Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals

Aurélien Appriou, Andrzej Cichocki, Fabien Lotte

Modern machine learning algorithms to classify cognitive and affective states from electroencephalography signals

Aurelien Appriou[1,2], Andrzej Cichocki[3,4,5], Fabien Lotte[1,2]

[1]Inria, LaBRI (CNRS/Univ. Bordeaux /INP), Talence, France, [2]RIKEN Brain Science Insitute, Wakoshi, Japan
[3]Skolkovo Institute of Science and Technology, Moscow, Russia, [4]Nicolaus Copernicus University, Torun, Poland
[5]Hangzhou Dianzi University, Hangzhou, China

Estimating cognitive or affective states from brain signals is a key but challenging step in the creation of passive brain-computer interface (BCI) applications. So far, estimating mental workload or emotions from EEG signals is only feasible with modest classification accuracies, thus leading to unreliable neuroadaptive applications. However, recent machine learning algorithms, notably Riemannian geometry based classifiers (RGC) and convolutional neural networks (CNN), have shown to be promising for other BCI systems, e.g., motor imagery-BCIs. However, they have not been formally studied and compared together for cognitive or affective states classification. This paper thus explores such machine learning algorithms, proposes new variants of them, and benchmarks them with classical methods to estimate both mental workload and affective states (Valence/Arousal) from EEG signals. We study these approaches with both subject-specific and subject-independent calibration, to go towards calibration-free systems. Our results suggested that a CNN obtained the highest mean accuracy, although not significantly so, in both conditions for the mental workload study, followed by RGCs. However, this same CNN underperformed in both conditions for the emotion data set, a data set with small training data. On the contrary, RGCs proved to have the highest mean accuracy with the Filter Bank Tangent Space classifier (FBTSC) we introduced in this paper. Our results thus contributed to improve the reliability of cognitive and affective states classification from EEG. They also provide guidelines about when to use which machine learning algorithm.

***Index Terms*—Brain-Computer Interfaces (BCI), Mental Workload, Emotions, Riemannian geometry, Deep Learning, EEG**

I. Introduction

BRAIN-Computer Interfaces (BCIs) enable their users to interact with computers by using brain activity only, usually measured with Electroencephalography (EEG) [1]. For example, BCIs can enable people with severe motor impairment to control a wheelchair with EEG only, e.g., by imagining left or right hand movements to make the wheelchair turn left or right [2]. Such BCIs are called active BCIs since users are actively sending commands to the system [3]. In contrast, the so-called passive BCIs [3] are not used for direct control, but to monitor users' mental states in real-time, to adapt an application accordingly. For instance, passive BCIs were used to estimate mental workload [4], i.e., the amount of cognitive resources currently engaged by subjects, or affective states [5], i.e., the emotions subjects currently feel.

A. Mental workload estimation from EEG

Passive BCIs were used to study mental workload during navigation tasks with different input devices [6], during visualization tasks [7] or plane piloting [8]. Workload estimation was also used to design applications that dynamically adapt to the users' states, e.g., video games with adaptive difficulty [9] or training applications with a sequence of exercises adapted to the cognitive capabilities of each learner [10]. In general, estimating mental workload from EEG is extensively used in the field of Neuroergonomics [11], [12] which consists in using neuroscience tools and results, e.g., BCIs, to assess the ergonomic qualities of interactive systems.

However, reliably estimating mental workload from EEG signals, over time, contexts and subjects is difficult [4]. For instance, in [4], discriminating low from high workload in 2s epoch of oscillatory EEG activity was possible with a classification accuracy of only about 69%, using a Filter Bank Common Spatial Patterns (FBCSP) algorithm [13] coupled with a Linear Discriminant Analysis (LDA) [14]. In [15], authors used a Bilinear CSP and a Linear Probabilistic Support Vector Machine (SVM) to classify 1.1s epochs of oscillatory EEG activity into four levels and obtained 93% classification accuracy. In [16], authors applied a SVM to classify two levels of workload, using N-back tasks. They obtained 84% classification accuracy on 0.5 to 1.5s long epochs. In [17], authors extracted features using a wavelet entropy, and then applied a Multi-Layer Perceptron (MLP) to classify workload data into 7 levels with 5s long epochs. They obtained 98% classification accuracy on their own data set, and 83% on the data set from [18]. Except the last study above, all classification accuracies have been obtained in offline analysis settings.

