Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Neurorobotics Report

MI BCI Analysis

Federico Pivotto, Alessandro Bozzon, Riccardo Simion, Riccardo Zerbinati

1 Introduction

In this paper we will report and discuss the results obtained during the analysis of EEG data collected during a 3-day Motor Imagery (MI) Brain-Computer Interface (BCI) experiment involving 8 healthy participants. The data was recorded using a 16-channel EEG amplifier (g.USBamp, g.Tec) at a sampling rate of 512 Hz. Electrodes were positioned according to the 10-20 international system. The placement and order of electrodes are illustrated in Figure 1.

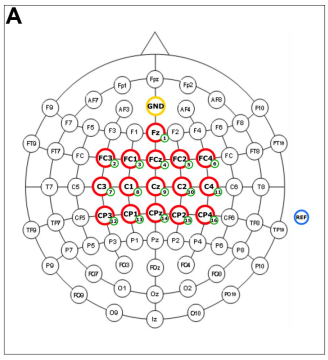

Figure 1: Electrodes scheme

Participants performed two MI tasks, imagining movements of both-hands or both-feet.

2 Analysis

2.1 Topoplot

Topoplots are a kind of plot that let us visualize the ERD/ERS response of the single areas of the brain that lies under the electrodes of the EEG cap. From this kind of plots, it is possible to have a quick look at how the brain of a subject responds to the execution of a given Motor Imagery (MI) task like both-feet or both-hands, in particular they are useful for looking at Event Related Synchronizations (ERS) or Desynchronizations (ERD) in correspondence of the electrodes that record the specific region of the brain that we know is correlated with the performed task, thanks to the scientific studies on the brain that were done in the past. One of the downsides of topoplots is that we do not have a high spectral specificity because all the frequencies of the μ-band (8-14 Hz) are considered at the same time, but every subject learns to modulate only a subset of those frequencies so it can be that some frequencies are not useful and could lead to some noise in the plots. Because of this problem, a Power Spectral Density (PSD) computation and spectrogram analysis was performed as it is possible to see in Section 2.2. To compute the ERD/ERS for a given subject, we need to have a reference value that is usually defined during the fixation or the cue period of the recording protocol. Then, we use the values coming from the continuous feedback period as the activity values so that we can compute the ERD/ERS using the formula:

$$ERD/ERS = 10 * log_{10}(Activity/Reference) \quad (1)$$

Now, as it is possible to see from the Figure 2 the plots for the reference value for both the MI tasks have a global value that is uniform and around the zero value, which confirms that they are correctly computed. For what concerns both-feet MI task, with the topoplots reported in the second row of the Figure 2, we expect that in correspondence of the surroundings of the *Cz* electrode a somewhat clear area of desynchronization should appear, meaning that the task was performed correctly. We can see that most of the subjects had a hard time performing it, with the subject *AI7* having the worst performance and that could be because it has difficulties in performing this specific MI task. The subject that performed the best in this MI task turned out to be the subject *AJ4* which showed a clearer desynchronization around the area of the *Cz* and *C2* electrodes. On average, the ERD/ERS value for both-feet MI task turned out to be not that significative and discriminative to determine what kind of MI task the subject was performing.

A completely different case was the ERD/ERS values obtained for both-hands MI task as it is possible to see in the last row of the Figure 2 reporting all the topoplots. For this specific MI task, the *C3* and *C4* electrodes areas are the ones where we expect a desynchronization. It is important to note that it is possible that only one of the two regions shows a desynchronization or that the intensity of the desynchronization is not the same for the two areas. Even though there are some problematic subjects, like the subject *AI7* or *AJ9* where it is not possible to see a clear desynchronization in correspondence of the *C3* or *C4* electrodes, on average the other subjects performed the task in a satisfactory manner that should enable us to detect useful features. Subjects *AJ1* and *AJ4* are the one

