Analysis of Multiday Motor Imagery BCI Data in Healthy Participants

Course in Neurorobotics
University of Padova

Simone Bozzetto, Davide Gasparini, Samuele Pinello

31 January 2026

Abstract

This work investigates motor imagery (MI)–based brain–computer interface (BCI) performance using EEG recordings from 3 days acquisition and from eight healthy participants. Two complementary analyses were conducted: a population-level grand average analysis to characterize event-related desynchronization/synchronization (ERD/ERS) patterns in the μ and β bands, and a subject-specific BMI decoding framework trained on offline calibration runs and evaluated on online runs. Grandd average results confirmed consistent sensorimotor modulations during hand and foot motor imagery tasks. The subject-specific decoding pipeline achieved high offline performance, while a systematic reduction was observed during online evaluation. Overall, the results emphasize the subject-dependent nature of MI-based BCIs and underline the challenges of translating offline decoding performance into reliable online control.

1 Introduction

In the last few years, motor imagery (MI)–based brain–computer interfaces (BCIs) have attracted considerable interest compared to other BCI paradigms. A BCI is a system that allows for direct communication between the brain and an external device. Among the various non-invasive BCI approaches, MI-based systems allow users to generate control commands without executing any movement and without relying on external stimuli (as is the case for the P300-based BCI). Motor imagery refers to the mental simulation of a movement without its physical execution. With this approach, no actual motor output is produced, but still MI engages cortical networks involved in motor planning and execution, particularly in the sensorimotor areas. This process modulates ongoing brain rhythms, especially in the μ and β frequency bands, leading to event-related desynchronization (ERD) and synchronization (ERS) phenomena. Moreover MI-based BCIs exploit induced brain activity, that are in the form of self-generated oscillatory modulations associated with voluntary mental tasks. Even though MI-based BCIs represent a promising and flexible control strategy, the major challenge remains the high inter-subject variability. Indeed, for this main reason the researchers focused on training models which are subject-dependent. In this project, we analyzed data collected during a three-day motor imagery BCI experiment involving eight healthy participants. Participants performed three different tasks: motor imagery of both hands, motor imagery of both feet, and rest. The experimental protocol included calibration and evaluation runs performed during different days. The main objectives of this work were two: first, we performed a grand-average analysis to characterize the sensorimotor patterns associated with the different motor imagery tasks at population level and at single subject level. Second, we implemented and evaluated a subject-specific decoding pipeline, including feature extraction, feature selection, classifier training during the calibration phase, and evaluation of the performance of the model during the online phase. All the analyses were conducted using MatLab. The results show on average that subjects were able to perform MI tasks with discretely good accuracy (62%), and for the personalized models' performances the average accuracy is 74%.

2 Materials and Methods

2.1 Dataset Descriptions

The dataset consists of EEG recordings from eight healthy participants collected over three experimental days. 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 which is a standard that defines the placement of EEG electrodes on the scalp based on percentage distances (10% and 20%) of the head size. The placement and order of electrodes are illustrated in Figure 1.
Each participant completed at least two recording days. During the first day each participant performed from 2 to 3 offline runs used for the calibration of the model without real feedback and 2 online runs with real feedback. Participants performed two motor imagery tasks—imagining movements of both hands or both feet—and a rest task. The training visual paradigm is shown in Figure 2 and represent the full length of a trial. The colour of the cue indicated which motor imagery task to perform (both hands, both feet, or rest).

Figure 2: Paradigm of the task.

Each trial began with the presentation of a circular interface on the screen. A fixation cross appeared in the center

Figure 1: International 10-20 system electrodes position used in this project. Red circles indicate the electrode positions, small green circles the electrode number, blue circle the reference electrode and the yellow circle the GND.

to prepare the participant and reduce eye movements. Then a directional cue was presented inside the circular interface. The direction and color of the cue indicated the task to be performed. For the calibration runs the feedback was automatically set in the correct direction associated with the cue, independently from the subject's brain activity. During the evaluation runs the feedback was instead controlled in real time by the classifier output. In particular, the circular ring that represent the feedback interface progressively moved toward the target direction as function of the classifier's posterior probability. A successful trial corresponds to the feedback reaching the target region, while failure occurs when the feedback did not reach the target in the allowed time window. This paradigm allowed participants to adapt their motor imagery skill across trials and supported subject-specific learning during the experiment.