B. Affective states estimation from EEG

Passive BCIs estimating affective states are called affective BCIs (aBCIs) [5]. Examples of aBCI work include studies that found neurophysiological responses to differentiate between frustration and boredom in e-learning [19] and between frustration and normal game play [20]. Moreover, Rani et al. [21] showed that both players' enjoyment and skills increased when tasks were adapted to their affective states rather than to their performances. In [22], players could modulate their affective states to influence games' parameters. Affective BCIs can also be used for automatic media recommendation [23], or real-world emotions detection [24].

Despite an increasing number of aBCIs studies, defining and clustering emotion dimensions remains challenging. There are multiple main approaches to define emotion classes [5]: the most popular one, and the one used in our study, is the circumplex model of Russell [25], which assumes that any affective state can be localized on a two-dimensional plane. The first axis of this plane is the valence, ranging from positive feelings to negative ones, and the second axis represents arousal, ranging from calm to excited. In [23], they used an SVM for a 3-class problem with 15s long epochs, for valence (low vs neutral vs high, 50% accuracy) and arousal (low vs neutral vs high, 62%). [26] studied valence and arousal (low vs high) as well, but with 6s EEG epochs. They respectively obtained 77% and 74% classification accuracy by using a combination of wavelet entropy and average wavelet coefficient coupled with an SVM. [27] used SVMs, Deep Neural Networks or Random Forest to classify data from the DEAP data set [28] (presented in the *Methods* section of this paper), using a 2-class problem with 60s long epochs. They respectively obtained 79%, 49% and 56% classification accuracy. However, the Deep Learning method was not described in this paper, making it difficult to assess the validity and superiority of this approach. [29] used a Logistic Regression (LR) to discriminate valence levels (low vs high), and obtained 71% accuracy on 6s epochs. In [28], which introduced the DEAP data set we use here, they applied a Naive Bayes Classifier for 2-class discrimination (low vs high) for both valence and arousal dimensions (60 s epochs). They obtained 57% and 62% of accuracy for the valence and arousal dimensions, respectively. Other studies used Deep Learning methods to optimize affective states classification (see supplementary material).

Most studies proved that classifying affective states from EEG remains really challenging since results hardly go over chance level accuracy. Some studies were even unable to obtain better than chance results when reproducing previous works with statistically rigorous evaluation methods [30]. Finally, confounding factors due to electromyography (e.g., facial muscles activity during emotion expression) have likely played a role in the performances obtained in many studies. However, other studies have obtained better accuracies when using different EEG patterns. For instance, [31] obtained 81% classification accuracy using evoked potentials for a 4-class valence/arousal classification. Note that this paper focuses on oscillatory activity, as this can classify cognitive and affective states from continuous EEG, without the need for stimulus-locked response, which evoked potentials do need.

C. Paper objectives

Thus, the classification accuracies obtained so far - mostly around 70% for workload, and around 60-65% for emotions in oscillatory-based studies - revealed the need for more robust and accurate EEG classification algorithms. Therefore, we propose here to study algorithms that proved efficient either in recent active BCI classification competitions [13], [32], notably Riemannian geometry classifiers, or in other fields of artificial intelligence, such as Deep Learning [33], [34]. Note that such algorithms have been mostly explored for EEG classification of motor tasks, but not systematically studied and compared for workload/affective states estimation. Here we formally study and compare these various algorithms as well as two new variants we proposed, for both workload, arousal and valence classification from EEG signals[1]. We also propose guidelines about which algorithm to use in which context. As baseline, we use two standard methods for studying workload levels/affective states classification: 1) Common Spatial Pattern (CSP) spatial filters with an LDA classifier and 2) the FBCSP [13], which is a CSP extension that won numerous active BCI competitions. Then, we studied two Riemannian approaches, manipulating and classifying EEG signals as covariance matrices: Minimum Distance to the Mean with Fisher geodesic filtering classifier (FgMDM) and Tangent Space Classifier (TSC). Such methods have recently won six international brain signals competitions [32]. We then propose to improve these Riemannian approaches by working on a bank of band-pass filters such as the ones used for FBCSP, instead of using a unique band-pass filter. We name these new approaches FBFgMDM and FBTSC. Finally, we used a Convolutional Neural Network (CNN), i.e., a Deep Learning algorithm, which recently obtained promising results for many machine learning problems [34]. We studied the CNN developed in [33], since it obtained promising results for motor imagery-based BCIs.