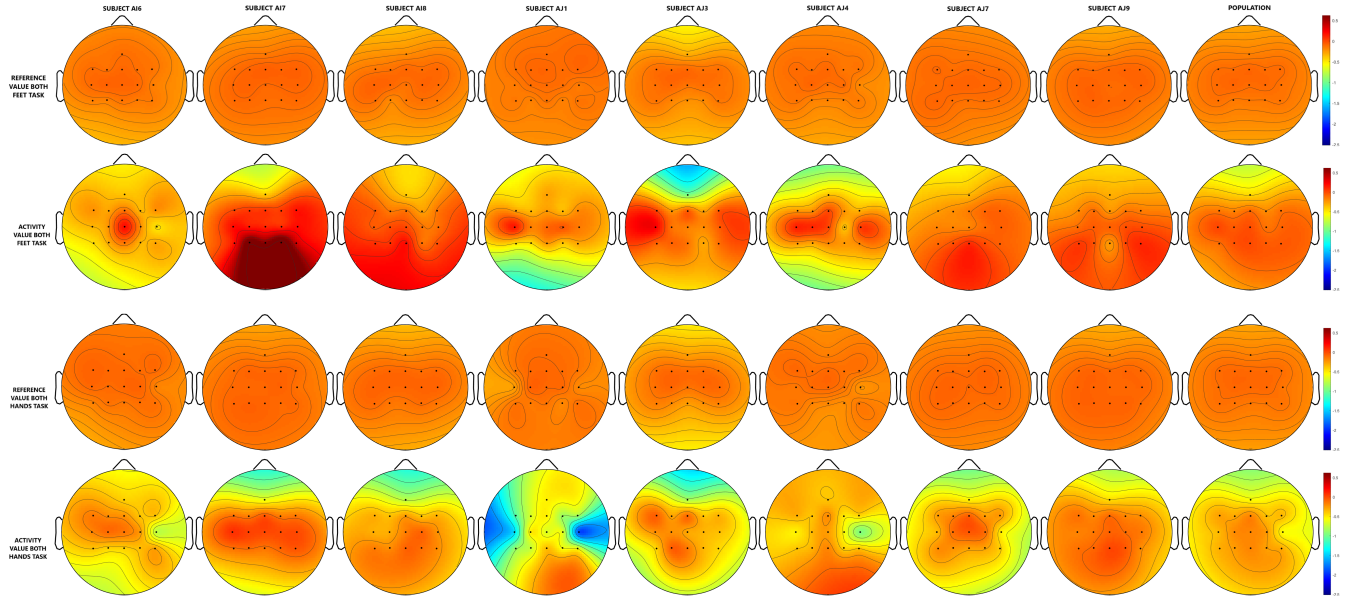

FIGURE 2: μ band topoplots

that had the best performance for this MI task, with subject *AJ1* that showed the clearest desynchronization on both *C3* and *C4* electrodes areas, suggesting that those channels and some of the frequencies of the μ rhythms band could be used for the creation of a classifier.

In general, from the topoplots reported in Figure 2, we can say that the subjects had a harder time performing the both-feet MI task rather than performing the both-hands MI tasks. It is possible that some training phases could solve the problem, however it is not certain but only a hypothesis. The most concerning subject is *AI7* that had poor performances in terms of ERD/ERS values for the areas that correspond to the *C3*, *Cz* and *C4* electrodes for both the MI tasks. In cases like this, we can have different reasons for the results:

- the user is not able to perform the two MI tasks or has difficulties in performing them so, in this case, the first thing to do is to try different MI tasks to see if there are some that are easier to perform for the subject;
- the user is able to perform the task, but it is not able to perform it in a satisfactory manner and some training could be useful to make the subject more accustomed to performing the MI tasks that, in turns, should correspond to an improvement in the ability of the subject in the modulation of particular frequencies of the μ rhythms that are correlated with the MI tasks.