2.2 Methodological Approach: Population-Level and Subject-Specific Analyses

Two complementary analytical frameworks were implemented: a population-level analysis and a subject-specific decoding analysis. The first one, based on the grand-average computation, was designed to characterize the common patterns associated with the different motor imagery tasks across participants. This approach aimed at identifying consistent spectral and spatial modulations and at the recognition of the most representative subjects. In the second approach the aim was the implementation of a classifier and the optimization of the classification accuracy for each individual participant.

2.3 Population-Level Analysis (Gran Average Analysis)

The population-level analysis was conducted to identify neurophysiological patterns that are consistent across participants during the execution of the different motor imagery tasks. EEG signals were pre-processed, divided in the μ and β frequency bands and segmented into task-related epochs in order to obtain insights in the event-related desynchronization (ERD) and synchronization (ERS) during the MI. Subsequently, we averaged across trials and subjects to obtain population-level representations. This approach allowed the identification of spatial and frequency-specific patterns associated with motor imagery of both hands, both feet, and rest, providing a physiological reference framework for subsequent subject-specific decoding analysis.

2.3.1 Pre-processing for the Grand Average Analysis

Signal pre-processing uses a fourth-order Butterworth filter, applied to the μ (8-12 Hz) and β ($13-30$ Hz) bands. From this operation, we obtained two different signals relative to the specific frequency band. This type of filter was chosen because it allows the isolation of relevant brain oscillations without introducing significant distortion. The filter effectively attenuates high-frequency noise and slow signal drifts, preserving the main features of neural activity related to motor imagery. A Laplacian mask was applied to increase the spatial resolution of the signal. The mask calculates the potential of each channel by subtracting the average of the surrounding channels, thus improving the identification of local cortical activity and reducing the influence of diffuse reference signals or global artifacts. The signal was then rectified by squaring it and a moving average was applied using a 1-second window. The ERD was computed using Equation 1 shown below, after the concatenation of the signal into trials.

$$ERD\% = \frac{A - R}{R} \times 100 \tag{1}$$

Where A is the activity period (continuous feedback) and R is the reference (fixation cross). The fixation data was extracted from the signal and used as baseline to the computation of the ERD considering the signal just on the fixation period. All the pre-processing steps were applied to every calibration and evaluation runs for each subject. The ERD was then averaged over trials for each subject and over subjects to obtain the average ERD.

2.3.2 Gran Average Analysis

To characterize MI-related neural activity, grand average analysis was performed on the entire population. Event-related spectral perturbations were computed to estimate ERD/ERS patterns during MI tasks relative to rest. Temporal dynamics and spatial distributions of power changes were analyzed with a focus on central electrodes (C3, Cz, C4) which overlie the sensorimotor cortex and are known to be strongly involved in motor imagery processes. This analysis aimed to identify consistent patterns of desynchronization associated with imagery of both hands and both feet.

2.4 Subject-Specific BMI Detection

The subject-specific BMI detection framework was designed to evaluate decoding performance at the individual level. Due to the well-known inter-subject variability in motor imagery–related brain activity, a personalized calibration procedure was implemented for each participant. This analysis was structured into two main phases: a calibration phase and an evaluation phase. During the calibration phase, only offline runs were considered in order to extract discriminative features and train a subject-specific classifier. During the evaluation phase, the trained model was applied to online runs to assess its generalization capability under real-time feedback conditions. The following sections describe the methodological steps used for signal processing, feature extraction and selection, classifier training, and performance evaluation.

2.4.1 Pre-processing for BMI Detection

For each calibration and evaluation runs of each subject, the signal pre-processing uses a fourth-order Butterworth filter, applied to a frequency band (8-30 Hz) that comprehend both the μ and β bands. The filter preserved the main features of neural activity related to motor imagery. The same Laplacian mask used for the Grand Average Analysis was applied to increase the spatial resolution of the signal. Then the Power Spectral Density (PSD) was computed on a subset of frequencies (4-48 Hz). The Fisher score was calculated for each run and subject for the computation of the features.

