Defining and Quantifying Users' Mental Imagery-based BCI skills: a first step

Fabien Lotte, Camille Jeunet

Defining and Quantifying Users' Mental Imagery-based BCI skills: a first step

Fabien Lotte[1,2] and Camille Jeunet[3,4]

[1] *Inria - France*
[2] *LaBRI - CNRS/Univ. Bordeaux/INP Bordeaux - France*
[3] *Defitech Chair in Brain-Machine Interfaces (CNBI), EPFL - Switzerland*
[4] *Univ. Rennes, Inria, IRISA, CNRS - France*

Abstract

Objective: While promising for many applications, Electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs) are still scarcely used outside laboratories, due to a poor reliability. It is thus necessary to study and fix this reliability issue. Doing so requires the use of appropriate reliability metrics to quantify both the classification algorithm and the BCI user's performances. So far, Classification Accuracy (CA) is the typical metric used for both aspects. However, we argue in this paper that CA is a poor metric to study BCI users' skills. Here, we propose a definition and new metrics to quantify such BCI skills for Mental Imagery (MI) BCIs, independently of any classification algorithm.

Approach: We first show in this paper that CA is notably unspecific, discrete, training data and classifier dependent, and as such may not always reflect successful self-modulation of EEG patterns by the user. We then propose a definition of MI-BCI skills that reflects how well the user can self-modulate EEG patterns, and thus how well he could control an MI-BCI. Finally, we propose new performance metrics, *classDis*, *restDist* and *classStab* that specifically measure how distinct and stable the EEG patterns produced by the user are, independently of any classifier.

Main results: By re-analyzing EEG data sets with such new metrics, we indeed confirmed that CA may hide some increase in MI-BCI skills or hide the user inability to self-modulate a given EEG pattern. On the other hand, our new metrics could reveal such skill improvements as well as identify when a mental task performed by a user was no different than rest EEG.

Significance: Our results showed that when studying MI-BCI users' skills, CA should be used with care, and complemented with metrics such as the new ones proposed. Our results also stressed the need to redefine BCI user training by considering the different BCI subskills and their measures. To promote the complementary use of our new metrics, we provide the Matlab code to compute them for free and open-source.

1 Introduction

While they are very promising for numerous applications, such as assistive technology or gaming, Electroencephalography (EEG)-based Brain-Computer Interfaces (BCIs) are still scarcely used outside laboratories [3]. This is mostly due to their poor reliability, as they often recognize erroneous mental commands from the user. One of the main current challenges for the community is thus to improve BCI reliability [3]. This is currently addressed at different levels, such as trying to design more robust EEG signal processing algorithms, or trying to improve BCI user training approaches, which have been shown to be inappropriate and a major cause of poor performances, both in theory and in practice [3, 17, 26, 28]. Improving these different aspects requires metrics to measure BCIs reliability and thus their performances. Indeed such performance metrics could identify what are the limitations of a given algorithm or training approach, which is a necessary first step towards fixing these limitations [3].

User performance metrics are particularly useful for studying and improving Mental Imagery (MI) BCI user skills acquisition. Appropriate performance metrics could indeed help to understand what users have successfully learned or still need to improve, which can then be used to guide them, i.e., to provide them with appropriate training tasks and feedback. In EEG-based BCI, the most used metric is online Classification Accuracy (CA), i.e., the percentage of mental commands that were correctly recognized by the BCI during online use [39, 42, 43]. Online CA, together with other machine learning evaluation metrics [39, 42, 43], have been successfully used to quantify the decoding performance of the BCI, i.e., how well the BCI recognizes the users' commands. However, CA is also used to study BCI users' performance and learning, i.e., how well users can modulate/self-regulate their EEG signals to control the BCI, and how much they learn to do so. For instance, CA is typically used to study how different kinds of feedback influence BCI users' training [21, 31, 35], or how different psychological factors influence BCI users' learning and performances [19].

In this paper, we argue and demonstrate that CA alone, as used in online MI-BCI, is not enough to study user performance and thus their MI-BCI skills. Indeed, this metric is notably discrete - an input data is either correctly or incorrectly classified - as well as classifier and training data dependent, since changing the training data or classifier will change the resulting performance, independently of the actual users' BCI skills, i.e., independently of how well they can modulate their EEG signals using MI. Overall, CA can tell us how well the machine can recognize the EEG patterns from the user, but was not designed to tell us how well the user is able to produce clear, stable and distinct EEG patterns. As such, CA might not provide us with this information, and we actually show in this paper that CA can actually fail to do so in practice. In order to fully understand BCI user skill acquisition, alternative or additional metrics are thus necessary. Moreover, to be able to quantify MI-BCI skills using such metrics, we need to define what those skills are. Therefore, in this paper, we propose a first definition of MI-BCI skills and propose simple and computationally efficient metrics to quantify them. We then compare them with the classically used online CA[1]. We show that using

[1]Preliminary results with a subset of the metrics proposed here, and on a single data set, with only two classes and a single session per subject, have been published as a conference paper in [27]

online (or simulated online) CA as metric may actually hide several relevant aspects of BCI skill acquisition. In particular, online CA may miss users' MI-BCI skills increase overtime or fail to identify that a mental task performed is actually no different than rest EEG. Our new metrics can overcome these limitations. Since performance metrics were also used in several papers to identified psychological and neurophysiological factors influencing BCI performance, we also studied whether our new metrics could confirm the previously identified influence of one of them: spatial abilities [19]. Our results showed that indeed some of our new metrics are also significantly correlated to spatial abilities, hence further confirming the importance of this factor.

This paper is organized as follows: The following section presents a brief survey of the common performance metrics used in BCI, notably those related to the classification accuracy, as well as their limitations. The next section introduces the materials and methods, notably the MI-BCI skills definition and the new metrics we propose. It also presents the data sets on which these measures are compared. Then the Results section compares the performances estimated with all metrics, which are then discussed in the Discussion section. The last section concludes the paper.

2 Performance metrics for BCI

2.1 Current common performance metrics

As indicated before, Classification Accuracy (CA) is the most used metric to quantify BCI performances. Typically, the classifier is trained on the EEG signals from the trials of the first BCI run(s) (calibration run(s)) and applied to classify the users' EEG signals from the trials of the subsequent runs. CA is defined as the percentage of these EEG trials that were correctly classified [42]. From the classification results, we can also obtain a more detailed information on the performances from the Confusion Matrix (CM), which informs about how many trials from each class were estimated to be from each one of the possible classes. The CM is defined as follows for a two class problem:

Table 1: Confusion matrix for two classes

		Estimated class	
		Class 1	Class 2
Real Class	Class 1	a	b
	Class 2	c	d

Here, the number in row i, column j is the number of trials from class i that was classified as belonging to class j. Thus, a and d correspond to correct classifications (the real and estimated classes are the same), and c and b to erroneous classifications. CA (in %) can thus be computed as follows:

$$CA = \frac{a+d}{a+b+c+d} \times 100 \tag{1}$$

or more generally, for any number of classes, if CM_{ij} is the $(i,j)^{th}$ element of the

confusion matrix CM, then:

$$CA = \frac{\sum_i CM_{ii}}{\sum_{i,j} CM_{ij}} \times 100 \quad (2)$$

From there we can also estimate the CA of each class, e.g., $\frac{CM_{ii}}{\sum_j CM_{ij}} \times 100$ is the percentage of trials from class i that were correctly classified.

In addition to the CA, other metrics have been proposed based on the CM, notably the Kappa coefficient [39, 42, 43], or the extended CM to support non-control states [2].

When the BCI is used to select items (e.g., letters to spell or a direction with a wheelchair), metrics that also take into account the selection time have been proposed, such as the widely used Information Transfer Rate (ITR) [49], the BCI utility metric [7], the correct number of spelled letters per minute (for spellers) [41] or the Rate of Information Gain (RIG) [16].

Metrics have also been proposed to study the BCI performances specifically for ERP-BCI applications such as ERP-Spellers, with the projected accuracy [4] or the classifier-based latency estimation (CBLE) [44]. For BCI-based cursor control, Fitt's law and its variants have also been used to estimate BCI performances [9].

Interestingly enough, Hill and colleagues proposed to quantify BCI user training and performances using an adaptive stair-case task, i.e., a game with adaptive difficulty [16]. The median difficulty level reached by the BCI user in this game has been shown to reflect well BCI control performances, and to be highly correlated to the RIG mentioned above.

Bauer et al also proposed to quantify the participant opportunity for learning in BCI according to the concepts of the Zone of Proximal Development (ZPD). They estimated the ZPD as the difference between the True Positive Rate (TPR) and the False Positive Rate (FPR) for various thresholds, in a binary BCI with a positive (active) and negative (rest) class [1].

It should be mentioned that to study user learning in BCI and in NF applications, the neurophysiological variations of EEG over time have been studied. Typically the power of the EEG from a given channel and frequency band is computed, and compared after and before training, or is correlated to the BCI class labels (commands), see, e.g., [6, 15, 47]. This can provide spatial and/or spectral topographies of the main changes in EEG due to training. No global metric of neurophysiological change due to BCI/NF training is available though.

Among all these metrics, the CA and possibly the ITR (when taking into account a specific application) are by far the most used performance metrics for BCI. Typically CA is the metric used in most papers to quantify BCI user performance and learning [19]. Unfortunately, as we will see below, this metric suffers from several limitations for this task. The other metrics mentioned here also share most of these limitations.

2.2 Limitations

CA and the other metrics mentioned above are very useful to quantify the decoding performance of a BCI and/or the performance with the applications controlled using

the BCI (e.g., spelling performance with an ERP-speller) [6, 43, 42, 39]. However, when it comes to studying how well users can voluntarily modulate their EEG signals to control the BCI, we argue that such metrics actually suffer from many limitations, and that dedicated metrics are needed.

First, these metrics are unspecific: they only provide the global performance, but not what is correctly classified or not, nor why it is so. Then, these metrics are typically discrete measures: a trial is either correctly classified or not, there is no middle ground. As such, even if the user produces a stronger EEG modulation than before, but not strong enough to make the trial correctly classified, metrics such as CA will not change.

All the metrics above measure the global performance of the BCI or of the BCI-controlled application. As such the obtained performance metrics depend on both the EEG signal processing tools, the user skill at BCI control and the application interface (e.g., different speller configurations or interaction techniques), among others. Thus, such metrics cannot unambiguously inform us about the users' MI-BCI skills and learning curve. It should be mentioned though that Kübler et al. proposed to evaluate BCI systems not only according to machine performance but also according to the user experience. They notably proposed to evaluate the overall usability of BCIs, by considering the system efficiency and effectiveness, as well as the satisfaction it gave users [22]. Note that the satisfaction is a subjective metric, and that both efficiency and effectiveness depend also heavily on the EEG signal processing tools and interface design. Metrics dedicated to quantify the users' skills at BCI control are thus still lacking.

The currently used performance metrics and notably CA are also strongly classifier and training data dependent. Changing the classifier type, its parameters, or the amount and/or quality of the training data will change the metric, independently of how well users can modulate their EEG activity using MI. Therefore, variations of these metrics might not always reflect users' proficiency at BCI control. In particular, such metrics and notably CA reflect the user's performance if and only if the classifier is a good classifier, i.e., a classifier that can classify successfully the EEG patterns that the user can modulate using MI to control the BCI. However in practice it is rarely the case, due to many factors. Classifiers are indeed sensitive to non-stationarities, and thus would lead to poor CA when applied on EEG data from a different distribution than that of the calibration run. This is likely to happen if users are trying out various strategies or are learning. This can also happen if their MI were of poor quality during the calibration runs, which is also likely to happen for naive users, unfamiliar with such MI. To take an extreme case, let us consider a hypothetical user who can produce very strong and very localized EEG desynchronization during motor imagery. With a properly trained classifier, such a user could reach 100% classification accuracy. However, if we give that user a dysfunctional, random classifier, trained on pure noise, then this user will reach a CA at chance level. In that last case, using CA as the metric to quantify the user's BCI skills would have made us conclude this user was unable to use an MI-BCI at all. Yet, using another classifier, adapted to the EEG signals that the user can modulate using MI, would have revealed how skilled this user is, and how well he could control an MI-BCI. This example highlights the need for metrics to quantify users' BCI skills independently of a single given classifier, as classifier-dependent metrics such as CA can be highly misleading. Finally, when based on a discriminative classifier such

as Linear Discriminant Analysis (LDA), the most used classifiers for BCI [3], CA does not reflect how well a given mental command can be recognized but rather how distinct the mental commands are from each other. Therefore, if users are unable to modulate their EEG signals for one class (e.g., left hand MI), they may still obtain very high CA as long they can modulate their EEG for the other class (e.g., right hand MI), since the EEG signals from the two classes would still be distinct. This leads to a last limitation: in MI-BCI, CA and other metrics usually consider the MI EEG signals only, but not the rest EEG signals. As illustrated just before, this prevents us from identifying whether the user's EEG patterns during MI are actually any different from rest EEG.

