Evaluation criteria for closed-loop adaptive dynamic discrete-continuous brain-computer interfaces: clinical study case with tetraplegic patient.

Félix Martel
Univ. Grenoble Alpes,
CEA, LETI, Clinatec
F-38000 Grenoble,
felix.martel@cea.fr

Tamara Dupuy
Univ. Grenoble Alpes,
CEA, LETI, Clinatec,
F-38000 Grenoble,
tamara.dupuy@phelma.grenoble-inp.fr

Alexandre Moly
Univ. Grenoble Alpes,
CEA, LETI, Clinatec,
F-38000 Grenoble,
alexandre.moly@cea.fr

Stéphan Chabardès
Centre Hospitalier Universitaire
Grenoble Alpes
F-38700, La Tronche,
schabardes@chu-grenoble.fr

Tetiana Aksenova
Univ. Grenoble Alpes,
CEA, LETI, Clinatec,
F-38000 Grenoble,
tetiana.aksenova@cea.fr

***Abstract*— Quantifying and evaluating properly the performances is a critical issue in BCI experiments. The choice of the most adapted metrics can be difficult because they are specific to the experimental paradigm, task control, and data. In the current study evaluation criteria for closed-loop adaptive dynamic and hierarchical discrete-continuous brain-computer interfaces are examined. The challenges such as imbalanced multi-class bias for discrete decoding (classification), online computing cost and dynamic analysis of results in time are considered. There are two main objectives of the study: identifying the best suited performances metrics according to the requirements and combining several levels of evaluation in the whole BCI system (decoder and patient oriented). The main novelty of the study is the combination of several common metrics with new temporal metrics and a time dynamic approach which allows adequately reflect the performance of adaptive dynamic and hierarchical discrete-continuous brain-computer interfaces BCI systems. Additional response time and blocking error patterns reveal complementary information for BCI system performance evaluation. Developed criteria are applied to performance evaluation of 8-dimensional exoskeleton control by tetraplegic patient for both decoder and user-centered metrics.**

Keywords—brain computer interface, performance, metrics, discrete, continuous, hybrid, self-paced, adaptive, online.

I. INTRODUCTION

Brain-Computer Interface (BCI) is a system that aims to establish a direct communication between brain and external effectors, by translating brain activity signals into useful control signals. Motor BCI are systems which extract motor commands imagined by the patient from his neuronal activity and send this control information to external real or virtual effectors that execute the movements [1]. By allowing people to use brain signals instead of muscles, such a system can enable a person with severe motor disabilities to restore or replace crucial motor behaviors [1] normally served by the motor system. Many studies have focused on motor BCI and have successfully demonstrated that it is possible for severely motor-impaired patients to control such effectors, e.g. robotic limbs [2], wheelchair [3], avatar [4] or exoskeleton [5].

Generally, BCIs system consists in a transducer and a control interface [6]. The transducer acquires brain signals and translates them into output commands. It includes signal acquisition, preprocessing and decoding. At the next level, the control interface converts transducer output into meaningful communication and controls to the effector, providing a sensitive feedback to the user. A decoding algorithm transforms the neural inputs into output variables according to the domain of interest. Output variable can be either discrete (with limited number of values), or continuous. Some BCIs are hybrid and combine both discrete and continuous decoders [7]. Whatever the case, the decoder first needs to be trained by having both the initial input data with the corresponding output data for supervised learning (training stage). After training the decoder, it is applied on the new incoming neural activity (testing stage). More efficient adaptive decoders allow being trained while being used, and are well suited to overcome non-stationarity challenge [8].

A critical question for BCI technology is to be able to quantify properly BCI performance. Performance measurements are necessary at both the training and the testing stages. In training stage, they play a critical role in achieving the optimal decoder by selecting the optimal solution and hyper-parameters to produce the best prediction. In testing stage, they are the measurement tools to estimate the effectiveness of the

Clinatec is a Laboratory of CEA-Grenoble and has statutory links with the University Hospital of Grenoble (CHUGA) and with the University Grenoble Alpes (UGA). This study was funded by CEA (recurrent funding) and the French Ministry of Health, Fondation Motrice, Fondation Nanosciences, Institut Carnot, Fonds de Dotation Clinatec. Fondation Philanthropique Edmond J. Safra is a major founding institution of the Clinatec Edmond J. Safra Biomedical Research Center.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

decoder with new data. Such performance measures are highly specific to the context of experimental paradigm, task control, and training data. It is a topic of growing literature where a large number of performance measures already exist. The choice of the most adapted ones is a critical issue [9].

This study took place in the context of Clinatec's BCI project[1] which aims to control complex effectors such as a full-body exoskeleton by a tetraplegic subject, from intracranial Electrocorticogram (ECoG) signals. Patients are implanted with two wireless 64-channel epidural electrocorticagraphy WIMAGINE® implants, specificly designed for long-term signal recording [5,10]. The decoder used in this project is an online adaptive hybrid dynamic decoder, used in self-paced BCI experiments. It decodes discrete (states) and continuous (movements) outputs, and can be trained while being used. Discrete and continuous decoders are combined as mixure of experts [7]. The discrete decoder is based on dynamic classification (Hidden Markov Model, HMM), which assumes that data are temporally dependent. It can decode a no-control (NC) state, when the user does not want to move, and several intentional control (IC) states, corresponding to the different kinds of movements available (arm, wrist, etc.). The continuous part of the decoder, a set of linear experts, predicts limbs translations or rotations. Continuous outputs correspond to the directed distance (or angle) between the current position and the target. Continuous decoder is based on a linear regression. Online adaptive decoder is updated in real time during closed-loop BCI experiments.

The main objective of building an efficient performance measurement system in the online software chain is to be able to evaluate and quantify the BCI system performance during the experiments. Based on these results, smart interpretation can be added to the training procedure to trigger the model update when necessary and to generate experimental tasks related to the lowest performance of the model [11]. Intrisic and unbiased indicators must then be identified to assess and quantify the BCI system in real-time. Different levels of evaluation must be considered when evaluating a whole BCI system. At the algorithmic level, the ability to interpret and translate correctly the brain activity, as well as the stability and convergence associated to this adaptive decoding model should be evaluated. At the user and experimental level, the patient ability to generate and control brain state should be assessed, as well as the overall experiment according to the task difficulty.

Constraints and requirements are firstly presented in this document, for each part of the multimodal performance measurement system. Then, known performance evaluation metrics from the literature are confronted to these requirements. Finally, the whole measurement system is built and applied to clinical data, to conclude on its relevance.