2.4.2 Calibration runs analysis: feature extraction, model training and performance assessment

The calibration phase was conducted using exclusively the offline runs recorded on the first experimental day. The objective of this phase was to extract subject-specific discriminative features and to train a classifier capable of distinguishing between the motor imagery tasks. The features selection was performed manually by inspecting the Fisher score matrix for each calibration run in order to keep only the most discriminative components, allowing to construct a model with reduced dimensionality and minimizing the risk of overfitting. The selected features were then used to train a subject-specific classifier. Model parameters were estimated exclusively on the calibration data to ensure a clear separation between training and evaluation phases. Performances were tested directly on the offline data to have an insight of the created model. Two complementary metrics were computed. First, the single-sample accuracy was evaluated by using a leave-one-out run cross validation. This metric reflects how well the extracted features separate the motor imagery classes at single sample level giving a measure of the quality and robustness of the trained model before proceeding with the online validation. Second, the trial accuracy was computed by aggregating classifier outputs over each trial and assigning a final class decision per trial. This metric gives us an idea of the suitability of the model for real-time BMI control.

2.4.3 Evaluation runs analysis: evaluation of the classifier and evidence accumulation framework

During the evaluation phase, the same features selected in the calibration phase were extracted for each subject and the trained model was evaluated by computing the single sample accuracy and the trial accuracy on the online runs. The classifier was directly applied to the unseen online data to assess its generalization capability under feedback conditions. To improve the decision stability and reduce the misclassification, an evidence accumulation framework was developed using the posterior probabilities estimated by the classifier during the online phase. First it was implemented an exponential accumulation framework in order to integrate the evidence over time. It was used the equation 2 shown below:

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

Where $D(t)$ represents the accumulated evidence for a class, α is the memory factor that controls the trade-off between past evidence and current posterior probability. At the beginning of each trial, the accumulated evidence was set to a neutral value of 0.5 and fixed decision thresholds were applied to determine the command activation (respectively set to 0.2 and 0.8). This exponential accumulation reduces high-frequency oscillations in the classifier output and increases temporal consistency of the decision signal. To further regularize the system response, a dynamic control mechanism was applied to the accumulated evidence. The control signal $C(t)$ was updated according to the equation 3:

$$C(t) = C(t-1) + \alpha_{control} \cdot (D(t) - C(t-1)) \tag{3}$$

Where $\alpha_{control}$ is an integration gain controlling the responsiveness of the system. The combined framework provides a more stable and physiologically plausible control signal, improving robustness and reducing spurious command activations during online BMI operation. Decision-level performance was evaluated based on the control signal obtained with the accumulation framework. For each trial, a command was chosen when the control variable passed the upper threshold for the first time. If the control variable did not cross the threshold, no decision was assigned. Online trial accuracy (Equation 4) was computed as the proportion of correctly classified trials among the trials in which a decision was taken.

$$Acc_{trial} = \frac{N_{correct}}{N_{total\_decisions}} \tag{4}$$

In addition to the trial accuracy, the decision rate was computed as the proportion of trials in which a command was correctly delivered. This metric, in particular, is informative about the system's ability to produce control outputs under real-time conditions. Finally, the time to deliver a command was evaluated as the latency between the trial onset and the first threshold crossing. Since the trials were measured in windows, the time to deliver a command was evaluated by converting them into seconds using the window shift parameter. The average and standard deviation of this latency were calculated across trials with valid decisions.

3 Results

3.1 Gran Average Analysis

Grand average analysis revealed a clear ERD in the μ and β bands over sensorimotor areas during MI tasks, noticing also the beta rebound. Hand imagery elicited lateralized ERD over C3 and C4, whereas foot imagery induced stronger desynchronization over Cz. These patterns were consistent across subjects, though with varying magnitudes.

To analyze and be able to make a comparison between different subjects of the population, the following metrics have been used: plot of averaged ERD/ERS over trial time to have a temporal visualization on the expected or not trend of the signal, topoplot for a spatial visualization of the ERD/ERS amplitude over the scalp. In order to have informative plots and average values over the entire population, it was necessary to remove the second and third calibration runs of subject 2, which resulted in being extremely noisy. Since the first calibration run was acceptable, it was decided to keep the subject with only one calibration run instead of removing it completely.

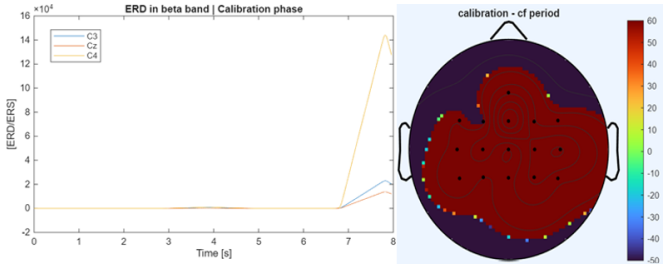

Figure 3: Subject 2 ERD/ERS signal and topoplot of the calibration phase before the removal of the noisy calibration runs.