For all these reasons, CA and other existing metrics may not be able to reveal some important aspects of MI-BCI user performance and learning. In other words, CA and related metrics are appropriate measures to study how well the BCI can decode the users' mental commands, but on their own, they are not enough to study how well users can modulate their EEG patterns using MI in order to control an MI-BCI and how well they are learning to do so. This thus calls for a definition of what MI-BCI skills can be and for new and specific metrics to quantify these skills. This is what we propose in this paper. We present our definition of MI-BCI skills and describe our new metrics in the following sections.

3 Materials and methods

3.1 Defining MI-BCI Skills

So far, whereas it is clear that MI-BCI control requires and involves learning [33, 28, 18] there is no formal or agreed upon definition of the skills to be learned. There is thus a need to go towards a definition and quantification of such skills that we denote here as MI-BCI skills. In order to try to conceptualize MI-BCI skills, let us first consider a simple analogy in which a human user also employs a given human-computer interface (HCI) to achieve a goal: Formula 1 racing. With Formula 1 racing, a human user - the pilot - has to drive a Formula 1 car in order to complete a race as fast as possible. The outcome of the race, which is the score of the pilot, is how fast they completed the race. This outcome is compared to that of other pilots to designate the winner. It is important to note that this outcome of the race depends on both the driver, and notably his driving skills, and on the Formula 1 car used (e.g., how fast it can go). Therefore, a poor race outcome can be due to either a bad (slow) car, that is thus unlikely to win irrespectively of how skilled the driver is, or can be due to a bad driver, which can also fail despite being provided with a fast car. The best outcome is to be expected with both a skilled driver and a fast car. The driver should thus train in order to acquire good Formula 1 racing skills. Note that this skill is not dependent on a single Formula 1. Naturally, knowing well the used Formula 1 will increase the likelihood of success for the driver. However, good Formula 1 drivers are not good with a single Formula 1, they are good drivers in general. For instance, Michael Schumacher won many races with many different types of cars, with different engines, tires, chassis and from different constructors[2].

[2]http://www.statsf1.com/en/michael-schumacher/victoire.aspx

In this paper, we argue that quantifying MI-BCI performance should follow a similar logic. In order words, in our analogy, the pilot should be replaced by the MI-BCI user, and the Formula 1 car by the classifier. This would mean that CA, i.e, the BCI use outcome, depends both on the users' skills at MI-BCI control and on the quality of the classifier used. As such, CA does not reflect BCI users' skills only. While training to use a given classifier is likely to improve CA, a skilled MI-BCI user should be able to reach high MI-BCI control with various classifiers, and not just with a single one. Note by the way that during MI-BCI use, classifiers are regularly retrained, or updated, see, e.g., [47, 34], which means that a good BCI user cannot be good with a single classifier, but has to be good with various classifiers. This ensures generalizable skills, and prevent overfitting to a specific classifier. If the BCI user is able to produce clear and stable brain activity patterns, then multiple types of classifiers would be able to recognize those patterns and could consequently be used to control the BCI. Therefore, for the user, controlling an EEG-based MI-BCI means producing EEG patterns that can be reliably translated into commands for an application using MI. Accordingly, we propose the following definition for the skills required to do so, i.e., MI-BCI skills:

"MI-BCI skills correspond to the ability of the user to voluntarily produce brain activity patterns that are distinct between mental tasks, and stable within mental tasks, so that they can be translated reliably and consistently into control commands. The more stable and distinct the brain activity patterns, the higher the MI-BCI skills."

Note that this definition being that of a skill, it is naturally focused on the user, and not on the machine (i.e., not on the classifier or other acquisition and processing hardware and software). From this definition, it follows that classification accuracy, which is the most commonly used metric to quantify MI-BCI performance and user learning [19, 23, 33], could successfully measure MI-BCI skill in some conditions, but may fail to do so properly in some other conditions. In particular, classification accuracy would reflect MI-BCI skills if and only if the classifier[3] is able to exploit and recognize the brain activity patterns mentioned above, and perfectly translate them into control commands. If the classifier exploits other brain activity patterns, that the user struggles to make distinct and stable, then classification accuracy will not reflect the true user MI-BCI skills. In other words, while high MI-BCI skills can mean high classification accuracy, low classification accuracy may not always mean low MI-BCI skills, but can rather mean an inappropriate classifier. There is no bijection between these two measures. We aim at demonstrating this point experimentally in the present paper, and at proposing metrics to quantify such MI-BCI skills, in order to provide new tools to study MI-BCI user training.

3.2 New Performance metrics 1: Run-Wise Cross-Validation

To address some of the limitations mentioned above, a possible approach (not new in itself but typically not used to study MI-BCI user training) would be to perform Run-

[3]In this discussion, we use the word classifier in a broad sense, i.e., including the feature extraction and filtering steps. In other words, the classifier refers here to all the processing pipeline translating EEG signals into a control command.

Wise Cross-Validation (RWCV). The idea is to use CV to estimate offline the CA of each run. With RWCV, the trials from the current run are divided into K parts, K-1 parts being used to train the classifier, and the last part to test the latter, the process being repeated K times, and the obtained CA averaged over the K testing parts. This also provides a run-wise confusion matrix and class-specific CV accuracies, as done with the standard CA. We will assess this approach in this paper. Interestingly enough, in [29], McFarland et al. explored classifier adaptation based on CV on each run, which proved to be the most efficient offline adaptation strategy among those tested for motor-imagery BCI. It suggests that EEG signals vary at the run scale, and thus that RWCV might also be a relevant tool to study varying performances over time.

Since training and testing are performed on each run, and for different parts of each run, it makes RWCV CA much less sensitive to training data and to non-stationarities. This metric remains non-specific and discrete though along with being classifier-specific. In addition, it still ignores the rest EEG. It is also computationally expensive. As such its use can be impractical, or even impossible if we target online performance evaluation for instance.

3.3 New Performance metrics 2: Stability and distinctiveness

To further improve on the metric mentioned above, there is thus a need for metrics that are also specific, continuous, that consider rest EEG signals, that are classifier independent and computationally cheap and that actually measure MI-BCI skills as we defined in Section 3.1. In other words, we need metrics that measure how distinct and stable the EEG patterns produced by the users with MI are, in a continuous and classifier independent way. A stable pattern would be a pattern that is not changing dramatically between trials, and thus with a small variance. A distinct EEG pattern would be both 1) an EEG pattern that is distinct from the rest EEG pattern, i.e., there is a specific signature to that pattern and 2) a pattern that is distinct from the EEG patterns of the other MI tasks, so that each can be associated to a distinct command. As such, it would make sense to design a set of metrics dedicated to estimating how stable and distinct the EEG patterns for each MI task actually are.

Interestingly enough, metrics quantifying these various properties can be defined using distances in a Riemannian geometry framework. Indeed, Riemannian geometry offers an efficient and simple way to measure distances between covariance matrices, such matrices being increasingly used to represent EEG patterns [5, 50].

3.3.1 Riemannian geometry in brief

Let us first consider matrix $X_i \in \mathbb{R}^{N_c \times N_s}$ of EEG signals from trial i, with N_c the number of channels and N_s the number of samples per trial. This can be a matrix of preprocessed EEG signals, for instance a matrix of EEG signals band-pass filtered in 8-30Hz for a motor imagery-based BCI experiment. The spatial covariance matrix C_i of this trial is defined as $C_i = \frac{1}{N_s} X_i X_i^T$, with T being transpose. Therefore, the diagonal elements of C_i represent the EEG band power for each channel (in the band in which the signals were band-pass filtered, e.g., 8-30Hz in our example above), and the off-diagonal elements, their covariations. Such spatial covariance matrices are

used - implicitly or explicitly - to represent EEG signals in numerous MI-BCI designs, notably those based on the Common Spatial Patterns (CSP) algorithm, and many others [5, 38, 45, 50]. The Riemannian distance $\delta_R(C_i, C_j)$ between covariance matrices C_i and C_j can be defined as:

$$\delta_R(C_i, C_j) = [\sum_{i=1}^{n} \log(\lambda_i)^2]^{1/2} \tag{3}$$

where the λ_i are the eigen values of $C_i^{-1}C_j$. This Riemannian distance is particularly interesting since it is affine invariant: it is invariant to full rank linear transformations, i.e., to variations such as normalization or channel displacement [5, 50]. As such, the Riemannian distance has been used successfully for robust EEG signal decoding, in various kinds of BCIs [5, 50]. In this paper, we show that this distance can also be a very relevant tool to quantify how distinct and stable the EEG patterns produced by a BCI user are.

3.3.2 Basic stability and distinctiveness metrics

How distinct the EEG patterns produced during two MI tasks are from each other could be quantified using the Riemannian distance between the average covariance matrices for each MI task. Then, the stability of a given EEG pattern can be defined using the average distance between each trial covariance matrix and the average covariance matrix for this task, which is a form of Riemannian standard deviation [50]. More formally, let us first define the Riemannian mean $\bar{C}$ of a set of covariance matrices C_i [50] as:

$$\bar{C} = \text{argmin}_C \sum_{i=1}^{N} \delta_R^2(C_i, C) \tag{4}$$

Note that there are efficient implementations to obtain such mean matrices [50]. We can also define the mean absolute deviation (i.e. a form of dispersion measure) σ_C of a set of covariance matrices C_i as:

$$\sigma_C = \frac{1}{N} \sum_{i=1}^{N} \delta_R(C_i, \bar{C}) \tag{5}$$

Distinctiveness From the definitions above, we propose to define the distinctiveness *classDis* of the EEG patterns from two MI classes A and B as:

$$classDis(A, B) = \frac{\delta_R(\bar{C}^A, \bar{C}^B)}{\frac{1}{2}(\sigma_{C^A} + \sigma_{C^B})} \tag{6}$$

where $\bar{C}^K$ and σ_{C^K} are respectively the mean and mean absolute deviation of the covariance matrices from MI class K. This equation can be seen as an extension of the Fisher criterion (see, e.g., [14]) to covariance matrices: the further apart from each other are the average EEG patterns from each class (measured by $\delta_R(\bar{C}^A, \bar{C}^B)$), with

respect to their variance (represented by $\frac{1}{2}(\sigma_{C^A} + \sigma_{C^B})$), the higher their distinctiveness.

However such a metric only works for 2 classes. Thus, still following the analogy with the Fisher criterion and its multiclass extensions, we can define a similar metric for the multiclass case. For a BCI with N_c MI commands/classes ($N_c > 2$), this metric would be defined as the ratio of the between class variance to the within class variance, as for the multiclass Fisher criterion [14]:

$$classDis(\{A_i\}) = \frac{\sum_{i=1}^{Nc} \delta_R(\bar{C}^{A_i}, \bar{\bar{C}}^A)}{\sum_{i=1}^{Nc} \sigma_{C^{A_i}}} \tag{7}$$

where $\bar{\bar{C}}^A$ is the average of the mean covariance matrices from all classes.

Similarly, we propose to define the distinctiveness *restDis* between the EEG patterns from one class and those from the rest state as:

$$restDis(A) = \frac{\delta_R(\bar{C}^A, \bar{C}^{rest})}{\frac{1}{2}(\sigma_{C^A} + \sigma_{C^{rest}})} \tag{8}$$

where $\bar{C}^{rest}$ and $\sigma_{C^{rest}}$ are respectively the mean and mean absolute deviation of the covariance matrices of the rest EEG.

Stability Finally, we can define the stability of the EEG patterns from one MI task as being inversely proportional to the mean absolute deviation of the covariance matrices from that task:

$$classStab(A) = \frac{1}{1 + \sigma_{C^A}} \tag{9}$$

As such, the lower the mean absolute deviation of the EEG patterns, as represented by the spatial covariance matrix, the higher the stability.

3.3.3 Considering more than the spatial EEG patterns

The metrics above only consider the spatial covariance matrices, i.e., the spatial EEG features, in a given frequency band (the band in which the EEG signals were band-pass filtered). While this is the most common way to represent EEG patterns, it could be interesting to study the EEG patterns more finely. In particular, if we assume that MI-BCI skills or learning could differ in different EEG spectral features, it would seem relevant to have metrics that measure users' MI-BCI skills in both spatial and spectral EEG patterns.

This could be achieved for instance by using large covariance matrices containing both the spatial (EEG channels) and spectral (EEG frequency bands) information as rows/columns. However, such large matrices may lead to numerical instabilities and statistical estimation issues (see, e.g., [51]). Thus, to avoid such problems, we propose the following simple metrics which have the same matrix dimensionality as with the purely spatial EEG patterns.

First we can estimate the covariance matrices $C_{i,F_j} = \frac{1}{N_s} X_i^{F_j} (X_i^{F_j})^T$ for various frequency bands, where $X_i^{F_j}$ is the matrix of EEG signals from trial i, filtered in frequency band F_j. We can then compute the distinctiveness $classDist_{F_j}(A, B)$ using covariance matrices C_{i,F_j} in equations 6 or 7 (for binary or multiclass problems respectively), for each frequency band F_j. Then we can compute the overall spatio-spectral distinctiveness as:

$$spaspecClassDis(A, B) = \sum_{j=1}^{N_f} classDist_{F_j}(A, B) \tag{10}$$

where N_f is the number of frequency bands considered. In a similar way, we can define the spatio-spectral class stability and rest distinctiveness by summing the values obtained by each of these metrics over different frequency bands F_j. This thus give us spaspecClassStab and spaspecRestDis, see also Table 2.

