Stephanie Brandl, Laura Frølich, Johannes Höhne, Klaus-Robert Müller, Wojciech Samek

Brain–computer interfacing under distraction: an evaluation study

Brandl, S., Frølich, L., Höhne, J., Müller, K.-R., & Samek, W. (2016). Brain–computer interfacing under distraction: an evaluation study. Journal of Neural Engineering, 13(5), 56012. https://doi.org/10.1088/1741-2560/13/5/056012

Brain-Computer Interfacing under Distraction: An Evaluation Study

Stephanie Brandl[†], Laura Frølich[‡], Johannes Höhne[†], Klaus-Robert Müller[†¶], and Wojciech Samek[§]

[†]Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany
[‡]Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[¶]Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea
[§]Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany

E-mail: stephanie.brandl@tu-berlin.de, klaus-robert.mueller@tu-berlin.de, wojciech.samek@hhi.fraunhofer.de

Abstract.

Objective

While motor-imagery based Brain-Computer Interfaces (BCIs) have been studied over many years by now, most of these studies have taken place in controlled lab settings. Bringing BCI technology into everyday life is still one of the main challenges in this field of research.

Approach

This paper systematically investigates BCI performance under 6 types of distractions that mimic out-of-lab environments.

Main results

We report results of 16 participants and show that the performance of the standard CSP+RLDA classification pipeline drops significantly in this "simulated" out-of-lab setting. We then investigate three methods for improving the performance: 1) artifact removal, 2) ensemble classification, and 3) a 2-step classification approach. While artifact removal does not enhance the BCI performance significantly, both ensemble classification and the 2-step classification combined with CSP significantly improve the performance compared to the standard procedure.

Significance

Systematically analyzing out-of-lab scenarios is crucial when bringing BCI into everyday life. Algorithms must be adapted to overcome nonstationary environments in order to tackle real-world challenges.

1. Introduction

Brain-Computer Interfacing (BCI) [1, 2] allows non-muscular communication between a human and a computer by detecting a user's intentions via brain signals, e.g. with an electroencephalogram (EEG), and translating them into control commands. This is particularly useful for people affected by diseases that lead to the loss of muscular control,

such as amyotrophic lateral sclerosis (ALS), brainstem stroke, multiple sclerosis and especially for people who suffer from locked-in syndrome. BCIs can not only be applied for communication but also for the control of external devices such as a wheelchair [3], for rehabilitation [4] and mental state monitoring [5].

Over the years, various improvements in BCI research have been presented. Integrating machine learning algorithms has helped to substantially reduce calibration time [6, 7, 8, 9] which has crucially enhanced usability of BCIs. Also, novel approaches in robust feature extraction [10, 11, 12, 13, 14], artifact detection [15, 16, 17] and adaptive methods [18, 19, 20] have particularly improved reliability.

Those approaches already work well in controlled lab environments. However, these environments are highly artificial and significantly differ from everyday life situations, where people have to handle various visual, auditory or other cognitive distractions. In order to fulfill its main purpose, to provide disabled people with a non-muscular communication pathway, BCI research has recently begun to leave those controlled lab environments. Since most algorithms may not work well in real world scenarios, it becomes mandatory to investigate and improve them.

First steps into the real world have been made [5]. Ambulatory BCIs, for instance, allow participants to walk indoors [21], outdoors [22] and on a treadmill [23] while using a P300 spelling device. Motor imagery-based BCIs have been investigated under the presence of speech [24] and also applied to control a pinball machine [25], a virtual helicopter [26], a quadcopter [27] and a tetris game [28]. Several patient studies have been carried out on stroke, tetraplegic and even locked-in patients [29, 30, 31, 32, 33, 34].

However, there still lacks, to the best of our knowledge, a study where data is recorded in systematic out-of lab scenarios and evaluated in detail. In this paper we close this gap by presenting and analyzing a motor imagery-based BCI study where 16 healthy participants were distracted by 6 different secondary tasks while performing a primary motor imagery task. The aim of those secondary tasks was to simulate a more realistic environment where participants e.g. listen to news, watch a flickering video, search the room for a particular number or handle vibrotactile stimulation. This study investigates BCI performance in environments different from the training environment and analyze the problems that occur in such scenarios.

Note that this paper represents an extension of a preliminary analysis of this study [35]. In particular, we investigate standard machine learning techniques, commonly used in BCI research, in semi-realistic scenarios. We identify three major problems which lead to poor performance in those out-of-lab scenarios

(i) Artifact contamination

(ii) Feature shifting

(iii) High cognitive workload

and discuss several new approaches which try to overcome those challenges.

The rest of this paper is organized as follows: In the next section, we present the BCI study and describe its setup in detail. In the third section, we evaluate BCI performance with standard

machine learning techniques and discuss the problems that appear. In the fourth section, we present novel approaches to tackle the identified problems, before we summarize and discuss our findings in the conclusion.

2. Experiments

The simulation of everyday life situations during a motor imagery task such as watching TV or listening to news, gives us the opportunity to investigate BCI performance in a more realistic environment. Since we conduct those experiments in-lab, we are able to systematically analyze scenarios and to draw conclusions for future experiments.

Table 1: Overview over distraction tasks.

	Task	**Purpose**
Clean	Motor imagery without distractions	Control task
Eyes-closed	Motor imagery with eyes closed	Investigation of α-rhythm
News	Motor imagery while listening to news sequences	Distraction + activation of auditory cortex
Numbers	Searching the room for one of the 26 letter-number combinations hanging on the wall while performing the motor imagery task	Distraction + muscular artifacts
Flicker	Motor imagery while watching a video with a flicker in gray shades at a frequency of 10 Hz	Investigation of SSVEP
Stimulation	Motor imagery plus vibtrotactile stimulation on both forearms with carrier frequencies of 50 and 100 Hz, modulated at 9, 10 and 11 Hz	Investigation of SSVSEP

2.1. Participants

We recorded EEG from 16 healthy participants (6 female; age range: 22-30 years; mean age: 26.3 years) of which only three had previously participated in another BCI experiment. Since all the instructions were in German, a certain level of language proficiency was required. Three of the participants are members of the TU Berlin Machine Learning Group, whereas the other volunteers were paid for their participation.

2.2. Experimental setup

The participants sat in an armchair at a distance of 1m from a 24" computer screen. During the experiment, the participants wore headphones to receive auditory instructions.
To record the EEG signals, we used a Fast'n Easy Cap (EasyCap GmbH) with 63 wet

Ag/AgCl electrodes and placed them at symmetrical positions according to the international 10–20 system [36] referenced to the nose. We furthermore used two 32-channel amplifiers (BrainProducts) to amplify the signals, which were sampled at 1000 Hz.
Including breaks and preparation, each experimental session lasted about three hours, of which signal recording took about 90 minutes. Before starting the main experiment, we recorded 8 baseline trials in which the participant had to either open or close both eyes within 15 seconds.
We divided the main experiment into 7 runs. Each run lasted about 10 minutes and consisted of 72 trials. Since the first run was used as a calibration phase, no distractions were added and no feedback was given. See Figure 1 for an overview of the setup. After the calibration phase, each run also consisted of 72 trials but was divided into 3×4 trials per secondary task. Each trial lasted 4.5 seconds and included one motor imagery task. The order of the secondary tasks was randomized such that it changed after a random sequence of two *left* and two *right* trials. See Figure 2 for an example of a run during the feedback phase. Auditory instructions (*left* and *right* commands) were given over the headphones at the beginning of each trial (since the experiment was conducted in German, the actual instructions were *links* and *rechts*). When the trial finished after 4.5 seconds, there was a *stop* command followed by a break of 2.5 seconds, after which the next trial started. Every three to four minutes the participant had the possibility to take a break.
Participants could choose their individual haptic motor imagery task involving a movement that only effects the hand itself. Most popular, also because it was the first and easiest example we gave them, was the imagination of squeezing a soft ball. Other strategies involved opening a water tap, piano playing or using a salt shaker.
To keep motivation levels high, we included auditory feedback after the calibration phase. Therefore, Laplacian filters [37] of the C3 and C4 electrode were calculated and a classifier based on Regularized Linear Discriminant Analysis (RLDA) [38] was computed using the spectral power of the signals in two broad bands (9-13 Hz and 18-26 Hz) as features. During the feedback phase, the classifier was applied to classify the motor imagery tasks and to provide auditory feedback. Due to the *eyes-closed* task we could not give any feedback via the screen. This means that the auditory *stop* command was followed by a *decision left* (*Entscheidung links*) or *decision right* (*Entscheidung rechts*) during the 2.5 seconds break.

