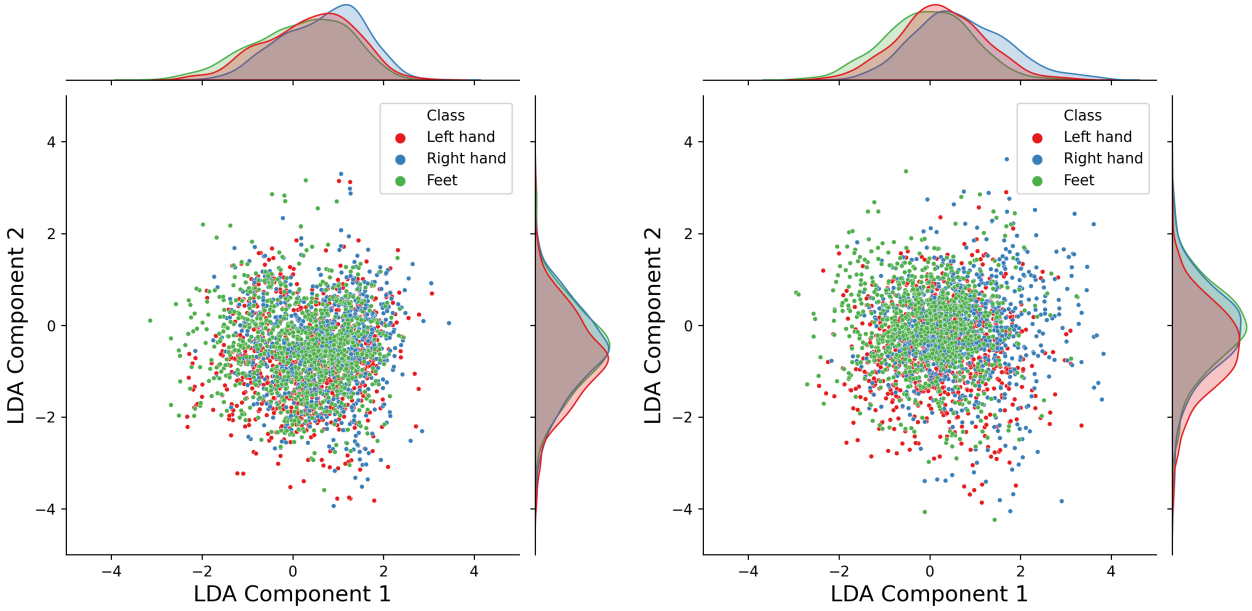


(b) With Euclidean alignment.

Figure 9: Neural patterns underlying the CSP filters learnt on all users in the training data set, with and without previous Euclidean alignment.

Figure 10 gives an indication of the overall challenge of transfer learning for MI-BCIs. Without alignment, the three classes almost completely overlapped when combining the data of all users. While class separation after alignment was still far from perfect, the class densities indicated that overlap between the right-hand class and the other two classes had been reduced on the first component and overlap between the left-hand class and the other two classes had been reduced on the second

component.



(a) Without Euclidean alignment.

(b) With Euclidean alignment.

Figure 10: Class separability of all user trials expressed through the learnt LDA components for the training data set, with and without previous Euclidean alignment. Class densities for the individual LDA components are plotted on their respective axes.

3.2 Evaluation Data Set

The evaluation data set was used to test the optimal BCI architecture estimated on the training data set on novel data. Additionally, it was used to investigate whether transfer learning could be performed online, that is during data collection when users were performing the actual motor imagery. As BCI use cases inherently involve online classification, we found this assessment especially important for evaluation of the general BCI efficacy. Within-user classification results are presented together with transfer learning results in Figure 11. For CSP and LDA visualizations corresponding to the best and worst-performing users in the evaluation data set, see Figure 17 and Figure 18 in the Appendix.

3.2.1 Transfer Learning Classification

The overall results as seen in Figure 11 showed a clear improvement of EA (0.442) over the baseline system (0.369) for offline transfer. EA also achieved similar performance to a classifier trained on user-specific data (0.44). For online classification, both transfer approaches were tested on a single experimental block made up of 30 trials only. Additionally, the alignment matrix needed for online EA was estimated on twelve calibration trials. The online within-user classifier was trained on two experimental blocks and tested on a third. Interestingly, the improvement from baseline to EA for offline transfer was not visible for online transfer learning. Instead, the online EA system performed at chance level for almost all participants, as is also indicated by its small standard deviation. This could potentially be caused by the estimated alignment matrix not being sufficient, or 30 trials being a too-small test set to account for variation between measurements. As EA constitutes a data transformation, an unrepresentative alignment matrix might distort the data and impair classification, rather than

improve it. It also has to be noted that the higher baseline accuracy is largely due to a single participant who achieved an accuracy of 0.7 for the baseline block and 0.33 for the EA block. For the majority of participants, both baseline and EA classifiers performed at chance level.

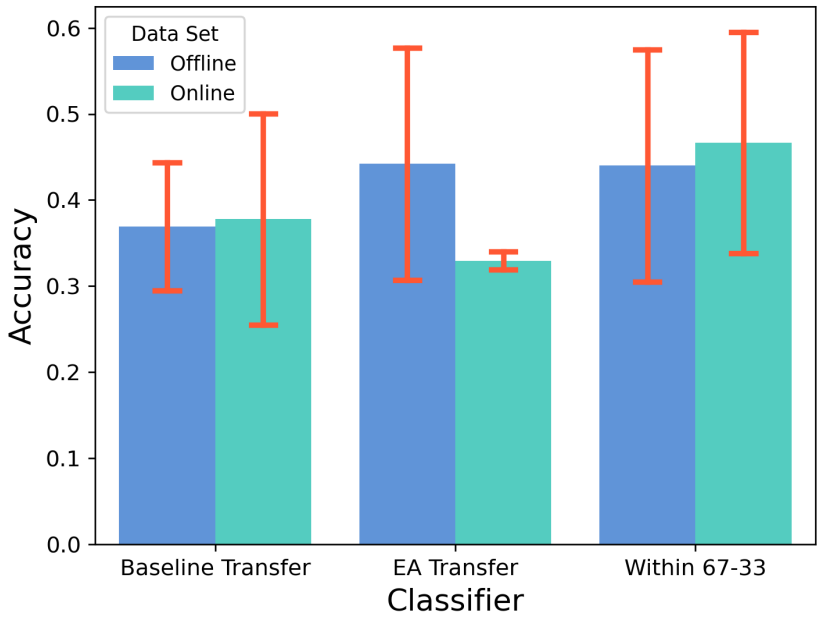


Figure 11: Classification accuracy averaged across users for different classifiers on the evaluation data set, estimated both offline and online. Error bars represent the standard deviation across users.