II. REQUIREMENTS

A. Study specific requirements

A large number of performance-measuring metrics exist in literature [12-17]. An appropriate measure must be chosen according to the context and objectives. Evaluation metrics can provide influenced results according to:

- The nature of the data: the distribution of the classes (balanced or unbalanced dataset), numbers of classes (binary or multi-classes), or sample temporal dependence (dependent or independent).
- The type of the decoder and the nature of the prediction: discrete or continuous output, static or dynamic decoding, distribution of the output (biased or unbiased classification), misclassification costs.
- The experimental methods: online or offline analysis, task difficulty, timing conditions, user abilities.

In the present study, the decoder combines several specificities and challenges. First, both discrete and continuous outputs are predicted. Moreover, self-paced experiments generate particularly imbalanced multi-class datasets (NC state is much more frequent than any other state). In addition, a focus on the dynamics of the discrete decoder must be added to the classic sample-wise methods because of the dynamic classification. Finally, although the performance of some metrics is excellent for evaluation and discrimination processes, they might nor be suited for online adaptive BCIs [18,19]. Computational cost is a restricting factor when dealing with performance metrics for online applications. Thus, there are two main novelties associated with this study: identify best suited performances metrics according to the present study requirements (online, hybrid discrete-continuous, dynamic, asynchronous, imbalanced multi-limb control), and combine several levels of evaluation in the whole BCI system (decoder and patient oriented). Performance evaluation criteria are additionally proposed to reflect the dynamics of classification errors: the error blocks frequency and duration.

B. Performance metrics requirements

In the present study, the discrete decoder is a multi-class classification problem. In the same time, most of the common metrics were originally developed and are applicable for binary classification. Only a few of them are extended for multi-class classification evaluation [18,20]. A multi-class problem can be divided into several binary problems, where the performance can be calculated for each individual class confronted to all the rest (one versus all), or one by one (one versus one). Overall performance measures can be calculated as the micro and macro average of the per-class performances. The macro-average is formed by taking the average of all the per-class measures, whereas a micro-average will aggregate the contributions of all classes to compute the average metric [20]. Even if an overall single metric is useful for comparing systems, the performance per class may be more relevant for analysis, for informativeness between classes. One of the challenges of quantifying self-paced BCI performance is to determine appropriate criteria to quantify imperfect NC state support. Indeed, NC period must have an important consideration, especially to avoid false activation error [12,21-24]. Due to the delay for effectors desactivation (up to a few seconds, depending on the effector), it is crucial to avoid unwanted activations of the effectors which are directly in contact with the patient. Thus, even in the multi-class case, a particular separate analysis must be done to obtain

[1] A clinical research protocol called "BCI and Tetraplegia. Further information available at *https://clinicaltrials.gov/ct2/show/NCT02550522*

informativeness between the IC and NC [22]. This complicates the comparison of methods, but also capture important aspects of the performance.

The class imbalance problem occurs when there are many more samples of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. Imbalanced datasets frequently appear in self-paced brain computer interface systems, as the no-control state is more probable than the intentional-control. Most of popular metrics are affected by the class distribution [19,20,25–27], biased towards the majority class [25]. One of the good approaches to deal with this issue is to choose and optimize performance metrics that are designed to best handle the imbalance by staying invariant [20]. A number of solutions to the class-imbalance problem were proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of resampling (random over-sampling, random under-sampling, SMOTE oversampling [28], or both [29]). At the algorithmic level, solutions include adjusting classes costs [17,26], decomposition into binary classes, boost algorithm [30]. The issue about imbalanced is much worse for multi-class problems, where fewer solutions are applicable [19,26].

Very few metrics take into account time characteristics when evaluating a model performance, however it is a crucial parameter in assessing the performance of a BCI especially in self-paced BCI [3,15,22,31].

Low BCI system latency is generally deemed to be crucial for motor BCIs. Evaluation of the response time of the BCI system in general and of the decoding model response time in particular is one of the important metrics of BCI system evaluation. This measure is usually biased by the patient, who also adds a response time according to the actual onset of intent activity after a cue and to the feedback received during the task. The exact time at which the user generates brain activity for a particular task command during experiment is usually unknown [15] and can impact the analysis when labelling correctly the references. Some study tried to determine it through IRM [22] to determine the real Expected User Intent but it is not suited for online applications, while other studies computed the response time based on the steepness of any evalu ation metrics [15].

Others temporal metrics can be useful to describe block error dynamic pattern [7,22]. For classification, a block error can be defined by all the consecutive samples corresponding to the same state mismatch. Comparing block error dynamic such as block error duration or block error frequency of apparition can capture important aspects when evaluating decoding model. But to the best of our knowledge very few studies deal with those dynamic approach.

Finally, reporting timing experiment during evaluation will allow cross-study comparisons [12]: by normalizing some metrics according to the task duration in order to be comparable for different users or different tasks, or by normalizing performance according to decision rate to be comparable between several BCI transducers.

TABLE I. REQUIREMENTS SUMMARY.

Requirements
Both discrete and continuous performances
Multi-class classification
Imbalance (especially towards NC state)
Dynamics considerations (response times, error dynamics)
User variability
Task variability
Online computations

C. User and task variability

Most of metrics concern the evaluation of the prediction in the first stage of the BCI system: the transducer. However, the overall experiment not only depends on intrinsic BCI system but also on the equipment, the user, or task asked. To be able to determine both the user and experiment impact, one must be able to quantify the abilities of the user (user variability), and to compare these performance across several different tasks [13,32] (task variability) by the use of appropriate metrics.

User oriented performance metrics may reveal important aspects of the user performance, showing how well the user is able to produce clear, stable, and distinct patterns [33]. User variability is a significant challenge [12] considering brain signals are non-stationarity [8], especially with changes in the user's mental states such as habituation, fatigue, or inattention.

Task variability is also a key point when evaluating a unique subject across several tasks. Intrinsic task setup such as difficulty and timing must be involved to the performance metrics in order to compare several experiments and several users. Most current methods use fixed level of task difficulty and cannot reveal the possible evolution of the BCI performance [32]. Moreover, when results are compared for cross-subjects or cross-session manner, data distribution may differ and can lead to metrics bias. Such changes must be considered. Study specific requirements are summerized in TABLE I.

III. METHOD

In this section, the known performance evaluation criteria are revised according to listed requirements. Additional criteria are proposed.

A. Discrete performance measures

Performance metrics for discrete classification from the literature are compared based on the criteria previously cited (TABLE II).

One of the most widely used tool to visualize classification performance is the confusion matrix [9,13-15], comparing the classification prediction with the actual class. Confusion matrices are widely used in BCI systems [34-36]. This metric can easily be adjusted in real-time [34] for online applications, and is applicable for multi-class problems. Informations can then be computed from the confusion matrix, such as the prior distribution of each class, and the model bias towards each predicted class. Normalizing the confusion matrix can be more convenient [14], as it represents the classification and misclassification rate of all classes with a better interpretability,

but information about the number of trials of each class is lost. Even if the confusion matrix contains all information in one-versus-one analysis, it is difficult to compare several ones. Therefore, most BCI studies usually use scalar performance measures for one-versus-all and global analysis directly derived from the confusion matrix [9,15].