2.3. Distractions

To study the effects of increased cognitive load and additional artifacts, we included 6 distractions in form of secondary tasks in addition to a primary motor imagery task in the experimental setup (see Table 1). We will now explain the design and motivation for those distractions.

(i) *Clean*
This condition serves as a control group without any distraction.

(ii) *Eyes Closed*
Participants performed motor imagery with their eyes closed. Here, we investigated the

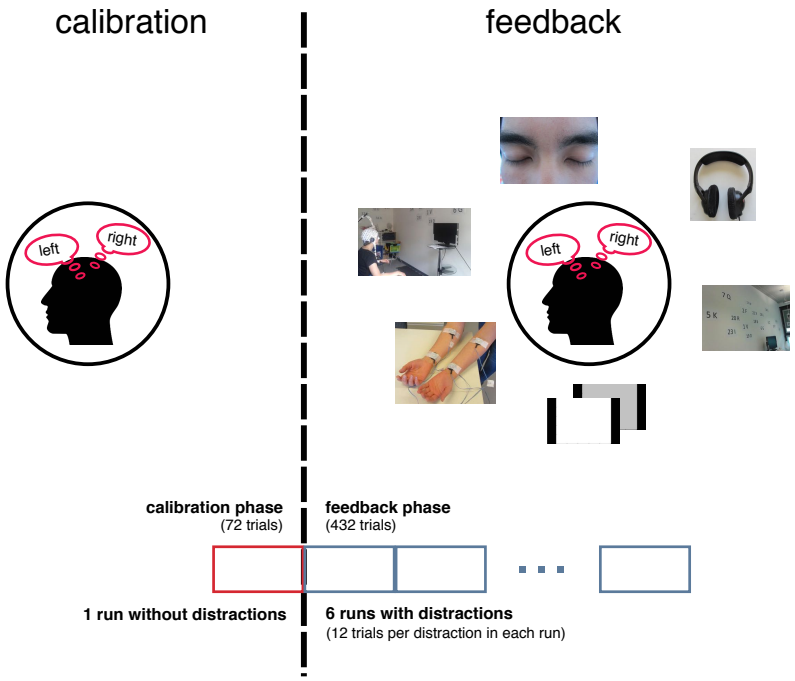

Figure 1: The experiment consisted of 7 runs, each containing 72 motor imagery trials. The calibration run did not contain secondary tasks, whereas each feedback run consisted of 12 trials per type of secondary task.

effect of a more prominent α-rhythm due to the closed eyes. Since the motor task related μ-rhythm appears within a similar frequency band (8-13 Hz), we expected an overlay with the α-rhythm. Because of this task we gave all instructions and feedback over headphones and not visually.

(iii) *News*

Sequences of a public newscast were played over the headphones containing current news and news from 1994. Each sequence was played once in each experiment, except for participant *od*, for whom some files were played twice. Here, we analyzed the influence of the cognitive distraction and of an activated auditory cortex on the motor imagery performance. The volume of the instructions was increased for this task.

In the experiment, we did not check if the participants were actively listening to the newscast. We therefore tried to contact most of them afterwards; only three participants remembered that they were actively listening (participants *od, njy, njz*) and one perceived it more as background noise *(nkm)*. The remaining participants either could not remember or could not be contacted.

(iv) *Numbers*

For this task, 26 sheets of paper with a randomly mixed letter-number combination were put up on the wall in front of the participants and also on the left and right side of the room. This means the participants needed to turn their head in order to see the sheets.

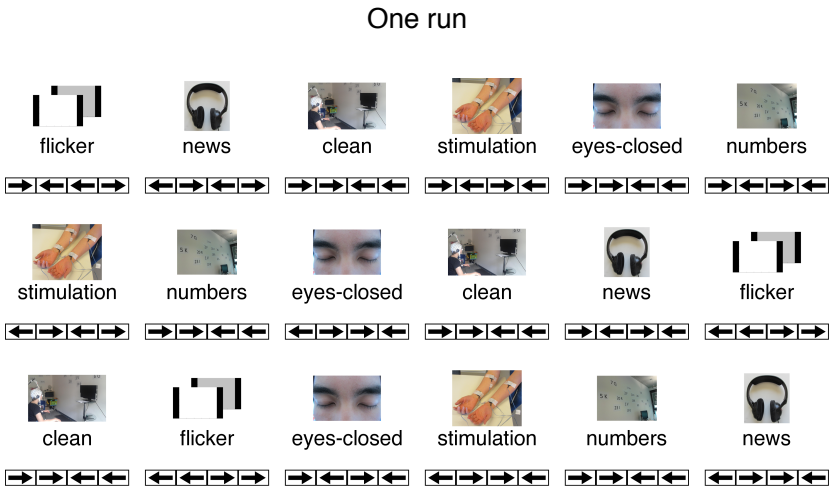

Figure 2: This represents an example of a run during the feedback phase. After 4 trials (2 left MI, 2 right MI), the secondary task changes randomly. In every run we record 12 trials of each secondary task.

For each trial a new window appeared on the screen asking the participants to search the room for a particular letter to match with a stated number. Each combination was shown 2-3 times to all participants. We counted the found letters and out of 72 trials, 59.7 combinations were found on average. This task investigates both, the effect of a high cognitive distraction and of additional muscular artifacts.

(v) *Flicker*
A flickering stimulus with alternating gray shades at a frequency of 10 Hz was presented on the screen. We included this task to analyze the influence of the *steady state visually evoked potential* (SSVEP) [39] .

(vi) *Stimulation*
We placed two coin vibration motors with a diameter of 3 cm on the insides of both forearms, one over the wrist and another one just below the elbow. To investigate the interference of *steady state vibration somatosensory evoked potential* (SSVSEP) [40, 41] on the motor imagery task, vibrotactile stimulation was carried out with carrier frequencies of 50 and 100 Hz, each modulated at 9, 10 and 11 Hz.

2.4. Data analysis

We downsampled the data to 100 Hz and selected a frequency band between 5 and 35 Hz and a time interval for the offline analysis of each participant individually as described in [42]. All frequency bands and time intervals are displayed in the second and third column of Table 2. All classification approaches throughout this paper are based on feature extraction using Common Spatial Patterns (CSP) [43] and RLDA [42, 9].

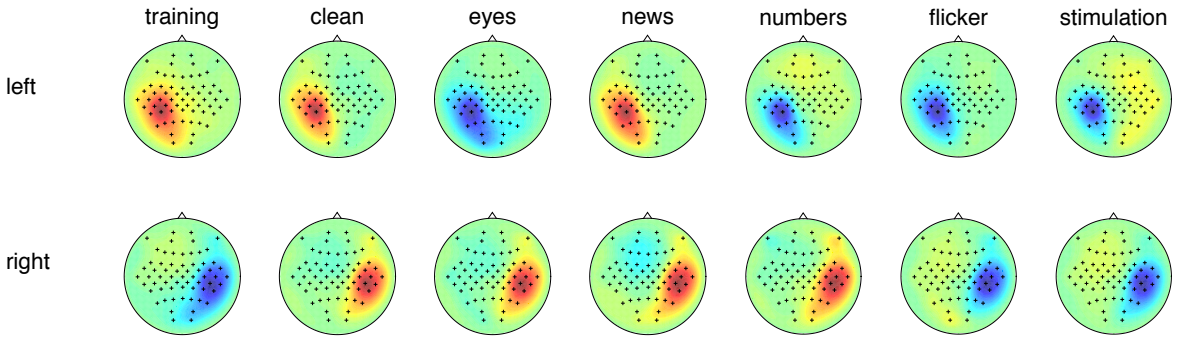

Figure 3: CSP Patterns for participant *od* for training and testing data

3. Evaluation of in-lab training

In this section, we present the results, obtained from classification based on the calibration data with three CSP filters per class. That is, we trained the classifier on data without secondary tasks and tested it on data including secondary tasks. Translating this to our systematic framework, we trained *in-lab* and tested in the *out-of-lab* setting. Since the resulting classification rates lead to the assumption of poor out of lab BCI performance, we investigate possible reasons for this outcome.

