Deep Learning personalised, closed-loop Brain-Computer Interfaces for multi-way classification

Pablo Ortega[1], Cédric Colas[2] & Aldo Faisal[3]

***Abstract*— Brain-Computer Interfaces are communication systems that use brain signals as commands to a device. Despite being the only means by which severely paralysed people can interact with the world most effort is focused on improving and testing algorithms offline, not worrying about their validation in real life conditions. The Cybathlon's BCI-race offers a unique opportunity to apply theory in real life conditions and fills the gap. We present here a Neural Network architecture for the 4-way classification paradigm of the BCI-race able to run in real-time. The procedure to find the architecture and best combination of mental commands best suiting this architecture for personalised used are also described. Using spectral power features and one layer convolutional plus one fully connected layer network we achieve a performance similar to that in literature for 4-way classification and prove that following our method we can obtain similar accuracies online and offline closing this well-known gap in BCI performances.**

I. INTRODUCTION

Brain-Computer Interfaces (BCI) are communication systems that allow our brain to directly communicate with the external world [1]. For many severely paralysed users, such as those suffering from Spinal Cord Injury, Multiple Sclerosis, Muscular Dystrophy or Amyotrophic Lateral Sclerosis it constitutes often the only way to interact meaningfully with their environment.

While invasive methods have made considerable progress [2], these remain experimental, exclude considerable number of potential end-users due to costs and medical risks, and require brain surgery and subsequent interventions to manage implanted neurotechnology [3]. This is why non-invasive approaches remain at the forefront of practical deployed neurotechnology for paralysed users, with electroencephalographic (EEG) recordings being the most prominent exponent. To date EEG decoding approaches have been mainly performed as either off-line classification challenges, or as clinical neuroengineering challenges in closed-loop settings with dedicated patient end-users. Furthermore, BCI technology is tailored to very specific tasks, equipment and end-users – and occasionally how studies were controlled for artefacts. This broad range of different approaches hinders the objective comparison across studies and algorithms. In these circumstances the Cybathlon provides with a unique and equal setting for evaluating and proposing different BCI approaches focusing on end-users requirements, specifically algorithms that work on-line responding to user's intent in short time. The BrainRunners game (figure 1) for the Cybathlon's BCI-race [4] requires four commands whose decoding accuracy determines the velocity of an avatar facing an equal number of obstacles indicated by a color. Currently, several machine learning techniques are used to decode user's brain signals into commands to control the devices that carry out the designed actions. Deep Learning (DL), as one of them and despite its successes in other fields, remains little explored for BCI use due to its high computational demands. In particular, the deeper the architecture the longer the training and decoding times, and the greater number of examples required.

Fig. 1. Snapshot from pilot's point of view of BrainRunners video-game. Right bottom corner shows the EEG set up on our pilot. Each avatar correspond to a user competing in the race. Each obstacle is indicated buy a different color: cyan for avatar rotation; magenta for jump; yellow for slide; and gray for no-input. The control is achieved by decoding different mental tasks associated with each desired command. Source: *Cybathlon BCI Race 2016*.

Several DL approaches have been proposed for BCI-EEG [5], [6], [7] but most of them limit to off-line analyses. DL for EEG signals decoding for clinical studies has also been recently used in [9], [10] but the architectures provided are computationally demanding and would render real-time decoding unsuitable.

Focused in on-line BCI use constraints we (1) investigated different Convolutional Neural Networks (CNN) architectures that led us to select a simple one –or SmallNet– made of one convolutional layer, one fully connected layer and a logistic regression classifier layer. To overcome the

[1]P. Ortega currently works on his PhD in the Department of Computing at the Imperial College London, United Kingdom. po215@ic.ac.uk

[2]C. Colas is currently doing an internship at the Brain and Spine institute in the Motivation, Brain & Behavior team in Paris, France. cedric.colas@icm-institute.org

[3]Dr. Faisal Dr Faisal is a Senior Lecturer in Neurotechnology jointly at the Dept. of Bioengineering and the Dept. of Computing at Imperial College London, United Kingdom. a.faisal@ic.ac.uk

reduced abstraction capabilities of our architecture, we then (2) explored different preprocessing strategies that reduced the complexity and size of the network input compared to the raw signal. In particular we analysed spectral power features preserving the spatial arrangement of electrodes and compared them to time series. Third, (3) we exploited topographical and spectral differences of EEG activities related to 8 different mental tasks that one volunteer found easy to perform to find a combination of four rendering better classification accuracies. Finally, based on results from previous steps (4) we carried out co-adaptive training which has been reported to require less time than off-line training [12]. Ideally we would have wanted to test the performance of every combination of architectures, imageries and features in the adaptive training setting. However the time for one subject to participate would have been excessively long and the variability of brain signals across time would not have made possible to compare results. To the best of our knowledge this is the first time a CNN has been tested in on-line conditions and four classes. In [8] a similar approach is taken but the CNN used consists of four layers for a binary classification and there is no on-line testing of the architecture proposed. Our main contribution consists on the design and implementation of a BCI based on a very simple CNN architecture achieving over random accuracies on 4 classes in real use conditions, establishing a baseline for DL real-time BCI implementations within the standarised framework of Cybathlon.

The remaining sections of this paper are organised in parallel to those four stages of our approach. We start describing the general aspects of data acquisition, and the methods to perform the described analysis. Continuing with the results of our proposed approach. And concluding with the discussion of the results and the limitations.

II. METHODS

As mentioned the four stages of our approach are highly dependent and also dependent on the state of mind of the subject. Because of the long time such a study would require and the different mental states subjects can present along it, it would be impractical to record on several ones all the required variations of the analysis, and even then, results would not be comparable as the architecture, imageries and features working for one subject may not work for the others or at different times. Instead we separated stages 1, 2 and 3 of our analysis by choosing a set of imageries, features and architecture that were working well and then did independent modifications over each variable fixating the others. Although limited by this constraint, this approach would allow us to give a first proof of concept for real-life use of a 4-class BCI system based on a CNN.