Precision measures the fraction of correctly predicted instances from the total instances of predictions, while Recall (also called Sensitivity) measures the probability of detection. Specificity is similar to recall, but focuses on the classification of the other classes. They are the most classic measures derived from confusion matrix used for BCI evaluation [37-39] but they are often represented by combined measures such as the harmonic mean Fscore, the geometric mean Gmean, arithmetic mean HF difference [9,14,15] (also called Individual Classification Success Index [17]) or Bookmaker informedness [25]. Such combination metrics have been developed to cumulate the advantage of each metric and to overcome their limitations. However, combinations of measures can be difficult to interpret correctly [17]: the range can be different and intermediate value can have an unclear meaning.

Accuracy is the percentage of samples that were correctly classified. It is one of the most widely used metric for measuring the performance of a classifier [15] and it is commonly used for online BCI systems [3,7,34,40-45]. The possible reason for its popularity is that it can be very easily calculated and interpreted. However, accuracy has several weaknesses which are less discriminability, less informativeness, and a strong bias to unbalanced class [18,27]. Cohen's Kappa is a measure for the agreement between the accuracy and chance level (i.e. the accuracy of a random classifier) [14,15,46]. Kappa is thus calculated by transforming accuracy value from the interval [po,1] to the interval [0,1]. When the classes are unbalanced, Cohen's Kappa is more recommended [14,15,27], but it fails to diagnose temporal dependence [46]. A new evaluation criteria $\text{Kappa}_{\text{temporal}}$, proposed in [46] is not based anymore on random classifier but on the accuracy of the persistent classifier, which predicts the same label as the previous state observation. It assesses the temporal dependence but is not suited for imbalanced datasets. Moreover, as states are quite stable in our data, bad performances can be expected compared to the persistent classifier.

Other metrics are derived from confusion matrix. From its construction Class Balanced Accuracy (CBA) can be either the recall either the precision [47]. By choosing the maximum of the row or column sum as the denominator, the metric provides the most informative and critical contribution of the off-diagonal elements compared to what can be achieved in other metrics. Jaccard coefficient is the ratio of the estimated and true classes' intersection over their union. It ranges from -1 to 1, 1 meaning a complete overlap and a perfect classification. The Matthews correlation coefficient (MCC) can be seen as a discretization of the Pearson correlation for binary variable, and more recently generalized to the multiclass case [48]. This metrics is more widely used in bioinformatics classification [48–50] than BCI. It ranges between [-1, 1], where the bounds represent perfect misclassification and perfect classification respectively. Its score is high only if a classifier is doing well on both the negative and the positive elements while a value of 0 indicates the classifier performed the classification randomly. MCC has been identified as the best null-biased metrics for binary classification [25,51] to best handle data imbalance, but multi-class imbalance has not been well studied yet.

TABLE II. METRICS AND CHALLENGES.

Metrics	Multi-class	Imba-lance	On-line	Dyna-mics
Precision	✓	X	✓	X
Recall	✓	✓	✓	X
Specificity	✓	✓	✓	X
Fscore	✓	X	✓	X
Gmean	✓	✓	✓	X
HF difference	✓	X	✓	X
Bookmaker informedness	✓	✓	✓	X
Accuracy	✓	X	✓	X
Kappa	✓	✓	✓	X
$\text{Kappa}_{\text{temporal}}$	✓	X	✓	✓
Class balanced ac.	✓	X	✓	X
Jaccard	✓	X	✓	X
Matthews Cor. Coef.	✓	✓	✓	X
ROC curve	✓	✓	X	X

Roc curve is obtained by plotting the recall versus False Positive Rate [9,15] and varying the detection threshold which separate different classes. The area under the ROC curve (AUC) is usually calculated, the larger the area is the better the performance. By sweeping this threshold, several combinations of misclassification ratio are implemented. However, although the performance of AUC was excellent for evaluation and discrimination processes, the computational cost increased for multi-class and not adapted for on-line analysis [18,19].

Finally, majority of metrics derived from confusion matrix can easily be extended to multi-class problem and can be adjusted and incremented for on-line experiment with a relatively low computational cost except Roc curve. None of confusion matrix derived metrics consider time consideration such as temporal depence or dynamic analysis except $\text{Kappa}_{\text{temporal}}$, which is not adapted for imbalanced datasets [46]. All these metrics present different aspects but they mainly focus on classification successes and not on errors, except for MCC, Bookmarker, and HF difference. These make a distinction between errors and consider different impact, which is relevant for no-control evaluation in self-paced.

Imbalance invariance has been studied recently [25] where several clusters of performance metrics have been identified as the best null-biased metrics: G-mean, MCC, and Bookmaker Informedness are reported as good discriminator and performed better than other evaluation metrics in optimizing classifier for classification problems especially for imbalanced dataset, while the use of accuracy, precision, or Fscore has been extensively discouraged. Kappa is also known to present good invariance face to imbalance.

In the present study, to verify their stability against imbalance, we have tracked metrics evolution before and after normalization of the confusion matrix.

In addition, a study was done on transitions from one state to another, to differentiate response times to actual decoding errors. Response time includes both the patient and the decoding model response times, when a new state is asked to the patient. To measure latencies in an online manner, a Maximum Response Window (MRW) has been computed in offline from the distribution of latencies. To obtain this MRW, the distribution of latencies was fitted with a gamma probability density function, taking the 80% threshold on the cumulative probability. The MRW can be computed in offline on past experiments, and used in online to differentiate latency samples from the others, and discard when incrementing the confusion matrix.

All the performance metrics mentioned above (TABLE II) are adapted for static sample-wise evaluation of performances. They do not reflect temporal dynamics of the decoder predictions. To complete this analysis with information about the error dynamics, a second confusion-like matrix is proposed. It counts the error blocks, defined as an uninterrupted sequence of samples of the same error (i.e. the same desired/predicted states pair). From the two matrices can be computed the mean error blocks durations and frequencies for each desired/predicted pair, as in (1) and (2):

$$Duration = \frac{n}{f \times N} \tag{1}$$

$$Frequency = \frac{f \times N}{(TP + FN)} \tag{2}$$

With N the number of coresponding error blocks, n the number of samples, fthe sampling rate, TP and FN the number of true positives and false negatives for this specific desired class (i.e. $TP + FN$ is the number of samples with this desired class). The global error block durations and frequencies can be computed by averaging these values for one desired class whatever the predicted class is, or for all the classes on the whole training.