3.4 Summary of the new metrics

Most of the metrics presented above are intuitive and computationally efficient metrics to quantify some aspects of users' skills at MI-BCI control. They are also training data and classifier independent, as well as robust to some non-stationarities given the affine invariance of δ_R. All these metrics are summarized in the Tables below. In particular, the core, basic metrics *classDist*, *restDist* and *classStab* are presented in Table 1, while their spatio-spectral variants are presented in Table 2.

Table 1: Basic distinctiveness and stability metrics

	Basic metric
Class distinctiveness (2 classes)	$classDis(A, B) = \frac{\delta_R(\bar{C^A}, \bar{C^B})}{\frac{1}{2}(\sigma_{C^A} + \sigma_{C^B})}$
Class distinctiveness (multiclass)	$classDis(\{A_i\}) = \frac{\sum_{i=1}^{Nc} \delta_R(\bar{C^{A_i}}, \bar{\bar{C^A}})}{\sum_{i=1}^{Nc} \sigma_{C^{A_i}}}$
Rest distinctiveness	$restDis(A) = \frac{\delta_R(\bar{C^A}, \bar{C^{rest}})}{\frac{1}{2}(\sigma_{C^A} + \sigma_{C^{rest}})}$
Stability	$classStab(A) = \frac{1}{1+\sigma_{C^A}}$

Regarding the distinctiveness metrics, it is important to realize that they do not depend nor use any classifier, and as such that what they measure may be very different from CA. Indeed, an unadapted classifier, i.e., a classifier trained on outdated data, may lead to poor CA, while an adapted one may lead to high CA, if class distinctiveness is high and if the classifier can exploit these EEG patterns that are highly distinct from each other. Class distinctiveness is thus a measure of MI-BCI skills that consider the potential BCI control that the user can get, provided that the BCI classifier is adapted to this user, i.e., uses the EEG patterns produced with MI that are distinct from each other. In the following, we compare the various metrics offline with CA and RWCV CA.

Table 2: Spatio-spectral variants of the core metrics, considering N_f frequency bands F_j

	spatio-spectral metric
spatio-spectral class distinctiveness	$spaspecClassDis(A,B) = \sum_{j=1}^{N_f} classDis_{F_j}(A,B)$
spatio-spectral rest distinctiveness	$spaspecRestDis(A) = \sum_{j=1}^{N_f} restDis_{F_j}(A)$
spatio-spectral stability	$spaspecClassStab(A) = \sum_{j=1}^{N_f} classStab_{F_j}(A)$

3.5 Data set and evaluation

To compare the performance metrics, we used two MI-BCI data sets, from our previous experiments. More specifically, we used 1) a 2-class motor imagery data set, in which N=20 users trained for a single session (i.e., one day) to use the BCI [17], and 2) a 3-class mental imagery BCI data set, in which 17 users trained for 6 sessions (i.e., 6 days) to use the BCI [20]. These two data sets enable us to study the various metrics with both short term (single session) and longer term (6 sessions) user training, as well as for binary and multiclass problems. They also enable us to see how the metrics behave for motor imagery as well as non-motor imagery tasks. It should be noted here that the BCI users in these two experiments were trained to control the BCI using as feedback a classical bar feedback, i.e., a bar extending towards the recognized class according to the classifier output. More precisely, the feedback bar had a length and a direction updated 16 times per second, according to the distance of the features, extracted from the last second of EEG data, to the separating hyperplane of a Linear Discriminant Analysis classifier (see [17] and [20] for details). Thus, users were trained to control the BCI using a feedback somehow related to the online classification accuracy. The two data sets are described here after.

3.5.1 Data set 1: Single session, 2-class motor imagery data set

This data set comprises the EEG signals of 20 BCI-naive participants, who had to learn to do 2 MI-tasks, namely imagining left- and right-hand movements. Participants first had to complete a calibration run, without feedback, followed by 4 feedback runs. Each run was composed of 20 trials for each of the two MI tasks (displayed in a random order). At the beginning of each trial, a fixation cross was displayed. Then, after 2s, a beep sound occurred. Then, at t = 3s, the instruction appeared as an arrow the direction of which indicates the MI task to be performed, i.e., an arrow pointing left indicated a left hand MI and an arrow pointing right a right hand MI. From t = 3.250s, a feedback was provided for 4s in the shape of a bar the direction of which indicating the mental task that had been recognized and the length of which representing the classifier output. More details about this data set can be found in [17].

3.5.2 Data set 2: Multi-sessions, 3-class mental imagery data set

Seventeen BCI-naive participants took part in this study, for 6 different sessions each (each on a different day). The three mental imagery-tasks that participants had to learn to perform were 1) left-hand motor imagery, 2) mental rotation of a 3D geometric figure and 3) mental subtraction of a 2 digit number from a 3-digit number (both displayed on screen). Each session comprised 5 runs. During each run, participants had to perform 45 trials (15 trials per task), each trial lasting 8s. At t=0s, an arrow was displayed with a left hand pictogram on its left (left hand MI task), the subtraction to be performed on top (mental subtraction task) and a 3D shape on its right (mental rotation task). At t=2s, a "beep" announced the coming instruction and one second later, at t=3s, a red arrow was displayed for 1.250s. The direction of the arrow informed the participant which task to perform, e.g., an arrow pointing to the left meant the user had to perform a left hand MI task. Finally, at t=4.250s, for 4s, a visual feedback was provided in the shape of a blue bar, the length of which varied according to the classifier output. Only positive feedback was displayed, i.e., the feedback was provided only when there was a match between the instruction and the recognized task. More details about this data set can be found in [20].

3.5.3 The impact of spatial abilities on BCI performances

Interestingly enough, both data sets come from experiments which aimed, inter alia, to identify cognitive and personality profiles that correlated to BCI performances [17, 20]. The typical online classification accuracy was used as a measure of BCI performance in these works. In both studies, we observed a strong and significant correlation between classification accuracy and user's spatial abilities, the latter being measured using a mental rotation test [46], see [17, 20] for details. We will thus also study in the present work whether a correlation between spatial abilities and our new metrics of MI-BCI skills can be observed as well.

3.5.4 EEG signal recording and processing

For both data sets, EEG signals were recorded using 30 active scalp electrodes (F3, Fz, F4, FT7,FC5, FC3, FCz, FC4, FC6, FT8, C5, C3, C1, Cz, C2, C4, C6, CP3, CPz, CP4, P5, P3, P1, Pz, P2, P4, P6, PO7, PO8, 10-20 system) using a g.USBAmp (g.tec, Austria). EEG data were sampled at 256 Hz.

For both data sets, EEG data was band-pass filtered in 8-30 Hz. For each trial, the MI EEG segment used was the 2s long segment starting 0.5s after the cue (left, right or up arrow), i.e., from second 3.5 to 5.5 of each trial. For the rest EEG signals, we used the 2s long segment immediately before the cue, i.e., from second 1 to 3 of each trial.

Regarding the CA and RWCV CA metrics, for both data sets we used Common Spatial Patterns (CSP) spatial filters [37], the spatially filtered signals band power as features, and a Linear Discriminant Analysis (LDA) [25] as classifier. More precisely, for the 2-class data set, we used 3 pairs of Common Spatial Pattern (CSP) spatial filters and a LDA classifier, as in [17]. For the multi-class data set, we used 3 sets of CSP filters, each optimized to discriminate EEG signals for a given class from those for the

other two classes. We optimized 2 pairs of spatial filters for each class, thus leading to 12 CSP filters. The resulting 12 band power features were classified using a multi-class LDA, built by combining three LDA in a one-versus-the-rest scheme, as in [20].

For the standard (here simulated online) CA, we trained the CSP and LDA classifiers on the EEG data from the calibration run and used it to classify the EEG data from the 4 subsequent runs, as done online in [17, 20]. For the multi-session data set (data set 2), to account for some of the between session-variability, the LDA classifiers' biases were re-calculated after the first run of sessions 2 to 6, based on the data from this first run, as done in [13], and as done during the online experiments that resulted in this data set [20].

For the spatio-spectral metrics (*spaspecClassDis*, *spaspecRestDis*, *spaspecClassStab*) we used the frequency bands 8-10Hz, 10-12Hz, 12-18Hz and 18-30Hz, which corresponds to low-alpha, high-alpha, low-beta and high-beta frequency bands.

Regarding the RWCV CA, we used 4-fold CV on each run. For *classDis*, *restDis*, *classStab*, *spaspecClassDis*, *spaspecRestDis*, and *spaspecClassStab* the trial covariance matrices were estimated using automatic shrinkage with the algorithm proposed in [24].

4 Results

We first present the results of the various metrics averaged over all subjects of each data set, to provide an overview of what they measure. Then, for each data set we focus on some specific subjects, to reveal more specific behaviors of each metric. Note that when presenting the different metrics on the same Figure, we used Z-scores as these metrics use different units (percentage for CA and RWCV CA, arbitrary distance units for *classDis*, *restDis* and *classStab*).

4.1 Average results

4.1.1 Data set 1: Single session, 2-class motor imagery data set

Figure 1 shows the average measures of distinctiveness between classes (MI tasks), i.e., CA, RWCV CA, *classDis* and *spaspecClassDis*, for each run of data set 1 (2-class motor imagery). CA displays some oscillations in performance, but no global increase in performance over runs. On the other hand, RWCV CA, *classDis* and *spaspecClassDis* reveal a clear continuous increase in distinctiveness between classes over runs. The 2-way ANOVA *Metric*Run* (*Metric*: CA, RWCV CA, *classDis*, *spaspecClassDis* - transformed to z-score to enable comparisons; *Run*: 2 to 5) for repeated measures showed a trend towards a metric*run interaction [$F(1,19) = 3.081$; $p = 0.095$; $\eta^2 = 0.140$], see also Figure 2.

Figures 3a and 3b show the class-specific performance metrics, i.e., class-wise CA, class-wise RWCA, *restDis* and *spaspecRestDis*. Here as well, CA does not show any obvious increase in performance over runs, while both RWCV CA, *restDis* and *spaspecRestDis* show some performance increase over runs, notably for class 2 (right hand motor imagery). The 3-way ANOVA *Metric*Class*Run* (*Metric*: CWCA, CW

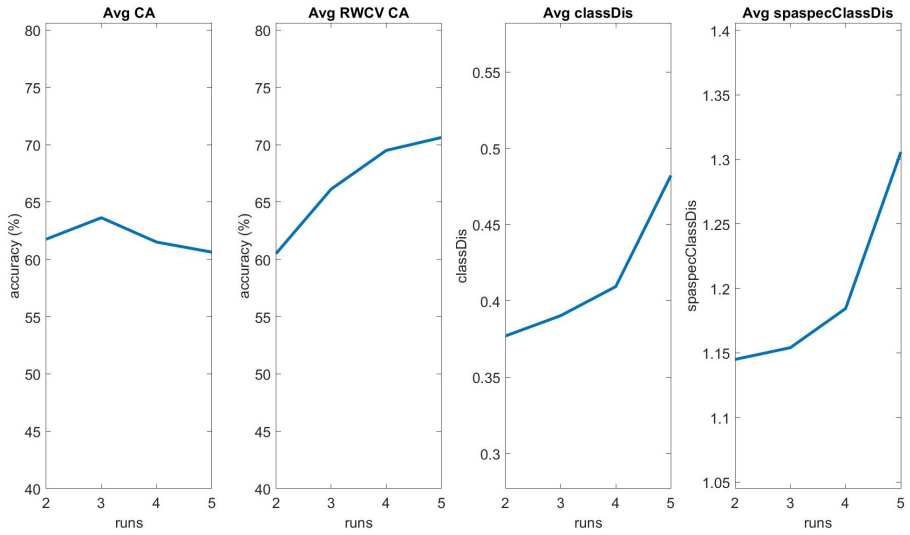


Figure 1: The average measures of distinctiveness between classes, across runs, for data set 1.

RWCVCA, *restDis* and *spaspecRestDis* (z-score); *Class*: left- vs. right-hand MI; *Run*: 2 to 5) for repeated measures did not show any significant effect though.

Concerning the stability metrics (*classStab* and *spaspecClassStab*, see Figures 4a, 4b and 4c), the 3-way ANOVA *Metric*Class*Run* (*Metric*: *classStab* and *spaspecClassStab*; *Class*: left-hand, right-hand MI and rest; *Run*: 2 to 5) revealed a significant metric*run interaction [$F(1,19) = 4.579$; $p < 0.05$; $\eta^2 = 0.194$] as well as a significant metric*class*run interaction [$F(1,19) = 4.412$; $p < 0.05$; $\eta^2 = 0.188$]. This interaction indicates that *classStab* and *spaspecClassStab* may not measure the same patterns of performance variation over runs, although on this data set, they increase and decrease at the same times.