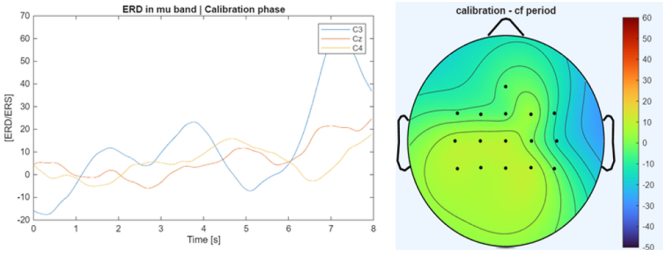

Figure 4: Subject 2 ERD/ERS signal and topoplot of the calibration phase after the removal of the noisy calibration runs.

On the other hand, there are two subjects (1 and 4) that are more representative of the task among the others. Subject 1 has a strong desynchronization in the μ band over the C4 electrode in both hands task, while in both feet one it has a strong beta rebound in the β band over the Cz electrode. This is also appreciable in the topoplots, where for both feet is clear the strong activation over the central area of the motor cortex.

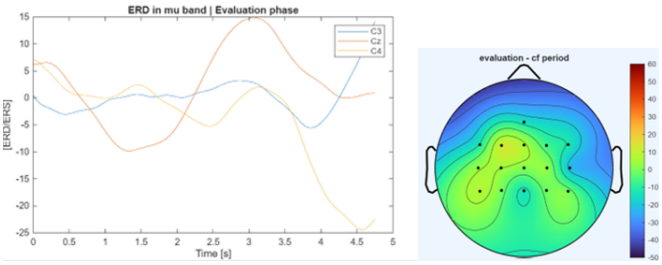

Figure 5: ERD analysis of subject #1 in the μ band during the evaluation phase during the Both-Hands task. Left: Temporal evolution of ERD/ERS on C3, Cz, and C4 channels. Right: Topographic map of power distribution.

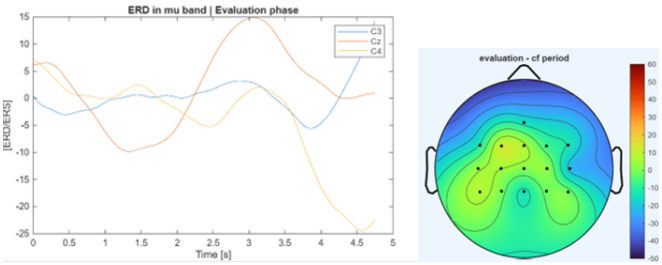

Figure 6: ERD analysis of subject #1 in the μ band during the evaluation phase during Both-Feet task. Left: Temporal evolution of ERD/ERS on C3, Cz, and C4 channels. Right: Topographic map of power distribution.

For subject 4, is also observable the beta rebound in both feet task and the μ desynchronization in both hands one. Therefore, these results are also retrievable in the topoplots.

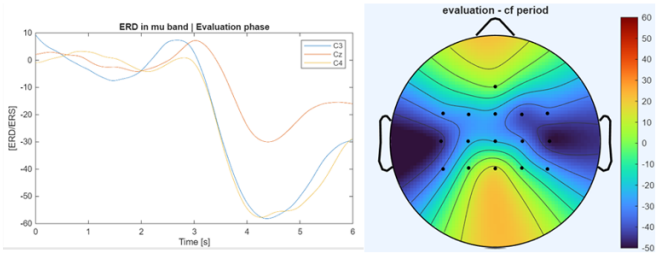

Figure 7: ERD analysis of subject #4 in the μ band during the evaluation phase during Both-hand task. Left: Temporal evolution of ERD/ERS on C3, Cz, and C4 channels. Right: Topographic map of power distribution.

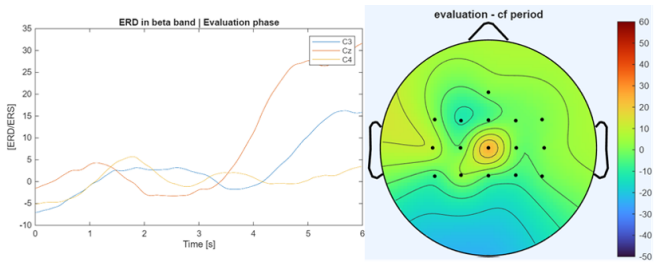

Figure 8: ERD analysis of subject #4 in the μ band during the evaluation phase during Both-feet task. Left: Temporal evolution of ERD/ERS on C3, Cz, and C4 channels. Right: Topographic map of power distribution.