A. Generalities

Data was recorded using a BrainVision ActiCHamp® (v. 1.20.0801) recorder with filters set to 0.1Hz (high-pass) and 50Hz (notch filter) at a sampling rate of 500Hz. 64 electrodes were placed using the 10-20 system using Fpz as reference. Electrooculogram (EOG) activity was recorded on the right eye to correct for ocular artifacts using independent component analysis (ICA) from the *MNE* Python toolbox [13], [14]. A 28 year old, right handed man volunteered throughout all the stages.

CNN architectures were implemented in Theano [15]. The input size to the convolutional layer (CL) depended on the preprocessing methods applied to the raw EEG. The BCI was built in an Intel i7-6700 CPU at 3.40GHz using a NVIDIA 1080GTX.

B. Architecture selection - Stage 1

The input feature selected for this stage was *pwelch-allF-grid* as explained in section II-D and same data-set was used for this analysis. Each example consisted of a tensor with the first dimension corresponding to 129 spectral power points and second and third dimensions to an approximate grid representation of the positions of the electrodes in a 2D projection. Architectures are presented in figure 2.

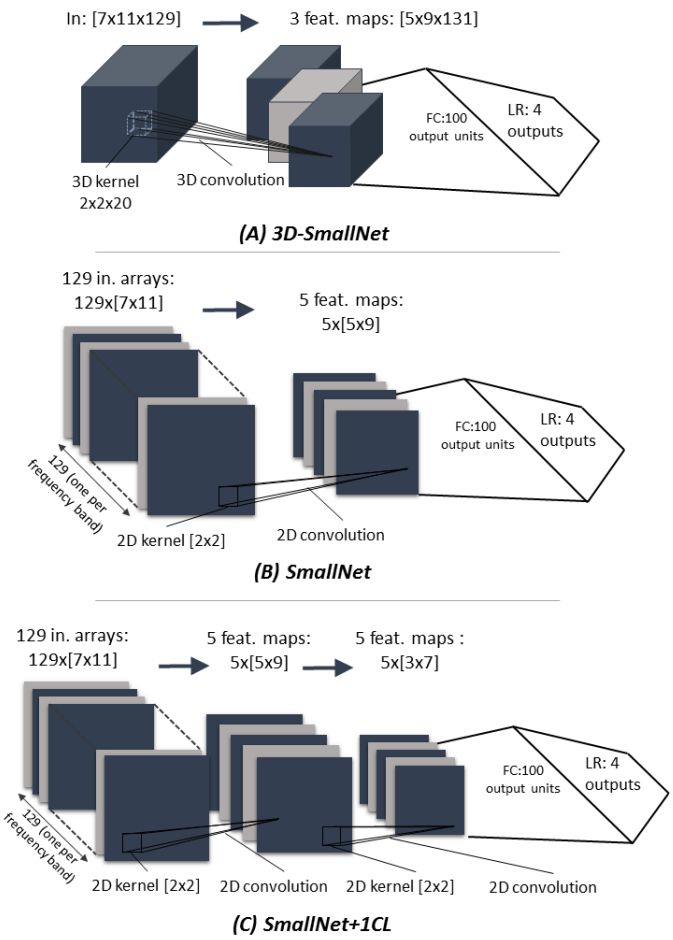

Fig. 2. Three different CNN architectures: A 3D-SmallNet (A) using 3D convolution instead of the 2D convolution used by SmallNet (B). A convolutional layer was added to Small-Net (C, SmallNet+1CL) and also a fully connected after the first convolutional one not presented in the figure (SmallNet+1FC).

This way, different complexities of CNN were tested with

the intention of finding the simplest one able to abstract enough information from our limited set of examples. In addition, SmallNet was tested using *tanh* or RELU activations functions. For each run, the weights were randomly initialised following a uniform distribution within the [-1,1] range.

C. Mental tasks evaluation - Stage 2

People control differently well BCIs due to different performance in mental tasks [16]. We devised this stage to find the mental tasks that were better classified by SmallNet. The subject made an informed decision and chose eight mental tasks he felt comfortable performing. Figure 3 depicts the different cognitive processes that each selected mental task should entrain. It also conveys the idea of the separability sought, not in the feature space but, in a qualitative manner, in categories of brain activities.

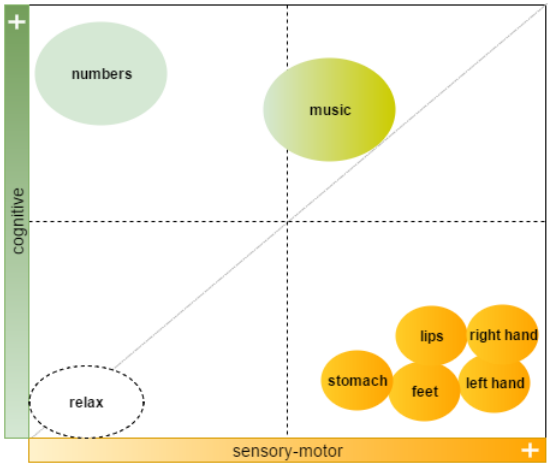

Fig. 3. Selected mental tasks represented in a qualitative space. *Numbers* should entrain higher cognitive processes related to arithmetics. *Music* can entrain auditive sensory activities and motor activities if the subject imagines himself signing the song. Long term memory as cognitive process also plays a role in recalling the sounds. *Motor imageries (stomach, lips, feet, right hand, left hand)* can entrain both sensory and motor activities depending on whether the subject imagines the movement or the sensations that produces. The cognitive side can be represented as muscular memory. *Relax* entrains idle activity in both types of categories.

The 8 mental tasks as chosen by the volunteer consisted in:

- **Music**. Recalling the 3 first seconds of Hendrix's *Little Wing*. Notice that no lyrics are included, thus no language related cortical processes are expected.
- **Lips**. Imagining contracting and relaxing the lips repeatedly.
- **Relax**. Leaving mind in blank and visualise a bright light.
- **Numbers**. Randomly choosing 3 numbers and subtracting the last one to the remaining cipher.
- **Left or right hand (RH or LH)**. Open and close either left or right hand, constituting each a different task.
- **Stomach**. Imagine contracting the abdominals.
- **Feet**. Imagine contracting the sole.