3.1. Classification on clean training

In Figure 3 and 4, we show patterns, ERD/ERS effects [44] and spectra of participant *od*. We selected the most discriminative CSP filter in both classes.
In Figure 3, we plotted CSP patterns for all secondary tasks and for the calibration phase. In Figure 4, each subplot depicts data of left hand MI (blue) and right hand MI (red). In the first two columns, we visualized ERD/ERS effects as time series, each row representing one secondary task. Columns 3-4 show spectra between 5 and 35 Hz. Below each subplot, we also added colorbars with r^2-values, representing the discriminancy between left and right hand MI. Since r^2-values vary much between ERD/ERS and spectra, there are two separate colorbars, displayed on the upper left and right. We further marked the background of the selected time interval and frequency band which we used for extracting the CSP filters in grey.
Changes due to the different character of the secondary tasks are visible in both, spectra and ERD/ERS effects. The fourth row with the *numbers* task shows a much less prominent peak in the spectrum, especially for the right hand CSP filter (column 4). For the *stimulation* task, the spectrum of the left hand CSP filter (column 3) shows several additional peaks, e.g. over 18 Hz, 20 Hz and 22 Hz. They represent the 2f peaks (first harmonics), since modulation frequencies (f) were at 9 Hz, 10 Hz and 11 Hz. This coincides with findings in earlier publications on SSVSEP [40].
In Table 2, average classification rates are summarized; each row represents one participant. Classification accuracies vary much between participants [between 49.42% and 90.97%] but also within a participants' recording [*njz:* $45.83\% - 83.33\%$]. Most of the volunteers par-

ticipated for the first time in a BCI experiment, so not everyone achieved classification rates significantly higher than chance level. Applying binomial tests ($\alpha = 5\%$) led to thresholds of 54.17% and 61.11% for the overall experiment (432 trials) and for each secondary tasks (72 trials), respectively. For classification accuracies above or equal to these thresholds, we assume actual BCI control. The topic of significant BCI control taking into account the number of trials has been addressed in [45].

We grouped the 16 participants in three categories. The first category contains the three participants which gained significant BCI control in all secondary tasks. The second category contains 7 participants that reached the overall threshold of at least 54.17%. The last category contains the remaining 6 participants below this threshold which means that we cannot assume any BCI control.

Out of the 16 participants, 4 did not reach the threshold of 61.11% in their best distraction task (*nkp,ma4,nkk,nks*). Especially in the *numbers* task with its dual challenge of contaminated data and high cognitive load, most participants only reached classification accuracies around 50%. Also for the *eyes-closed* task, only 6 participants reached the threshold of 61.11%. Whereas most users gained their highest classification rates in the *flicker* task.

3.2. Why poor performance?

Most participants did not reach the significance threshold in at least one task. We therefore need to find out how to improve the overall performance. We found several possible explanations. One is that the distraction tasks influence the EEG recordings in a way that leads to major feature shifts between calibration data (without secondary tasks) and testing data (with secondary tasks). Some of the tasks might cause too many artifacts. These could contaminate data in a way that makes it impossible to identify actual neural activity. It is also worth considering that some tasks are more cognitively demanding than others. This could increase the difficulty to fully concentrate on the motor imagery task. This leads to three possible explanations for the poor BCI performance which we will tackle in the following sections.

(i) artifact contamination

(ii) major feature shifts between distraction tasks

(iii) participants are too distracted to focus on the motor imagery task, so the data is not separable

Apart from those three reasons there is also the probability of BCI illiteracy which is the lack of ability to control a BCI. This phenomenon affects an estimated proportion of $15\% - 30\%$ of the participants and has already been discussed in broad variety in the literature [46, 47, 48, 49]. In the *clean* task which represents a classical motor imagery task, 11 out of 16 participants reached the significance threshold of 61.11% which means, BCI control failed for 31.25% of the participants. This proportion lies slightly above the $15\% - 30\%$ range and could be explained by the fast change of scenery. The secondary task changed every 4 trials. We will investigate the subject of non-separability of the data in Section 3.5.

Table 2: Mean classification accuracies for all distractions and all participants. One row represents one participant, the first column shows the participant codes. For each experiment, the tasks with highest (bold) and lowest (red) performance rates are highlighted. Participants with former BCI experience are marked as *, lab members as $^+$.

	band [Hz]	ival [ms]	**overall**	**clean**	**eyes**	**news**	**num**	**flicker**	**stim**	#
od$^{\star,+}$	[5,13.5]	720-4050	90.97	**95.83**	**95.83**	93.06	72.22	**95.83**	93.06	
obx	[10.5,14.5]	700-3960	82.87	88.89	87.50	81.94	70.83	**91.67**	76.39	1
nko	[10.5,13.5]	1170-4420	82.13	93.06	83.33	80.56	62.50	**94.44**	78.87	
njz	[10.5,14]	930-3540	71.30	**83.33**	81.94	75.00	45.83	77.78	63.89	
nkq*	[7.5,13]	640-3740	63.26	**73.61**	59.72	61.97	47.89	66.67	69.44	
nkt	[9.5,13]	1020-2640	61.34	66.67	62.50	66.67	51.39	**70.83**	50.00	
nkr	[8.5,15.5]	810-4200	61.11	63.89	61.11	62.50	51.39	**65.28**	62.50	2
njy$^+$	[6.5,12.5]	500-2720	60.42	62.50	54.17	65.28	50.00	**69.44**	61.11	
nkm	[19.5,25.5]	620-3170	60.42	**68.06**	52.78	65.28	56.94	55.56	63.89	
nkn$^+$	[7.5,12.5]	910-2380	58.00	62.50	52.78	61.11	49.30	**65.28**	56.94	
nkl	[11.5,16.5]	730-3950	52.55	45.83	48.61	54.17	54.17	**61.11**	51.39	
nkp	[16.5,22.5]	1140-3900	51.62	51.39	**55.56**	50.00	50.00	52.78	50.00	
nku	[8.5,13.5]	600-3620	51.62	**61.11**	52.78	47.22	50.00	48.61	50.00	
ma4*	[8.5,13.5]	600-3620	51.16	56.34	**58.33**	48.61	49.30	41.67	52.78	3
nkk	[13.5,18.5]	590-3910	50.00	48.61	**55.56**	43.06	51.39	51.39	50.00	
nks	[5,9]	1370-3990	49.42	47.22	47.14	45.83	47.89	**54.17**	**54.17**	
∅			62.39	**66.68**	63.10	62.64	53.81	66.41	61.53	

1: BCI control during all distraction tasks (at least 61.11% in each task)
2: BCI control during most distraction tasks (at least 54.17% in overall experiment)
3: no BCI control

3.3. Artifacts

Since different artifact types influence data in different ways, the impact of their contamination and the methods for their removal also differ. We wanted to quantify the extent of contamination in each secondary task to investigate differences that might explain classification performances in the different tasks. Additionally, we investigated whether the removal of groups of artifacts could improve classification. We performed these analyses by decomposing the calibration data with Independent Component Analysis (ICA) and classifying the resulting independent components as different artifact types. We used the implementation of Extended Infomax in EEGLab [50] to perform the ICA and an automatic classifier of independent components of EEG data, IC_MARC [51], to classify ICs. IC_MARC assigns a probability to each independent component of representing neural activity, eye blinks, heart beat artifacts, lateral eye movements, or muscle contractions. In addition to these five well defined classes, a class referred to as "mixed" is also included. This

class contains artifact types other than those already mentioned and independent components that include several types of activity or noise. We assigned independent components to the class for which the highest probability was predicted by IC_MARC. Figure 5 shows two randomly selected examples of each class from the *clean* task. The samples from the blink, neural, heart beat, and lateral eye movements are good examples of what we would expect in these classes. The top sample from the muscle class is more similar to what we would expect from the lateral eye class, while the bottom sample from the mixed class represents a typical muscular artifact. The two samples from the mixed class do not clearly belong to another class, as expected.

We grouped these artifact classes into five groups:

1) Muscular (muscle artifacts).
2) Ocular (blinks and lateral eye movements).
3) Non-neural (blinks, heart beats, lateral eye movements, muscle, and mixed artifacts).
4) Muscular and mixed (muscle and mixed artifacts).
5) Non-mixed artifacts (blinks, heart beats, lateral eye movements, and muscle artifacts).

Figure 6 shows the percentage of data variance explained by each artifactual group for all secondary tasks. Displayed values represent mean and standard deviation over all participants. The percentage of variance explained by the artifact groups is quite similar for all tasks, except for the *numbers* task. There, the ocular artifacts, non-neural independent components, and non-mixed artifacts explain more than twice as much variance as in the other tasks. We might expect that a BCI system would perform similarly across the secondary tasks whose artifact distributions are similar. If this is the case, one classifier should work well across the *not-numbers* tasks (*clean, eyes-closed, news, flicker, stimulation*). The *numbers* task on the other hand requires a separate classifier. The artifact distribution of unseen data might be informative enough to distinguish between these two cases in order to select the appropriate classifier. This would make the 2-step approach introduced later a suitable method for out-of-lab BCI systems.