Moving the attention to the actual population average, some consistent results have been obtained: the average

temporal analysis shows the desynchronization in the mu band and the beta rebound, also the spatial one shows the general activation of the right motor cortex areas, even with some disturbances.

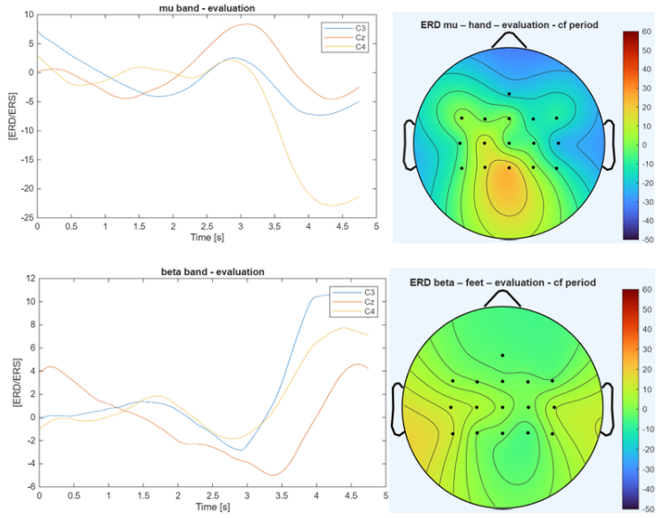

Figure 9: Grand Average ERD analysis in the μ band during the evaluation phase during Both-hand (top) and Both-Feet (bottom) task. Left: Temporal evolution of ERD/ERS on C3, Cz, and C4 channels. Right: Topographic map of power distribution.

3.2 BMI Decoding Performance

During the calibration files the selected features were analyzed to obtain insights into the subject-specific discriminative patterns.

Subject	[Frequency (Hz), Channel]
Subject 1	[18, 9], [20, 9], [22, 9]
Subject 2	[14, 11], [14, 7]
Subject 3	[14, 11], [14, 7]
Subject 4	[12, 7], [12, 11], [10, 7], [10, 11]
Subject 5	[14, 7], [12, 7], [12, 6]
Subject 6	[12, 7], [14, 7], [12, 8], [14, 8]
Subject 7	[12, 11], [10, 11]
Subject 8	[12, 8], [12, 12], [12, 16], [14, 16]

Table 1: Selected Frequency-Channel pairs for each subject.

Across subjects, the most common selected features involved the channels positioned over the sensorimotor cortex, in particular the central electrodes C3, C1, Cz, C4, and frequencies within the μ and β bands (going from 10Hz to 20Hz). These features are consistent with the motor imagery, where task-related ERD/ERS modulations are expected over sensorimotor areas. However, variability across subjects was observed both in the spatial distribution of selected channels and in the relative contribution of frequency bands. This variability confirms that a subject-specific calibration procedure is required, as no single feature configuration was universally optimal across the population. According to the observation of the second subject in the Grand Average Analysis, also in the calibration phase we had an extremely noisy signal in two of the calibration runs. These two runs were excluded from the calibration procedure, the selection of the features and the training of the model was performed only using the remaining calibration run that presented clean data.

3.2.1 Calibration Performance

The offline single-sample accuracy, shown in figure 10 demonstrated that the subject-specific classifiers were able to discriminate between the two motor imagery classes during the calibration phase. All subjects achieved performance above the chance level (50%), confirming that the selected features contained meaningful information to perform class selection.