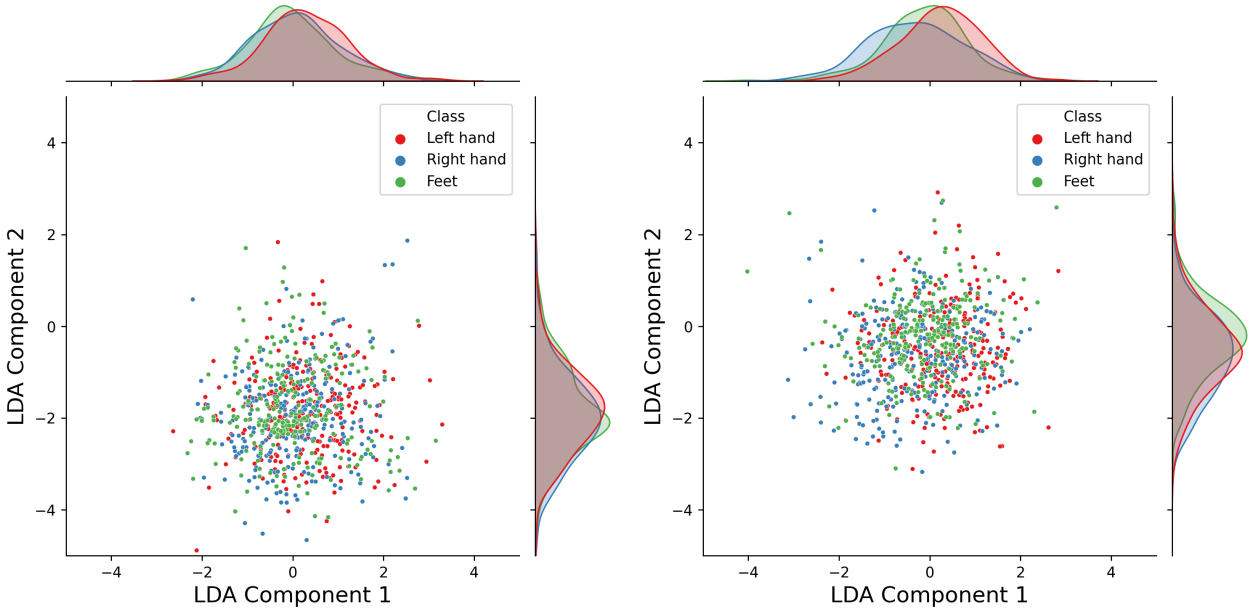
Euclidean alignment again seemed to improve class separability as indicated by the class density curves seen in Figure 12. Here the first component seemed to separate the right-hand class more strongly from the other two classes and the second component increased separability of the feet class and the two hand classes.

3.3 Benchmark Data Set

3.3.1 Within-User Classification

While the performance of MI-BCIs is heavily dependent on one's choice of feature extraction and classification approaches, other external factors, such as the type of measurement device and user characteristics, also play a large role. To better estimate how these factors influence performance we also tested our BCI architecture set on a second, external benchmark data set. As determined in Section 3.1.1 we employed CSP with regularization and an LDA classifier without covariance shrinkage. Due to differences in experimental setup and the number of channels, we kept the imagery window fixed at three seconds and experimentally investigated the frequency band and number of CSP components again. Similar to the training data set, we also explored the effect of fewer experimental trials, fewer channels and different numbers of motor imagery classes on performance. All results in terms of averaged accuracy and standard deviation across users are presented in Table 3.

Similar to results for the training data set, accuracy for a specific parameter configuration varied significantly between users. For the set of default parameters, user accuracies ranged from 0.426 to 0.93. Overall, standard deviations for performance between users ranged from 0.12 to 0.202. Standard deviations generally decreased as the number of trials increased. Furthermore, they increased



(a) Without Euclidean alignment.

(b) With Euclidean alignment.

Figure 12: Class separability of all user trials expressed through the learnt LDA components for the evaluation data set, with and without previous Euclidean alignment. Class densities for the individual LDA components are plotted on their respective axes.

concurrently with the number of CSP components as well as the number of MI classes.

Different types of frequency bands were explored first. Specifically, bandpass filters from 0.5-100 Hz (noise filtering only), 8-13 Hz (mu-band), 10-12 Hz (sub-band of the mu-band), 18-25 (beta-band) and 8-25 (combination of the mu- and beta-band) were tested. As expected, the lowest performance was found for the minimal bandpass filter from 0.5-100 Hz with an averaged overall accuracy of 0.66. The highest accuracy of 0.725 was observed for the combined mu- and beta-bands, whereas the use of singular bands resulted in accuracies of 0.681 and 0.678, respectively. In contrast to the training data set, the 10-12 Hz sub-band did not outperform the more general mu-band and yielded an accuracy of 0.676. The large increase in accuracy from single frequency bands to combined bands indicates that both the mu- and beta-band contain unique information for class separability. This stands in contrast to the results for the training data set where the mu-band was found to be superior and the addition of the beta-band instead decreased performance. For the benchmark data set we conducted all further investigations using an 8-25 Hz bandpass filter. The corresponding accuracy using otherwise default parameters (0.725) was used as a reference point for other parameter experiments.

In regards to MI classes, the majority of analyses presented in this section focused on the same three motor imagery classes found in our training data set (left hand, right hand and feet) by removing trials related to motor imagery of the tongue. Further reducing the data set to a binary classification of only the left hand and right hand gave an accuracy of 0.784. For the full four-class data set accuracy was found to be 0.659, which broadly coincides with the accuracy of 0.644 reported by Naeem et al. (2006) for their CSP approach (using eight of the nine total users and a one-second imagery window).

The benchmark data set was recorded using a 22-channel medical-grade EEG device, rather than the eight-channel OpenBCI EEG device used for the training data set. Thus, we investigated how the performance would change when only a subset of channels was used (see Figure 7 for a comparison to the training data set). Of the eight OpenBCI electrode locations, all seven situated above the motor

Bandpass	Classes	Channels	N-CSP	Trials	Accuracy	SD
0.5-100	3	22	8	72	0.660	0.137
8-13	3	22	8	72	0.681	0.179
10-12	3	22	8	72	0.676	0.173
18-25	3	22	8	72	0.678	0.133
8-25	3	22	8	72	0.725	0.159
8-25	**2**	22	8	72	0.784	0.124
8-25	**4**	22	8	72	0.659	0.161
8-25	3	**3**	3	72	0.580	0.127
8-25	3	**7**	7	72	0.653	0.120
8-25	3	22	**2**	72	0.646	0.152
8-25	3	22	**4**	72	0.694	0.162
8-25	3	22	**16**	72	0.722	0.161
8-25	3	22	**22**	72	0.717	0.164
8-25	3	22	8	**20**	0.615	0.202
8-25	3	22	8	**30**	0.685	0.173
8-25	3	22	8	**45**	0.718	0.165

Table 3: Average within-user classification accuracy on the benchmark data set. Each row corresponds to a specific set of parameters, where the parameter being varied is highlighted in bold. The default parameter set, acting as a comparison for other sets, is highlighted with a grey background. N-CSP refers to the number of CSP components used during feature extraction.