Figure 7 shows the median power of each independent component class as a function of frequency for the independent components from the *calibration* data. These independent components were the ones used to clean the other conditions in Section 4.1. The power spectra were first calculated for each epoch using the default settings of the function *periodogram* in Matlab R2014b at 100 evenly spaced frequencies between 5 and 33 Hz. The median over epochs for each participant was then calculated, followed by the median over participants for each independent component class. Since data was band-pass filtered during pre-processing, the spectra are flat below 8 Hz and above 30 Hz. It is reassuring that the neural independent components' power peaks at around 7-15 Hz since this band contains the motor-imagery related μ-band. The low amplitudes of the blink and lateral eye independent components' spectra relative to neural components is not surprising since for these artifacts, activity typically lies in frequency bands lower than 8 Hz (blinks' power peaks at 3 Hz and drops off before 10.5 Hz while lateral eye movements exhibit most power at frequencies below 6 Hz [52, p. 1237], [53]). Fequency information of heart beats mainly lies between 15-32 Hz [54]. The

high power seen at lower frequencies than 15 Hz for heart beats indicates that independent components classified as heart beat artifacts probably also contain other types of activity. Muscle artifacts are active at high frequencies, from about 20-300 Hz [55]. This expectation is reflected in the plot. Since mixed artifacts may contain many types of activity, we do not have any expectations for how the power spectrum for mixed components may look.

3.4. Feature shifts

In Figure 8 we plotted training and testing features for the *numbers* task (2 best CSP filters) for participants *obx* and *njy* respectively. Since we trained both tasks on calibration data without secondary tasks, this plot shows how differently data shifts between training (no secondary tasks) and testing (with secondary tasks). For participant *obx*, training features differ from testing features, but they are still separable. Whereas for participant *njy* test set features shift in a way that makes it impossible to separate them with the trained classifier. The corresponding classification rates (70.83% and 50%) support that finding.
With one CSP filter per task, we also performed a 6-fold cross-validation to classify the different secondary tasks against the (*clean*) task for both hands. Average classification rates over all 16 participants are visualized in the form of boxplots in Figure 9. While classification rates for the *news* and *flicker* task against *clean* are mostly around chance level, it is clearly observable that it is much easier to classify *stimulation, eyes-closed* or *numbers* against *clean* where the median accuracies lie between 90% and 95%.
This means, we can indeed assume major feature shifts in the data, especially in the *eyes-closed*, *numbers* and *stimulation* tasks which significantly complicates classification. Including an adaptation step into the classification process could solve this problem if we assume that the data is separable at all.

3.5. Non-discriminativity

To find out whether the data is separable at all, we computed one classifier for each secondary task and only tested on the same secondary task on which we trained. We therefore conducted a 6-fold cross-validation on each secondary task (60 training trials, 12 testing trials). Average classification rates are displayed in Table 3.
In comparison with the results from Table 2 where we computed one classifier for all tasks (see Section 3.1), the overall classification rates improved for most participants. While the overall classification rate for the *news* task hardly changed, the performance for both the *numbers* task and the *stimulation* task, averaged over all participants, improved significantly by almost 7% and almost 5%, respectively. Both significance tests were carried out with one-sided Wilcoxon signed rank test to a significance level of $\alpha = 5\%$. This improvement is not surprising considering the detection of feature shifts in both tasks in Section 3.4.
Several participants (*nku, nkp, nkl, nkk*) could still not reach the threshold of 61.11% which shows that their data is not even separable into left and right hand motor imagination. Classification in the *flicker* task for participant *nkn* did not work at all considering the accuracy of 27.78%. However, for participants *nkq* and *nko* classification rates improved by $8\% - 9\%$.

Table 3: Mean classification accuracies for 6 classifiers. One row represents one participant, the first column shows the participant codes. The results that improved compared to Table 2 are highlighted in (**bold**). Participants which are assigned to a higher group compared to Table 2 are marked in **bold** (lower in red), those with former BCI experience are marked as *, lab members as $^{+}$.

	overall	**clean**	**eyes-closed**	**news**	**numbers**	**flicker**	**stimulation**
od$^{\star,+}$	**95.83**	**98.61**	**100.00**	**98.61**	**83.33**	**98.61**	**95.83**
obx	**85.65**	**91.67**	83.33	**86.11**	**81.94**	**95.83**	75.00
nko	**91.31**	**95.83**	**84.72**	**91.67**	**90.28**	94.44	**90.91**
njz	**72.92**	75.00	**84.72**	**81.94**	**68.06**	**68.06**	59.72
nkq*	**71.53**	70.83	**69.44**	**72.73**	**60.61**	**80.56**	**75.00**
nkt	**65.97**	63.89	**72.22**	58.33	**65.28**	**79.17**	**56.94**
nkr	55.56	48.61	48.61	51.39	**66.67**	52.78	**65.28**
njy^{+}	**63.66**	59.72	**68.06**	65.28	41.67	**73.61**	**73.61**
nkm	58.33	63.89	52.78	56.94	**66.67**	54.17	55.56
nkn^{+}	50.13	55.56	**55.56**	54.17	42.42	27.78	**65.28**
nkl	49.31	**52.78**	48.61	51.39	40.28	52.78	50.00
nkp	50.93	**54.17**	44.44	**52.78**	48.61	**54.17**	51.39
nku	**52.78**	52.78	**54.17**	**54.17**	48.61	50.00	**56.94**
ma4*	**59.24**	43.94	48.61	**59.72**	**71.21**	**68.06**	**63.89**
nkk	**51.62**	48.61	50.00	**58.33**	**55.56**	50.00	47.22
nks	**53.16**	**62.50**	**50.00**	**48.61**	43.94	52.78	**61.11**
∅	**64.24**	64.90	**63.45**	**65.14**	**60.95**	65.80	**65.23**

4. Evaluation of new strategies for out-of-lab

In the last section, we found that artifacts highly contaminate the data, especially in the *numbers* task. Another problem is that testing data heavily shifts from calibration data [19, 56, 10]. However, if we compute task-specific classifiers, we could separate left from right hand motor imagination for most participants. Since standard machine learning methods may fail in case of feature shifts and artifact contamination, we need to further investigate other methods such as adaptation or artifact removal. In this section, we propose three strategies to improve classification.

(i) Artifact removal. Since we discovered in Section 3.3 that data is highly artifact-contaminated, we remove the classified artifacts before classification.

(ii) Classifier ensemble. Instead of dividing the dataset into the different secondary tasks, we apply all 6 classifiers from Section 3.5 and average the classifiers output.

(iii) 2-step classification. We first identify the type of distraction task during the motor

imagination before applying the respective classifier.

4.1. Improvement via artifact reduction

Table 4 shows the mean classification rates over participants for each secondary task when each artifact group is removed from data. To remove artifact groups, the independent components from each artifact group were extracted from both the calibration and test data. When testing whether the performances differ from the baseline performances with a one-sided Wilcoxon signed rank test, no p-value is below 0.05 (not corrected for multiple hypothesis tests).

Although removing artifacts does not cause any significant performance differences, some qualitative trends are interesting. From the first row of Table 4, we see that the best artifact group to remove is that containing ocular artifacts. This improves the classification performance for all secondary tasks except *stimulation* and *clean*. The most difficult groups to remove are the muscular and muscular and mixed groups. Both groups cause a decrease in classification performances in 4 secondary tasks. The performance in the *news* task is improved by removing all the artifact groups. Similarly, the *numbers* task is also improved when any artifact group, except the muscle group, is removed. However, the improvements are not statistically significant. These results are consistent with previous investigations in which removing artifacts did not improve BCI performance significantly [57].

Table 4: Mean classification accuracies for all distractions and removed artifact groups averaged over participants. For each experiment, the artifact group with improved performances compared to Table 2 is highlighted in **bold**. The overall (rightmost) column from Table 2 is reproduced for baseline comparison.