4.1.2 Data set 2: Multi-sessions, 3-class mental imagery data set

As further described here-after, it is interesting to observe that, on this data set, spatio-spectral metrics revealed different performance variation dynamics than those revealed by the purely spatial metrics.

In particular, regarding the measures of distinctiveness between classes (MI tasks), i.e., CA, RWCV CA, *classDis* and *spaspecClassDis* (See Figure 5), *spaspecClassDis* revealed a continuous increase in performances over all sessions except the last one, whereas *classDis* rather revealed an increase of performance over the first 3 runs followed by a decrease in performance for the subsequent runs. In contrast, CA showed a continuous decrease of performance over sessions. The 2-way ANOVA *Metric*Session* (*Metric*: CA, RWCV CA, *classDis*, *spaspecClassDis* (z-scores); *Session*: 1 to 6) for repeated measures showed a significant metric*run interaction [$F(1,17) = 4.502$;

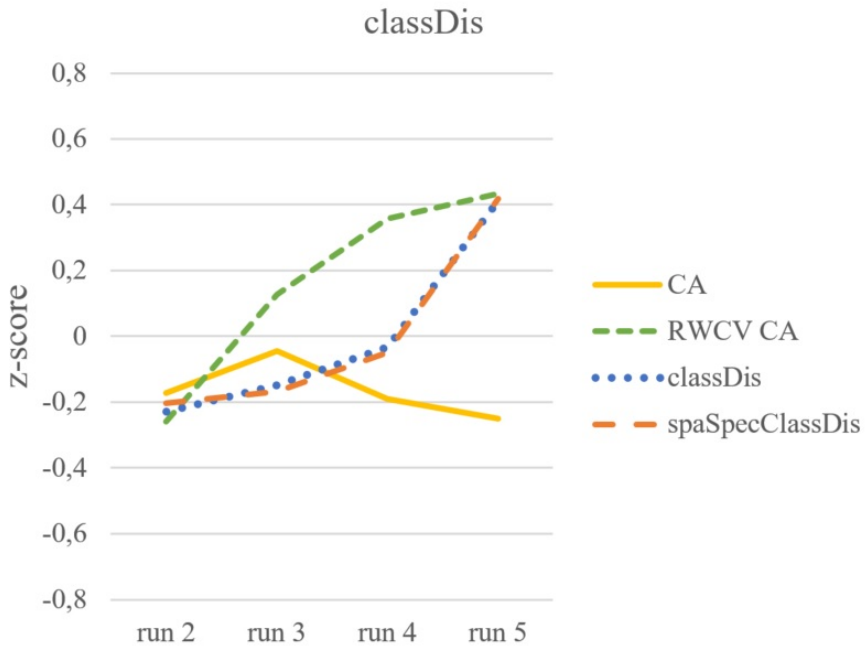


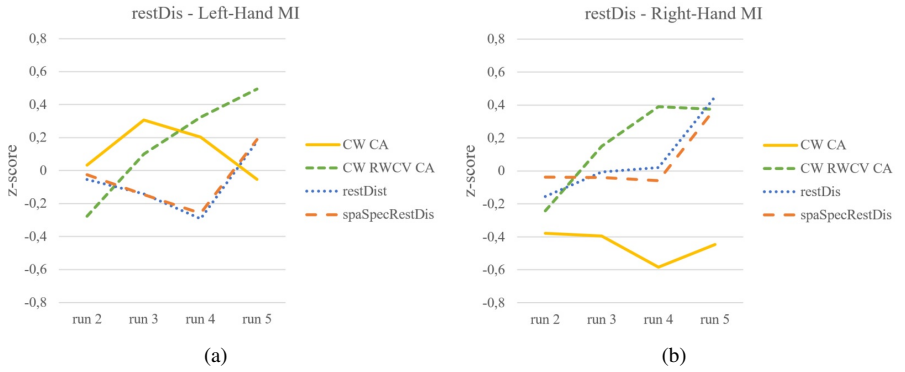
Figure 2: Z-scores for each of the metrics of distinctiveness between classes (CA, RWCV CA, *classDis* and *spaspecClassDis*) as a function of the run for the 2-classes data set (data set 1). This graph reveals improvements across runs for all the metrics but CA.

$p < 0.05$; $\eta^2 = 0.220$], see also Figure 6.

Regarding class-specific distinctiveness measures, i.e., class-wise CA, class-wise RWCV CA, *restDist* and *spaspecRestDist* (see Figures 7a, 7b and 7c), the 3-way ANOVA *Metric*Class*Session* (*Metric*: CW CA, CW RWCV CA, *restDis* and *spaspecRestDis* (z-score); *Class*: left-hand MI, mental subtraction and mental rotation; *Session*: 1 to 6) for repeated measures showed a significant metric*class interaction [$F(1,17) = 12.474$; $p < 0.01$; $\eta^2 = 0.438$].

Finally, regarding class stability (see Figures 8a, 8b, 8c and 8d), *spaspecClassStab* revealed a continuous positive increase in stability over the first 4 sessions, then a decrease, whereas *classStab* revealed a continuous decrease of stability over sessions. The 3-way ANOVA *Metric*Class*Session* (*Metric*: *classStab* and *spaspecClassStab*; *Class*: left-hand MI, mental subtraction, mental rotation and rest; *Session*: 1 to 6) revealed a significant metric*class interaction [$F(1,17) = 25.675$; $p < 0.001$; $\eta^2 = 0.616$] as well as a significant metric*session interaction [$F(1,17) = 24.692$; $p < 0.001$; $\eta^2 = 0.607$]. The latter indicates that *classStab* decreases along the 6 sessions while *spaspecClassStab* first increases before decreasing (inverted U-shaped curve).

While it was not the case with the previous data set, here significant differences are



(a) (b)

Figure 3: Z-scores for each of the metrics of class-specific distinctiveness (CA, RWCV CA, *restDis* and *spaspecRestDis*) as a function of the run and of the class (a: left-hand MI (class 1), b: right-hand MI (class 2)) for the 2-classes data set (data set 1).

revealed between purely spatial and spatio-spectral metrics. Thus, which metric should we believe and how should we interpret such results? To find out what makes the spatio-spectral metrics differ from the spatial ones, we computed the average *classDis* and *classStab* for each frequency band separately, and averaged them across subjects, within each band. The results are represented in Figure 9 for *classDis* and Figure 10 for *classStab*.

As can be seen on Figure 9, different performance variations over sessions seem to occur, on average, in the different bands. In particular, with *classDis*, there is a continuous positive increase in performance, which may be due to learning, that can be observed across sessions around the alpha band (8-10Hz, 10-12 Hz) and low beta one (12-18Hz), whereas in the high beta band (18-30Hz), after a sharp performance increase in the first session, performances tended to decrease over the subsequent sessions This could explain the inverted U-shape performance obtained with spatial *classDis*, with performances first increasing over the first 3 sessions, and then decreasing, which could be the result of a simultaneous increase in performance in alpha and decrease in performance in high beta. Similarly, when looking at *classStab*, we can see a positive increase in stability over sessions in alpha, no obvious change in low beta, and a decrease in performances in high-beta.

Altogether these results seem worth considering: they suggest that even when using the classifier output in a broad EEG band (here 8-30Hz) as feedback, users can increase their performance in a more specific band only, which was mostly in alpha for this protocol and subjects. This then stresses the need to monitor the progresses in each band, in order to 1) be able to observe such performance increase that may reflect learning, if any and 2) possibly restrain the feedback/training tasks or the EEG features to focus on that frequency band in which there may seem to be a more efficient learning. The metrics we propose in this paper enable us to do exactly that kind of monitoring.

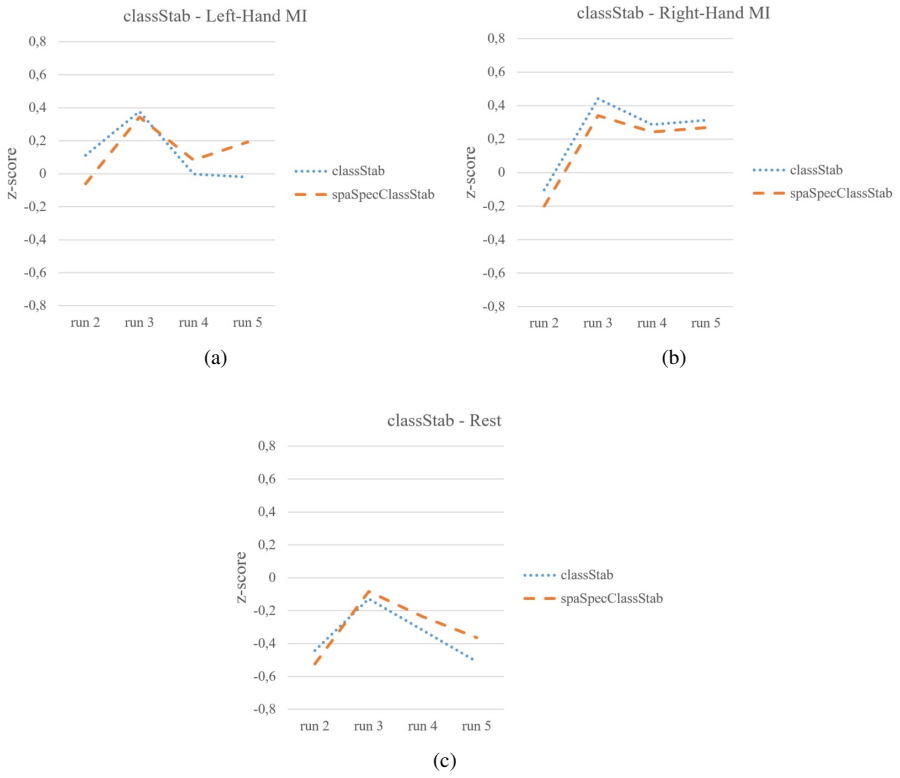


Figure 4: Z-scores for each of the metrics (*classStab* and *spaspecClassStab*) as a function of the run and of the class (a: left-hand MI, b: right-hand MI, c: rest) for the 2-classes data set (data set 1).

4.2 Some subject specific results

As stated earlier, we observed a high inter-subject variability, therefore it is interesting to further investigate the different patterns observed in terms of metrics' evolution across the runs, for individual subjects. It will enable the analysis of the behavior of the different metrics and provide insights on their pros and cons.

4.2.1 Data set 1: Single session, 2-class motor imagery data set

In this data set, for instance, all the distinctiveness measures for subject S4 could reveal a clear performance improvement over time, possibly due to learning. However, the same metrics for subject S5 did not show any performance improvement over time with the online CA, whereas both RWCV CA, *classDis* and *spaspecClassDis* revealed a clear performance increase over runs, which might me due to some form of learning (see Fig. 11). Metrics for subject S9 (Fig. 12) revealed another interesting phe-

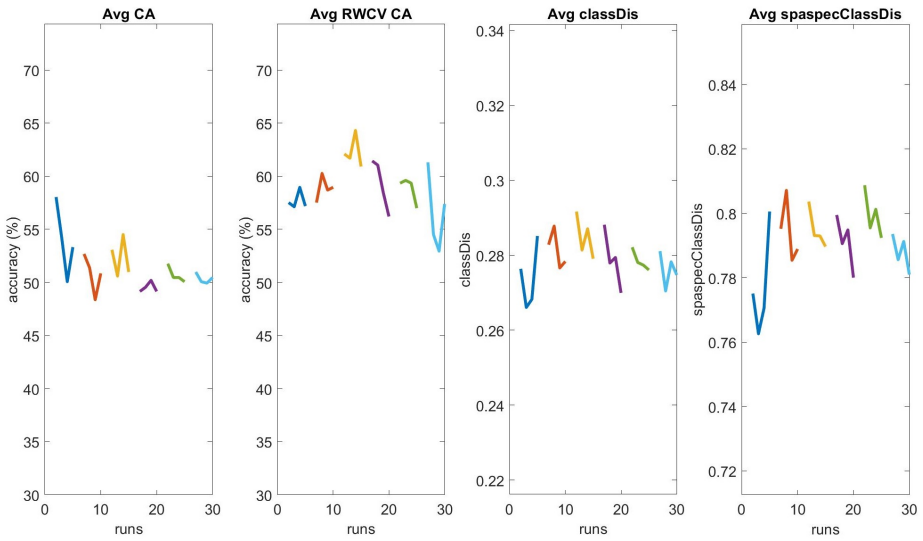


Figure 5: The average measures of distinctiveness between classes, across runs and sessions, for data set 2. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

nomenon. While both CA and RWCV CA did not show any performance increase, *classDis* and *spaspecClassDis* did. However, both *restDis* and *spaspecRestDis* revealed that class 1 (left hand motor imagery) actually became increasingly more similar to rest EEG over the runs (*restDis* and *spaspecRestDis* for class 1 sharply decreased from run 2), and thus that the increased *classDis* was probably due to the BCI discriminating rest vs right-hand MI rather than left- vs right-hand MI. CA cannot identify such a phenomenon since it ignores rest EEG.