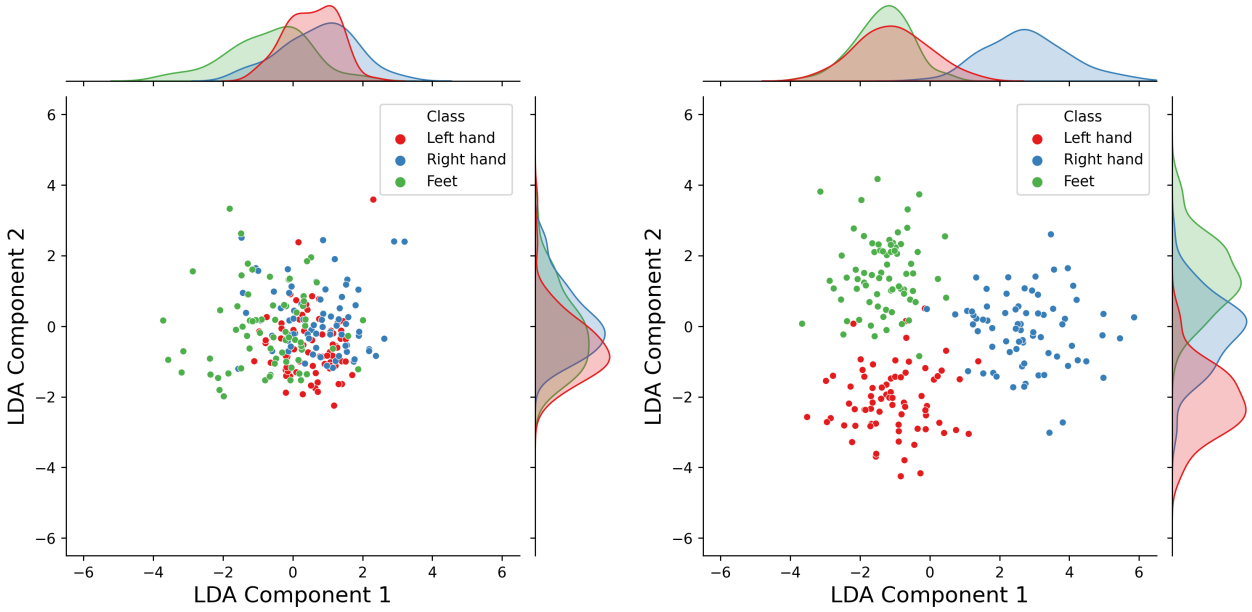
cortex were also present in the benchmark data set (see Figure 4). A subset consisting of these seven electrodes yielded an accuracy of 0.653, compared to 0.725 for all electrodes. While this constitutes a noticeable decrease, performance with seven channels was nevertheless high, indicating that lower-channel BCIs are feasible. Further reducing channels to three locations, C3, Cz and C4, decreased accuracy significantly to 0.544 indicating that the four additional motor cortex positions (FC1, FC2, CP1 and CP2) add considerable additional information.

Investigating the impact of the number of CSP components used during feature extraction indicated that eight components, as estimated on the training set, were sufficient for the benchmark data set. Extending the number of components to 16 and 22 actually slightly decreased performance. Reducing the number of components expectedly also reduced accuracy, with four components yielding 0.694 and two components resulting in an accuracy of 0.646.

Compared to the training data set which consisted of 45 trials per class, the benchmark data set contained 72 examples. We again simulated the effect of training on fewer trials by averaging performance across ten random subsets of the data. Accuracy was found to decrease as the number of trials decreased, with 45 trials resulting in an accuracy of 0.718, 30 trials in 0.685 and 20 trials in 0.615. Figure 7 illustrates a comparison to the training data set. Interestingly, for both data sets the gains in performance due to more trials seemed to diminish as the number of trials increased. For the benchmark data set there was only a marginal increase in performance from 45 to 72 trials.

In regards to class separability, the majority of users in the benchmark data set showed clearer

separation compared to the training and evaluation set, potentially due to the superior data quality associated with medical-grade EEG. While the MI classes were almost perfectly separated for the user with the highest accuracy, as seen in Figure 13, classes were also somewhat separated for the worst-performing user. Refer to Figure 20 in the Appendix for a visualization of the CSP filters learnt for these users.



(a) User with the lowest accuracy.

(b) User with the highest accuracy.

Figure 13: Class separability expressed through the learnt LDA components for the users with the lowest and the highest classification accuracy in the benchmark data set. Class densities for the individual LDA components are plotted on their respective axes.

3.3.2 Transfer Learning Classification

For the benchmark data set, transfer learning was conducted within the data set, that is transfer systems were trained on all but one user and tested on the left-out user. The overall results for the benchmark data set are shown in Figure 14, where results for the training data set are added for comparison. EA achieved an accuracy of 0.574 and outperformed the transfer baseline (0.452). However, while the performance of the EA system was similar to within-user classifiers for the training data set, the within-user classifiers outperformed EA by a larger margin for the benchmark data set. The within-user classifier trained on the least user-specific data (50-50) achieved an accuracy of 0.697, indicating that there was a considerable advantage of using user-specific data for the benchmark data set. Overall, Figure 14 also shows the large differences in performance between the training and the benchmark data set, likely a combination of the larger number of electrodes and the superior data quality associated with using medical-grade EEG.

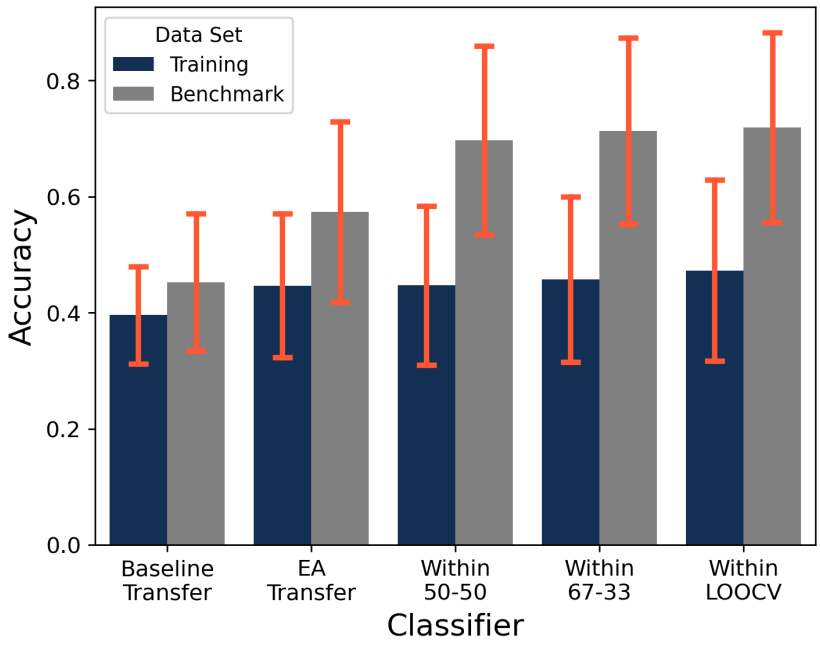
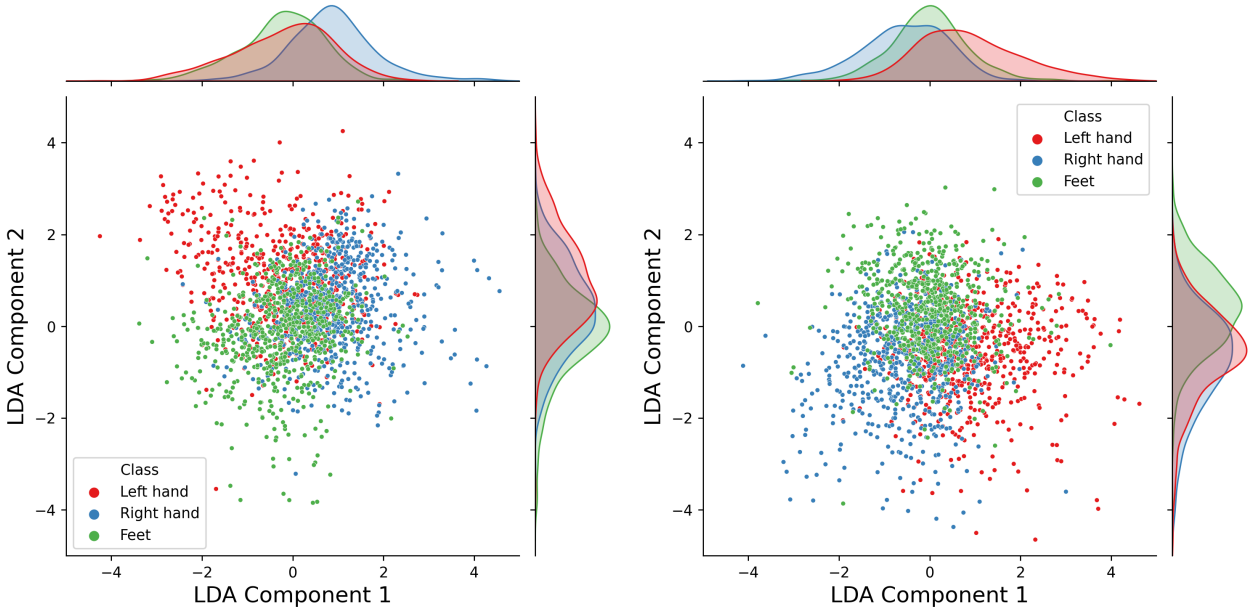