	Muscular	Ocular	Non-neural	Muscular and mixed	Non-mixed artifacts	overall	baseline
overall	62.30	**62.69**	62.20	61.83	62.01	62.20	62.39
clean	66.46	65.67	**66.97**	65.94	64.72	65.95	66.68
eyes-closed	62.50	**63.80**	60.75	60.68	**63.19**	62.19	63.10
news	**62.81**	**63.76**	**63.51**	**63.94**	**63.42**	**63.49**	62.64
numbers	53.56	**55.29**	**56.86**	**54.69**	**55.12**	**55.10**	53.81
flicker	**67.10**	**66.58**	64.93	65.28	65.10	65.80	66.41
stimulation	61.35	61.01	60.14	60.49	60.48	60.69	61.53

4.2. Improvement via classifier ensemble

Instead of choosing a task-specific classifier for each trial, we propose an ensemble approach, where we applied all 6 classifiers to all trials and averaged over the output to determine

Table 5: Mean classification accuracies for classifier ensembles. The results that improved compared to Table 2 are highlighted in (**bold**). Participants which are assigned to a higher group compared to Table 2 are marked in **bold**, those with former BCI experience are marked as *, lab members as $^\#$.

od$^{\star,+}$	obx	nko	njz	nkq*	nkt
97.92	**88.89**	**93.66**	**75.93**	55.56	**71.99**
nkr	njy $^+$	nkm	nkn $^+$	nkl	nkp
73.24	**64.58**	**65.28**	**58.92**	50.23	**52.78**
nku	**ma4** *	nkk	**nks**	∅	
52.55	**55.63**	49.07	**54.69**	**66.31**	

whether the user performed left or right hand motor imagination. We therefore also conducted a 6-fold cross-validation on each secondary task. Since we recorded 72 trials per secondary task, we trained on 60 trials and tested on 12 trials, equivalent to Section 3.5.
Average classification rates for all 16 participants can be found in Table 5. Compared to the results from Table 2, we could significantly improve classification accuracies by 4.5% (one-sided Wilcoxon signed rank test, $\alpha = 5\%$).
Ensemble methods benefit from accuracy and diversity of individual classifiers. Because of the feature shifts we detected in Section 3.4, we can assume that data of the different secondary tasks does not arise from the same two Gaussian distributions (one per MI-class). Hence corresponding classifiers are indeed diverse. Although separate classifiers do not classify all participants significantly above chance level, most of them are accurate on average. Combining those two findings explains why ensemble CSP improves accuracy for participants of category $\#1$ and $\#2$.

4.3. Improvement via 2-step classification

Calculating one classifier for each distraction indeed yields higher classification rates but if we think about applying this concept to real world situations, we might not have that much prior knowledge about the scenarios the BCI is used in. Therefore, we propose a 2-step classification approach which combines classifying the respective secondary task before separating *left* from *right* hand MI.
In Section 3.4 we identified three secondary tasks with major feature shifts: *eyes-closed, numbers* and *stimulation*. Since features can shift in different dimensions we have to treat those tasks separately. The *numbers* task reaches the lowest accuracy when applying the original CSP approach (see Table 2). Therefore, in the first step, we separated the *numbers* task from the *not-numbers* task (*clean, eyes-closed, news, flicker, stimulation*). After categorizing a trial to one of these groups, we applied one of two *left* vs. *right* classifiers (one for *numbers* and one for *not-numbers*) to decide whether this trial consisted of a *left* or

right hand motor imagination. For this approach, we only considered the 6 runs including secondary tasks and conducted a 6-fold cross-validation. This means that for the first step we used 360 training trials and 72 testing trials. In the second step we used 300 training trials for *not-numbers* and 60 training trials for *numbers*. The number of testing trials depends on the accuracy in the first step and can theoretically vary between 1 and 72 for both classes.
Results are summarized in Table 6 where the average classification rates for both steps are listed. The overall classification rate is the weighted average performance of the *numbers* and *not-numbers* task. The weighting compensates for the different number of trials in both tasks. Please note that a 3-step approach where we additionally separated the *stimulation* task led to lower performance than the 2-step approach.
As already discussed in Section 3.4, the different secondary tasks are easily separable, results of the first step in Table 6 show that classification rates are mostly between 94% and 100%. Except for participant *nkm* where we are only able to classify 86% of the tasks correctly. Since only 12 out of 72 trials in each run belong to the *numbers* task, this result leads to the conclusion that the tasks are not really distinguishable for this participant. However, classification accuracies significantly improved (one-sided Wilcoxon signed rank test, $\alpha = 5\%$) compared to the original CSP results in Table 2 and for most participants (except *nkk,nkl,nkp,nkr*) we reached higher classification rates with this 2-step approach. One reason for this may also be the amount of training data we used to train the *not-numbers* classifier. Whereas we used between 60 and 72 training trials for the previous approaches, here we could use now up to 300 training trials.

In Figure 10, we display the comparison of the results of all three approaches to our original CSP approach (see Section 3.1) together with the p-values of the one-sided Wilcoxon signed rank tests. Each circle represents one participant. For the artifact removal we display the removal of each artifact group in different colors.

5. Conclusion

In this paper we presented a motor imagery-based BCI study where participants had to handle 6 different secondary tasks in addition to the motor imagery task. The idea behind those tasks was to simulate a semi-realistic environment and to systematically analyze the influence of different scenarios on the motor imagery performance. We first trained on data without secondary tasks and classified on data with secondary tasks using three CSP filters per class and RLDA for classification. Results between participants varied a lot and only 10 out of 16 participants reached an overall accuracy higher than chance level. Especially classification results in the *eyes-closed, numbers* and *stimulation* tasks are much lower compared to the *clean, news* and *flicker* tasks. We investigated three possible reasons for this poor BCI performance, namely

(i) artifact contamination

(ii) feature shifts between training (no secondary task) and testing (with secondary tasks)

Table 6: Mean classification accuracies for 2-step classification. The results that improved compared to Table 2 are highlighted in (**bold**). Participants which are assigned to a higher group compared to Table 2 are marked in **bold**, those with former BCI experience are marked as *, lab members as $^{+}$.

		1st step	2nd step	
	overall	**cond**	**not-numbers**	**numbers**
od $^{\star,+}$	**96.53**	100.00	99.17	83.33
obx	**90.28**	99.31	91.92	82.19
nko	**93.19**	96.71	93.82	90.00
njz	**77.55**	97.45	78.71	72.00
nkq *	**77.93**	99.53	80.28	66.20
nkt	**76.85**	99.77	79.11	65.75
nkr	58.80	99.07	57.26	66.22
njy $^{+}$	**66.20**	96.53	70.54	46.84
nkm	**66.90**	86.34	70.66	50.62
nkn $^{+}$	57.75	96.95	59.08	51.90
nkl	46.99	99.31	47.90	42.67
nkp	49.07	95.37	48.56	51.19
nku	**52.08**	98.84	52.65	49.32
ma4 *	**61.27**	98.83	60.45	65.28
nkk	48.61	94.68	48.12	50.57
nks	**57.75**	98.83	58.43	54.29
∅	67.36	97.34	68.54	61.77

(iii) non-separability of the data.

We found a high amount of artifacts in the data, especially for the *numbers* task, where 77% of data variance can be explained by non-neural artifacts. We also found severe feature shifts for the *stimulation, eyes-closed* and *numbers* tasks by classifying each secondary task against the *clean* task. To investigate non-separability, we trained one separate classifier for each secondary task. Since we could improve classification accuracies for most participants, we can conclude that, applying the correct classifier, left and right hand motor imagination is indeed separable for most participants. Thinking about real-world scenarios, the problem however is, that we do not always know which task a user is carrying out while controlling the BCI.

We therefore proposed three different approaches to improve performance.

(i) Artifact removal. After decomposing calibration data with ICA, we classified the independent components (ICs) with IC_MARC [51] and assigned one artifact type to each IC. After removing 5 different groups of artifacts, we used RLDA based on CSP for

classification. Results are displayed in Table 4.

(ii) Ensemble CSP. For each of the 6 secondary tasks, we trained one classifier based on CSP and RLDA. We applied all 6 classifiers to the data and used the averaged output as a classification decision, Table 5 shows the results.

(iii) 2-step CSP. We classified in two steps. During the first step, we separated the noisiest secondary task (*numbers*) from the remaining 5 tasks (*not-numbers*). In the second step we used two different classifiers, one for *numbers* and one for the remaining *not-numbers* tasks to separate the data into *left* and *right* hand MI, results can be found in Table 6.