2.2 Spectrogram

A spectrogram is a visual representation of the frequency spectrum of a signal over time. In our case it is particularly useful for analysing the modulation of specific frequency bands of interest, such as μ-band and β-band, during the recorded Motor Imagery (MI) tasks. The spectrograms obtained display the time-frequency representation of EEG data for the two MI tasks taken into account: both-hands and both-feet, allowing us to observe Event-Related Desynchronization (ERD) and Synchronization (ERS). The colour scale used allows to indicate the ERD/ERS values, with negative values in red for the ERD and positive values in yellow for the ERS. The analysis relies on Power Spectral Density (PSD) estimation, computed using Welch's method to ensure a detailed spectral resolution. This step is essential as clear ERD/ERS might appears only after averaging and subjects might learn to modulate only few sub-bands (not the whole μ-band).

Let's consider the two most representative subjects: the best-performing *AJ1* and the poorest-performing subject *AI7*. Starting with subject *AJ1*, focusing on the both-hands MI task since out of the two MI tasks performed by all the subjects is the one that on average is performed in a more satisfactory way. It can be noticed in Figure 3 a strong ERD of the channels *C3*, *Cz* and *C4* in the μ-band. The fact that all three channels are involved means that the MI task has been carried on correctly, in fact:

- *C3*: represents the lateralized activation for the right hand;
- *Cz*: central involvement, typical for bilateral motor tasks;
- *C4*: represents the lateralized activation for the left hand.

Furthermore the amplitude and the persistence of the ERD in the μ-band indicates a good predisposition of the subject to effectively carry out the MI task.

Let's now consider subject *AI7*, which is the one with some of the poorest performances obtained during our analysis. As can be noticed in Figure 4, the spectrogram highlights a large quantity of ERDs, which could mean that the subject was not able or trained enough to complete the MI task, since the brain activity is so dominant that it masks any changes in synchronization or desynchronization.