Figure 14: Classification accuracy averaged across users for different classifiers on the training and benchmark data set. Error bars represent the standard deviation across users.

Compared to the combined user data for the training data set, classes were already somewhat separated even without EA, as seen in Figure 15. Similar to the training and evaluation data set, applying EA seemed to improve class separation.



(a) Without Euclidean alignment.

(b) With Euclidean alignment.

Figure 15: Class separability of all user trials expressed through the learnt LDA components for the benchmark data set, with and without previous Euclidean alignment. Class densities for the individual LDA components are plotted on their respective axes.

Lastly, we also investigated transfer between the benchmark data set and the two OpenBCI data sets. As they differ in measurement devices, experimenters, experiment location and participant instructions, transfer ability should be limited. Table 4 summarizes the resulting accuracies. As can be seen, all baseline transfer from and to the benchmark data set resulted in chance-level accuracy. However, using EA when transferring from the training data set to the benchmark data set yielded an accuracy of 0.53. Additionally, transfer from the benchmark data set to the training and evaluation data sets resulted in accuracies of 0.43 and 0.42, respectively. These results give a clear indication that the EA transfer system, despite the differences between the two types of data sets, managed to extract motor imagery information which is shared between the data sets.

Transfer Source	Transfer Destination	Transfer Type	Accuracy
Training	Evaluation	Baseline	0.36
Training	Evaluation	EA	0.44
Training	Benchmark	Baseline	0.33
Training	Benchmark	EA	0.53
Benchmark	Evaluation	Baseline	0.33
Benchmark	Evaluation	EA	0.42
Benchmark	Training	Baseline	0.33
Benchmark	Training	EA	0.43

Table 4: Classification accuracies for baseline and EA transfer between the training, evaluation and benchmark data sets.

4 Discussion

This study aimed to address barriers to the large-scale adoption of BCIs in two ways. First, a low-cost EEG device with dry electrodes was used to measure brain activity, affording faster setup and easier access to BCIs for users without a research budget. Second, transfer learning was employed to reduce the need for user-specific data by training BCI systems on the combined data of previous users.

4.1 Low-Cost EEG Devices

In regards to the efficacy of low-cost EEG devices, classification accuracies for the training and evaluation data set, collected using an OpenBCI Ultracortex headset, were significantly lower compared to accuracies on a benchmark data set collected using medical-grade EEG. For the training and evaluation data sets, the highest accuracies observed were 0.473 and 0.44 compared to 0.725 on the benchmark data set. Figure 14 provides a visual comparison of the training and benchmark data set across different BCI architectures. Contrasting the LDA components learnt on both data sets as seen in Figure 10 and Figure 15 also shows the increased class separability for the benchmark data set. In order to better understand this discrepancy in performance between the data sets, we first aim to disentangle the different factors potentially affecting performance.

While the capabilities of the measurement device are essential, other factors can also lead to performance loss. Even before recording EEG data, the ability of users to successfully produce motor imagery will affect the final performance scores. As such, the quality of MI instructions is important and, when comparing performance between studies, it is crucial to consider whether participants were naive (i.e. first time BCI users) or experienced MI practitioners. In this study, we aimed to minimize the impact of user differences by collecting a sufficiently large sample. Between the training and evaluation set, 29 participants were recorded while performing motor imagery. For comparison, most participant samples in the BCI literature, including the benchmark data sets popularized through the BCI competitions, include around ten participants. Additionally, sampled participants in this study and the benchmark data set were naive MI practitioners, as would be the case for new users in applied scenarios as well.

Outside of user characteristics, other factors influencing the performance of BCIs include the measurement device used, the type, number and locations of electrodes, the number of experimental trials used during training and one's choice of BCI pipeline components. To establish whether performance loss was due to our BCI pipeline, we compared the performance of our system to accuracies reported by other studies for the same benchmark data set. Whereas the full benchmark data set contained trials belonging to four motor imagery classes, we only used the three MI classes also present in the training data set to allow for comparison. When testing our BCI on all four classes of the benchmark data set within users, an accuracy of 0.659 was achieved, similar to the accuracy of 0.644 reported by Naeem et al. (2006). He and Wu (2019) further reduced the benchmark data set to a binary classification of left and right hand. They reported an accuracy of 0.678 using a similar CSP-LDA approach between users, whereas our architecture yielded a binary accuracy of 0.677 for between-user classification. These concurrent performance scores on the benchmark data set indicate that our choice of feature extraction and classification methods, namely regularized CSP and LDA without covariance shrinkage, are not the cause for the decrease in performance on the training and evaluation data sets.

The number of experimental trials available for training also affects the BCI's ability to adequately separate MI classes. Whereas the training data set consisted of 45 trials per MI class, the benchmark data set featured 72 examples per class. Figure 7a shows the effect of fewer trials for both data sets. For the benchmark data set, the additional trials when expanding from 45 trials to 72 trials did not affect performance much. For the training data set, the performance also remained similar when reducing the number of trials to 30. For both data sets, a large decrease in accuracy was visible when only using 20 trials per class. Based on these results, we conclude that 30 trials per class are adequate when using a low-cost OpenBCI device, and 45 class trials are sufficient when using medical-grade EEG devices. Additionally, the transfer learning classifiers used in this study pooled together examples from all training data set users, essentially increasing the number of trials per class by a factor of 20.

In regards to electrodes, the benchmark data set was collected using 22 electrodes, whereas the training data set used eight electrodes. Between the two data sets, seven electrodes were placed over the same measurement locations, allowing for direct comparison. As seen in Figure 7b, reducing the number of electrodes used with the benchmark data set from 22 to seven led to a drop in performance from 0.725 to 0.653. However, as the training data set performance using the same set of electrodes was 0.473, it appears that the large decrease in performance when comparing the benchmark and training data set was not only due to fewer measurement electrodes. Furthermore, the majority of additional electrode locations used for the benchmark data set are not available for the OpenBCI headset used in this study. Access to additional locations over the motor cortex, such as C1 and C2, would require designing a custom frame for the OpenBCI electrodes.