While the artifact removal did not lead to significant improvement (see Figure 10), both, ensemble CSP and the 2-step approach did. It is clearly visible that for both, the ensemble CSP and the 2-step CSP, classification performances improve particularly for those participants already reaching significant BCI control in the original CSP approach ($\#1$ and $\#2$ in Table 2). It is worth mentioning again that the 2-step approach clearly benefits from the increased number of training trials compared to the original and the ensemble CSP approach.
We also have to note that most of the participants were confronted for the first time with a BCI system. Imagining a movement is relatively abstract and some participants may improve by engaging in more feedback training before going "out-of-lab".
After first steps have been made to leave the controlled lab environment, this study systematically and quantitatively analyzes how different scenarios influence BCI performance. The challenges we identified, especially the ones of dual character in the *numbers* task need to be considered for future studies and are worth being further analyzed. Also training the BCI users more detailed beforehand could lead to a better understanding and higher performance rates.
Although two of the proposed approaches significantly improve classification accuracy we should take into account that when applying them to everyday life, additional challenges might occur. For example, whereas in our study we exactly know under which secondary task a participant is performing the MI, it is not always clear in which situation end-users find themselves when controlling a BCI in their everyday life environment. However, this issue might be tackled with a large amount of training data in various possible real-world scenarios.
We also have to consider the number of EEG channels we used. Preparation of a gel-based EEG cap is time-consuming and challenging outside of the lab since electrodes dry out over time and users need to wash their hair afterwards. Reducing the number of channels and considering dry EEG systems would also further improve usability of real-world applications.
Future studies will also explore whether harvesting a database of significantly larger numbers of participants and tasks may allow an invariant and user independent decoding [58], e.g., using deep neural networks.

Acknowledgment

This work was supported in part by the Adaptive BCI (FKZ 01GQ1115) and by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the

Ministry of Education. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein. Correspondence to SB, KRM and WS.

References

[1] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, MA, 2007.

[2] Bernhard Graimann, Brendan Z Allison, and Gert Pfurtscheller. *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer, 2010.

[3] Tobias Kaufmann, Andreas Herweg, and Andrea Kübler. Toward brain-computer interface based wheelchair control utilizing tactually-evoked event-related potentials. *Journal of neuroengineering and rehabilitation*, 11(1):7, 2014.

[4] Janis J Daly and Jonathan R Wolpaw. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, 2008.

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

[6] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

[7] Matthias Krauledat, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Müller. Towards zero training for brain-computer interfacing. *PLoS ONE*, 3(8):e2967, 2008.

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

[9] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[10] Wojciech Samek, Carmen Vidaurre, Klaus-Robert Müller, and Motoaki Kawanabe. Stationary common spatial patterns for brain-computer interfacing. *Journal of Neural Engineering*, 9(2):026013, 2012.

[11] M. Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.*, 24(4):610–619, 2013.

[12] Wojciech Samek, Motoaki Kawanabe, and Klaus-Robert Müller. Divergence-based framework for common spatial patterns algorithms. *IEEE Reviews in Biomedical Engineering*, 7:50–72, 2014.

[13] Motoaki Kawanabe, Wojciech Samek, Klaus-Robert Müller, and Carmen Vidaurre. Robust common spatial filters with a maxmin approach. *Neural Computation*, 26(2):1–28, 2014.

[14] Stephanie Brandl, Klaus-Robert Müller, and Wojciech Samek. Robust common spatial patterns based on bhattacharyya distance and gamma divergence. In *Proc. of Int. Winter Workshop on Brain-Computer Interface*, pages 1–4, 2015.

[15] Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. Emg and eog artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3):480–494, 2007.

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

[17] Irene Winkler, Stephanie Brandl, Franziska Horn, Eric Waldburger, Carsten Allefeld, and Michael Tangermann. Robust artifactual independent component classification for bci practitioners. *Journal of neural engineering*, 11(3):035013, 2014.

[18] Peter Sykacek, Stephen J Roberts, and Maria Stokes. Adaptive BCI based on variational bayesian kalman filtering: an empirical evaluation. *IEEE Transactions on Biomedical Engineering*, 51(5):719–727, 2004.

[19] Pradeep Shenoy, Matthias Krauledat, Benjamin Blankertz, Rajesh PN Rao, and Klaus-Robert Müller. Towards adaptive classification for bci. *Journal of neural engineering*, 3(1):R13, 2006.

[20] Carmen Vidaurre, Claudia Sannelli, Klaus-Robert Müller, and Benjamin Blankertz. Machine-learning based co-adaptive calibration. *Neural computation*, 23(3):791–816, 2011.

[21] Fabien Lotte, Junya Fujisawa, Hideaki Touyama, Rika Ito, Michitaka Hirose, and Anatole Lécuyer. Towards ambulatory brain-computer interfaces: A pilot study with p300 signals. In *Proc. of the Int. Conf. on Advances in Computer Enterntainment Technology*, pages 336–339, 2009.

[22] Maarten De Vos, Katharina Gandras, and Stefan Debener. Towards a truly mobile auditory brain–computer interface: exploring the p300 to take away. *International journal of psychophysiology*, 91(1):46–53, 2014.

[23] Thierry Castermans, Matthieu Duvinage, Mathieu Petieau, Thomas Hoellinger, CD Saedeleer, Karthik Seetharaman, Ana Bengoetxea, Guy Cheron, and Thierry Dutoit. Optimizing the performances of a p300-based brain–computer interface in ambulatory conditions. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 1(4):566–577, 2011.

[24] Hayrettin Gürkök, Mannes Poel, and Job Zwiers. Classifying motor imagery in presence of speech. In *Proc. of Int. Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[25] Michael Tangermann, Matthias Krauledat, Konrad Grzeska, Max Sagebaum, Benjamin Blankertz, Carmen Vidaurre, and Klaus-Robert Müller. Playing pinball with non-invasive bci. In *NIPS*, pages 1641–1648. Citeseer, 2008.

[26] Alexander J Doud, John P Lucas, Marc T Pisansky, and Bin He. Continuous three-dimensional control of a virtual helicopter using a motor imagery based brain-computer interface. *PloS one*, 6(10):e26322, 2011.

[27] Karl LaFleur, Kaitlin Cassady, Alexander Doud, Kaleb Shades, Eitan Rogin, and Bin He. Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface. *Journal of neural engineering*, 10(4):046003, 2013.

[28] Gabriel Pires, Mario Torres, Nuno Casaleiro, Urbano Nunes, and Miguel Castelo-Branco. Playing tetris with non-invasive bci. In *Serious Games and Applications for Health (SeGAH), 2011 IEEE 1st International Conference on*, pages 1–6. IEEE, 2011.

[29] C Neuper, GR Müller, A Kübler, N Birbaumer, and G Pfurtscheller. Clinical application of an eeg-based brain–computer interface: a case study in a patient with severe motor impairment. *Clinical neurophysiology*, 114(3):399–409, 2003.

[30] Robert Leeb, Doron Friedman, Gernot R Müller-Putz, Reinhold Scherer, Mel Slater, and Gert Pfurtscheller. Self-paced (asynchronous) bci control of a wheelchair in virtual environments: a case study with a tetraplegic. *Computational intelligence and neuroscience*, 2007, 2007.

[31] Kai Keng Ang, Cuntai Guan, Karen Sui Geok Chua, Beng Ti Ang, Christopher Wee Keong Kuah, Chuanchu Wang, Kok Soon Phua, Zheng Yang Chin, and Haihong Zhang. A large clinical study on the ability of stroke patients to use an eeg-based motor imagery brain-computer interface. *Clinical EEG and Neuroscience*, 42(4):253–258, 2011.

[32] Robert Leeb, Serafeim Perdikis, Luca Tonin, Andrea Biasiucci, Michele Tavella, Marco Creatura, Alberto Molina, Abdul Al-Khodairy, Tom Carlson, and José dR Millán. Transferring brain–computer interfaces beyond the laboratory: successful application control for motor-disabled users. *Artificial intelligence in medicine*, 59(2):121–132, 2013.

[33] Johannes Höhne, Elisa Holz, Pit Staiger-Sälzer, Klaus-Robert Müller, Andrea Kübler, and Michael Tangermann. Motor imagery for severely motor-impaired patients: evidence for brain-computer interfacing as superior control solution. *PLOS ONE*, 9(8):e104854, 2014.

[34] Andrea Kübler, Elisa M Holz, Angela Riccio, Claudia Zickler, Tobias Kaufmann, Sonja C Kleih, Pit Staiger-Sälzer, Lorenzo Desideri, Evert-Jan Hoogerwerf, and Donatella Mattia. The user-centered design as novel perspective for evaluating the usability of bci-controlled applications. *PloS one*, 9(12):e112392, 2014.

[35] Stephanie Brandl, Johannes Höhne, Klaus-Robert Müller, and Wojciech Samek. Bringing bci into everyday life: Motor imagery in a pseudo realistic environment. In *Proc. of the Int. IEEE/EMBS Neural*

Engineering Conference (NER), pages 224–227, 2015.

[36] H.H. Jasper. The ten twenty electrode system of the international federation. *EEG Clin. Neurophysiol.*, 10:371–375, 1958.