The experimental paradigm for data acquisition was devised to acquire time-locked examples of the brain activity related to each mental task that could be as free of other stimuli as possible: A fixation cross appeared for 1s and its disappearance indicated when the mental task should be performed. For 5 seconds the mental task was performed and only the 3 first seconds were used for analysis. After each 16 trials the subject was given the option to have a resting period of any desired length. 100 examples were recorded for each imagery, giving a total of 800, and an approximated experimental time of 2 hours that were split in two sessions the same day. Instructions were randomly ordered using a uniform distribution so that an equal number of them were present in both sessions.

The input feature used was the same as in previous stage. ICA correction was performed on the data from each session separately to avoid non-convergence issues due to data discontinuity. Epochs were extracted from continuous EEG data 0.2 seconds before to 3 seconds after the cue.

The 8-choose-4 combinations of imageries (70 combinations) were used to train the same number of SmallNet models using a 4-fold strategy resulting in 4 measures of test classification error for each combination of imageries. This k-folding strategy was used to avoid any effect on inhomogeneous chronological presentation of examples.

The random initialisation of weights led to starting points more beneficial for some executions than others, leading to unstable results. To overcome this, we run the previous procedure three times, gathering enough data to perform statistical analysis on differences of test error means.

For the training, a batch size of 5 examples, a learning rate of 0.03, and 150 epochs were used. The rest of the parameters for SmallNet were given by the *pwelch-allF-grid* preprocessing, i.e. filter shape $[3, 3]$ and number of kernels 3, as explained in next section.

D. Preprocessing strategies - Stage 3

Figure 4 shows the preprocessing pathways that were followed by the raw data aiming to reduce input dimensionality and enhance features that would help SmallNet better discriminate among mental tasks. In total 28 different preprocessing strategies were analysed.

Energy features extracted spectral power from the signal at different frequencies or frequency bands. CSP or PCA were tested as filtering techniques.

Finally energy features belonging to each EEG channel were spatially organised in two types of images. Both represented a 2D projection of the 3D position of the channel on the scalp. The third branch (blue) is reserved only for time, where there is no spatial reorganisation. At this point each of the coloured paths convey a selection of different filter shapes according to the shape of the resulting input-feature.

In this case 20 videos of different races using the BrainRunners game were recorded and presented to the subject. During each video the subject performed his four preferred mental activities. 18 obstacles composed each race and were homogeneously distributed on each video.

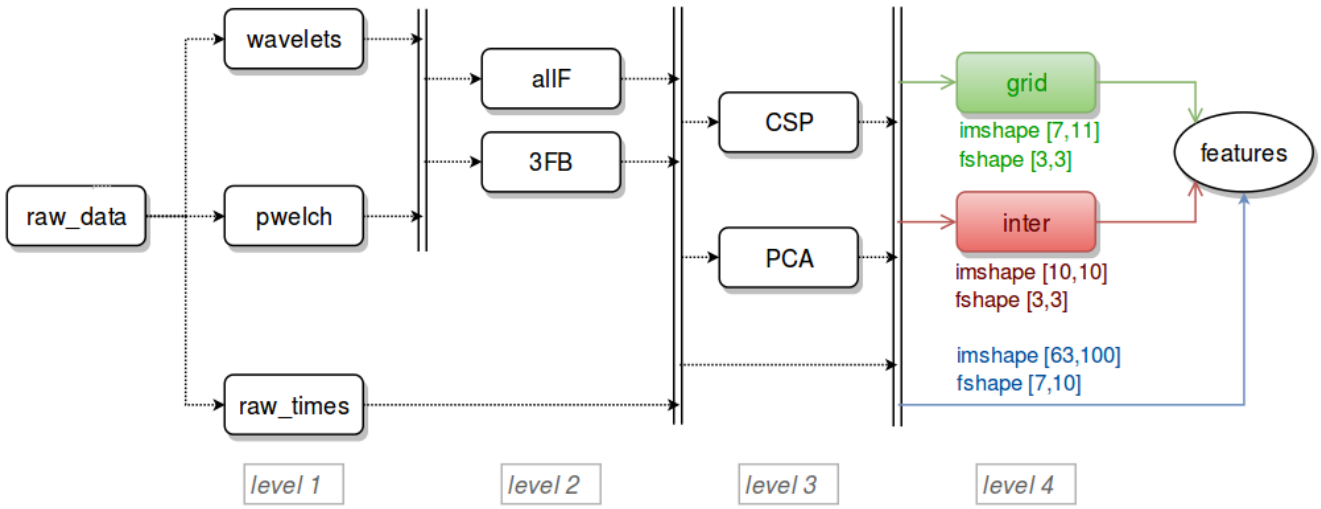

Fig. 4. Preprocessing pathways. The preprocessing strategies are divided in 5 processing levels. Starting with raw EEG time series examples ($[63, 600]_{chn \times samps}$) the first preprocessing level consisted on extracting energy features or downsampling the time series (*raw*), either wavelets, Welch periodograms or none are computed. In the second level, if any of the first two is computed the next step consist in averaging the power spectrum in three frequency bands or keeping all the frequencies. In level 3 common spatial patterns, principal component analysis or none of them are applied in order to enhance signal to noise ratio. Finally, spectral features are placed in a tensor with the first two dimensions representing the 2D projection of scalp localisation and the third the frequency or frequency band the power is related to. In the case of the grid each electrode is represented by a point. The interpolation technique performed as in [9] interpolate the values of the 2D projection and find the interpolated values for the positions in the grid. Otherwise, voltage time-series are just ordered in a [channel] x [time] matrix. The CNN input and filter shape varies depending on the preprocessing strategy as represented by each color in the last stage.

Examples were extracted the same way as in real time. Epochs of $1.2s$ were extracted with an overlap of 75%. Finally artifacts were corrected using previously computed ICA matrices. Following this approach 8736 examples were extracted.