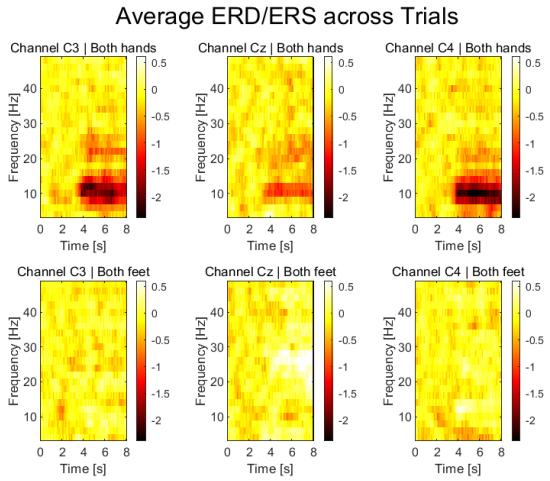

FIGURE 3: *AJ1* spectrogram

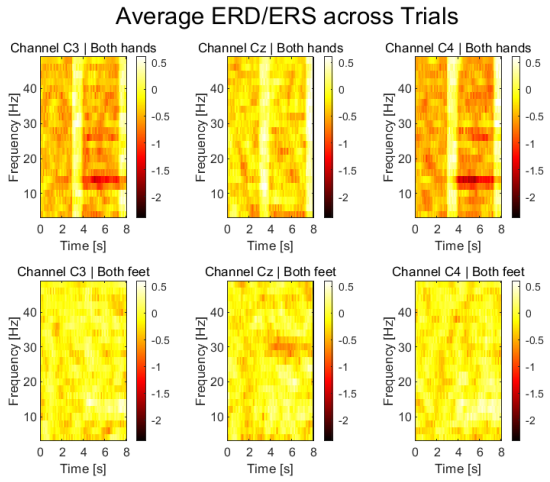

FIGURE 4: *AI7* spectrogram

3 Feature Selection

A feature map is a visual representation that highlights the relevance of different EEG frequencies and channels. In our study, feature maps have been utilized to distinguish Motor Imagery (MI) tasks. The procedure used for feature selection in our analysis was qualitative rather than quantitative. Specifically, we analysed the feature maps of each subject and visually identified the features based on the prominence of colors in the maps. While we did not define a strict reference value for the Fisher score, we carefully examined the maps to identify features that stood out visually. It is important to note that this qualitative selection process may not apply well to poorly performing subjects, such as *AI7* and *AJ7*, where clear feature patterns were more challenging to observe. For each subject, we analysed the offline feature maps individually. Our focus was on identifying features that consistently appear across the various feature maps, aiming to strike a balance that allowed us to select those that were stable. Given the MI tasks under analysis, specifically both-hands and both-feet, we concentrated on the key electrodes *C3*, *Cz* and *C4*. When these electrodes did not exhibit strong signals of interest, we extended our focus to neighbouring electrodes, such as *C1* and *C2*, which occasionally displayed relevant activity. At the population level, we adopted a broader strategy to ensure robustness across subjects. This involved identifying features that remained stable across all offline feature maps of all participants. While this approach was intentionally coarse, it enabled us to account for inter-subject variability while retaining features that were critical for analysing the MI tasks.

3.1 Feature Map

Let's consider the best and the worst subject. In the feature map of *AJ1* (Figure 5), we observe prominent activations in specific channels, particularly *C3*, *Cz* and *C4*, within the μ-band frequency range. The Fisher score indicates a clear differentiation in these channels, signifying a strong Event-Related Desynchronization (ERD). The presence of these well-defined activations implies that *AJ1* demonstrates effective task performance, with well-modulated brain activity in the μ-band. The amplitude and persistence of these signals across the feature maps highlight *AJ1*'s ability to correctly engage in the MI task for both-hands and both-feet.

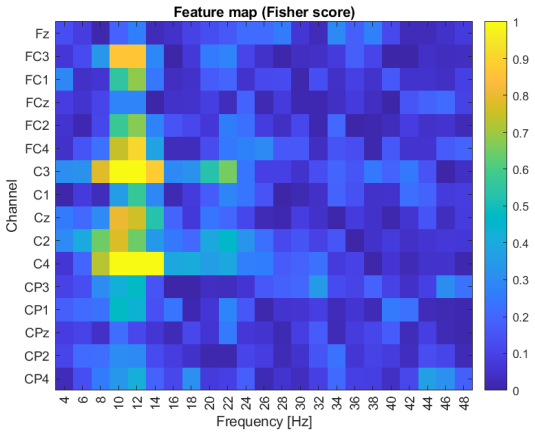

FIGURE 5: *AJ1* feature map

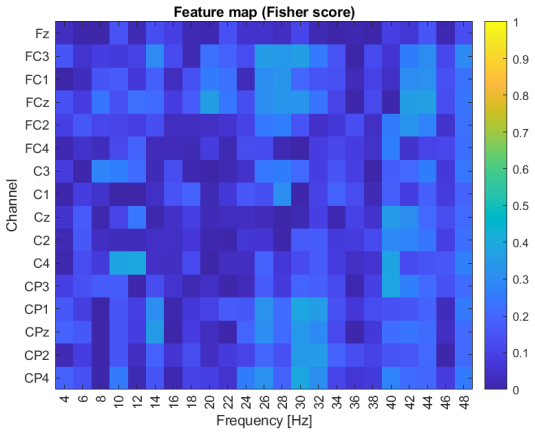

FIGURE 6: *AJ7* feature map

For subject *AJ7* (Figure 6), the feature map reveals a more dispersed pattern of activations with less focus on

the key channels for the two MI tasks. The signal in the μ-band appears inconsistent, with lower Fisher scores across the key channels. This suggests that *AJ7* struggled to engage effectively the brain regions necessary for the MI tasks. The widespread but low-intensity activations might indicate a lack of specific motor-related brain activity or insufficient training to complete the task. The brain activity appears dominated by noise or unrelated processes, making it challenging to detect clear synchronization or desynchronization patterns.

4 Models

In our experiment, a model was trained for each subject and the whole population, intended as the concatenation of all subjects. These models were trained and evaluated to classify Motor Imagery (MI) tasks in two classes, both-hands vs. both-feet. Both training and evaluation phases utilizes the Power Spectral Density (PSD) computed for the subject, respectively of the available offline and online recordings. While training is in charge of learning a model given the train set, the evaluation assesses the model performance on the test set, i.e. data never seen before, to measure its ability to generalize. The visual paradigm is shown in Figure 7.