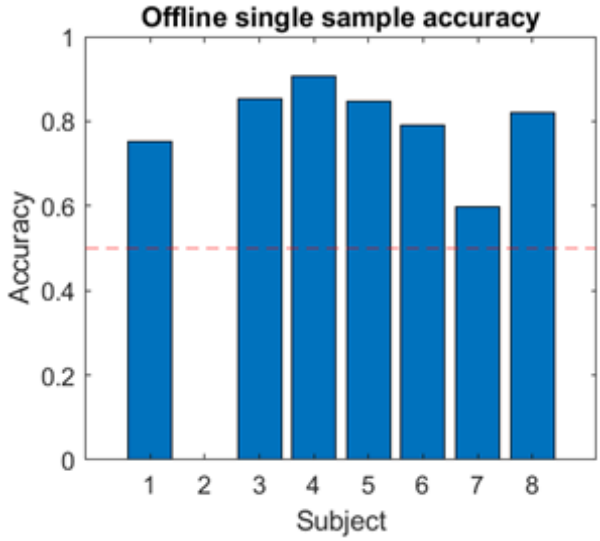

Figure 10: offline single sample accuracy of single subject.

Accuracy values ranged approximately between (60%) and (90%), indicating inter-subject variability. Subject 4 showed the highest discriminative capability, approaching (90%) accuracy. Subject 2 was excluded from this analysis because only one calibration run was available. As a leave-one-run-out cross-validation scheme was adopted, at least two runs were required to ensure proper training–testing separation. Therefore, single-sample accuracy could not be computed for this subject. In Figure 11 it is shown the offline trial accuracy that exceeds the single-sample accuracy reaching high values as almost the unit.

3.3 Online performance

During the evaluation phase, all single-sample accuracies, shown on Figure 11, were above chance level, confirming that the calibrated models preserved discriminative capability when applied to unseen online data. Accuracy values were distributed between approximately (70%) and (80%) for most participants.

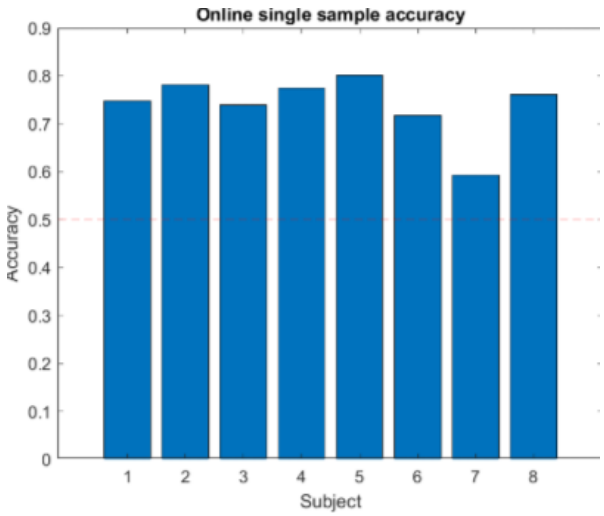

Figure 11: online single sample accuracy of single subject.

Compared to the offline calibration phase, a reduction in performance can be observed in several subjects. This decrease is expected and can be due to distributional shifts between calibration and online runs. On the other hand, the online trial accuracy (in Figure 12) showed more variability among subjects going from (53%) to (88%) but remaining above the chance level.

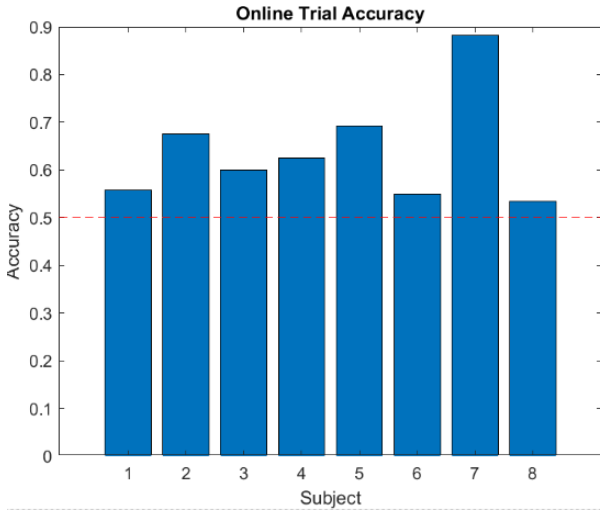

Figure 12: online trial accuracy of single subject.