Two runs of a 4-folding strategy were used to sample the measures of interest. Due to the instability caused by random initialisation we proceeded to acquire more data only on the 4-th fold, running the algorithms eight times, thus rendering a total of 10 runs for the 4th-fold if we consider the two 4th-folds results of the first two runs.

E. Co-adaptive training design - Stage 3

The last stage sought to validate and analyse the results in real conditions. And secondly, investigate an adaptive on-line training providing feedback to the user. The importance of this kind of co-adaptation has been already addressed emphasizing the relevance of a BCI being able to adapt to different states of mind and fatigue along its use [17].

Validation stage. To evaluate the on-line use the following strategy was devised. First, a recording session of 20 videos, similar to that in the previous section, was used in the same manner to train SmallNet (Fig. 5). A second session, immediately after, used that model for on-line decoding during five races made of 20 pads. Both the race time, and two decoding accuracies were used to analyse the results. One decoding accuracy (acc_1) considered the label corresponding to the pad where the EEG data was generated and the other (acc_2) considered that the decoded label arrived in the correct pad. They could differ if a long decoding time prevented the decoded label for one data portion to arrive on the pad it

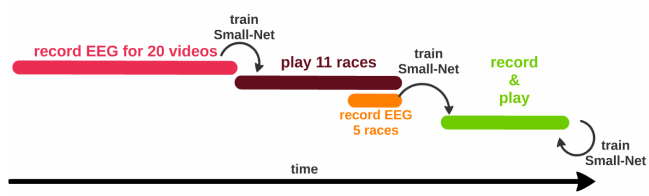

Fig. 5. Non-adaptive (warm colours) and adaptive training (green) strategies. At the beginning of the session EEG activities are recorded during 20 videos. These were used to train the SmallNet and this model used to play for 11 races. EEG data in playing conditions from the last 5 races were recorded and used to retrain the model to start the adaptive training. After each adaptive training race data was recorded, appended to those 5 races and the model retrained.

was generated. The same previous training parameters were used in this stage.

Co-adaptive training. An old model was used to decode the commands for the first race, meanwhile the data generated was stored and used to train and validate another model for posterior races. This second model was updated after each race appending to previous examples the new data generated. A limit of 2000 was used to train the model dropping data from the oldest race each time a newer one was appended.

III. RESULTS

A. Architecture selection - Stage 1

A first analysis conducted using a 5-fold strategy allowed us to discard 3D-SmallNet due to its long training time, making it impractical for an adaptive on-line approach. For the rest of the architectures, a 5-fold strategy was run 5 times per architecture to control for the effect of different initialisations on accuracy results (Table I). A Friedman test

showed that only SmallNet was significantly better than its RELU variant. Because there was no clear benefits of adding more layers, we used SmallNet as it was the one requiring the shortest training time.

TABLE I
TEST ACCURACY (TA) OF RELEVANT ARCHITECTURES

	mt_4	TA_{avg}	TA_{std}
1	SmallNet	58.21	1.34
2	SmallNet+1CL	53.52	8.56
3	SmallNet+1FC	58.89	1.92
4	SmallNet-RELU	53.84	1.49

B. Mental tasks evaluation - Stage 2

Figure 6 shows the test errors (TE) for each mental task combination. A Kruskall-Wallis test revealed significant differences ($\alpha = 0.01$). Table II shows the number of different TE for some pairwise comparisons, corresponding to those groups with lower TE mean than the group they are compared to. Combination 19 was the one preferred by the user. Above rank 30th no more differences were present.

TABLE II
PAIRWISE COMPARISONS OF IMAGERIES COMBINATIONS

	im1	im2	im3	im4	# sig. diff.
1	RH	feet	lips	numbers	18
2	RH	feet	relax	numbers	16
3	RH	feet	relax	lips	16
4	RH	relax	lips	stomach	11
5	RH	feet	stomach	numbers	11
6	RH	feet	music	numbers	10
7	RH	relax	stomach	numbers	10
8	RH	relax	lips	numbers	10
9	LH	feet	relax	lips	8
10	RH	LH	relax	lips	6
18	LH	feet	relax	stomach	3
19	**RH**	**feet**	**relax**	**music**	**3**

To keep our user-centred design and given that there was no best combination in absolute terms, i.e. one significantly better than all the rest, we let the volunteer choose the one which he felt more comfortable when playing: **RH-feet-relax-music**, which presents some advantages over 3 groups and is not statistically different from any combination above it.

C. Preprocessing strategies - Stage 3

Figure 8 presents test errors (TE), execution and preprocessing times. In addition, same figure shows the time SmallNet requires to run 400 epochs of training.

Differences among groups were significant and we proceeded to analyse pairwise comparisons (Fig. 7). As we can see spectral frequencies (not band grouped) are ranked in the top eleven positions and information processing steps (PCA and CSP) do not seem to offer an improvement.

Again, differences among some groups are observed but there is not a single preprocessing strategy better than all the others. Although it is clear that time series perform significantly worse than the rest, justifying the use of spectral features preprocessing in later analytical stages. To clarify the situation we fixed one variance degree of freedom by using always the 4th-fold to test the model and we rerun the training 10 times. This way only initialisation effects affected how well each feature represented the information. Notice that except CSP and PCA, which are computed for the training data and then applied to the test, the rest of strategies do not depend on the quality of the data they are applied to.

A similar to the previous analysis allowed us to determine that features including only 3 frequency bands performed worse than features with higher frequency resolution (i.e. *allF*). However among *allF* there was not a clear winner as differences were not statistically significant among those including CSP, PCA or none of them. Differences were neither found among the two different steps to construct the topological images of brain activity. As a result, we chose *pwelch-allF-grid* as feature given its stable behaviour during early stages of system assessment.

D. Co-adaptive training - Stage 4

Given the off-line results of our analysis we tested the on-line system with **SmallNet**, **RH-feet-lips-numbers**, and **pwelch-allF-grid**.