Finally, analyses of Subject S19's data (Fig. 13) showed decreasing class discriminability with CA, RWCV CA, *classDis* and *spaspecClassDis*. However, the data revealed some continuous performance increase over runs with both *restDis* and *spaspecRestDis*, for both classes. One possible hypothesis to explain this phenomenon could be that this subject learned to modulate their EEG signals so that they differ from rest EEG, but may have more trouble generating consistently distinct patterns between both the MI tasks. Such a phenomenon has also been observed with simultaneous EEG-fMRI in [52], in which some subjects showed modulations of brain activity during MI with respect to rest signals, but no lateralization of the patterns. The *restDis* and/or and *spaspecRestDis* metrics could thus be a cheap and easy way to identify this phenomenon in EEG.

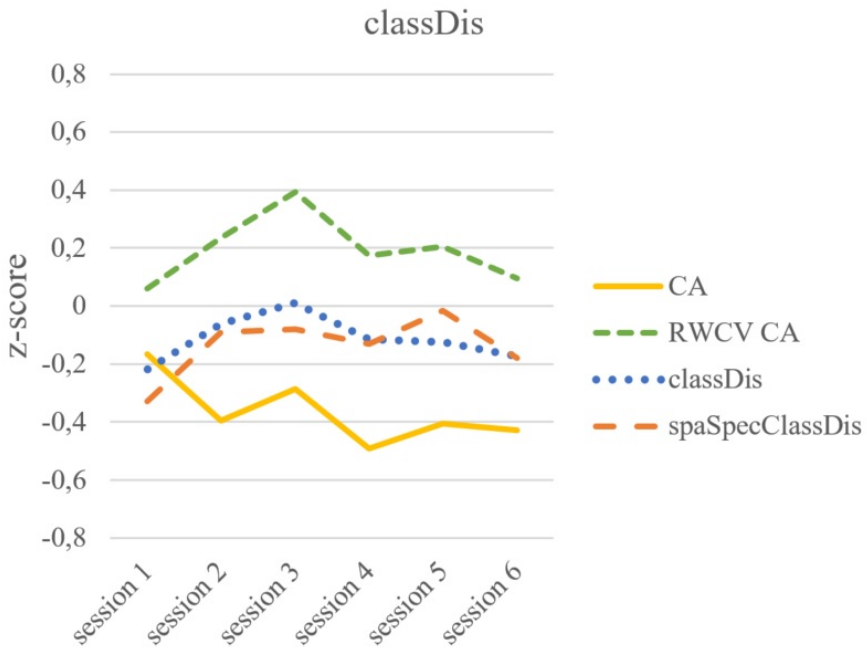


Figure 6: Z-scores for each of the metrics of distinctiveness between classes (CA, RWCV CA, *classDis* and *spaspecClassDis*) as a function of the run for the 3-classes data set (data set 2).

4.2.2 Data set 2: Multi-sessions, 3-class mental imagery data set

Regarding the multi-sessions data set, relevant individual results include for instance those of subject s11. With this subject, all metrics of class distinctiveness measured a continuous increase in performance over sessions, see Figure 14.

However, the same metrics did not measure similar performance dynamics on the data of subject s9, see Figure 15. On this subject, the CA metric suggested that the distinctiveness between classes actually decreased continuously across runs and sessions, whereas both RWCV-CA, *classDis* and *spaspecClassDis* actually measured the exact opposite: a continuous increase in performances across runs and sessions. Note here that, again, CA does not actually measure how distinct the EEG patterns from each class are, but how well the classifier trained on the data from the calibration run can distinguish these patterns. Contrary to CA, both RWCV-CA, *classDis* and *spaspecClassDis* are training data independent, and actually measure how distinct the class EEG patterns from each run are. As such, they can reveal what could be interpreted here as a strong learning effect (see the Discussion in Section 5 for more details about this).

A last interesting case is that of subject s3, presented in Figure 16. With this sub-

Figure 7: Z-scores for each of the metrics of class-specific distinctiveness (CA, RWCV CA, *restDis* and *spaspecRestDis*) as a function of the run and of the class (a: left-hand MI, b: mental subtraction, c: mental rotation) for the 3-classes data set.

ject, the distinctiveness between classes actually did not really show a specific performance increase over sessions. However, looking at the class-specific distinctiveness does reveal a gradual performance increase, that may suggest some learning: the EEG patterns from each class actually gradually became more distinct from rest EEG patterns over runs and sessions, as measured by *restDis* and *spaspecRestDis*. Only *restDis* and *spaspecRestDis* can reveal such performance increase since only them actually consider rest EEG data, which other metrics based on classification accuracy usually ignore.

4.3 Do the new metrics confirm the influence of spatial abilities on MI-BCI performances?

As indicated earlier (see Section 3.5.3), for both data set 1 and data set 2, the original online experiments revealed a significant correlation between online CA (average peak

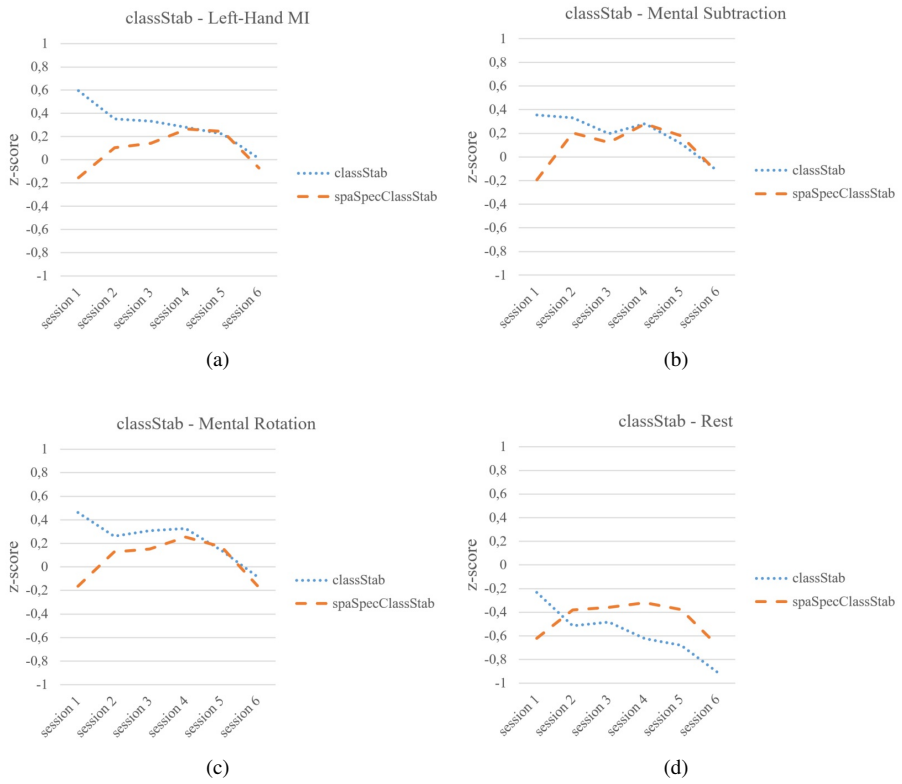


Figure 8: Z-scores for each of the stability metrics (*classStab* and *spaspecClassStab*) as a function of the run and of the class (a: left-hand MI, b: mental subtraction, c: mental rotation, d: rest) for the 3-classes data set.

CA over the runs for data set 1 and average mean CA over the sessions for data set 2) and spatial abilities [17, 20]. Participants' spatial abilities were assessed from the mental rotation score they obtained after completing the Mental Rotation Test of Vandenberg [46]. In the present paper, we thus also computed these correlations with our new metrics. We performed partial correlations in order to control for the gender effect associated with mental rotation scores [46]. Regarding data set 1, a significant correlation was revealed between mental rotation scores and average RWCV CA [r = 0.500, $p < 0.05$], but not with the other metrics: CA [r = 0.373, $p = 0.116$], *classDis* [r = 0.125, $p = 0.609$], *spaSpecClassDis* [r = 0.069, $p = 0.780$]. On the other hand, regarding data set 2, significant correlations were revealed between mental rotation scores and all the metrics averaged over sessions: CA [r = 0.519, $p < 0.05$], RWCV CA [r = 0.535, $p < 0.05$], *classDis* [r = 0.609, $p < 0.05$] and *spaSpecClassDis* [r = 0.557, $p < 0.05$].

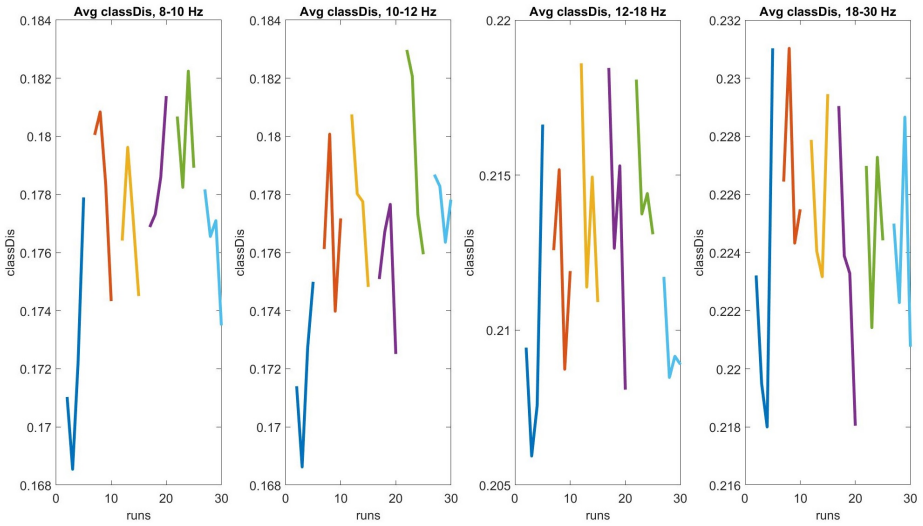


Figure 9: grand average spatial *classDis* in various frequency bands, over sessions and runs, for data set 2. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

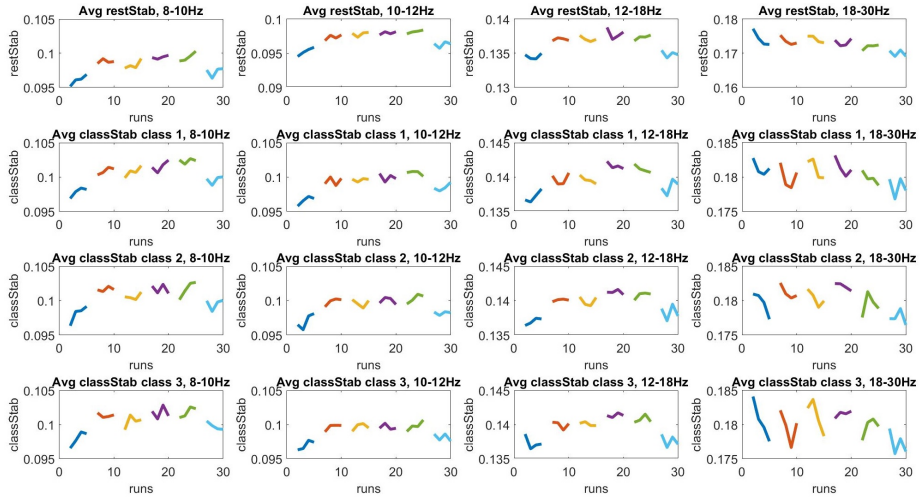


Figure 10: grand average spatial *classStab* in various frequency bands, over sessions and runs, for data set 2. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

5 Discussion

Globally, average results showed either a trend towards significance (data set 1) or a significant (data set 2) metric*run interaction. This suggested that some metrics (here