B. Continuous performance measures

To analyze only continuous decoding and not have redundancy with discrete results, only samples once correct state prediction is reached are considered. The aim is to evaluate only the period of time when the patient focuses on the movement of the limb that is actually moving. In our case, continuous outputs correspond to the directed distance between the current effector position and the target. At each timestep, the desired movement is the optimal movement from the current position to the target position (straight line, with the maximum speed). A maximum speed is imposed at 0.1m/s for 3D movements and 1rad/s for wrists rotations, for both the security and the comfort of the patient. The evaluation of multi-dimensional movement is realized either considering each individual task when reaching each target, either considering each degree of freedom (DoF) over the whole past data. Those two evaluations allow to assess which limb needs to get improved and along which degree of freedom. This is critical information especially for generating more adapted training targets in the purpose of automation of the whole BCI system. Per-task evaluation is meaningful during online experiment to assess in real-time what kind of movement tasks are well performed and which ones need some improvement. An evaluation along each DoF is also computed during per-task evaluation. Additionally, an evaluation per DoF is computed along several dataset to give overall performance about movement preferential direction.

The correlation coefficient is one of the widely used metrics in the literature to compare predicted and desired movements in BCI systems [12,16,52-55]. The most commonly used is Pearson correlation coefficient (PCC). It measures how strong the linear relationship between both variables is. PCC is sufficiently informative, easily computed and interpreted. It should be used for Gaussian data, while for non-Gaussian data the rank correlation is recommended [9,15]. Other measures are commonly used for regression evaluation such as the Normalized root-mean-squared error (NRMSE) [16], which depends on the total scaling of the dataset, or the Coefficient of determination (COD) that can be interpreted as the squared correlation [16]. They are the most common measures, but other measures of continuous trajectories can be applied [56] such as movement direction changes, movement offset or movement variability.

In the present study, PCC is used as result reference with literature, as continuous performance measure evaluation per-experiments and per-task. For per-task evaluation, the variation is no more computed according to the mean value during the task but rather according to mean and standard deviation along the whole dataset, to have comparable measures from one task to another, with identically centered data.

Another metric introduced here for continuous movement evaluation is the normalized dot product (DP). This scalar measuring the angular difference between two vectors is very rarely used to measure BCI performances in the literature. This measure is used to quantify continuous decoding performances between the desired and predicted movement vectors $\boldsymbol{u}_i$ and $\widehat{\boldsymbol{u}}_i$ at sample i, along all DoFs (3). The product is then averaged over samples ($1<i<N$) to give a reliable indicator on the global static decoding (4)

$$DP_i = \frac{\sum_{per\ DoF}(u_{iDoF} \times \widehat{u_{iDoF}})}{max(\|u_i\|, \|\widehat{u}_i\|)} \tag{3}$$

$$DP = \frac{1}{N}\sum_i DP_i \tag{4}$$

where the normalization with the maximum between desired and predicted magnitude forces the metrics within [-1,1]. This measure is computed for each singular DoF along all datasets, and for each task command. For per-task evaluation, only a single measure over all DoFs is given to evaluate the global movement. When the patient becomes too close to reaching the target, desired directed distance becomes close to zero and the noise in the prediction makes the DP computation senseless. The corresponding samples below threshold are removed from the analysis.

C. Experiment and user skill measures

Existing metrics may not be able to reveal important aspects of the user performance and learning evolution. Appropriate performance metrics could help to understand what users have successfully learned or still need to improve, which can be used to guide them and provide appropriate training tasks and feedback in order to inform them about their progress. New metrics are proposed [33,34] in order to quantify users' learning skills. When comparing two kinds of brain signal patterns for two different task-classes A and B, we defined $\mathbf{a}_i=(a_i^1,a_i^2\ldots a_i^f\ldots a_i^F)^T \in \mathbb{R}^F$ summarizing brain signal information for each sample ($1<i<N_A$) concerned by class A and along all the features F, and as well $\mathbf{b}_i=(b_i^1,b_i^2\ldots b_i^f\ldots b_i^F)^T \in \mathbb{R}^F$ for each sample ($1<i<N_B$) concerned by class B. Distinctiveness is one of the metrics to quantify how distinct a pattern would be from other class pattern. The distinctiveness can be computed between two classes for binary classification and can be generalized to multi-class problem (5). In such case, for each task-class A_k ($1<k<K$), the vector $(\boldsymbol{a}^k)_i \in \mathbb{R}^F$ and mean vector $\boldsymbol{a}^k \in \mathbb{R}^F$ are defined as well.

$$Distinctiveness\ (A,B)=\frac{distance(\bar{a},\bar{b})}{\frac{1}{2}(\sigma_A+\sigma_B)} \tag{5}$$

$$\bar{a}=\frac{1}{N_A}\sum_{i=1}^{N_A} a_i\ ,\ \sigma_A=\sqrt{\frac{1}{N_A\text{-}1}\sum_{i=1}^{N_A} distance(a_i,\bar{a})} \tag{6}$$

$$distance\ (a,b)=\sqrt{\sum_{f=1}^{F}(a^f\text{-}b^f)^2} \tag{7}$$

$$\overline{Distinctiveness}\ (A^k)=\frac{\sum_{l=1,l\neq k}^{K}(\bar{a^k},\bar{a^l})}{\sigma_{A^k}+\sum_{l=1,l\neq k}^{K}\sigma_{A^l}} \tag{8}$$

It is also relevant to consider learning behavior according to time, i.e how fast the user produces the patterns and how long he can maintain them. Stability metrics is defined in [23] and is based on standard deviation measure (9)

$$Stability\ (A)=\frac{1}{1+\sigma_A} \tag{9}$$

However, those metrics can have high computational cost depending on the number of features used to describe brain signal into vectors. This must be considered when analyzing results especially because high dimension can bias results.

The distinctiveness is computed between each class (5) confronted to each other, as well as averaged for each class for an analysis one-versus-all (8). Stability measure (9) is also computed for each class. As feature dimension is particularly huge in this study (64*10*15=9600 for the three kinds of dimensions: spatial, temporal, and spectral), projection over each feature dimension (electrodes, time, or frequency) is performed to study dimensionality, compare bias, and reduce the computation load.

IV. Experimental setup

A. Data description

Data used in this study are from Clinatec's BCI project, from ten 8-dimensional control experiments [5]. These experiments consists in reaching targets with both arms of the exoskeleton (3D for each arm), as well as rotating the wrists (1D for each arm). During the experiments targets are shown by lighting LEDs which the patient has to reach with the corresponding arm (each arm has its own workspace that do not overlap, thus each LED can be reached by one arm only).

Neural signals are recorded at 586 Hz by two wireless 64-channels ECoG implants placed on each hemisphere, on the dura mater. Only half of the channels (32 per implant, selected in a checkerboard pattern) are used due to data transfer limitations.