Figure 9 shows the results for the on-line session. Acc_1, and acc_2 are the validation accuracies measured as explained in methods, while acc_{test} is the test accuracy of the model used to decode during playing conditions. First, the model trained with videos (which yielded 54.5% of test accuracy) was used to decode brain signals for the first 11 races. Starting at race 7, examples extracted during actual playing were saved and later used (after race 11) to train a new different model only using on-line data. Note that from races 7 to 11 data is only appended but the model used is still the one trained with data from videos. This is to gather enough examples to start training the model which is randomly initialised at the beginning.

Several tests were carried out analysing aspects related to the on-line function. First, no differences existed between the two methods to measure accuracies: acc_1 and acc_2. In terms of accuracy there is no differences between the adaptive and the non-adaptive training ($p = 0.38$). Indeed both accuracies yielded by both strategies (adaptive and non-adaptive) correlate similarly to the time invested on each race. Their Pearson's coefficients (accuracy, time) being, -0.385 and -0.388 respectively, thus both carrying a similar reduction in time when accuracy is increased (Figure 10).

Finally, the aspect that concerns us most is how well test accuracies predict the on-line behaviour of the system. While test accuracies for adaptive training were equal to those achieved on-line, the accuracies achieved during off-line training were systematically higher than those achieved in using the off-line trained model in real time. This supports the idea that off-line reported results often overestimate the achievable real time performance.

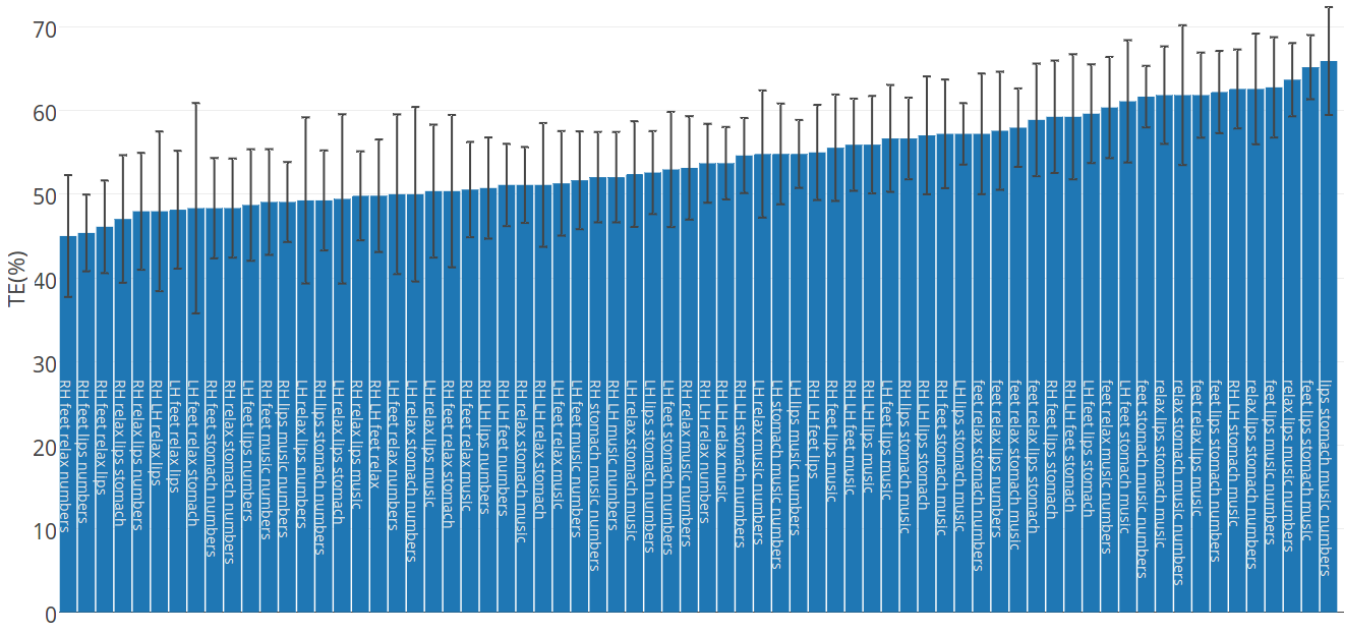

Fig. 6. Test Error (TE) mean and standard deviation of the $[4-folds]x[3-runs]$ i.e. 12 samples per combination. TE refers to the percentage errors SmallNet makes classifying the test-set of examples. Random guess for four classes correspond to 75% TE.

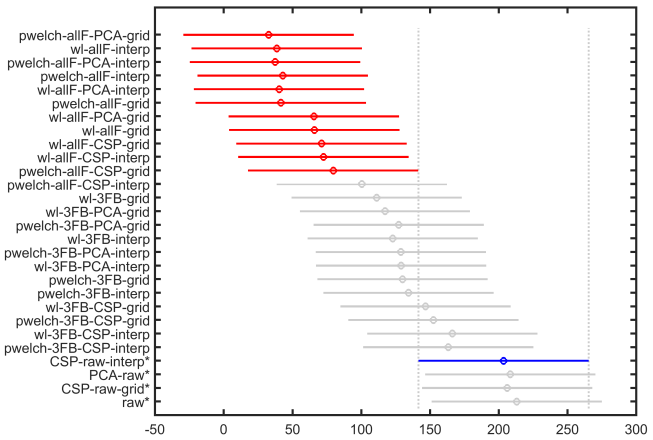

Fig. 7. Pairwise comparison of preprocessing strategies. Horizontal axis, mean ranks. Data set extracted from videos recordings using Rh-feet-relax-music imageries. (*)Rejected strategies with significant higher test error compared to top-11 results.

IV. DISCUSSION

A. Architecture selection - Stage 1

a) Discussion: For the first time, we have showed the capabilities of a simple CNN to distinguish among four different brain activities in a single example basis, achieving significant over random accuracies ($> 25\%$). We tested variations departing from a very simple CNN architecture in order to find if the increase in computing complexity could led to higher accuracies. For our particular set-up and data-set we found that SmallNet being the simplest model provided similar accuracies than more complex ones justifying its use.

b) Limitations: Finding optimal hyperparameters for a CNN is a complex task. In our case the search of these hyperparameters would have been affected as for each imagery combination and feature they may have been different. As mentioned in the beginning we decided to constraint our search space and thus our results should be considered from this perspective.