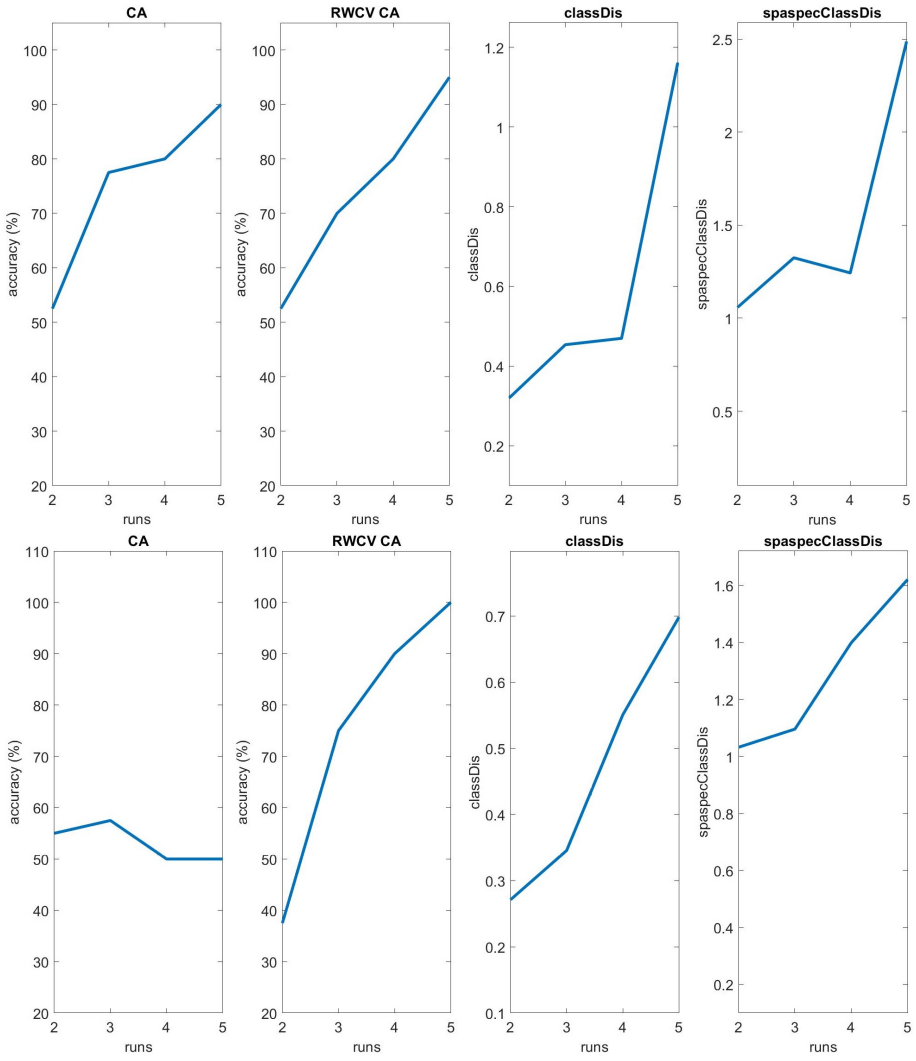


Figure 11: Examples of 2 subjects for which, either CA measured a learning effect like the other metrics (top - subject S5), or did not whereas the other metrics did (bottom - subject S4)

RWCV CA, *classDis* and *spaspecClassDis*) revealed continuous increase in performance while another (CA) did not. Such continuous increase in performance may arguably be due to learning, with subjects MI-BCI skills gradually improving with practice. In the BCI literature, studies usually consider that learning occurred when CA gradually increased over runs and/or sessions [23, 33, 19]. It is worth noting, however, that a continuous increase of performance over time, whatever the performance

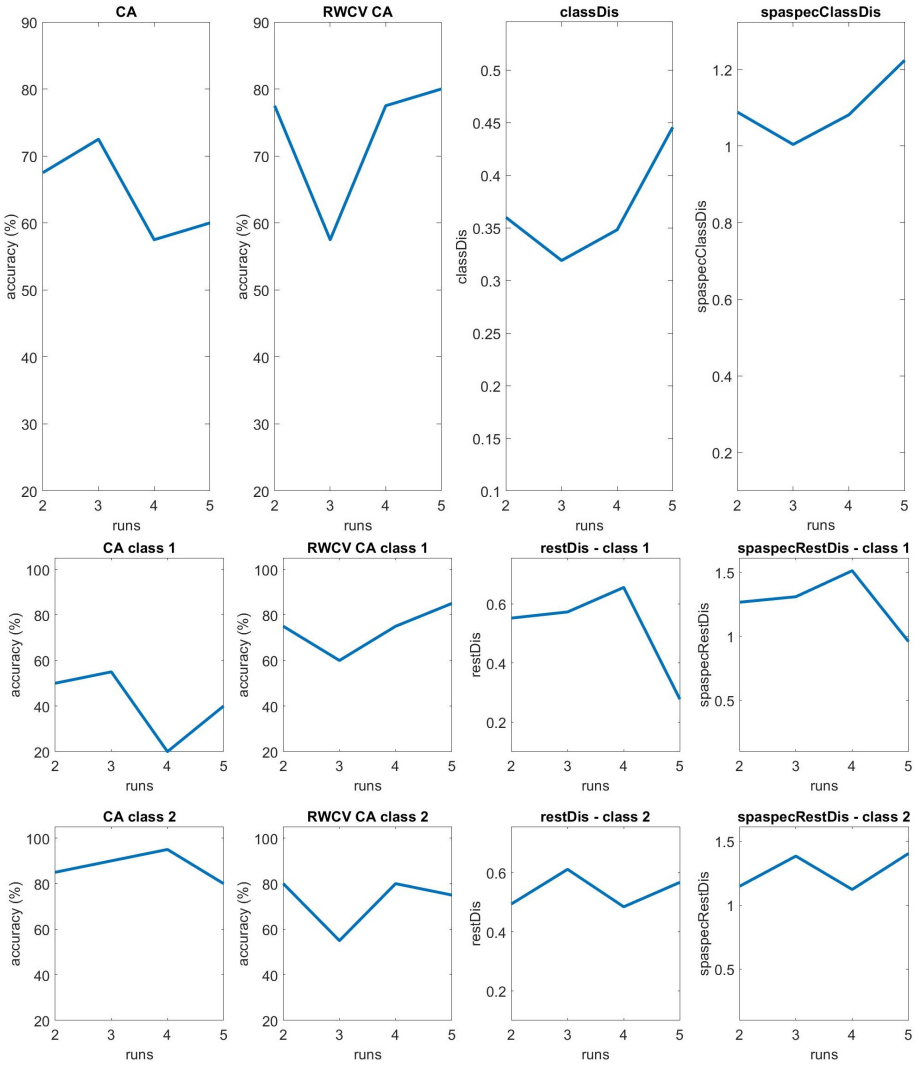


Figure 12: Subject S9, for which class 1 (left hand motor imagery) became like rest

metrics (i.e., both CA, RWCV CA and *classDis*), might also be due to some alternative factors other than learning. For instance, electrode impedance might decrease over time due to sweat or due to the gel moving and making better contact with the scalp. This would in turn increase the EEG signal-to-noise ratio and thus possibly increase CA or *classDis*, independently of any learning from the subject. However, we would like to stress that on data set 2, which was based on six sessions, over six different days, we also observed increase in class distinctiveness over sessions (i.e., over days), both on several individual subjects and on average over all subjects. Observing a continu-

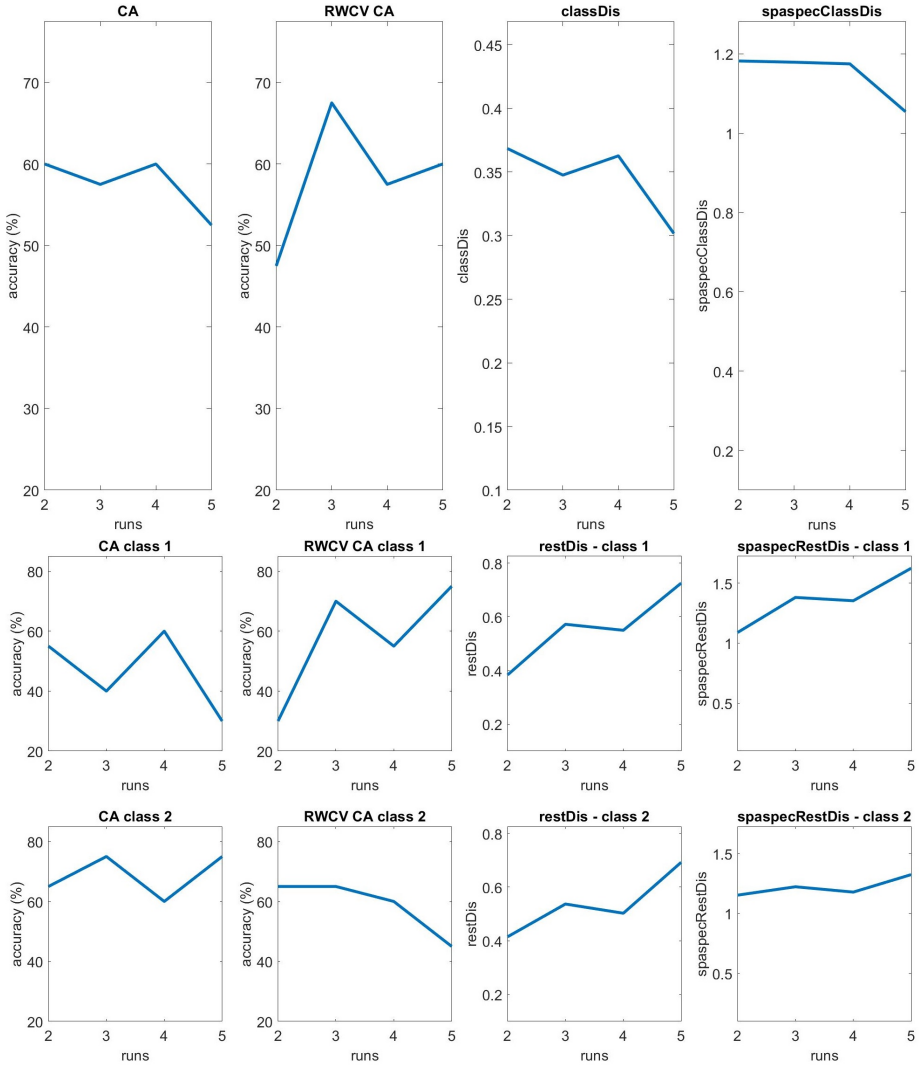


Figure 13: Subject S19 produced EEG patterns increasingly more different than rest, but not distinct from each other.

ous reduction of impedance over several separate days would seem extremely unlikely. Therefore, while we cannot rule out for sure alternative explanations and prove this is learning, learning remains the most likely explanation for a continuous increase of a performance metric (either *classDis* or *restDis*) over days. Thus, in the following discussion, we consider that such continuous increase in performance reflect learning. This is indeed the most likely hypothesis, but we keep in mind that this is not the only one.

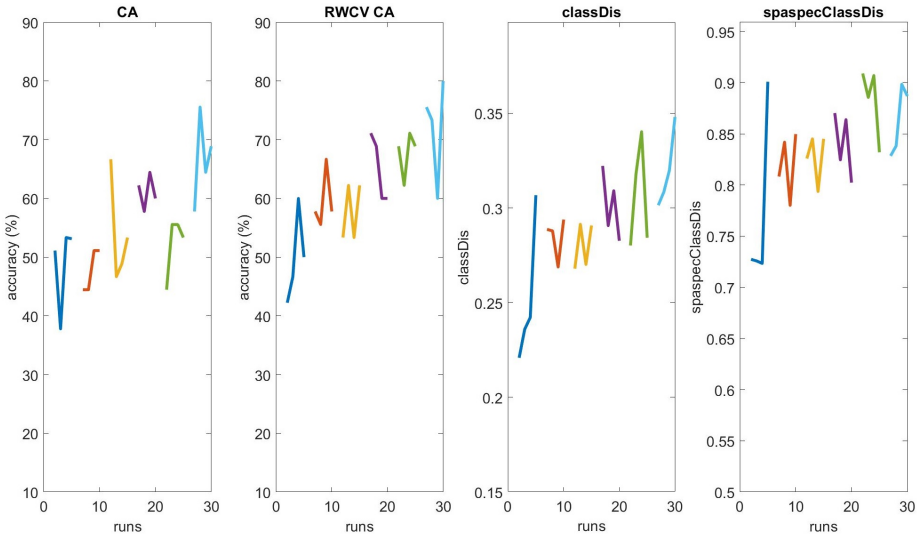


Figure 14: Subject s11, data set 2, for which all 3 metrics (CA, RWCV-CA, classDis) all measured a continuous increase in performance across sessions. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

Overall, RWCV CA, *classDis* and *spaspecClassDis* seemed to reveal some form of user learning that CA seemed to have missed. This is all the more interesting given the fact that the feedback was based on the CA metric. Indeed, participants were asked to make the bar feedback, that is proportional to the classifier output and thus related to the online CA, as long as possible in the correct direction. Despite such feedback being based on a possibly incomplete metric (as it may miss some learning), most of the participants demonstrated to be able to modulate their EEG patterns increasingly better, sometimes leading to the improvement of the performance metrics. This result is promising for the future as it suggests that with a better feedback, e.g., a feedback directly related to our new metrics, the ability of the participants to learn to modulate efficiently their EEG patterns, in order to improve their BCI control, may be enhanced.