In this paper, we first present the workload and emotion EEG data sets used, before describing each machine learning algorithm. We perform two evaluation studies: 1) a subject-specific study, with each algorithm trained on data specific to each subject, and then tested on other data from the same subject. This is the standard way current BCIs are designed, given the large between-subject variability [14]; 2) a subject-independent study, with each algorithm trained on all data recorded from all subjects except that of the target subject, on which algorithms are tested. This is much more challenging, but if successful, would enable BCI-based monitoring without requiring any calibration for new subjects.

II. Methods

A. Mental workload EEG data set

The data set used comes from [4]. Signals from 28 EEG electrodes (active electrodes in a 10/20 system without T7, T8, Fp1, and Fp2) were recorded from 22 users [4]. To induce mental workload variations, N-back tasks were used: the user had to indicate whether a letter displayed on screen was the same one as the letter displayed N letters before, in a stream of successively displayed letters. Here, 2-sec trials from a 0-back task were labeled as "low" workload, while those from a 2-back were labelled as "high" workload. In total, 720 trials were available for each workload level and user. See the supplementary material for more information.

As introduced previously, we studied both subject-specific and subject-independent calibrations. For subject-specific calibration, the first half of each user's trials was used for training and the second half for testing. For the subject-independent

[1]Preliminary results on mental workload data only, and with a few existing algorithms only, was published as an extended conference abstract in [35]

calibration, the training set comprised all trials of all users except the current user used for testing, i.e., around 21*1440 = 30240 training trials. To allow for comparisons between calibration types, the testing set of each user was the same testing set as with subject-specific calibration, i.e., the second half of the trials (720 testing trials) from this user.

B. Emotion EEG data set

The data set used for studying emotions was the "DEAP" database [28]. It used music-video clips to influence two types of emotion dimensions - valence and arousal, according to the circumplex model of Russell [25]. The data set contains 40 trials, corresponding to 40 music video-clips, recorded on 32 participants. EEG were recorded using 32 electrodes (placed according to the international 10-20 system). Valence and arousal levels were measured using Russell's valence-arousal scale directly after each videos, by clicking on a 1-9 continuous scale. This self-assessment system on a continuous scale makes the classes definition more complex: in DEAP [28] as well was in our study, 5 was kept as a threshold to split trials into two classes - low and high - for both "emotion-arousal" and "emotion-valence" data sets, making classes unbalanced. All the classifiers used were able to deal with unbalanced classes, except the CNN for which we up-sampled the minority class by randomly duplicating trials from this class in order to obtain balanced classes.

For the subject-specific study, given the low number of trials, we performed a "leave-one-out" cross-validation. Thus, we used 40 models for each subject, each model being trained on 39 trials and tested on 1 trial. For the subject-independent study, we kept all trials of all subjects to compose the training set, except the current subject used for testing (i.e., 31*40 = 1240 trials for the training set). The testing set of each subject was composed of all trials of this subject, i.e., 40 trials.

C. Machine learning algorithms explored

The existing algorithms we evaluate here were all studied on EEG-based motor imagery classification, a widely used BCI design, and obtained impressive results. Since both motor imagery, workload and emotions lead to change in EEG oscillatory activity, it is likely that methods that proved effective for motor imagery can prove effective for workload or emotion classification as well. However, to the best of our knowledge, such methods have never been tested and compared together neither on workload nor on emotions data sets nor with subject-independent calibration. We thus propose this evaluation in this paper. We also propose some new variants of some of these algorithms. Altogether, we studied 7 algorithms. First, CSP and LDA were used as a baseline since they are widely used by the BCI community [14]. We then explored the FFBCSP and LDA [13], a CNN [33], and four different methods based on Riemannian geometry: two existing ones, namely the Fisher geodesic Minimum Distance to the Mean classifier (FgMDM) and the Tangent Space Classifier (TSC) [32], and two new extensions we propose here to better exploit the spectral information, namely the Filter Bank FgMDM and Filter Bank TSC. For the workload data set, we assess performances using classification accuracy, i.e., the percentage of test trials correctly classified. For the emotion data set, we used balanced accuracy, i.e. the average of recall obtained on each class, since the classes were unbalanced.