[37] Dennis J McFarland, Lynn M McCane, Stephen V David, and Jonathan R Wolpaw. Spatial filter selection for eeg-based communication. *Electroencephalography and clinical Neurophysiology*, 103(3):386–394, 1997.

[38] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[39] Jian Ding, George Sperling, and Ramesh Srinivasan. Attentional modulation of ssvep power depends on the network tagged by the flicker frequency. *Cerebral cortex*, 16(7):1016–1029, 2006.

[40] Shozo Tobimatsu, You Min Zhang, and Motohiro Kato. Steady-state vibration somatosensory evoked potentials: physiological characteristics and tuning function. *Clinical neurophysiology*, 110(11):1953–1958, 1999.

[41] Anne-Marie Brouwer and Jan BF Van Erp. A tactile p300 brain-computer interface. *Frontiers in neuroscience*, 4:19, 2010.

[42] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Proc. Magazine*, 25(1):41–56, 2008.

[43] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8(4):441–446, 1998.

[44] Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[45] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, and Gert Pfurtscheller. Better than random: A closer look on bci results. *International Journal of Bioelectromagnetism*, 10(EPFL-ARTICLE-164768):52–55, 2008.

[46] Benjamin Blankertz, Claudia Sannelli, Sebastian Halder, Eva M Hammer, Andrea Kübler, Klaus-Robert Müller, Gabriel Curio, and Thorsten Dickhaus. Neurophysiological predictor of smr-based bci performance. *Neuroimage*, 51(4):1303–1309, 2010.

[47] Moritz Grosse-Wentrup, Bernhard Schölkopf, and Jeremy Hill. Causal influence of gamma oscillations on the sensorimotor rhythm. *NeuroImage*, 56(2):837–842, 2011.

[48] Eva Maria Hammer, Sebastian Halder, Benjamin Blankertz, Claudia Sannelli, Thorsten Dickhaus, Sonja Kleih, Klaus-Robert Müller, and Andrea Kübler. Psychological predictors of smr-bci performance. *Biological psychology*, 89(1):80–86, 2012.

[49] Silvia Marchesotti, Michela Bassolino, Andrea Serino, Hannes Bleuler, and Olaf Blanke. Quantifying the role of motor imagery in brain-machine interfaces. *Scientific reports*, 6, 2016.

[50] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[51] Laura Frølich, T. S. Andersen, and Morten Mørup. Classification of independent components of eeg into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015.

[52] Ernst Niedermeyer and Fernando Henrique Lopes da Silva. *Electroencephalography : basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, Philadelphia, 2005.

[53] T. Gasser, L. Sroka, and J. Mocks. The transfer of EOG activity into the EEG for eyes open and closed. *Electroencephalogr Clin Neurophysiol*, 61(2):181–193, Aug 1985.

[54] Joe-Air Jiang, Chih-Feng Chao, Ming-Jang Chiu, Ren-Guey Lee, Chwan-Lu Tseng, and Robert Lin. An automatic analysis method for detecting and eliminating ECG artifacts in EEG. *Computers in Biology and Medicine*, 37(11):1660 – 1671, 2007.

[55] S. D. Muthukumaraswamy. High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front Hum Neurosci*, 7:138, 2013.

[56] Paul von Bünau, Frank C. Meinecke, Franz Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103:214101, 2009.

[57] Laura Frølich, Irene Winkler, Klaus-Robert Müller, and Wojciech Samek. Investigating effects of different

artefact types on motor imagery bci. In *2015 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[58] Siamac Fazli, Sven Dähne, Wojciech Samek, Felix Bießmann, and Klaus-Robert Müller. Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based brain-computer interfaces. *Proceedings of the IEEE*, 103(6):891–906, 2015.

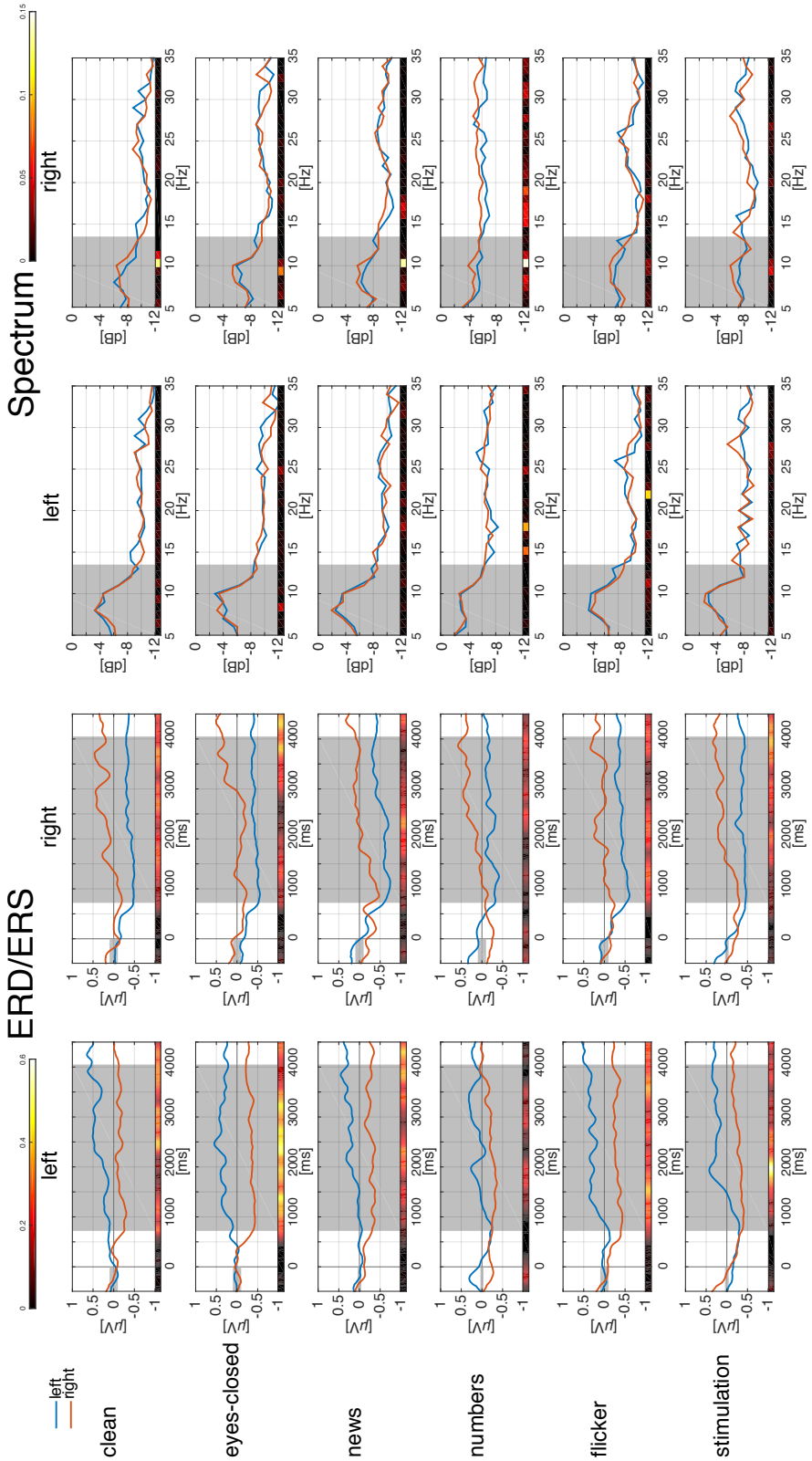

Figure 4: ERD/ERS (columns 1-2) and spectra (columns 3-4) of participant *od* for all 6 secondary tasks when CSP trained on *clean*. Each row represents one secondary task.

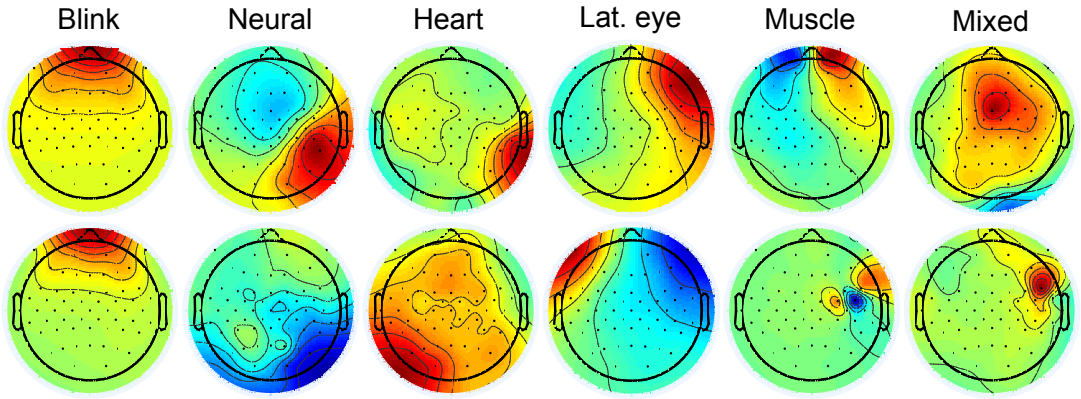

Figure 5: Examples of scalp maps of classified independent components from each class used by IC_MARC.