Lastly, the recording devices for the two data sets also differed in their sampling rate and electrode type. While the OpenBCI headset's sampling rate of 125 was half that of the medical-grade EEG used for the benchmark data set, all BCIs tested in this study used a bandpass filter with an upper bound of 25 Hz or less. Following the Nyquist-Shannon theorem, a sampling rate of 125 should be sufficient to digitize all information contained in the used frequency bands (Shannon, 1949). Therefore, it appears that the largest factor for performance loss is the OpenBCI's dry electrodes. During data collection, we also encountered difficulties with electrodes flatlining before or during experiments. While the conductive gel used for medical-grade EEG reduces the ease of BCI use, it appears that it is necessary to prevent a large decrease in BCI performance.

While several studies have been introduced which use the OpenBCI Ultracortex headset for MI classification, direct comparisons are difficult. Many studies investigated binary classification, often using hand MI and a rest condition, rather than two types of MI (Saragih et al., 2022; Peterson et al., 2022; Shen et al., 2022). The work by Sterk (2022) employs, to our knowledge, the BCI paradigm most similar to the one introduced in this study. Also classifying MI of the left hand, right hand and feet, they reported a final accuracy score of 0.751 for their best-performing model. However, as they sampled two experienced MI practitioners, namely the researchers themselves, it is questionable how representative their results are for OpenBCI MI studies in general. Nevertheless, their best-performing model consisted of a Long Short-Term Memory neural network trained on trials split into smaller segments, essentially increasing the number of features available. Applying their approach to the data sets introduced in this study would be worthwhile to confirm whether their observed accuracies are due to a superior classifier or an unrepresentative sample.

Overall, the accuracies observed for our low-cost EEG device indicated that meaningful information relating to motor imagery was learnt from the EEG data, but the performance was considerably lower than performance for a medical-grade EEG device. This loss in performance seemed largely due to the different types of electrodes used, as well as fewer electrodes being available over the region of interest.

4.2 Transfer Learning

As a second focus point, this study investigated a transfer learning approach to enable the use of BCI systems between different users. Here we used Euclidean alignment (EA) to normalize experimental data for each user and facilitate the training of a BCI on combined user data. A BCI employing EA was then compared to a baseline transfer systems which pooled users together without normalization. For offline transfer learning, the EA system outperformed the baseline system across all three data sets. For the training and evaluation data sets, the EA system also achieved similar performance to classifiers which were trained on user-specific data. Additionally, EA also enabled transfer between data sets which were initially recorded using different types of EEG devices (see Table 4). Whereas the baseline transfer system performed at chance level when transferring between the training and benchmark data set, the EA system achieved an accuracy of 0.53 when trained on the training data set and tested on the benchmark data set. When trained on the benchmark data set, testing on the training and evaluation data set yielded accuracies of 0.43 and 0.42, close to the maximal performance observed for within-user classification for these data sets. Visual investigation of the learnt CSP filters, as seen, for example, in Figure 9 also indicated that EA resulted in filters which more clearly corresponded to hemispheric differences of the left and right-hand class. Similarly, class separation as expressed through LDA components also seemed to improve through EA (see, Figure 10).

In addition to offline transfer learning on the three data sets, the evaluation data collection was also used to investigate online transfer learning while users were performing motor imagery. This

involved estimating the alignment matrix used to normalize new trials on a set of twelve calibration trials. In contrast, the alignment matrix for offline transfer was estimated on all available data for each user. Additionally, both the baseline and EA transfer system were used for online classification during one experimental block each (30 trials). As such, the test set size was smaller compared to offline transfer learning tested on all of a user's experimental data. Figure 11 compares the different BCI systems. While the baseline transfer system and a classifier trained on user-specific data performed similarly between the offline and online scenarios, EA performed at chance level for the online scenario. Potential explanations could be that the estimated alignment matrix was not sufficient for online alignment, or that the test size of one block was too small to adequately assess performance. He and Wu (2019) simulated online EA on the benchmark data set by systematically increasing the number of experimental trials used in estimating the alignment matrix. Averaging this procedure over 30 repetitions, they found that performance stabilized when twelve or more trials were used. Simulated online learning on our training data set also indicated that twelve calibration trials were sufficient for stable alignment. However, in both cases simulation runs were repeated and averaged. As such, it seems that, on average, twelve calibration trials are sufficient to perform online alignment, but the alignment matrices estimated during the online evaluation were not. As the eventual use-cases for BCIs are online in nature and the majority of transfer learning investigations in the BCI literature are conducted offline, it seems especially important to further investigate which factors influence successful online alignment. If a large number of trials are needed to reliably learn an alignment matrix, the main advantage of BCI transfer learning, a reduced need for user-specific data, does not hold anymore. Therefore, it is essential for future research to establish how to reliably learn an alignment matrix. Another potential approach would be to adaptively re-estimate the alignment matrix as new user-specific examples become available. In this way, BCIs could be used even when little user-specific data is available and consequently improve over time.

4.3 Future Directions

Aside from further investigating how low-cost EEG devices and transfer learning influence BCI performance, several other avenues exist to build on the BCI architecture introduced in this study. Here we estimated the optimal frequency band for our architecture using hyperparameter estimation. Feature Bank CSP (FBCSP) represents a popular extension to the CSP feature extraction approach which optimizes the use of different frequency bands between users. Extending our architecture by using FBCSP could allow further tailoring of BCIs to individual users. Recent work has also focused on adaptive BCIs, that is systems which continue the learning process during use as new data becomes available (Lotte et al., 2018). Especially for an architecture employing transfer learning, where little user-specific data is available initially, adaptive continued learning holds great promise.

Furthermore, flatlined channels represented a significant challenge to adequate BCI performance in this study, particularly when transfer learning was involved. Our current approach involved determining flatlined channels before data collection and then retraining the transfer learning systems used for online classification without these flatlined channels. This approach could feasibly be extended to detect flatlined channels during data collection and retrain classifiers accordingly.

Motor imagery as a paradigm also has the distinct advantage of not requiring external stimuli to produce the neural patterns it aims to detect. In the current study, motor imagery was cued, that is users were instructed to perform motor imagery. Consequently, the period during which motor imagery was present in the EEG signal was known and classification involved distinguishing different types of motor imagery. Expanding this architecture to detect motor imagery in continuous EEG, requiring the classification of periods where no motor imagery is present, would greatly increase its

applicability to real-life scenarios.

In conclusion, this study showed that a low-cost EEG device can be used to conduct motor imagery classification. However, a more affordable, easier-to-use BCI came at the cost of a significant decrease in performance compared to a BCI employing medical-grade EEG. Additionally, Euclidean alignment was shown to enable successful transfer learning between users and across different data sets, but several challenges remain when conducting transfer learning in online scenarios. Further research is needed to determine the extent of user-specific data required for reliable online Euclidean alignment. Additionally, replicating the findings of this study using medical-grade EEG would confirm whether the online transfer challenges encountered in this study arise due to an interaction with the employed low-cost EEG device, or are inherent to online transfer in general. The resulting insights will then contribute to the overall goal of making BCIs more applicable to their intended uses outside of research laboratories.

References

Ahn, M., & Jun, S. C. (2015). Performance variation in motor imagery brain–computer interface: a brief review. *Journal of neuroscience methods*, *243*, 103–110.

Arvaneh, M., Guan, C., Ang, K. K., & Quek, C. (2013). Eeg data space adaptation to reduce intersession nonstationarity in brain-computer interface. *Neural computation*, *25*(8), 2146–2171.