1) Common Spatial Patterns (CSP)

CSP is a widely used algorithm for binary EEG classification, for oscillatory activity-based BCI. It has been shown that changes in both workload [16] and emotions [5] induce changes in EEG oscillatory activity.The CSP algorithm optimizes spatial filters, i.e., a linear combination of the original EEG signals. It is done such that the variance of a spatially filtered signal, i.e. the band power of this signal, is maximized for one class and minimized for the other class. Formally, CSP optimizes spatial filter w by either maximizing or minimizing:

$$J_{CSP}(w) = \frac{w\mathbf{X}_1\mathbf{X}_1^T w^T}{w\mathbf{X}_2\mathbf{X}_2^T w^T} = \frac{w\mathbf{C}_1 w^T}{w\mathbf{C}_2 w^T} \quad (1)$$

where T denotes transpose, $\mathbf{X}_i$ is the band-pass filtered training signal matrix for class i (with the samples as columns and the channels as rows) and $\mathbf{C}_i$ the spatial covariance matrix from class i. In practice, the covariance matrix $\mathbf{C}_i$ is defined as the average covariance matrix of each trial from class i [14]. The spatial filters w that maximize or minimize $J_{CSP}(w)$ are the eigenvectors corresponding to the largest and lowest eigenvalues, respectively, of the Generalized Eigen Value Decomposition of matrices $\mathbf{C}_1$ and $\mathbf{C}_2$. In this study, we used six filters, corresponding to the three largest and three lowest eigenvalues, as recommended in [14]. Once these filters are obtained, we use as CSP features $f = \log(w\mathbf{X}\mathbf{X}^T w^T)$, i.e., the band power of the spatially filtered signals. We used these features as input to an LDA classifier. The CSP requires EEG signals to be band-pass filtered in a specific narrow frequency band. The Alpha rhythm (8-12Hz) being known to vary according to both workload [4] and emotions [28], we applied CSP after band-pass filtering in 8-12 Hz.

2) Filter Bank Common Spatial Patterns (FBCSP)

The FBCSP is an algorithm that optimizes both spatial and spectral filters. To do so, FBCSP first filters EEG signals into multiple frequency bands using a filter bank. Here we used nine band-pass filters in 4Hz-wide bands (in 4-8 Hz, 8-12 Hz, ..., 36-40 Hz) as in [13]. Then, for each band-passed signals, CSP is used to optimize two spatial filter pairs. From the resulting 36 features (9 bands $\times$ 4 CSP filters per band), the four most relevant ones were selected using minimal Redundancy Maximal Relevance (mRMR) [36], and used as input to an LDA. The FBCSP algorithm proved its efficiency when winning the Fifth International BCI competition [13].

3) Riemannian Geometry

Riemannian approaches represent EEG trials as covariance matrices, which are symmetric positive definite (SPD) matrices, and manipulate them with an appropriate geometry, the Riemannian geometry [32], [37]. Classifiers based on such geometry are called Riemannian Geometry Classifiers (RGC).

First, in a Riemannian $manifold$ we can estimate $intrinsic$ non-Euclidean distances between two SPD matrices, i.e. two points (here $\mathbf{C}_1$ and $\mathbf{C}_2$), using the Riemannian distance:

$$\delta^2(\mathbf{C}_1, \mathbf{C}_2) = \sum_n log^2 \lambda_n(\mathbf{C}_1^{-1}\mathbf{C}_2), \quad (2)$$

where $\lambda_n(\mathbf{M})$ is the n^{th} eigenvalue of matrix $\mathbf{M}$. The set of tangent vectors to point $\mathbf{G}$ on the $manifold$ defines the manifold tangent space at $\mathbf{G}$. More generally, any SPD matrix $\mathbf{C}_i$ can be projected onto the tangent space at point $\mathbf{G}$ using:

$$\mathbf{S}_i = Log_{\mathbf{G}}(\mathbf{C}_i) = \mathbf{G}^{1/2} logm(\mathbf{G}^{-1/2}\mathbf{C}_i\mathbf{G}^{-1/2})\mathbf{G}^{1/2}, \quad (3)$$

$\mathbf{S}_i$ being the projection of $\mathbf{C}_i$ onto the tangent plane, and $logm(\cdot)$ denotes the logarithm of a matrix.

In this paper, we studied two existing RGCs - the Mean classifier (FgMDM) and the Tangent Space classifier (TSC) - and introduced two new ones - the Filter Bank TSC (FBTSC) and the Filter Bank FgMDM (FBFgMDM).