Data are labellised by the decoding software. Five states are decoded (No-control, left hand movements, right hand movements, left wrist rotation, right wrist rotation), with a total of 8 degrees of freedom (respectively, 0, 3, 3, 1, 1). As the control is asynchronous, a lot of idling/resting time is in the datasets, creating a strong imbalance towards the NC state.

All the performance computations have been done afterwards in MATLAB2017b, from the dataset logged during the ten experiments, accumulating 3h38 of data. However, the computations were made in a pseudo-online manner, simulating the passage of time and using only "past" data.

B. Features extraction and decoding algorithm

Every 100ms, an epoch is generated using the last 1s time interval. After applying a complex continuous wavelet transform (CCWT) (Morlet mother wavelet) to obtain frequency information from 10 to 150 Hz, with a 10 Hz step for each electrode, absolute values are calculated and are downsampled to 10 Hz. A 64x10x15 features tensor is thus obtained at each processing step. From this features tensor, the model can predict, both the state (desired movement) and the direction of the limb movement. The discrete decoding part is ensured by dynamic classification based on Hidden Markov Models (HMM) [7]. This method assumes that successive samples are temporally dependent and that a selection of state is conditioned on previous states. Based on probabilities adjusted from state transition, the limb moving is the one with maximum probability. The continuous decoding part is ensured by a supervised training algorithm: Recursive Exponentially Weighted Multi-way Partial Least Square (REW-NPLS) [57], which is a generalization of the regression algorithms from the PLS-family. Both discrete classifier and continuous regression are updated online, in real time.

V. Results

Fig. 1 shows the distribution of the response times when a change of state is asked by the supervisor, with the fitted gamma probality density function and its 80% threshold, i.e a 5s maximum latency.

By setting this threshold as MRW, the latency can be measured in online experiments. Results obtained are shown in TABLE III, with the mean and stardard deviation of the latency of transition towards each state.

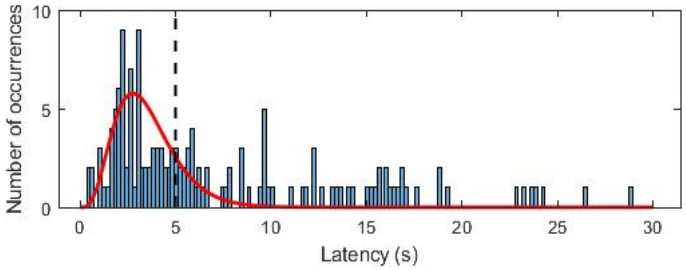

Fig. 1. Latency distribution on all the experiments. Occurences are in blue and the fitted gamma distribution is in red. The dash line is the threshold set at 80% of the cumulative gamma distribution.

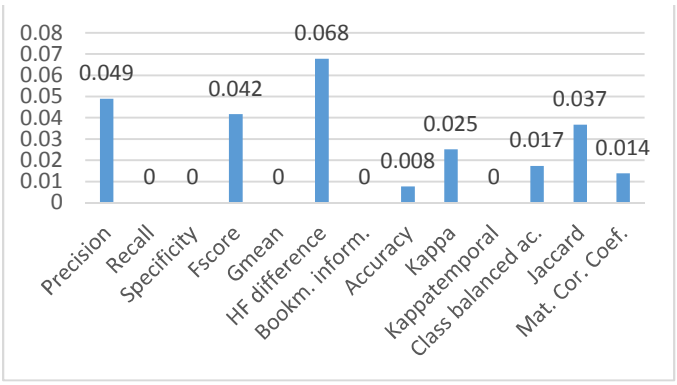

Fig. 2. Evolution of the discrete metrics when normalizing the confusion matrix.

TABLE III. LATENCY RESULTS WHEN TRANSITIONING TO EACH STATE.

State	Idle	Left Arm	Right Arm	Left Wrist	Right Wrist
Latency	1.29 ± 1.17	3.08 ± 1.50	3.93 ± 0.81	0.91 ± 1.33	0.74 ± 1.08

TABLE IV. CONFUSION MATRIX ON ALL THE DATASETS. LATENCY SAMPLES ARE ALREADY DISCARDED.

		Desired class				
		Idle	Left Arm	Right Arm	Left Wrist	Right Wrist
Predicted class	Idle	37285	2596	6184	1835	606
	Left Arm	3514	26955	303	1203	1
	Right Arm	2037	539	25683	110	0
	Left Wrist	196	890	34	4730	74
	Right Wrist	480	59	3985	226	6273

TABLE V. DISCRETE PERFORMANCE METRICS.

Metrics	Idle	Left Arm	Right Arm	Left Wrist	Right Wrist	Macro-Average
Precision	0.769	0.843	0.905	0.798	0.569	0.777
Recall	0.857	0.868	0.710	0.584	0.902	0.784
Specificity	0.864	0.947	0.97	0.990	0.960	0.946
Fscore	0.810	0.856	0.796	0.674	0.698	0.767
Gmean	/	/	/	/	/	0.774
HF difference	0.626	0.711	0.615	0.382	0.471	0.561
Bookmaker informedness	0.720	0.815	0.680	0.574	0.862	0.730
Accuracy	0.861	0.928	0.895	0.964	0.957	0.921
Kappa	0.702	0.807	0.727	0.656	0.676	0.713
$Kappa_{temporal}$	-121.4	-189.3	-248.7	-119.6	-135.2	-162.9
Class balanced ac.	0.769	0.843	0.710	0.584	0.569	0.695
Jaccard	0.681	0.748	0.661	0.509	0.536	0.627
Matthews Cor. Coef.	0.704	0.807	0.736	0.665	0.697	0.722

TABLE VI. ERROR BLOCK FREQUENCY, IN NUMBER OF BLOCK PER MINUTE OF THE GIVEN DESIRED CLASS.

		Desired class				
		Idle	Left Arm	Right Arm	Left Wrist	Right Wrist
Predicted class	Idle	/	0.816	1.099	1.728	0.656
	Left Arm	0.544	/	0.066	1.106	0.082
	Right Arm	0.231	0.019	/	0.069	0
	Left Wrist	0.122	0.418	0.016	/	0.328
	Right Wrist	0.245	0.057	0.837	0.277	/

TABLE VII. ERROR BLOCK DURATION, IN SECONDS.

		Desired class				
		Idle	Left Arm	Right Arm	Left Wrist	Right Wrist
Predicted class	Idle	/	6.04	9.23	7.34	7.58
	Left Arm	8.78	/	7.58	7.52	0.1
	Right Arm	11.98	53.90	/	11.00	0
	Left Wrist	2.18	4.05	3.40	/	7.4
	Right Wrist	2.67	1.97	7.81	5.65	/

TABLE VIII. CONTINUOUS PERFORMANCE MEASURES. PER-DOF RESULTS ARE AVERAGED ALONG ALL THE DATASET, WHILE PER-TASK RESULTS ARE AVERAGED ALONG ALL THE TASKS.