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2011). Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, *59*(4), 920–928.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., ... Flor, H. (1999). A spelling device for the paralysed. *Nature*, *398*(6725), 297–298.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Müller, K.-r., & Nikulin, V. (2007). Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. *Advances in neural information processing systems*, *20*.

Blankertz, B., Muller, K.-R., Curio, G., Vaughan, T. M., Schalk, G., Wolpaw, J. R., ... others (2004). The bci competition 2003: progress and perspectives in detection and discrimination of eeg single trials. *IEEE transactions on biomedical engineering*, *51*(6), 1044–1051.

Blankertz, B., Muller, K.-R., Krusienski, D. J., Schalk, G., Wolpaw, J. R., Schlogl, A., ... Birbaumer, N. (2006). The bci competition iii: Validating alternative approaches to actual bci problems. *IEEE transactions on neural systems and rehabilitation engineering*, *14*(2), 153–159.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Congedo, M., Barachant, A., & Bhatia, R. (2017). Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, *4*(3), 155–174.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Daly, J. J., & Wolpaw, J. R. (2008). Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, *7*(11), 1032–1043.

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., & Dutoit, T. (2013). Performance of the emotiv epoc headset for p300-based applications. *Biomedical engineering online*, *12*(1), 1–15.

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, *70*(6), 510–523.

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R., & Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, *22*(9), 1305–1312.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1–13. doi: 10.3389/fnins.2013.00267

Grosse-Wentrup, M., & Buss, M. (2008). Multiclass common spatial patterns and information theoretic feature extraction. *IEEE transactions on Biomedical Engineering*, *55*(8), 1991–2000.

Hart, P. E., Stork, D. G., & Duda, R. O. (2000). *Pattern classification*. Wiley Hoboken.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

He, H., & Wu, D. (2019). Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, *67*(2), 399–410.

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., & Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, *11*(1), 20–31.

Kang, H., & Choi, S. (2014). Bayesian common spatial patterns for multi-subject eeg classification. *Neural Networks*, *57*, 39–50.

Koles, Z. J., Lazar, M. S., & Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background eeg. *Brain topography*, *2*(4), 275–284.

Krusienski, D. J., McFarland, D. J., Principe, J. C., & Wolpaw, E. (2012). Bci signal processing: feature extraction. *Brain-Computer Interfaces: Principles and Practice, eds JR Wolpaw and EW Wolpaw (New York, NY: Oxford University Press)*, 123–146.

LaRocco, J., Le, M. D., & Paeng, D.-G. (2020). A systemic review of available low-cost eeg headsets used for drowsiness detection. *Frontiers in neuroinformatics*, 42.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, *10*(5), 603–621.

Ledoit, O., & Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, *30*(4), 110–119.

Lotte, F. (2015). Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, *103*(6), 871–890.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, *15*(3), 031005.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, *4*(2), R1.

Lotte, F., & Guan, C. (2010). Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. *IEEE Transactions on biomedical Engineering*, *58*(2), 355–362.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Marshall, D., Coyle, D., Wilson, S., & Callaghan, M. (2013). Games, gameplay, and bci: the state of the art. *IEEE Transactions on Computational Intelligence and AI in Games*, *5*(2), 82–99.

Maskeliunas, R., Damasevicius, R., Martisius, I., & Vasiljevas, M. (2016). Consumer-grade eeg devices: are they usable for control tasks? *PeerJ*, *4*, e1746.

Miller, K. J., Schalk, G., Fetz, E. E., Den Nijs, M., Ojemann, J. G., & Rao, R. P. (2010). Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proceedings of the National Academy of Sciences*, *107*(9), 4430–4435.

Muller-Putz, G. R., & Pfurtscheller, G. (2007). Control of an electrical prosthesis with an ssvep-based bci. *IEEE Transactions on biomedical engineering*, *55*(1), 361–364.

Naeem, M., Brunner, C., Leeb, R., Graimann, B., & Pfurtscheller, G. (2006). Seperability of four-class motor imagery data using independent components analysis. *Journal of neural engineering*, *3*(3), 208.

Neuper, C., & Pfurtscheller, G. (2001). Evidence for distinct beta resonance frequencies in human eeg related to specific sensorimotor cortical areas. *Clinical Neurophysiology*, *112*(11), 2084–2097.

OpenBCI. (2022, Sep). *Ultracortex mark iv: Openbci documentation.* Retrieved from https://docs.openbci.com/AddOns/Headwear/MarkIV/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, *60*(4), 389–443.

Peterson, V., Galván, C., Hernández, H., Saavedra, M. P., & Spies, R. (2022). A motor imagery vs. rest dataset with low-cost consumer grade eeg hardware. *Data in Brief*, *42*, 108225.

Pham, D. T. (2001). Joint approximate diagonalization of positive definite hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, *22*(4), 1136–1152.

Ramoser, H., Muller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, *8*(4), 441–446.

Sajda, P., Gerson, A., Muller, K.-R., Blankertz, B., & Parra, L. (2003). A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on neural systems and rehabilitation engineering*, *11*(2), 184–185.

Saragih, A. S., Basyiri, H. N., & Raihan, M. Y. (2022). Analysis of motor imagery data from eeg device to move prosthetic hands by using deep learning classification. In *Aip conference proceedings* (Vol. 2537, p. 050009).

Sawangjai, P., Hompoonsup, S., Leelaarporn, P., Kongwudhikunakorn, S., & Wilaiprasitporn, T. (2019). Consumer grade eeg measuring sensors as research tools: A review. *IEEE Sensors Journal*, *20*(8), 3996–4024.

Shannon, C. (1949, jan). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21. Retrieved from https://doi.org/10.1109/jrproc.1949.232969 doi: 10.1109/jrproc.1949.232969

Shen, X., Wang, X., Lu, S., Li, Z., Shao, W., & Wu, Y. (2022). Research on the real-time control system of lower-limb gait movement based on motor imagery and central pattern generator. *Biomedical Signal Processing and Control*, *71*, 102803.

Shinners, P. (2011). *Pygame - python game development.* http://pygame.org/.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, *38*(4), 387–401.

Sterk, H. (2022). *Think & play: Classifying left and right hand motor imagery eeg signals by using dry electrodes for real-time bci gaming* (Unpublished master's thesis). Utrecht University.

Steyrl, D., Kobler, R. J., Müller-Putz, G. R., et al. (2016). On similarities and differences of invasive and non-invasive electrical brain signals in brain-computer interfacing. *Journal of biomedical science and engineering*, *9*(08), 393.

Subha, D. P., Joseph, P. K., Acharya U, R., Lim, C. M., et al. (2010). Eeg signal analysis: a survey. *Journal of medical systems*, *34*(2), 195–212.

Sun, S., & Zhou, J. (2014). A review of adaptive feature extraction and classification methods for eeg-based brain-computer interfaces. In *2014 international joint conference on neural networks (ijcnn)* (pp. 1746–1753).

Sutton, S., Braren, M., Zubin, J., & John, E. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188.

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., ... others (2012). Review of the bci competition iv. *Frontiers in neuroscience*, 55.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial.* Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Wu, D., Xu, Y., & Lu, B.-L. (2020). Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(1), 4–19.

Yger, F., Berar, M., & Lotte, F. (2016). Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(10), 1753–

1762.

Yohanandan, S. A., Kiral-Kornek, I., Tang, J., Mshford, B. S., Asif, U., & Harrer, S. (2018). A robust low-cost eeg motor imagery-based brain-computer interface. In *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 5089–5092).

Zanini, P., Congedo, M., Jutten, C., Said, S., & Berthoumieu, Y. (2017). Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, *65*(5), 1107–1116.