B. Mental tasks evaluation - Stage 2

a) Discussion: We showed that the mental tasks composing the combination are determinant in this accuracy. In particular, we offered evidence suggesting the advantages of including a more diverse range of brain activities not only consisting of motor imageries. This is to say, none of the top-10 most distinguishable combinations included only motor imageries. More interestingly, only one combination of this top-10 included *LH-RH* together, combination which has been extensively used for binary BCI implementation.This does not defy previous research supporting consistency of only motor imageries (e.g. [19]) as these results are only valid for the volunteer under study, but it challenges the assumption of using only motor imageries for BCI implementations. We consider BCIs should be user-specific and that the brain activities should be regarded as another design parameter in the system following the Graz-BCI approach [18]. Ideally BCI systems should be able to adapt to any mental task that a subject chooses to perform to control it which is in alignment with [8].

In conclusion, two main aspects should be considered in future works: first, the quality and consistency of each brain activity along time, represented by its absolute effect on test accuracy; and second, how distinguishable the combination of 4 of these are in the feature space, as it may happen that

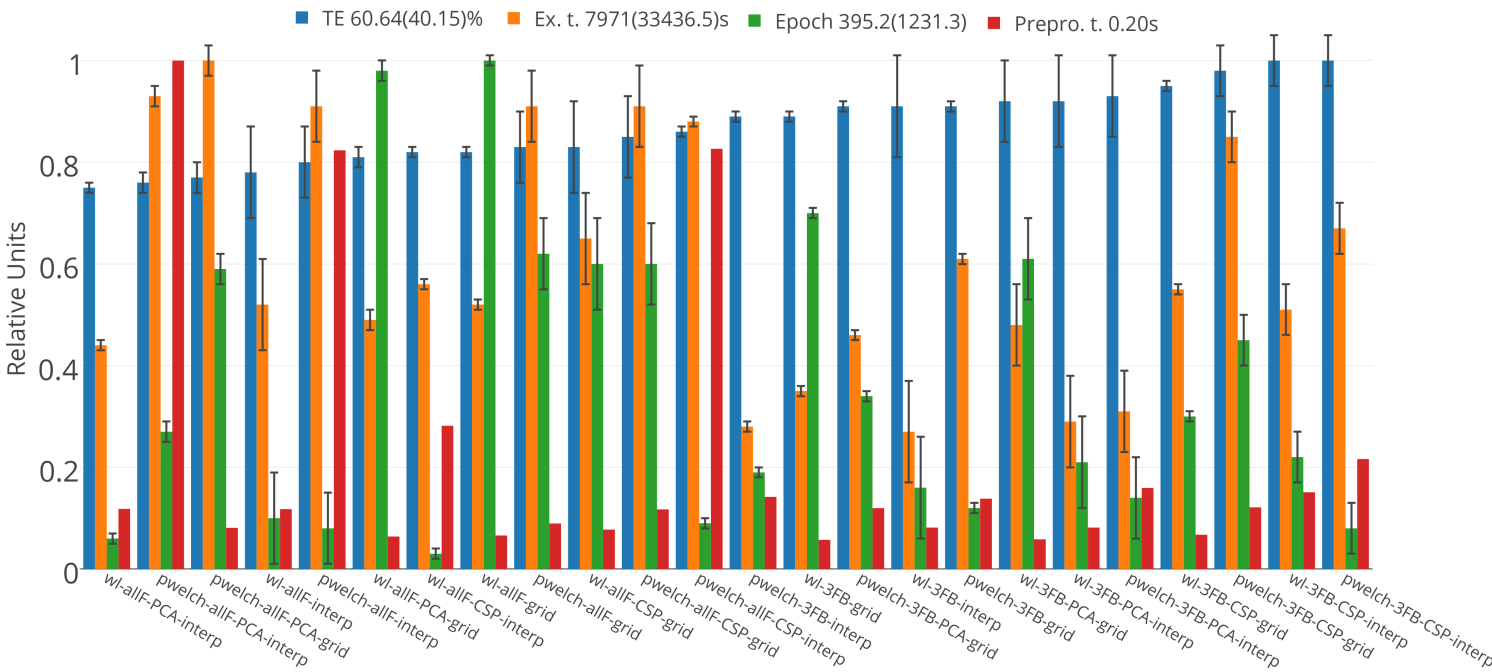

Fig. 8. Relative comparison of parameters of interest of preprocessing strategies using data from 4th-fold runs. Values in Y-axis are relative to values in the legend, mean and standard deviation, in brackets, respectively. Ordered in test error (TE) ascending order. Ex. t., execution time. Prepro. t. preprocessing time. I.e. the processing time required to convert one example of input raw data to each type of input to SmallNet.

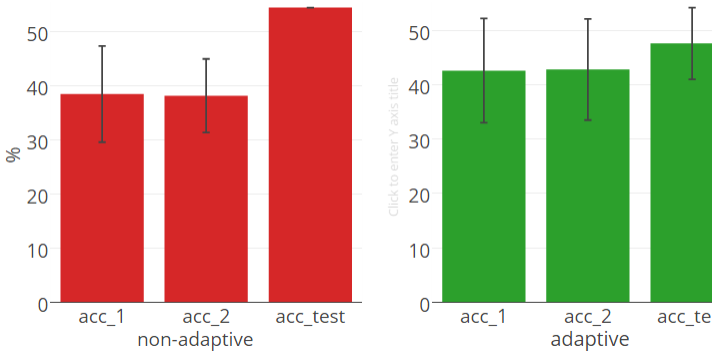

Fig. 9. Time VS validation accuracy. The figure shows the results in decoding accuracy (acc_2) during the game and the time required to finish the race. A linear fit is applied to show the correlation between both values.