		Left Arm			Right Arm			Left Wrist	Right Wrist
		x	y	z	x	y	z		
PCC	Per-DoF	0.15	0.27	0.07	0.15	0.29	0.06	0.73	0.64
	Per-Task	0.16	0.24	0.12	0.20	0.15	0.10	0.72	0.61
DP	Per-DoF	0.02	0.11	0.01	0.02	0.12	0.01	0.68	0.56
	Per-Task	0.06	0.06	0.06	0.08	0.08	0.08	0.68	0.56

TABLE IX. DISTINCTIVENESS BETWEEN EACH PAIR OF STATES: IDLE (ID), LEFT ARM (AL), RIGHT ARM (AR), LEFT WRIST (WL), AND RIGHT WRIST (WR). EACH LINE CORRESPONDS TO THE PROJETION TO DELETE ONE OF THE TENSOR DIMENSIONS, RESPECTIVELY NONE, TIME, FREQUENCY, AND SPATIAL.

	Id/AL	Id/AR	Id/WL	Id/WR	AL/AR	AL/WL	AL/WR	AR/WL	AR/WR	WL/WR
All	7.29	9.99	6.89	9.02	12.65	5.84	13.72	15.13	4.60	14.77
Elec./Freq.	5.06	6.93	4.68	6.17	8.98	4.05	9.59	10.49	3.21	10.10
Time/Elec.	2.41	2.85	2.20	2.61	3.88	2.11	4.38	4.60	1.49	4.66
Time/Freq.	2.46	2.86	1.01	1.34	0.77	2.02	1.32	2.58	1.86	0.74

TABLE X. STABILITY COMPUTED FOR EACH CLASS.

State	Idle	Left Arm	Right Arm	Left Wrist	Right Wrist
Stability	00.37	0.040	0.040	0.037	0.038

Once the latencies are detected, the corresponding samples can be discarded to obtain the confusion matrix presented in TABLE IV. Based on this confusion matrix, discrete performance metrics can be computed. Results are compiled in TABLE V. Fig. 2 shows the evolution of each previously computed metric when the confusion matrix is normalized to verify stability of each metric towards imbalance between the classes.

TABLE VI and TABLE VII are summerizing the error block analysis results, giving the frequency of the error blocks per minute of desired class training, as well as the mean duration of each block in seconds. Numbers in grey might not be relevant, as maximum one error block occurred in our dataset.

TABLE VIII shows the different metrics for continuous performance measurements. Per-Task values are averaged along all the tasks for concisness, which loses a lot of information. For online performance measurements, each task performance must be kept.

User variability results are presented in TABLE IX and TABLE X. Distinctiveness is computed for each pair of states, in four different spaces of different dimensions (space*frenquencies*time, and each projecting space deleting one of the dimension).

VI. DISCUSSION

All these measures give us complementary information about the performances of both the decoder and the patient during a training, they can all be computed in an online manner. The results presented above are average along all the dataset for conciseness. However, their evolution during a training is an important aspect.

A. Discrete performance measures

Dissociating response time errors and real error seems to be a critical issue when evaluating self-paced BCI system to avoid bias, but it is not well spread in literature. However, it is impossible to dissociate the patient reaction time from the decoder's latency. Response time is computed in online incremental manner thanks to a Maximum Response Window (MRW) determined with past data distribution. Response time results provide important information about experiments, e.g. transitioning towards the NC and wrist control states seems to be easier than towards the arm control.

Confusion matrix based performance metrics have been chosen according to their invariance towards imbalance which is a key issue as most of them can be biased. Despite the metrics suggested in literature, our analysis reveal that only Recall, Specificity, Gmean and Bookmarker metrics seem truly invariant towards imbalance dataset whereas MCC and Kappa are more affected.

Block error frequency and duration are both critical metrics to complete the analysis by giving information about the errors dynamics. In our case, single long errors are preferred to frequent short ones, as it is less disturbing for the patient, espetialy taking into account the delay for effectors desactivation. Given this information, it might be interesting to target the less performing state transitions and train them more intensively than the others. To the best of our knowledge very few studies deal with those dynamic approach.

B. Continuous performance measures

Pearson Correlation Coefficient (PCC) and normalized Dot Product (DP) gave quite similar results. Surprisingly, limbs which present better performance results are not the same than previous discrete conclusion. By choosing continuous data only when correct discrete state is reached, redundancy between results is removed. A limb can present a poor response time but once activated can have a good control. This emphasizes the need for distinct analysis between discrete and continuous performance evaluation.

Per-task and per-DoF evaluation for PCC or DP measures are complementary measures during the analysis. While DP per-task evaluation is more meaningful during online experiment to evaluate a whole movement, the PCC per-DoF evaluation brings better discriminability along each movement. Per-task evaluation is then crucial for online evaluation and has been thought for being easily implemented in an online incremental way. Those results, either at different level of evaluation either with different metrics, allow to assess which limb needs to get improved, along which degree of freedom, or or which movement in particular. This is critical information especially for generating more adapted training targets in the purpose of automating of the whole BCI training.

C. Experiment and user skill measures

Distinctiveness and stability give us an insight of the patient's performance. In our dataset, distinctiveness measures confirm that the user has difficulties to differentiate arm and wrist limb from the same side while he has efficient ability to dissociate left and right sides. However, distinctiveness and stability metrics are based on distance measurements which can be biased in high-dimension case. Moreover, a high number of dimension induces a big computational load, which can be too

heavy depending on the hardware. Projecting to reduce the dimensionality allows to avoid this bias, and to shorten the computations. Electrodes are in our case separated equally between the two hemispheres, and each hemisphere controls movements for the opposite side (mainly). By averaging all electrodes features, left and right information is first lost, as well as most limbs separation information. When evaluating user abilities to dissociate class tasks, electrodes dimension is thus a critical one and must not be removed from the analysis. However the time dimension seems to be not as important for distinctiveness and can be deleted.

User abilities are crucial assessment measures especially to unbiased decoder performance results and dissociate real decoder performance from those involved by the patient. Those two levels of evaluation are complementary and necessary to understand each independently. This is critical information specially to control the trigger of the model updating when necessary in the purpose of automatization of the whole BCI system.

D. Conclusion and perspectives

A lot of performance measures exist in the literature. However, depending on the context, some metrics might be more adapted than others. The present study has developed a whole performance measurement for a BCI system using an online adaptive hybrid dynamic decoder. This performance system is combining different levels of evaluation to obtain a complementary and complete performance assessment, meeting the requirements fixed by the experimental setup. The main novelty is the addition of an error dynamics analysis to the discrete decoder evaluation, and the use of the dot product for continuous performance measurements.