Zhu, D., Bieger, J., Garcia Molina, G., & Aarts, R. M. (2010). A survey of stimulation methods used in ssvep-based bcis. *Computational intelligence and neuroscience*, *2010*.

Appendix

A Hyperparameter Estimation

This section discusses the effect of different hyperparameters on performance for the training data set in more detail. In regards to general parameters, a 10-12 Hz bandpass filter corresponding to a sub-band of the larger mu frequency band yielded the highest performance (0.473). A BCI using the entire mu band (8-13 Hz) also performed well with an accuracy of 0.447. Interestingly, the best performance using the beta-band (18-25 Hz) was significantly lower with an accuracy of 0.391, indicating that there is only little discriminatory information in the beta band, as an accuracy of 0.33 for a three-class problem would constitute chance-level performance. A combination of both bands (8-25 Hz) resulted in accuracies of up to 0.443 but did not outperform the singular mu-band classifiers. Potentially, the beta-band carries more noise which impairs classification ability when it is added to the mu-band.

In regards to the regularization of the CSP feature extraction approach, using covariance shrinkage estimated through the Ledoit-Wolf lemma proved superior to no regularization. While differences were less pronounced, the highest result for no regularization was 0.469, compared to 0.473 with shrinkage, indicating that there is some benefit to regularized feature extraction.

For the number of CSP components, eight was observed to be consistently superior. Both two and four components performed less well across all configurations of the other general parameters, indicating that eight components are indeed necessary to capture all spatial patterns inherent in the data. For two and four components the highest achieved accuracies were 0.409 and 0.452, respectively.

Performance differences for imagery window lengths were less pronounced than expected. Given that this parameter essentially governs the number of time points for use in feature extraction, one would expect that two seconds of imagery would show a clearer decrease in performance compared to four seconds. However, accuracies for all three imagery lengths were generally quite close for different settings of the other parameters. For two, three and four seconds of imagery the highest accuracies were 0.447, 0.452 and 0.473, respectively, indicating that participants consistently performed motor imagery throughout the imagery period.

In regards to classifiers, LDA performed well in general with accuracies for automatically estimated shrinkage, fixed shrinkage values, as well as no shrinkage, all being above 0.46. The highest performance was found when using no covariance shrinkage as a regularization method. A possible explanation could be that covariance shrinkage performed during CSP feature extraction negatively interacts with further regularization during LDA. An architecture which used shrinkage for LDA but not for CSP also achieved a high accuracy of 0.461, indicating that regularization in general is worthwhile.

Performance of the Random Forest classifier was homogeneous, with accuracies ranging from 0.42 to 0.457 and overall oscillating around 0.44. Additionally, performance generally increased as the maximal tree branch depth and the number of decision trees used increased.

For the Support Vector Machines, the highest accuracy of 0.466 resulted from a linear kernel in combination with a regularization parameter of C = 0.75 and C = 1. In regards to kernels, the linear kernel clearly outperformed the polynomial, radial basis function (rbf) and sigmoid kernels across all levels of regularization. Generally, a smaller regularization value of C = 0.25 performed worst, with only minute differences between C = 0.5 and C = 0.75 and the best performance for C = 1.

	LDA		
	Shrinkage	Accuracy	SD
	None	0.473	0.156
	Auto	0.460	0.155
	0.25	0.467	0.157
	0.5	0.464	0.157
	0.75	0.460	0.147
	RF		
Depth	Trees	Accuracy	SD
10	10	0.420	0.143
10	100	0.446	0.158
10	500	0.447	0.152
20	10	0.429	0.137
20	100	0.441	0.159
20	500	0.449	0.152
100	10	0.431	0.134
100	100	0.442	0.158
100	500	0.457	0.147

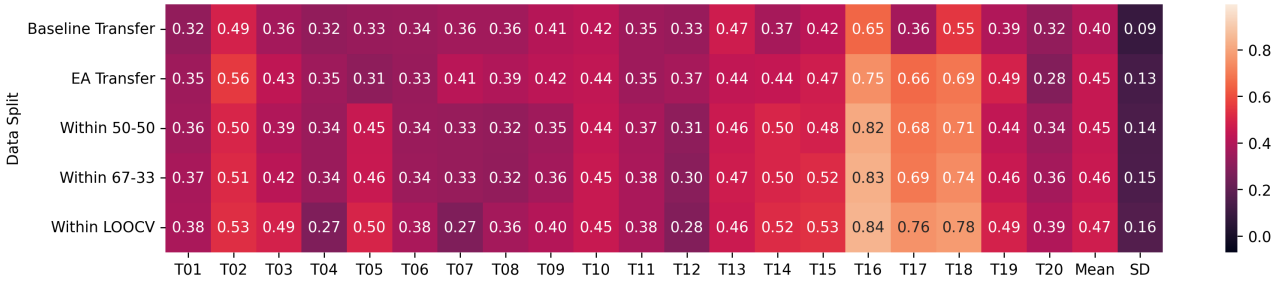
	SVM		
C	Kernel	Accuracy	SD
0.25	linear	0.451	0.157
0.25	poly	0.398	0.153
0.25	rbf	0.352	0.232
0.25	sigmoid	0.379	0.213
0.5	linear	0.465	0.146
0.5	poly	0.407	0.149
0.5	rbf	0.437	0.168
0.5	sigmoid	0.417	0.183
0.75	linear	0.466	0.147
0.75	poly	0.412	0.153
0.75	rbf	0.444	0.164
0.75	sigmoid	0.411	0.177
1	linear	0.466	0.147
1	poly	0.425	0.146
1	rbf	0.450	0.160
1	sigmoid	0.423	0.167

Table 5: Average within-user classification accuracy on the training data set for different classifiers. Each row corresponds to a specific set of classifier parameters. LDA, SVM and RF refer to Linear Discriminant Analysis, Support Vector Machine and Random Forest, respectively.

B Classification Results

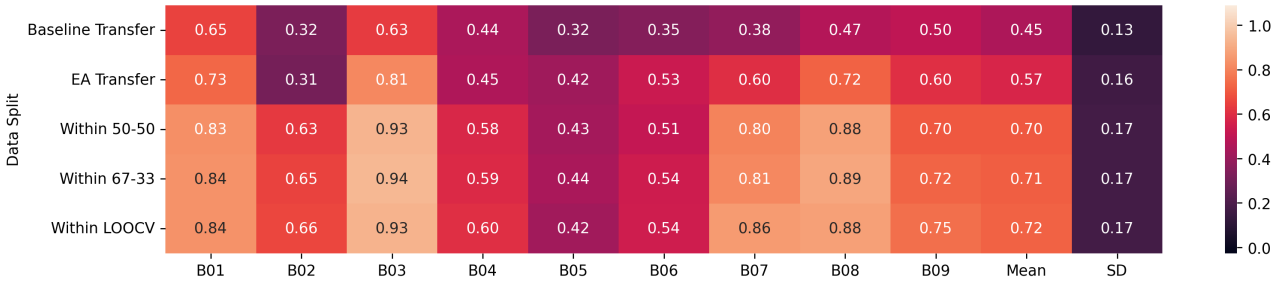
Trials	Channels	Classes	Accuracy	SD
20	8	3	0.413	0.169
30	8	3	0.464	0.145
45	8	3	0.473	0.156
45	3	3	0.415	0.127
45	7	3	0.472	0.146
45	8	2	0.588	0.121

Table 6: Classification accuracy on the training data set for subsets of fewer trials, channels and classes.