There is also a clear drop in terms of accuracy due to the difficulty of maintaining a stable control under feedback conditions. Although Subject 7 achieved the highest online trial accuracy, this result must be interpreted considering the extremely low decision rate (14%) in respect to the unit of all the other subjects). The system generated decisions in only (14%) of the trials, meaning that in most cases the control signal never exceeded the decision threshold. Comparing the calibration and evaluation results we can clearly see a reduction in performance, in particular at trial level. This brings a low generalization which is still the challenge of the real-time BMI control. Lastly, the average time required to deliver a command was computed and it provides a measure of the system's ability to choose rapidly the action to perform. It can be shown in Figure 13

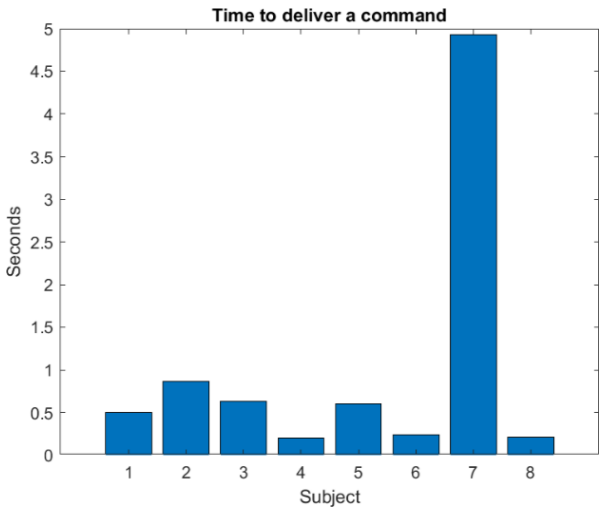

Figure 13: Average time to deliver a command

For most subjects, the average time-to-command remained below 1 second, indicating a relatively fast progress over the decision. This suggests that the accumulation framework allowed stable and timely command generation in most participants. Subject 7 had the highest latency which confirmed the extremely low decision rate and the difficulty of this subject to deliver a command.

3.4 Overall performance overview

The grand-average analysis across subjects shown below in Figure 14 confirms the trends observed at the individual level.

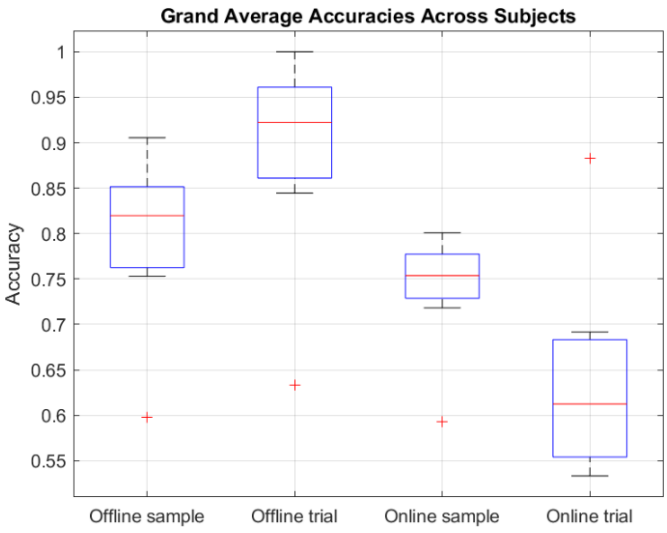

Figure 14: Grand average accuracies across subject.

Offline performance was better than the online performance, highlighting the impact of real-time conditions. Moreover, the reduction from offline to online accuracy emphasizes the generalization gap between calibration and evaluation phases. Overall, the results indicate that the proposed decoding and accumulation framework provides reliable performance maintaining acceptable robustness in online operation.

4 Conclusions

This work investigated both the patterns of motor imagery at the population level and the performance of a subject-specific BMI decoding framework. At the population level, the grand average analysis confirmed the presence of ERD/ERS patterns in the μ and β bands over the sensorimotor areas that also supported the use of these features for the decoding part. At decoding level, the calibration phase demonstrated that subject-specific models were able to achieve robust discrimination between motor imagery tasks under controlled conditions. However, a clear performance reduction was observed during online evaluation. The evidence accumulation and control dynamics framework contributed to stabilizing decision outputs, improving reliability at the trial level. In particular, two subjects highlight the practical challenges of MI-based BCIs. Subject 2 presented severely noisy calibration runs, requiring the exclusion of corrupted data to ensure model reliability. On the other hand, Subject 7 achieved high online trial accuracy but exhibited an extremely low decision rate, generating commands in only a small fraction of trials. This behavior illustrates that accuracy alone is not sufficient to assess BMI effectiveness. Overall, the results confirm that MI-based BCIs remain strongly subject-dependent systems, requiring individualized calibration and careful balancing between stability and responsiveness. The performances could surely increase with more subject training and model fine tuning.