Existing methods:

FgMDM [38]: FgMDM projects training matrices $\mathbf{C}_i$ onto the tangent space at point $\mathbf{G}$ (the mean of all training data) using Eq. (3), to obtain matrices $\mathbf{S}_i$. Then, a Fisher geodesic filter is obtained by optimizing an LDA classifier on $\mathbf{S}_i^{vec}$, the vectorized upper-triangular elements of $\mathbf{S}_i$, to discriminate classes using such vectors. This results in a matrix of weights $\mathbf{W}$ = LDA($\mathbf{S}_i^{vec}$). The projected SPD matrices $\mathbf{S}_i$ from both the training & the testing sets are then filtered with weights $\mathbf{W}$, using $\hat{\mathbf{S}}_i = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_i^{vec}$, where $\hat{\mathbf{S}}_i$ denotes the geodesic filtered SPD matrices from $\mathbf{S}_i$. Then, these filtered matrices $\hat{\mathbf{S}}_i$ are projected back onto the manifold using equation:

$$\hat{\mathbf{C}}_i = Exp_{\mathbf{C}}(\hat{\mathbf{S}}_i) = \mathbf{G}^{1/2} expm(\mathbf{G}^{-1/2}\hat{\mathbf{S}}_i\mathbf{G}^{-1/2})\mathbf{G}^{1/2}, \quad (4)$$

where $\hat{\mathbf{C}}_i$ are the filtered SPD matrices projected onto the manifold and $expm(\mathbf{M})$ denotes the exponential of matrix $\mathbf{M}$. Finally, this approach uses a Minimum Distance to the Mean classifier to classify testing geodesic filtered matrices $\hat{\mathbf{C}}_i$. To do so, during the training step, the class centroids $\mathbf{G}_k$ of each class k are computed by averaging the geodesic filtered covariance matrices $\hat{\mathbf{C}}_i^k$ from each class k. During testing, the Riemannian distances between the testing geodesic filtered matrix $\hat{\mathbf{C}}_j$ and each class centroid $\mathbf{G}_k$ are first calculated, using Eq. (2). The matrix $\hat{\mathbf{C}}_j$ is assigned class label k for which the centroid $\mathbf{G}_k$ is the nearest. In our study, FgMDM was applied on EEG band-pass filtered in 8-12Hz, as for the CSP.

TSC: TSC first projects all training SPD matrices $\mathbf{C}_i$ onto the tangent space at point $\mathbf{G}$ (the mean of all training matrices). Then, it uses any classifier such as LDA, SVM or Logistic Regression (LR) on the vectorized upper-triangular elements of the projected matrices [39]. We used LR with L_2 regularization (with the default $C = 1.0$ in scikit-learn [40]), As for FgMDM, TSC used data filtered in 8-12Hz.

New methods:

Filter Bank FgMDM (FBFgMDM): Contrary to FgMDM which exploits EEG signals in a single frequency band, this method applies FgMDM in multiple bands separately, and combines the resulting distances to exploit additional spectral information. This should possibly improve classification performances, as FBCSP did to improve CSP. To do so, FBFgMDM first filters EEG signals in multiple bands using a filter bank, as for FBCSP. Here we used the same bands as the FBCSP, i.e., 4-8 Hz, 8-12 Hz, ..., 36-40 Hz. Then for the EEG signals in each frequency band j, this method first uses a regular FgMDM, i.e., it computes the Riemannian distances $\delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij})$ between a geodesic filtered SPD matrix $\hat{\mathbf{C}}_{ij}$ and each class centroid $\mathbf{G}_{kj}$. Then, from all nine bands j, the four most useful ones for classification are selected with mRMR feature selection [36] on the Riemannian distances $\delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij})$ used as features, on the training set. For testing, we compute the squared Riemannian distances for the four bands selected using mRMR only and sum them:

$$\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i) = \sum_{j\in\Omega} \delta^2(\mathbf{G}_{kj}, \hat{\mathbf{C}}_{ij}), \quad (5)$$

where Ω is the set of frequency bands selected with mRMR. We thus obtain k new distances $\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i)$ to each class k for each trial i. The classification prediction results in choosing the class y_i for which the summed squared distance to the centroid is the smallest, i.e., $y_i = \text{argmin}_k(\gamma^2(\mathbf{G}_k, \hat{\mathbf{C}}_i))$.