(a) Training data set.

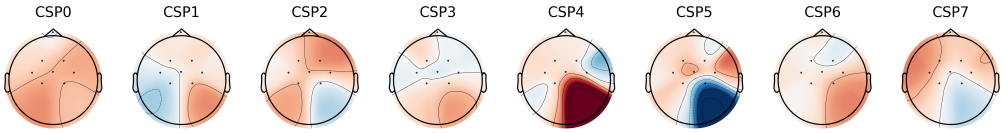
(b) Evaluation data set.



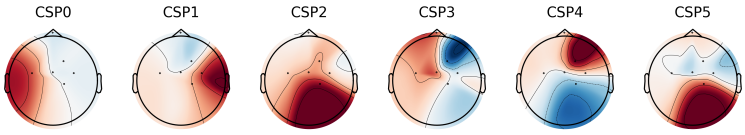
(c) Benchmark data set.

Figure 16: User classification accuracies for all data sets.

C CSP and LDA Visualization

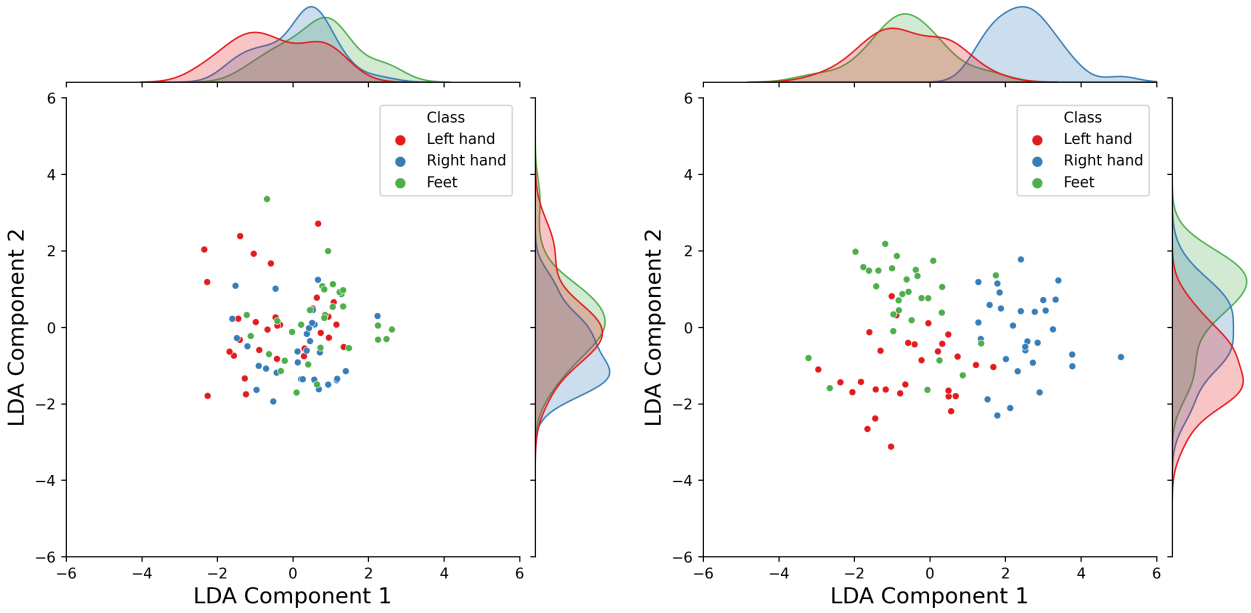


(a) User with the lowest accuracy.



(b) User with the highest accuracy.

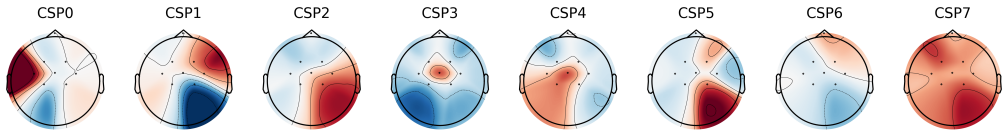
Figure 17: Neural patterns underlying the learnt CSP filters for the users with the lowest and the highest classification accuracy in the evaluation data set. Two channels flatlined for the user with the highest accuracy resulting in only six components being estimated.



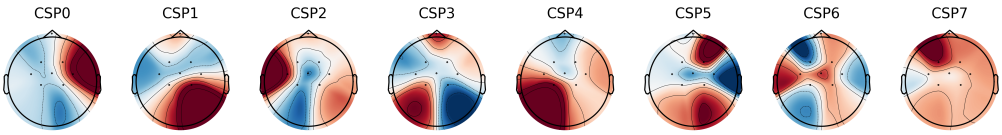
(a) User with the lowest accuracy.

(b) User with the highest accuracy.

Figure 18: Class separability expressed through the learnt LDA components for the users with the lowest and the highest classification accuracy in the evaluation data set. Class densities for the individual LDA components are plotted on their respective axes.

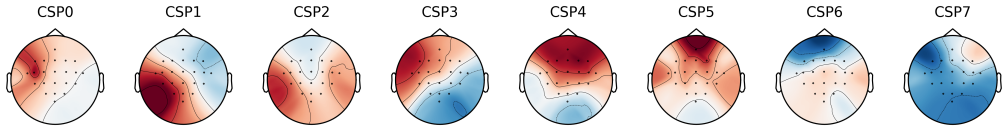


(a) Without Euclidean alignment.

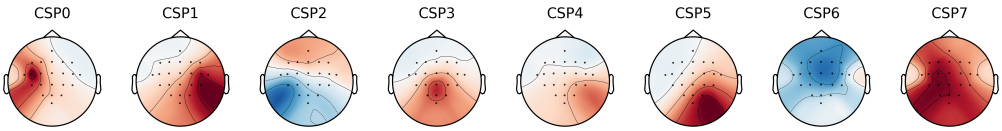


(b) With Euclidean alignment.

Figure 19: Neural patterns underlying the CSP filters learnt on all users in the evaluation data set, with and without previous Euclidean alignment.

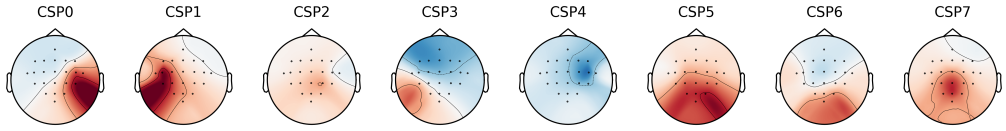


(a) User with the lowest accuracy.

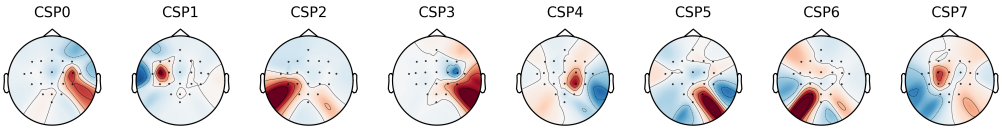


(b) User with the highest accuracy.

Figure 20: Neural patterns underlying the learnt CSP filters for the users with the lowest and the highest classification accuracy in the benchmark data set.



(a) Without Euclidean alignment.



(b) With Euclidean alignment.

Figure 21: Neural patterns underlying the CSP filters learnt on all users in the benchmark data set, with and without previous Euclidean alignment.