This result also raises the question of whether different aspects of "self-regulation learning" might have been involved. Indeed, two aspects may have taken place here. The first aspect would be voluntary learning, when users learned from BCI feedback how to self-regulate their EEG patterns by identifying the best mental strategies. This learning aspect thus involved cognitive processes underlain by specific neurophysiological activities, and would be the most typical aspect of learning that is targeted by feedback training in BCI. An additional aspect of learning may have been involved here though. This aspect of learning might have occurred in parallel, due to habituation or repeated practice of the mental imagery tasks, which may have induced neuronal plasticity. Indeed, the feedback provided to users was not directly related to our new met-

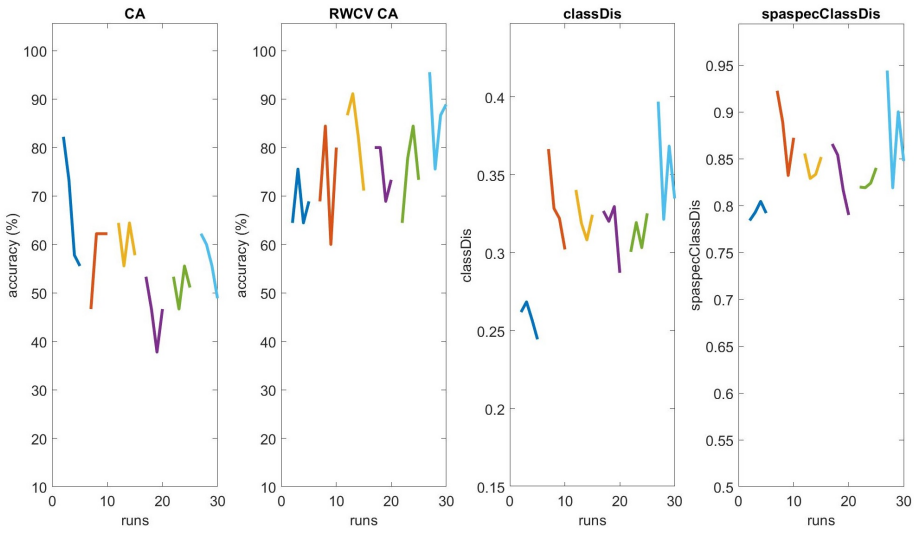


Figure 15: Subject s9, data set 2, for which CA suggested a continuous decrease in distinctiveness between classes whereas both RWCV-CA, *classDis* and *spaspecClassDis* - being training data independent - actually measured a continuous increase in distinctiveness across sessions. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

rics of performance, yet users managed to improve such metrics over time. This might suggest a form of learning that occurred without a dedicated feedback. In the future, it would be interesting to study to which extent each of these two aspects of learning is involved. Note that our new metrics are nonetheless not completely orthogonal to the online classifier output. As such, the online feedback might still carry some useful information that users might have used to increase their performances as measured by our metrics.

On the other hand, these results also suggested that different performance metrics can reveal different aspects of MI-BCI users' skills and learning. Notably, they first showed that CA may sometimes be unable to measure that users can modulate their EEG patterns using MI increasingly well, whereas metrics such as RWCV CA, *classDis* and *spaspecClassDis* can reveal performance increase over runs and sessions, most likely related to user learning. They even revealed fast performance increase, and thus possibly fast learning effects, in several subjects, with continuous progress over runs, over a single day of training in data set 1. This can have profound implications for the study of BCI user training. For instance, the present results may explain why in [23], it has been concluded that most BCI studies - and notably those based on machine learning - do not actually involve human learning (defined as continuous CA increase). Indeed, in most of the studies surveyed in [23], CA was used as the performance metric. As such, human learning might have occurred, but CA might not have

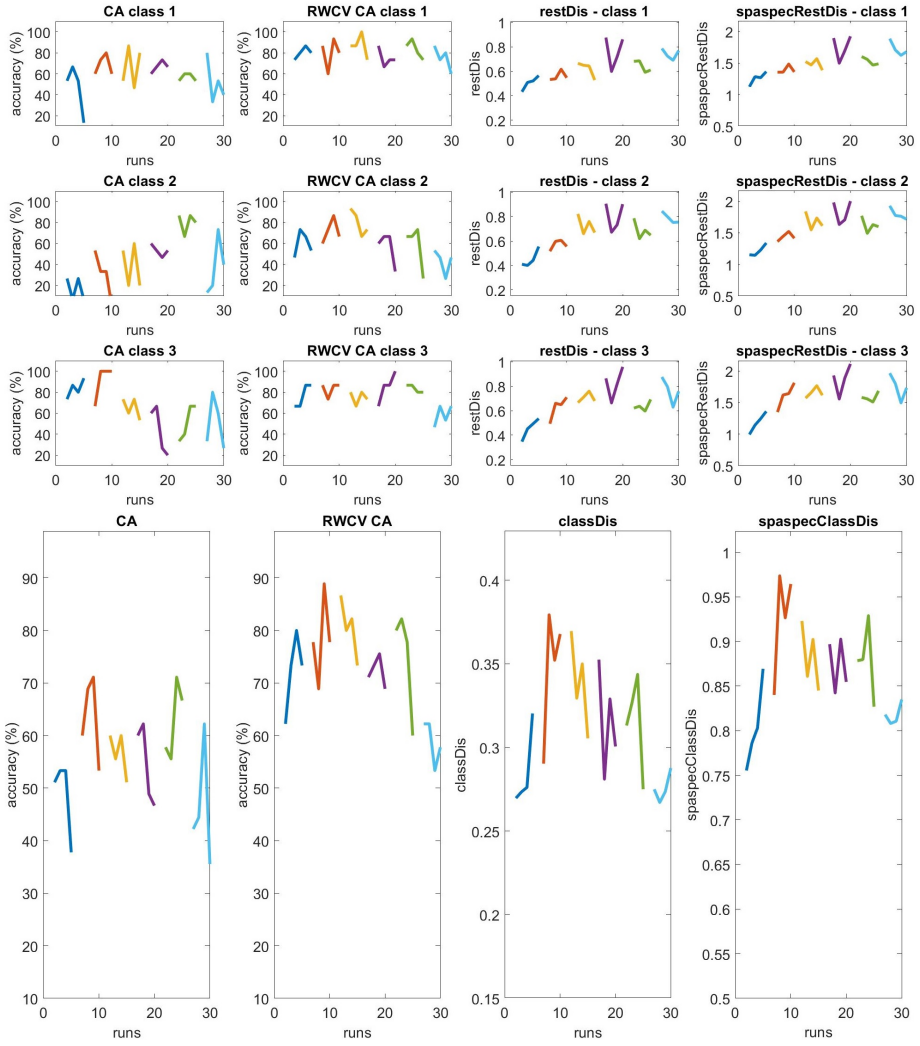


Figure 16: Subject s3, data set 2, for which there was no increase in distinctiveness between classes (as seen by all metrics - bottom figure), but for which the distinctiveness with the rest EEG patterns (*restDis* and *spaspecRestDis* - top Figure) increased across sessions, which only *restDis* and/or *spaspecRestDis* could measure. Different sessions (i.e., days) are displayed in different colors, and the consecutive runs from the same session are linked to each other.

been able to measure it. It thus seems necessary to re-analyze EEG data from previous studies with complementary performance metrics such as the ones proposed here, to obtain complementary measures that may reflect whether human learning could have

occurred.

The performance increase over runs and sessions, and thus possible learning effects, that were revealed by the new metrics also stress the need for co-adaptive BCI systems, and explain the success of these approaches, see, e.g., [8, 48]. On a related note, it would be relevant to compare the new metrics proposed in this paper with online classification accuracies obtained with adaptive classifiers. Indeed, adaptive classifiers being continuously updated, they are much less sensitive to the initial training data. The classification accuracy they provide is thus more likely to provide a more faithful account of the BCI user's skills. In the data sets studied here, subjects were trained online using a fixed classifier though, as in most BCI studies.

When comparing the spatial metrics to the spatio-spectral ones, it appeared that the latter were more likely to reveal performance increases and thus possible learning effects, due to different users' performance variations in different frequency bands. We would thus advocate for reporting these spatio-spectral metrics, or at least to report the spatial ones in various frequency bands.

The *restDis* and *spaspecRestDis* metrics also highlighted the need to consider rest EEG when evaluating MI-BCI users' skills. Not doing so may prevent us from realizing that the user is not able to perform one of the MI tasks. If the target BCI application actually requires the user to perform real MI, e.g., for stroke rehabilitation, this aspect should be monitored. On the other hand, as seen on the multi-sessions data set (data set 2), *restDis*/*spaspecRestDis* also showed that some subjects actually improved *restDis* and *spaspecRestDis* across sessions but not *classDis* nor *spaspecClassDis*, which can only be observed using such metrics. For such types of subjects, this means that it might be more efficient to use one of the mental tasks as a brain switch[4], since that mental task leads to EEG patterns that are distinct from those of rest EEG, but not so much from those related to the other mental tasks. Alternatively, this might mean that it might be worth specifically training the user to make the different patterns more distinct from each other.

It should be mentioned that our new metrics are relative metrics and not absolute ones. Indeed, the distance between matrices depends on the dimension of these matrices. Matrices with larger dimensions would tend to be further away from each other, the same way as vectors with larger dimensions would tend to be further away from each other [12]. Thus, although our metrics can be very useful to compare training procedures or feedbacks between each other, when using the same EEG cap and channels, they cannot be used to compare experiments with different number of channels between each other. Thus, our metrics do not aim at replacing classification accuracy, but rather at complementing it. Indeed using our metrics should maximize the chance to measure and compare MI-BCI skills improvement over time (i.e., possible learning effects), which classification accuracy, as used online, is likely to miss, as well as to understand in a more refine way what skills the user has learned or not. For instance, what mental commands the user is mastering increasingly well, or how stable their EEG patterns are should be studied with the metrics proposed here, i.e., *classDis*, *restDis*, *classStab*, and/or *spaspecClassDis*, *spaspecRestDist*, *spaspecClassStab*. In the

[4]A brain switch is a single-class BCI, which uses a single mental task to send a command when the EEG patterns evoked by this task becomes different than that of rest EEG signals, see, e.g., [32]

future, it might be worth considering using such metrics as dedicated feedback in initial training tasks for BCI, to train explicitly distinctiveness and stability. Then, once the users managed to produce stable and distinct EEG patterns, as measured using such metrics, then can be trained for BCI control using a classifier, as currently done.

It is interesting to note that for the data sets analyzed, RWCV CA and *classDis*/*spaspecClassDis* generally measured consistent performance dynamics: they usually both measured a positive performance increase at the same time, when CA may not have. This seems to stress the importance of metrics that are training data independent to measure MI-BCI users' skills and learning. Indeed, contrary to CA, neither RWCV CA nor *classDis*/*spaspecClassDis* depend on the training data from the calibration run used to obtain the online classifier. As such, if users managed to improve the distinctiveness of their EEG patterns, possibly in a direction that the online classifier will miss, both RWCA CV and *classDis*/*spaspecClassDis* will be able to measure it. Such metrics should be thus be preferred to study user MI-BCI skills and learning. Again, CA tells us how well the classifier can recognize the EEG patterns produced by the user, but not necessarily how well the user can produce these patterns and make them as stable and distinct as possible. While RWCV CA is thus also very useful to study the distinctiveness between the EEG patterns produced by the user, it cannot tell us how such patterns differ from rest EEG patterns nor how stable they are. On the other hand, that is something that our new metrics *restDis* and *classStab* can measure, and were designed to do. Thus, our new metrics not only address the limitations of current measures such as CA, but also enable us to look at other aspects of MI-BCI users' skills that current metrics cannot see.

When studying possible correlations between our new metrics and spatial abilities, as measured using mental rotation test scores, significant correlations were obtained, in particular for data set 2. This result seems to further confirm the relevance of spatial abilities as a major predictor of performance for MI-BCIs [19]. As argued in [19], spatial abilities corresponding to the ability of producing, manipulating and transforming mental images [36], it makes sense that they influence MI-BCI performance.

Altogether, these new metrics led to new ways to look at the data and at MI-BCI users' skills. They suggested that MI-BCI control skills are multidimensional, and cannot be summarized by using only CA. In particular, our metrics suggested that MI-BCI user training did not only influence the distinctiveness between classes, but also the distinctiveness between each class and rest EEG signals, or the stability of each class EEG patterns, possibly in a different way in different frequency bands. All those metrics thus seem to reflect different aspects/components of MI-BCI control skills, which we denote as "subskills" in the following. Our results with the new metrics also suggested that different users might acquire these various MI-BCI control subskills in different ways. For instance, some users managed to improve the distinctiveness of each class with rest EEG patterns over time but not the distinctiveness between classes, whether it was the opposite for some other users. In turns, this calls for new ways to analyze and improve MI-BCI user training, by taking into account such MI-BCI subskills. For instance, different neuropsychological factors might be needed to predict performances for each of the different BCI subskills (*restDis*, *classDis*, *classStab* - possibly in different frequency bands). Along the same lines, since MI-BCI skills seem to gather several subskills, it might mean that BCI training tasks and feedbacks should

also target and reflect such subskills. For instance, it might be worth designing and studying training tasks dedicated to improve each one of the subskills. Additionally, it might be necessary to provide BCI users with a dedicated feedback for each of the subskills, to inform them about their progress in each of them. These results thus pave the way for many promising ways to refine and improve MI-BCI user training.

Naturally, the proposed metrics are not solving all performance evaluation issues of MI-BCI, and can still be improved. First, contrary to metrics such as CA, there is no closed-form solution to estimate the chance level of these metrics. If one is interesting by chance-level performance, they should resort to permutation tests. These new metrics also implicitly assume that the covariance matrix variability (as measured by the mean absolute deviation) around the mean is uniform (same variability in all direction). If that is not the case in practice, it may distort the distinctiveness and stability metrics. Finally, these metrics do not consider the time dimension. It would be relevant in the future to consider metrics estimating how fast the user can produce the EEG patterns using MI, and how long they can maintain them. There are thus still a lot of room for improvement and/or extensions of these metrics. Nonetheless, our results show that these new metrics still provide some new and essential insights into MI-BCI users' skills and learning, and highlighted major limitations of current metrics for this topic. As such, we advocate that such new metrics should be considered in addition to existing ones when studying MI-BCI users' skills and learning in the future.