The measure of distinctiveness and stability is at the moment limited to the discrete part. Adapting it to continuous movements requires supplementary study, but could lead to interesting results on the differentiation of the directions of movements for a given limb.

The next step towards an adaptive and smart BCI system is the study of convergence of the models. Evaluating the convergence in online experiments while the model is trained would be essential to be able to develop an automated experiment supervisor capable of estimating the learning potential of the model, thus triggering and stopping the model update according to the current measured performances.

Most of BCI system are evaluated based on fixed task difficulty level. Using the Fitt's law [13,53,58-60] would also allow to adjust the task difficulty automatically to capture performance changes at all possible levels. For instance, an adaptive staircase procedure [32] developed according to Kaernbach's formula [61] could be used to vary the difficulty of each task.

ACKNOWLEDGMENT

The authors are grateful to all members of the CEA-LETI-CLINATEC, and especially to T. Costecalde for supervising and carrying out the experiments with the patient, S. Cokgungor for supervising all the software developments, G. Charvet, J.-C. Royer, and Prof. A.-L. Benabid for managing the project.

REFERENCES

[1] M. A. Lebedev and M. A. L. Nicolelis, "Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation", *Physiological Reviews*, vol. 97, no. 2, pp. 767–837, Apr. 2017.

[2] B. Wodlinger, J. E. Downey, E. C. Tyler-Kabara, A. B. Schwartz, M. L. Boninger, and J. L. Collinger, "Ten-dimensional anthropomorphic arm control in a human brain−machine interface: difficulties, solutions, and limitations ", J. of Neural Eng., vol. 12, nº 1, p. 016011, 2015.

[3] S. He et al., "A P300-based threshold-free brain switch and its application in wheelchair control", IEEE Trans. Neural Syst. Rehabil. Eng. 25, 715–725, 2017.

[4] G. Schalk et al., "Two-dimensional movement control using electrocorticographic signals in humans", J. Neural Eng. 5, 75, 2008.

[5] A. L. Benabid et al., "An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration", The Lancet Neurology, vol. 18 (12), 1112-1122, 2019

[6] M.-C. Schaeffer and T. Aksenova, "Data-driven transducer design and identification for internally-paced motor Brain Computer Interfaces: a review", J. Neural Eng.

[7] M.-C. Schaeffer and T. Aksenova, "Switching Markov decoders for asynchronous trajectory reconstruction from ECoG signals in monkeys for BCI applications", J. Physiol.-Paris 110, 348–360, 2016.

[8] M. Spüler, W. Rosenstiel, and M. Bogdan, "Principal component based covariate shift adaption to reduce non-stationarity in a MEG-based brain-computer interface", EURASIP J. Adv. Sig. Process., 129, 2012.

[9] M. Billinger et al.,"Is it significant? Guidelines for reporting BCI performance", in Towards Practical Brain-Computer Interfaces, 2012.

[10] C. S. Mestais et al., "WIMAGINE: Wireless 64-Channel ECoG recording implant for long term clinical applications", IEEE Trans. Neural Syst. Rehabil. Eng. 23, 10–21, 2015.

[11] A. Eliseyev and T. Aksenova, "Personalized adaptive instruction design (PAID) for brain–computer interface using reinforcement learning and deep learning: simulated data study", *Brain-Computer Interfaces*, vol. 6, no. 1–2, pp. 36–48, Apr. 2019.

[12] D. E. Thompson et al., "Performance measurement for brain–computer or brain–machine interfaces: a tutorial", J. Neural Eng. 11, 035001, 2014.

[13] M. R. Mowla, J. E. Hugginsb, and D. E. Thompson, "Evaluation and performance assessment of the brain-computer interface system", Brain-Computer Interface Handbook, Chapter 33, 2018

[14] E. Thomas, M. Dyson, and M. Clerc, "An analysis of performance evaluation for motor-imagery based BCI", J. Neural Eng. 10, 031001, 2013.

[15] A. Schlogl, J. Kronegg, J. Huggins, and S. Mason, "Evaluation criteria for BCI research", Toward Brain-Computer Interfacing, 2007.

[16] M. Spuler, A. Sarasola-Sanz, N. Birbaumer, W. Rosenstiel, and A. Ramos-Murguialday, "Comparing metrics to evaluate performance of regression methods for decoding of neural signals", Conf Proc IEEE Eng Med Biol Soc, vol. 2015, pp. 1083–1086, Aug. 2015.

[17] V. Labatut and H. Cherifi, "Accuracy measures for the comparison of classifiers", ArXiv Prepr. ArXiv12073790, 2012.

[18] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations", Internat. Jour. of Data Mining & Knowledge Manag. Proc., vol. 5, no. 2, pp. 01–11, Mar. 2015.

[19] T. Kautz, B. M. Eskofier, and C. F. Pasluosta, "Generic performance measure for multiclass-classifiers", Pattern Recognition, vol. 68, pp. 111–125, Aug. 2017.

[20] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[21] R. Leeb et al., "Self-Paced (Asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic", Comp. Intel. and Neurosc., vol. 2007, 2007.

[22] S. Mason, J. Kronegg, J. Huggins, M. Fatourechi, and A. Schlögl, "Evaluating the performance of self-paced brain computer interface technology", Neil Squire Soc., Vancouver, BC, Canada, Tech. Rep, 2006.

[23] R. Ortner, B. Z. Allison, G. Korisek, H. Gaggl, and G. Pfurtscheller, "An SSVEP BCI to control a hand orthosis for persons with tetraplegia.", IEEE Trans Neural Syst Rehabil Eng, vol. 19, no. 1, pp. 1–5, Feb. 2011.

[24] J. J. Williams, A. G. Rouse, S. Thongpang, J. C. Williams, and D. W. Moran, "Differentiating closed-loop cortical intention from rest: building an asynchronous electrocorticographic BCI", J. of Neur. Eng., vol. 10, no. 4, p. 046001, Aug. 2013.

[25] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix", Pattern Recognition, vol. 91, pp. 216–231, Jul. 2019.

[26] P. Branco, L. Torgo, and R. P. Ribeiro, "Relevance-based evaluation metrics for multi-class imbalanced domains", in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2017, pp. 698–710.

[27] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets", Seventh International Conference on Machine Learning and Applications, San Diego, CA, USA, 2008, pp. 777–782.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", J. of Art. Int. research, vol. 16, pp. 321–357, 2002.

[29] L. Hakim, B. Sartono, and Saefuddin, "Bagging based ensemble classification method on imbalance datasets", Internat. J. of Comp. Sc. And Netw., vol 6, 2017.