The following analysis focuses on three representative subjects: the best-performing *AJ1*, a middle-performing *AJ4*, and the poorest-performing subject *AJ7*. Additionally, a population-level analysis is conducted to highlight differences between individual and group performance. Metrics used for the model comparison are: single-sample accuracy, trial accuracy and average decision time. In particular, these last two metrics gives an idea of how effective and precise would be a BMI system in delivering the right discrete commands to a robot, according to an evidence accumulation framework applied on the output of the classifier, i.e. decoder of the subject neural activity.

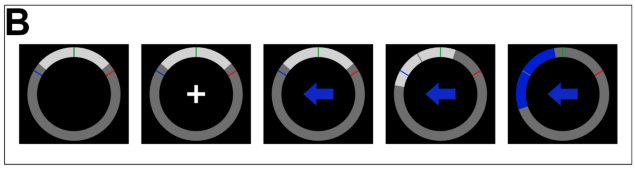

FIGURE 7: Visual paradigm for training and evaluation

4.1 Training Phase

Each model was trained on the continuous feedback of the corresponding train set, considering the subject-specific most discriminative features and the ground truth labels. The firsts were qualitatively and manually selected looking at the feature maps, with more focus on the channels of interest for the MI tasks we are analysing, i.e. both-hands and both-feet.

Since the subject neural activity is not the same for everyone, a Quadratic Discriminant Analysis (QDA) classifier was trained in order to take into account for different covariance matrices among the two classes, based on the selected features (2, 3 or 4). Their computation is critical for creating good models.

Once trained, the classifier output becomes the input of the exponential smoothing evidence control framework (2), with integration parameter $\alpha = 0.97$ and decision $D(t) = 0.5$ reset after each trial, to deliver the decision to the robot when the threshold 0.2 or 0.8, respectively for both-hands and both-feet, is crossed.

$$D(t) = \alpha D(t-1) + pp(t)(1-\alpha) \tag{2}$$

At the beginning, the exponential smoothing pushes evidence accumulation, then it stabilizes toward the end to be conservative, thus reducing oscillations and encouraging discrete command delivery.

4.1.1 Subjects Comparison

Looking at Table 1, individual subjects showed variable performance during training. *AJ1* achieved high single-sample accuracy, i.e. 95.4167%, with nearly perfect classification for both-hands and both-feet. Trial accuracy was similarly strong with fast decision times averaging $2.0615s$. *AJ4* demonstrated balanced but moderate performance, with single-sample accuracy of 82.0312% overall and trial accuracy of 74.4444% requiring $2.6528s$ for decisions. *AJ7*, the poorest performer, struggled during training, achieving 62.0313% single-sample accuracy and and never being able to make reliable trial classifications, with decision times as slow as $3.9375s$.

At the population level, training showed signs of underfitting, with single-sample accuracy of 49.8201% and similar trial accuracy. This reflects a loss of discriminability power when aggregating features across all subjects, potentially due to inter-individual variability in brain signals. The population classifier emphasizes the need for individualized feature selection to improve separability among the two classes.

4.2 Evaluation Phase

Similarly to the training phase, each model was evaluated on the continuous feedback of the corresponding test set, considering the same features adopted during training. In particular, evaluation focuses on model generalization and real-time performance on data never seen before by the model. Predictions were made using the exponential smoothing evidence accumulation framework with the same parameters used in training phase, enabling stable control signals and oscillations avoidance.