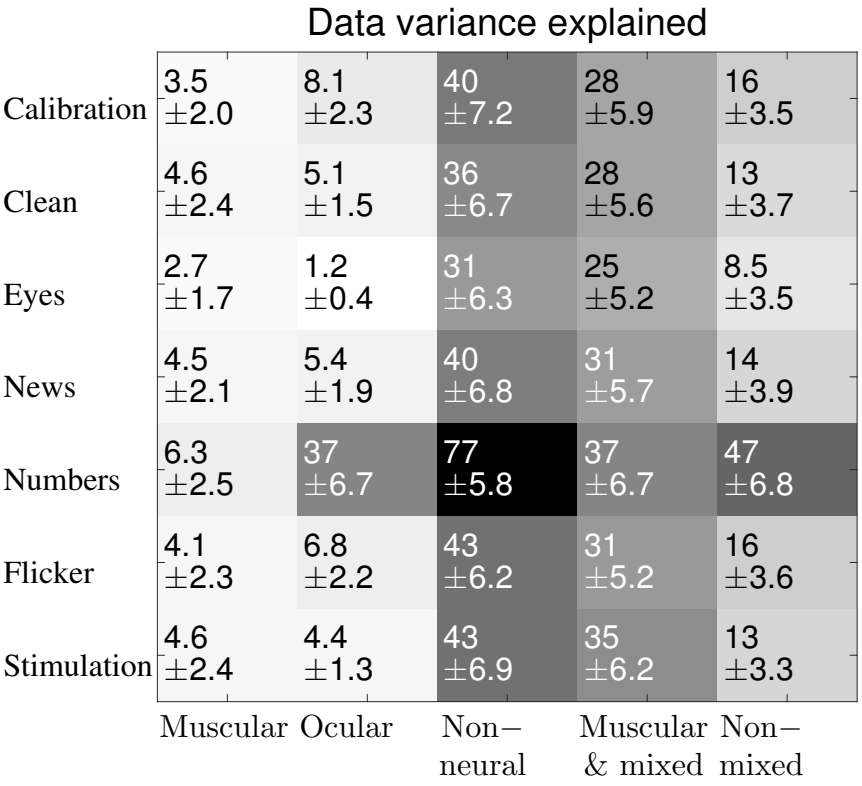

Figure 6: Percentage of data variance explained by each artifact group in each condition. The numbers represent the mean over participants plus/minus its standard deviation.

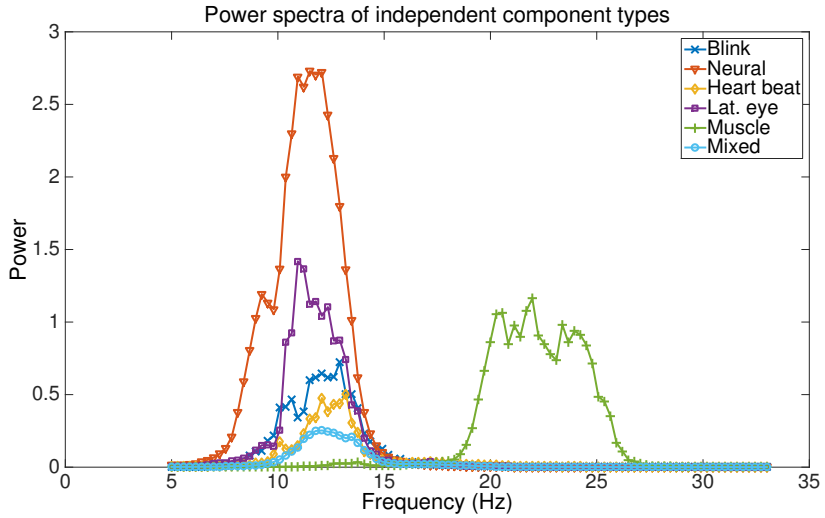

Figure 7: Power spectra of independent components from the six classes (blinks, neural, heart beats, lateral eye, muscle, and mixed). The power spectra were calculated for each epoch independently. Then the median was first taken over epochs for each participant, and then over participants.

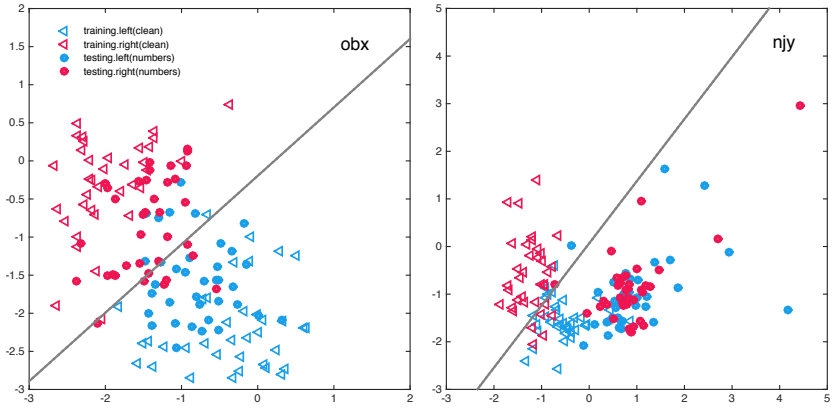

Figure 8: Features of participants *obx* and *njy* of the classifications between left and right hand motor imagery (two best CSP filters) where CSP was only trained on *clean* and tested on *numbers*.

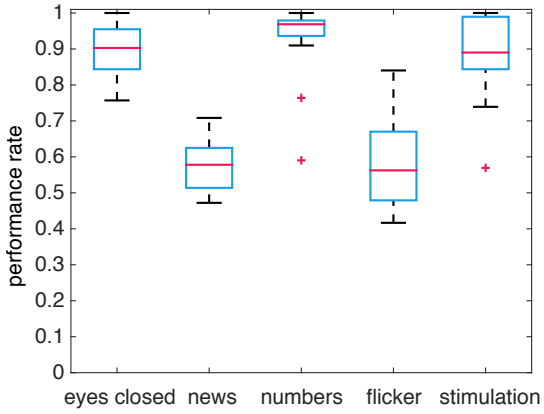

Figure 9: Mean classification accuracies across all 16 participants under different distraction tasks against *clean* motor imagery for both hands.

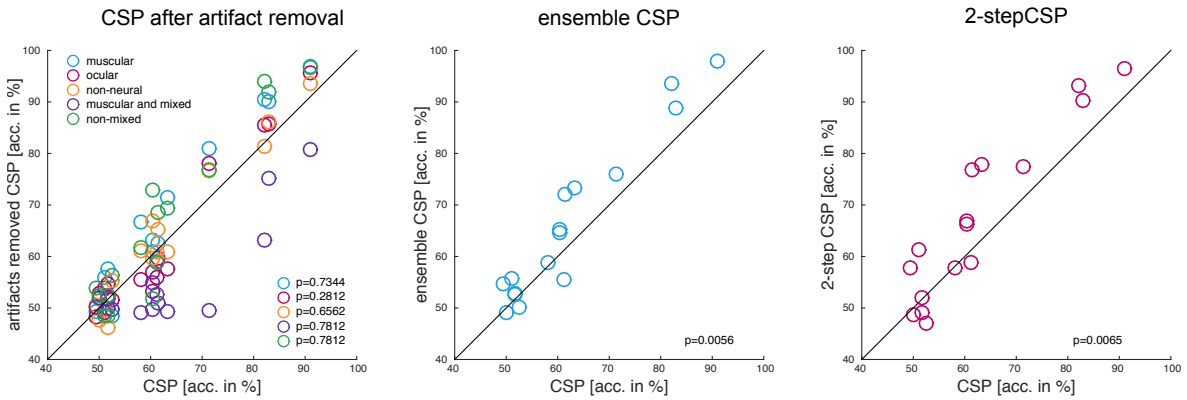

Figure 10: Accuracies of the three new approaches compared to the accuracy of CSP trained on *clean*. One circle represents the accuracy of one participant.