[30] S. Wang and X. Yao, "Multiclass imbalance problems: analysis and potential solutions", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119–1130, 2012.

[31] N. D. Skomrock et al., "A characterization of Brain-Computer Interface performance trade-offs using support vector machines and deep neural networks to decode movement intent", Front Neurosci, vol. 12, Oct. 2018.

[32] 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, vol. 11, no. 2, p. 026018, Apr. 2014.

[33] F. Lotte and C. Jeunet, "Defining and quantifying users' mental imagery-based BCI skills: a first step", J. of Neural Eng., vol. 15, no. 4, p. 046030, Aug. 2018.

[34] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Adaptive multi-degree of freedom Brain Computer Interface using online feedback: towards novel methods and metrics of mutual adaptation between humans and machines for BCI", PLoS ONE, vol. 14, no. 3, p. e0212620, 2019.

[35] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain-Computer Interface classification by Riemannian geometry", IEEE Transac. on Biomed. Eng., vol. 59, no. 4, pp. 920–928, Mar. 2012.

[36] O. AlZoubi, I. Koprinska, and R. A. Calvo, "Classification of Brain-Computer Interface data", in Seventh Australasian Data Min. Conf. (AusDM 2008), Glenelg, South Australia, 2008, vol. 87, pp. 123–131.

[37] Y. Atum, I. Gareis, G. Gentiletti, R. Acevedo, and H. Rufiner, "Genetic feature selection to optimally detect P300 in Brain Computer Interfaces", Conference proceedings : ... Annual Internat. Conf. of the IEEE Eng. in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, vol. 2010, pp. 3289–92, Aug. 2010.

[38] T. H. Falk, K. Paton, S. Power, and T. Chau, "Improving the performance of NIRS-based brain-computer interfaces in the presence of background auditory distractions", 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 517–520, 2010.

[39] I. Gareis, G. Gentiletti, R. Acevedo, and L. Rufiner, "Feature extraction on Brain Computer Interfaces using discrete dyadic wavelet transform: preliminary results", Journal of Physics: Conference Series, vol. 313, p. 012011, Sep. 2011.

[40] J. Faller et al., "Non-motor tasks improve adaptive brain-computer interface performance in users with severe motor impairment", Front. Neurosci., vol. 8, 2014, doi: 10.3389/fnins.2014.00320.

[41] S. Saeedi, R. Chavarriaga, and J. d R. Millán, "Long-term stable control of motor-imagery BCI by a locked-in user through adaptive assistance", IEEE Transactions on Neural Syst. and Rehab. Eng., vol. 25, no. 4, pp. 380–391, Apr. 2017.

[42] D. C. Irimia, R. Ortner, M. S. Poboroniuc, B. E. Ignat, and C. Guger, "High classification accuracy of a motor imagery based Brain-Computer Interface for stroke rehabilitation training", Front. Robot. AI, vol. 5, 2018.

[43] G. Hotson et al., "Individual finger control of a modular prosthetic limb using high-density electrocorticography in a human subject", J. of Neural Eng., vol. 13, no. 2, p. 026017, Apr. 2016.

[44] C. Vidaurre, A. Schlögl, A. Schlöogl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "A fully on-line adaptive BCI", IEEE Transactions Biomed. Eng., vol. 53, no. 6, pp. 1214–1219, Jun. 2006.

[45] J. Faller et al., "A co-adaptive brain-computer interface for end users with severe motor impairment", PLOS ONE, vol. 9, no. 7, p. e101168, Jul. 2014.

[46] I. Žliobaitė, A. Bifet, J. Read, B. Pfahringer, and G. Holmes, "Evaluation methods and decision theory for classification of streaming data with temporal dependence", Mach. Learn., vol. 98, no. 3, pp. 455–482, Mar. 2015.

[47] Mosley, L. "A balanced approach to the multi-class imbalance problem", Graduate Theses and Dissertations, 13537, 2013.

[48] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient", Computational Biology and Chemistry, vol. 28, no. 5–6, pp. 367–374, Dec. 2004.

[49] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme.", Biochim Biophys Acta, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[50] Chicco, D. "Ten quick tips for machine learning in computational biology". BioData Mining 10, 35, 2017.

[51] S. Boughorbel, F. Jarray, and M. El-Anbari, 'Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric', PLOS ONE, vol. 12, no. 6, p. e0177678, Jun. 2017.

[52] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Prediction of arm movement trajectories from ECoG-recordings in humans", J. of Neurosc. Methods, vol. 167, no. 1, pp. 105–114, Jan. 2008.

[53] J. D. Simeral, S.-P. Kim, M. J. Black, J. P. Donoghue, and L. R. Hochberg, "Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array", J. Neural Eng., vol. 8, no. 2, p. 025027, Mar. 2011.

[54] D. Sussillo, S. D. Stavisky, J. C. Kao, S. I. Ryu, and K. V. Shenoy, "Making brain–machine interfaces robust to future neural variability", Nature Communications, vol. 7, p. 13749, Dec. 2016.

[55] J. M. Carmena, M. A. Lebedev, C. S. Henriquez, and M. A. L. Nicolelis, "Stable Ensemble Performance with Single-Neuron Variability during Reaching Movements in Primates", J. Neurosci., vol. 25, no. 46, pp. 10712–10716, Nov. 2005.

[56] I. S. MacKenzie, T. Kauppinen, and M. Silfverberg, "Accuracy measures for evaluating computer pointing devices", Proceedings of the ACM Conf. on Hum. Fact. in Comp. Syst. - CHI 2001, pp. 9-16.

[57] A. Eliseyev et al., "Recursive Exponentially Weighted N-way Partial Least Squares Regression with Recursive-Validation of Hyper-Parameters in Brain-Computer Interface Applications", Scientific Reports, vol. 7, no. 1, p. 16281, Dec. 2017.

[58] R. D. Flint, Z. A. Wright, M. R. Scheid, and M. W. Slutzky, "Long term, stable brain machine interface performance using local field potentials and multiunit spike", J Neural Eng, vol. 10, no. 5, p. 056005, Oct. 2013.

[59] E. A. Felton, R. G. Radwin, J. A. Wilson, and J. C. Williams, "Evaluation of a modified Fitts law brain–computer interface target acquisition task in able and motor disabled individuals", J. of Neural E., vol. 6, no. 5, p. 056002, Oct. 2009.

[60] V. Gilja et al., "A high-performance neural prosthesis enabled by control algorithm design", Nature Neuroscience, vol. 15, no. 12, pp. 1752–1757, Dec. 2012.

[61] C. Kaernbach, "Simple adaptive testing with the weighted up-down method", Attention, Perception, & Psychophysics, vol. 49, no. 3, pp. 227–229, 1991.