4.2.1 Subjects Comparison

Looking at Table 2, individual subjects showed a general decrease of many percentage points with respect to train set, typically it is an expected behaviour in many machine learning applications. *AJ1* maintained excellent performance during evaluations achieving 75.7997% single-sample accuracy and 84.1667% trial accuracy. Decision times were slightly slower, i.e. $2.9068s$, but still efficient for real-time applications. *AJ4* demonstrated consistent generalization, with 73.4956% single-sample accuracy and 69.1667% trial accuracy, improving decision speed to $2.388s$. *AJ7*, however, showed no significant improvement,

	Single-sample accuracy [%]			Trial accuracy [%]			
Subject	*both-feet*	*both-hands*	*overall*	*both-feet*	*both-hands*	*overall*	**Avg decision time [s]**
AJ1	96.1458	94.6875	95.4167	96.6667	93.3333	95	2.0615
AJ3	88.1944	86.5278	89.8611	73.3333	88.8889	81.1111	2.5083
AI8	83.5417	93.7153	88.6285	71.1111	71.1111	71.1111	2.9931
AJ4	69.6528	94.4097	82.0312	55.5556	93.3333	74.4444	2.6528
AI6	67.9167	91.0069	79.4618	35.5556	80	57.7778	3.0729
AJ9	65.3472	91.3889	78.3681	42.2222	46.6667	44.4444	3.3382
AI7	99.8611	0.590278	50.2257	100	0	50	1.8944
AJ7	40.1042	83.9583	62.0313	0	0	0	3.9375
average	76.3454875	79.53558475	78.25305	59.3055625	59.1666625	59.2361	2.8073375
population	98.7169	0.923295	49.8201	98.4848	0	49.2424	1.9585

Table 1: Single-sample and trial accuracies on train set

	Single-sample accuracy [%]			Trial accuracy [%]			
Subject	*both-feet*	*both-hands*	*overall*	*both-feet*	*both-hands*	*overall*	**Avg decision time [s]**
AJ1	87.5339	68.1407	75.7997	95	73.3333	84.1667	2.9068
AJ3	69.7498	86.495	76.8088	68.3333	80	74.1667	3.0333
AI8	68.67	87.7408	77.2977	73.3333	75	74.1667	3.9115
AJ4	85.6559	63.4093	73.4956	71.6667	66.6667	69.1667	2.388
AI6	65.109	88.3262	77.88	40	88.3333	64.1667	3.8427
AJ9	46.2523	89.1254	72.5096	20	21.6667	20.8333	2.988
AI7	99.2273	0	47.0407	85	0	42.5	1.8833
AJ7	18.4253	96.2219	60.5202	0	0	0	7.6063
average	67.5779375	72.4324125	70.1690375	56.6666625	50.625	53.64585	3.5699875
population	99.4405	0.430152	47.4481	90.2083	0	45.1042	1.8977

Table 2: Single-sample and trial accuracies on test set

with 60.5202% single-sample accuracy and no reliable trial classifications, with slower decision times of 7.6063s.

The population model consistently underperformed, achieving 47.4481% single-sample accuracy and 53.64585% trial accuracy during evaluation. Decision-making remained inefficient, with unstable control signals and poor evidence accumulation. The results suggest that the variability in individual neural patterns can not be effectively captured by a generalized model, leading to significant drops in performance compared to subject-specific models. However, there are still subjects that were not able to send the correct commands in most of the trials, as *AJ9*, *AI7* and *AJ7*: based on trial accuracy, *AJ1*, *AJ3*, *AI8*, *AJ4* and *AI6* achieved the best performances, since acquired BMI skills are translated into great control in an acceptable time for practical application.

4.3 Results

All previous findings underscore the importance of individualized training and feature optimization strategies. Particularly, it is well demonstrated by the population model which may require substantial adaptations or personalization to achieve practical usability.

The mean results on test set shown in Table 2 give us a general idea about the neural activity behaviour of a randomly chosen subject, although the analysis is restricted to only 8 subjects. The results show that in average a quite good classifier is obtained for both-feet and both-hands tasks on the single-sample. However, once it is time to accumulate evidence to deliver commands, potentially to a robot, there is a great uncertainty leading to wrong decisions about half of the times a trial is analysed.

Additionally, the average decision time, which is 3.5s, gives an interesting insight regarding BMI systems. That time could arise problems in some real-time systems where a quick reaction is required. Important examples we may think and which have an impact on people health, are the teleoperation systems for surgical applications in which a precise and fast action is fundamental to reduce at the minimum the risk of creating damages to human body.