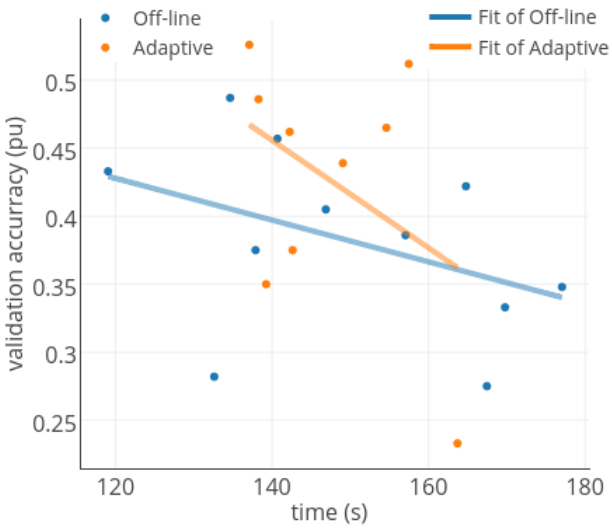

Fig. 10. On-line playing accuracies. Playing results for the model trained with off-line data (non-adaptive) are presented in red. Those ion green corresponding to the model trained with data gathered in actual playing conditions (adaptive).

by itself one imagery conveys overall good accuracies but when combined with others they would overlap too much to be easily distinguishable.

b) Limitations: First, the experiment we proposed under controlled conditions may not be as representative of the state of mind during the game as we would expect. During the experiment, where only instructions and black screens were presented, the volunteer can focus on imageries more comfortably than during the game or videos where visual stimulation is richer and execution of activities, in correct pad and in pads transitions is trickier. Secondly, we should consider the relevance of our results over a long period of time. We recorded and assessed brain activities one day and presumed they would not have changed days after when the BCI was used. Conducting this assessment before each BCI session would however increase set-up time beyond acceptable levels. Finally we only used one subject and it would be interesting to contrast these results with a greater population.

C. Preprocessing strategies - Stage 3

a) Discussion: Our first study of input-features evidenced that spectral energy features were, regardless of the data fold and the initialisation of the network weights, better

than any of the *raw-time* features. Spectral energy features in channel space have been classically used to characterise and study brain activities in several frequency bands, as they have been found to enhance statistical differences. Our second level study using data from only the 4th-fold presented enough evidence to reject frequency band features as worse performers than those including higher spectral resolution, and a posterior analysis studying differences among three band frequencies and high frequency resolution supported this point. Note that classically the frequency bands have been identified as those were populations under study presented a peak of power that changes either in intensity, frequency or channel location for groups of brain activities. However, each subject present their particular peak frequencies and intensity values, thus averaging across three frequency bands might hinder changes in peaks across brain activities.

Same analysis showed that CSP, which has been extensively used for LH-RH binary BCIs does not convey any relative improvement for any preprocessing strategy. This is not as surprising if we consider that this technique applies for binary source location and we extended it doing 4 one-vs-others CSP filters. On the one hand, this increased the input size four times, challenging the ability of the SmallNet to make an abstraction of the information with few units and from few higher dimensional examples.

As we can see the feature space we can research is enormous and the interactions complex. Nonetheless we can use other parameters, different to the test error, to select them as the time they need to converge to the best test accuracy, the time they require to preprocess the raw data and the time they would need to decode an example, or how much they overfit to the data (indicated by the difference between train and test error). In particular the interpolation technique led to fewer number of epochs and did not present statistical differences in test error results compared to PCA nor the grid technique. Having a consistent measure of how fast a CCN converge is very useful to have an idea of how much time it will take to train and therefore to select those features that lead to shorter training times and reduce the time that the user needs to invest prior to the use of the system. However we finally considered that *pwelch-allF-grid* was a good choice given its simplicity and low computational demands, as the interpolation slowed down the execution without offering clear advantages in accuracies.

b) Limitations: It is important to notice that all results and conclusions presented here are to be considered only under the framework of the SmallNet use as decoder, thus results are not generalisable. It may happen that a strategy is very good at extracting relevant information but yields (as for CSP) a greater input dimensionality exceeding the capacity of SmallNet to learn. Indeed, the results are also relative to the way each feature extraction method has been implemented. We tried to keep their parameters as equivalent as possible, though modifying them would certainly affect results. In particular *wl-allF* are still computed for frequency bands which is inherent to the Daubechies wavelet decomposition employed. Although 9 bands are included thus having higher spectral resolution compared to *wl-3FB*.

D. Co-adaptive training validation - Stage 4

a) Discussion: We validated the system using the volunteer preferred imageries (*RH-feet-relax-music* in the top-19) and *pwelch-allF-grid* input-feature (in the top-11 of features results for all folds), since our off-line analyses did not showed relevant differences across parameters. In particular we found that adaptive test accuracies offered a more reliable prediction of validation accuracies during playing. A reasonable explanation for this is that adaptive training uses EEG brain activity recordings that are produced in playing conditions, the same conditions that are present when the model is used to decode. Conversely, training the model with videos EEG recordings yielded better test accuracies in data sets with more examples but less representative of the brain activity during actual playing. We demonstrated that for a CCN based BCI, adaptive training can achieve same performance as off-line training while engaging the participant since the beginning.

b) Limitations: To this respect, one of our main concerns is the relevance of the ICA matrix along time. For on-line playing we computed it only once at the beginning. Therefore we consider that how good this ICA matrix is at correcting eye artifacts may change along time and not be as representative for future eye activity sources as time passes. A possible solution would be to recompute it after each race in the adaptive training just before training the model. However, differences in convergences of the matrices across races may introduce some differential distortions in the back projected data that hinder the ability of the network to effectively learn. Finally, as more races are used in the adaptive training more examples are available for the net to learn and more time the training takes. We already discussed in the introduction the importance of shorter set-ups and we consider relevant to study the effect of limiting the number of examples to train the network in order to stablish a limit in training time after each race making the user experience smoother. Another concern is the stability of the model along races. Although overall performance is good from race to race there the transition in accuracies were not smooth. We think it is important to understand how models can be transferred in a race to race basis and find if this is caused by user focus rather than model or recording instabilities. Finally and as previously mentioned, our major concern relates to how off-line results can be translated to on-line ones. Our off-line analyses gave us some clues about what architecture, imageries and features we could use. Following them we selected a set that were presenting early stable results on-line. However, in a later experiment, the results for **SmallNet**, **RH-feet-lips-numbers**, and **wl-allF-grid** gave worst performance though being similar in off-line results to the set reported. The user being manifestly fatigued during that session could have been the cause as both off-line and on-line results were bad ($TE \approx 65\%$). This reveals the complexity of the interactions and human limitations of such

a study searching optimal settings. Nonetheless our on-line results for 4 classes are comparable to those reported in [19] where only half the participants achieved around 65% of accuracy in a more time-locked experiment compared to our user-driven real-time set. In [20] using a real-time binary BCI the drop in on-line accuracy compared to off-line was $\sim 16\%$ in average compared to 29.9% for our non-adaptive training and 8.8% for adaptive. To the best of our knowledge, CNN and DL designed for BCI have been only applied off-line and its real-life use would correspond to a non-adaptive training with the consequent drop in accuracy shown.