6 Conclusion

In this paper, we argued that CA (online or simulated online), the most used metric to quantify BCI performance, should not be used alone to study MI-BCI users' skills and learning. We indeed identified many limitations of CA for this purpose. We proposed a first definition of MI-BCI skills and proposed new metrics, based on Riemannian distance, to quantify them. Our new metrics can measure how distinct the EEG patterns produced by the user are between each class, how each of them differ from rest EEG patterns, and how stable they are. These metrics all are classifier and training data independent. This makes them theoretically more appropriate to study MI-BCI user skills than CA, which heavily depends on the classifier and training data used - whereas MI-BCI users' skills should be measured independently from them.

An evaluation and comparison of these metrics indeed confirmed that in practice online CA may hide some continuous performance increase that may reflect learning effects and cannot identify how different an MI class is from rest EEG. They also revealed that different users seem to learn MI-BCI skills in different ways, thus highlighting various MI-BCI subskills. We therefore conclude that, when studying MI-BCI users' skills and learning, CA should be used with care, and should be complemented with metrics such as the ones proposed. Our results also stress the need to redefine MI-BCI user training by considering the different MI-BCI subskills and their measures. In order to ease the adoption of such metrics by the community, we provide their Matlab code for free and open-source on the following webpage: `http://sites.google.com/site/fabienlotte/code-and-softwares/metricsofperformance`

Naturally, this study would benefit from being replicated on other data sets, with

different protocols and classifiers, to further confirm and validate its outcome. Nonetheless, this study and metrics open many promising perspectives. In particular it would be interesting to re-analyze the relationship between users' profile, notably neurophysiological, personality and cognitive profile, and these new performance metrics (so far done by looking for correlation with online CA only, see [19] for a review), which could reveal new predictors of performance, and thus new ways of improving BCI user training. These metrics could also be used as the basis to design new feedbacks, and in particular explanatory feedbacks [40]. Indeed, these metrics being based on simple distance measures, they could be computed online, using incrementally estimated average covariance matrices. In contrast, the RWCV CA metric cannot be used online, notably due to its computational cost. The *classDis*, *restDis* and *classStab* metrics could thus be provided as online feedback, to tell users whether they should improve the distinctiveness with rest, with another class, or the stability of their patterns, for instance. These concepts being abstract and unusual for BCI users, a considerable work is needed in terms of user-centered design and human-computer interaction to find out the most consistent, intuitive and pleasant ways to provide such an explanatory feedback. To do so, such features might be combined with engaging and rich tools for real-time visualization of EEG patterns, such as those presented in [30, 10, 11]. These metrics revealing what seemed like fast learning effects, they could also be used as a cheap, possibly online way (faster and more convenient than CV) to identify when to update and retrain classifiers. Finally, it would be relevant to further refine these metrics, for instance by defining sub-metrics, for subsets of EEG channels, over specific brain areas, to study brain area specific learning processes. Overall, we are convinced that MI-BCI user training should be further studied, and we hope these new metrics could be a new way to look at it.

Acknowledgments

This work was supported by the French National Research Agency with the REBEL project (grant ANR-15-CE23-0013-01), the European Research Council with the BrainConquest project (grant ERC-2016-STG-714567) as well as by the EPFL/Inria International Lab.

References

[1] R. Bauer and A. Gharabaghi. Estimating cognitive load during self-regulation of brain activity and neurofeedback with therapeutic brain-computer interfaces. *Frontiers in behavioral neuroscience*, 9:21, 2015.

[2] L. Bianchi, L. R. Quitadamo, G. Garreffa, G. C. Cardarilli, and M. G. Marciani. Performances evaluation and optimization of brain computer interface systems in a copy spelling task. *IEEE Transactions on neural systems and rehabilitation engineering*, 15(2):207–216, 2007.

[3] R. Chavarriaga, M. Fried-Oken, S. Kleih, F. Lotte, and R. Scherer. Heading for new shores! overcoming pitfalls in BCI design. *Brain-Computer Interfaces*, pages 1–14, 2016.

[4] K. Colwell, C. Throckmorton, L. Collins, and K. Morton. Projected accuracy metric for the p300 speller. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(5):921–925, 2014.

[5] M. Congedo, A. Barachant, and R. Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, pages 1–20, 2017.

[6] J. L. Contreras-Vidal. Identifying engineering, clinical and patient's metrics for evaluating and quantifying performance of brain-machine interface (bmi) systems. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 1489–1492. IEEE, 2014.

[7] B. Dal Seno, M. Matteucci, and L. T. Mainardi. The utility metric: a novel method to assess the overall performance of discrete brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(1):20–28, 2010.

[8] J. Faller, R. Scherer, U. Costa, E. Opisso, J. Medina, and G. R. Müller-Putz. A co-adaptive brain-computer interface for end users with severe motor impairment. *PloS one*, 9(7):e101168, 2014.

[9] E. Felton, R. Radwin, J. Wilson, and J. Williams. Evaluation of a modified fitts law brain–computer interface target acquisition task in able and motor disabled individuals. *Journal of neural engineering*, 6(5):056002, 2009.

[10] J. Frey, R. Gervais, S. Fleck, F. Lotte, and M. Hachet. Teegi: Tangible EEG interface. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 301–308. ACM, 2014.

[11] J. Frey, R. Gervais, T. Lainé, M. Duluc, H. Germain, S. Fleck, F. Lotte, and M. Hachet. Scientific outreach with Teegi, a tangible EEG interface to talk about neurotechnologies. In *CHI'17 Interactivity-SIGCHI Conference on Human Factors in Computing System*, 2017.

[12] J. H. K. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[13] E. V. Friedrich, C. Neuper, and R. Scherer. Whatever works: A systematic user-centered training protocol to optimize brain-computer interfacing individually. *PloS one*, 8(9):e76214, 2013.

[14] K. Fukunaga. *Statistical Pattern Recognition, second edition*. ACADEMIC PRESS, INC, 1990.

[15] J. H. Gruzelier. EEG-neurofeedback for optimising performance. i: a review of cognitive and affective outcome in healthy participants. *Neuroscience & Biobehavioral Reviews*, 44:124–141, 2014.

[16] N. J. Hill, A.-K. Häuser, and G. Schalk. A general method for assessing brain–computer interface performance and its limitations. *Journal of neural engineering*, 11(2):026018, 2014.

[17] C. Jeunet, E. Jahanpour, and F. Lotte. Why standard brain-computer interface (BCI) training protocols should be changed: An experimental study. *Journal of Neural Engineering*, 13(3):036024, 2016.

[18] C. Jeunet, F. Lotte, and B. N'Kaoua. *Human Learning for Brain–Computer Interfaces*, pages 233–250. Wiley Online Library, 2016.

[19] C. Jeunet, B. N'Kaoua, and F. Lotte. Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates. *Progress in brain research*, 2016.

[20] C. Jeunet, B. N'Kaoua, S. Subramanian, M. Hachet, and F. Lotte. Predicting Mental Imagery-Based BCI Performance from Personality, Cognitive Profile and Neurophysiological Patterns. *PLoS ONE*, page 20, July 2015.

[21] T. Kaufmann, J. Williamson, E. Hammer, R. Murray-Smith, and A. Kübler. Visually multimodal vs. classic unimodal feedback approach for smr-bcis: a comparison study. *Int. J. Bioelectromagn*, 13:80–81, 2011.

[22] A. Kübler, E. M. Holz, A. Riccio, C. Zickler, T. Kaufmann, S. C. Kleih, P. Staiger-Sälzer, L. Desideri, E.-J. Hoogerwerf, and D. Mattia. The user-centered design as novel perspective for evaluating the usability of bci-controlled applications. *PLoS One*, 9(12):e112392, 2014.

[23] A. Kübler, D. Mattia, H. George, B. Doron, and C. Neuper. How much learning is involved in BCI-control? In *Int. BCI Meeting*, 2010.

[24] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[25] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi. A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 4:R1–R13, 2007.

[26] F. Lotte and C. Jeunet. Towards improved BCI based on human learning principles. In *Proc. Int BCI Winter Conf*, 2015.

[27] F. Lotte and C. Jeunet. Online classification accuracy is a poor metric to study mental imagery-based BCI user learning: an experimental demonstration and new metrics. In *International Brain-Computer Interface Conference*, 2017.

[28] F. Lotte, F. Larrue, and C. Mühl. Flaws in current human training protocols for spontaneous brain-computer interfaces: Lessons learned from instructional design. *Frontiers in Human Neuroscience*, 7(568), 2013.

[29] D. McFarland, W. Sarnacki, and J. Wolpaw. Should the parameters of a bci translation algorithm be continually adapted? *Journal of Neuroscience Methods*, 199(1):103 – 107, 2011.

[30] J. Mercier-Ganady, F. Lotte, E. Loup-Escande, M. Marchal, and A. Lécuyer. The mind-mirror: See your brain in action in your head using EEG and augmented reality. In *IEEE Virtual Reality (VR)*, pages 33–38. IEEE, 2014.

[31] J. Mladenović, J. Frey, M. Bonnet-Save, J. Mattout, and F. Lotte. The impact of flow in an EEG-based brain computer interface. In *7th International BCI conference*, 2017.

[32] G. Müller-Putz, V. Kaiser, T. Solis-Escalante, and G. Pfurtscheller. Fast set-up asynchronous brain-switch based on detection of foot motor imagery in 1-channel EEG. *Medical and Biological Engineering and Computing*, 48(3):229–233, 2010.

[33] C. Neuper and G. Pfurtscheller. *Brain-Computer Interfaces*, chapter Neurofeedback Training for BCI Control, pages 65–78. The Frontiers Collection, 2010.

[34] G. Pfurtscheller and C. Neuper. Motor imagery and direct brain-computer communication. *proceedings of the IEEE*, 89(7):1123–1134, 2001.

[35] L. Pillette, C. Jeunet, B. Mansencal, R. N'Kambou, B. N'Kaoua, and F. Lotte. PEANUT: Personalised Emotional Agent for Neurotechnology User-Training. In *7th International BCI Conference*, 2017.

[36] S. E. Poltrock and P. Brown. Individual differences in visual imagery and spatial ability. *Intelligence*, 8(2):93–138, 1984.

[37] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

[38] L. Roijendijk, S. Gielen, and J. Farquhar. Classifying regularized sensor covariance matrices: An alternative to CSP. *IEEE Trans Neur. Syst. Rehab.*, 24(8):893–900, 2016.

[39] A. Schlögl, J. Kronegg, J. Huggins, and S. G. Mason. *Towards Brain-Computer Interfacing*, chapter Evaluation criteria in BCI research, pages 327–342. MIT Press, 2007.

[40] J. Schumacher, C. Jeunet, and F. Lotte. Towards explanatory feedback for user training in brain-computer interfaces. In *Proc. IEEE SMC*, pages 3169–3174, 2015.

[41] M. Spüler. A high-speed brain-computer interface (BCI) using dry EEG electrodes. *PloS one*, 12(2):e0172400, 2017.

[42] E. Thomas, M. Dyson, and M. Clerc. An analysis of performance evaluation for motor-imagery based BCI. *J Neur Eng*, 10(3):031001, 2013.

[43] D. E. Thompson, L. R. Quitadamo, L. Mainardi, S. Gao, P.-J. Kindermans, J. D. Simeral, et al. Performance measurement for brain–computer or brain–machine interfaces: a tutorial. *Journal of neural engineering*, 11(3):035001, 2014.

[44] D. E. Thompson, S. Warschausky, and J. E. Huggins. Classifier-based latency estimation: a novel way to estimate and predict bci accuracy. *Journal of neural engineering*, 10(1):016006, 2012.

[45] R. Tomioka and K.-R. Müller. A regularized discriminative framework for eeg analysis with application to brain-computer interface. *Neuroimage*, 49(1):415–432, 2010.

[46] S. G. Vandenberg and A. R. Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2):599–604, 1978.

[47] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Co-adaptive calibration to improve bci efficiency. *Journal of neural engineering*, 8(2):025009, 2011.

[48] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural computation*, 23(3):791–816, 2011.

[49] J. R. Wolpaw, H. Ramoser, D. J. McFarland, and G. Pfurtscheller. EEG-based communication: improved accuracy by response verification. *IEEE transactions on Rehabilitation Engineering*, 6(3):326–333, 1998.

[50] F. Yger, M. Berar, and F. Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Trans Neur. Syst. Rehab.*, 2017.

[51] F. Yger, F. Lotte, and M. Sugiyama. Averaging covariance matrices for eeg signal classification based on the csp: an empirical study. In *Proc. EUSIPCO*, pages 2721–2725, 2015.

[52] C. Zich, S. Debener, C. Kranczioch, M. G. Bleichner, I. Gutberlet, and M. De Vos. Real-time EEG feedback during simultaneous EEG–fMRI identifies the cortical signature of motor imagery. *Neuroimage*, 114:438–447, 2015.