V. CONCLUSION

In a context where a consistent and standard benchmark for testing is lacking we have presented a rational approach to the design and implementation of a BCI fulfilling the Cybathlon competition requirements. We identified two usual weaknesses in BCI and DL as decoding technique: (1) in most BCI conceptions users performance evoking a richer range of brain signals is only summarily considered; and (2) in general most research involving DL focus on off-line evaluations of decoding accuracy, leading to complex architectures whose impact in real life conditions is rarely discussed. We took these improvement opportunities to design an experimental protocol based on a CNN selected by its simplicity and the low computing resources it required so that it could be used in real time and would not require long set-up times. In particular, we showed that more complex architectures did not provide any advantage in terms of accuracy which may suggest that our approach is limited by the number of examples and thus more complex architectures do not offer an advantage.

Our results confirmed the value of considering different brain activities categories in order to increase accuracy. We also found evidence suggesting that preprocessing methods rendering high frequency-resolution topographical-energy features of brain activity improve the capacity of our small architecture to distinguish among categories. We found that the adaptive training strategy yielded a more realistic representation of actual playing accuracies, though the accuracy achieved ($\sim 55\%$) is still far from what users need. the exploitation of different imageries and features is a reasonable source of accuracy improvement.

Finally, we have not discussed confusion matrices nor features maps after convolution. We consider that further research should explore the latter to find and guide the selection of efficient frequency based features. We also notice that using on-line generated data to conduct a similar approach for imageries selection would give a better understanding of what imageries would perform better in playing conditions, thus a new experiment including this new constrain should be devised.

ACKNOWLEDGMENT

We thank BrainProducts for the loan of the 160 channel ActiCHamp EEG recorder. We also appreciate and thank the organisation of Cybathlon 2016. The support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged during the last stages of this work.

REFERENCES

[1] Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., ... & Vaughan, T. M. (2000). Brain-computer interface technology: a review of the first international meeting. *IEEE transactions on rehabilitation engineering*, 8(2), 164-173.

[2] Bouton, C. E., Shaikhouni, A., Annetta, N. V., Bockbrader, M. A., Friedenberg, D. A., Nielson, D. M., ... & Morgan, A. G. (2016). Restoring cortical control of functional movement in a human with quadriplegia. Nature, 533(7602), 247-250.

[3] Makin, T. R., de Vignemont, F., & Faisal, A. A. (2017). Neurocognitive barriers to the embodiment of technology. Nature Biomedical Engineering, 1, 0014.

[4] Riener, R., & Seward, L. J. (2014, October). Cybathlon 2016. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2792-2794). IEEE.

[5] Yang, H., Sakhavi, S., Ang, K. K., & Guan, C. (2015, August). On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC) (pp. 2620-2623). IEEE.

[6] Lu, N., Li, T., Ren, X., & Miao, H. (2016). A Deep Learning Scheme for Motor Imagery Classification based on Restricted Boltzmann Machines. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*.

[7] Tabar, Y. R., & Halici, U. (2016). A novel deep learning approach for classification of EEG motor imagery signals. *Journal of Neural Engineering*, 14(1), 016003.

[8] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2016). EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. *arXiv preprint arXiv:1611.08024*.

[9] Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2015). Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. *arXiv preprint arXiv:1511.06448*.

[10] Stober, S., Sternin, A., Owen, A. M., & Grahn, J. A. (2015). Deep Feature Learning for EEG Recordings. *arXiv preprint arXiv:1511.04306*.

[11] Huggins, J. E., Moinuddin, A. A., Chiodo, A. E., & Wren, P. A. (2015). What would brain-computer interface users want: opinions and priorities of potential users with spinal cord injury. *Archives of physical medicine and rehabilitation*, 96(3), S38-S45.

[12] Vidaurre, C., Sannelli, C., Mller, K. R., & Blankertz, B. (2011). Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural computation*, 23(3), 791-816.

[13] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, M. Hmlinen, MNE software for processing MEG and EEG data, *NeuroImage*, Volume 86, 1 February 2014, Pages 446-460, ISSN 1053-8119.

[14] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, M. Hmlinen, MEG and EEG data analysis with MNE-Python, *Frontiers in Neuroscience*, Volume 7, 2013, ISSN 1662-453X.

[15] Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., ... & Belopolsky, A. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint* arXiv:1605.02688.

[16] Allison, B. Z., & Neuper, C. (2010). Could anyone use a BCI?. In Brain-computer interfaces (pp. 35-54). Springer London.

[17] Myrden, A., & Chau, T. (2015). Effects of user mental state on EEG-BCI performance. *Frontiers in human neuroscience*, 9, 308.

[18] Neuper, C., & Pfurtscheller, G. (2001). Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International journal of psychophysiology*, 43(1), 41-58.

[19] Friedrich, E. V., Scherer, R., & Neuper, C. (2013). Long-term evaluation of a 4-class imagery-based braincomputer interface. *Clinical Neurophysiology*, 124(5), 916-927.

[20] Escolano, C., Murguialday, A. R., Matuz, T., Birbaumer, N., & Minguez, J. (2010, August). A telepresence robotic system operated with a P300-based brain-computer interface: initial tests with ALS patients. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE (pp. 4476-4480). IEEE.