Filter Bank TSC (FBTSC): FBTSC also exploits more spectral information than TSC, by using a filter bank. FBTSC indeed projects matrices $\mathbf{C}_{ij}$, band-pass filtered in bands 4-8 Hz, 8-12 Hz, ..., 36-40 Hz, to the tangent space using Eq. (3). Then, the probabilities that the vectorized upper-triangular elements of the projected SPD matrix $\mathbf{S}_{ij}$ belongs to class k is calculated using standard classification algorithms with probabilistic outputs, such as LDA or LR. Here we used LR that directly provides such probability with its softmax function. Since we did so for nine frequency bands, in two classes k, we ended up with nine pairs of probabilities. From these pairs of probabilities, the four most relevant are selected using mRMR on the training set. Finally, we multiplied the probabilities associated to each class k, for the selected bands only, to end up with two probabilities, using $\mathcal{P}_{ki} = \prod_{j\in\Omega} \mathcal{P}_{kij}$, where $\mathcal{P}_{ki}$ is the probability of trial i to be part of class k, and $\mathcal{P}_{kij}$ the probability of a projected SPD matrix $\mathbf{S}_{ij}$, band-pass filtered in frequency band j, to be part of class k. The classification prediction results in choosing the class y_i for which $\mathcal{P}_{ki}$ is the highest, i.e., $y_i = \text{argmax}_k(\mathcal{P}_{ki})$.

4) Convolutional Neural Networks (CNN)

Shortly, a CNN is a feedforward neural network with at least one convolutional layer. This type of network flows information uni-directionally from the input to the hidden layers and finally to the output. A recent study presented a new type of CNN dedicated to motor task classification in EEG: the Shallow ConvNet [33]. The shallow ConvNet architecture consists in a 3-layer CNN with parameters that have been experimentally tested and validated by their authors [33]. The first layer is a convolutional layer along the temporal dimension, while the subsequent one is a convolutional layer along the spatial dimension, i.e., over EEG electrodes. The first temporal convolution aims at optimizing band-pass filters, and the spatial convolution aims at optimizing spatial filters. Then, signals are squared, a mean pooling is performed (to compute signals band power) and the CNN ends by a fully connected linear classification layer. Overall this CNN thus processes EEG data similarly to the FBCSP and LDA. In contrast to FBCSP, all these filters are optimized simultaneously though, which made it outperform the FBCSP on motor EEG signals

[33]. The Shallow ConvNet uses minimally preprocessed EEG signals as input, so we filtered them in 4-40 Hz.

III. Results

Figure 1 summarizes the mean performance obtained on each data set. As a reference, the statistical chance levels [41] were estimated at 50.47% for the mental workload study (1440 trials and 22 subjects) and 52.27% for the affective state study (40 trials and 32 subjects). Note that for statistical tests (ANOVA), we checked the data sphericity, and used Greenhouse-Geisser (GG) correction in ANOVA if needed.

data set	study	CSP	FBCSP	FgMDM	TSC	FBFgMDM	FBTSC	CNN
emotion-valence	subject-specific	57.5904	59.1921	58.8734	59.4658	61.0144	61.0934	46.32
	subject-independant	52.5	55.2344	47.9688	49.1406	48.4375	48.75	48.0469
emotion-arousal	subject-specific	58.2586	59.1315	60.0404	60.0404	60.3008	60.5969	40.1531
	subject-independant	55.7031	55.3125	56.25	55.7813	51.6406	53.2812	47.5
workload	subject-specific	67.0041	68.5089	69.9429	68.4964	70.3385	68.7242	72.7296
	subject-independant	58.0465	60.0742	58.3072	58.3072	61.2989	60.1805	63.7357

Fig. 1. Mean classification accuracy for each algorithm. The best performance of each study is in green, the worst in red.

A. Workload study

Performances obtained by each algorithm on this data set are reported on Figure 2. We performed a 2-way ANOVA with repeated measures to evaluate the performances of factor *Algorithm* according to factor *Calibration Type* (subject-specific vs subject-independent). It revealed a main effect of *Algorithm* [GG(1,22)=0.517, p=0.001], and *Calibration Type* [F(1,22)=33.308, $p \leq 0.0001$], but not for *Calibration Type*Algorithm* [GG(1,22)=0.558, p=0.618].

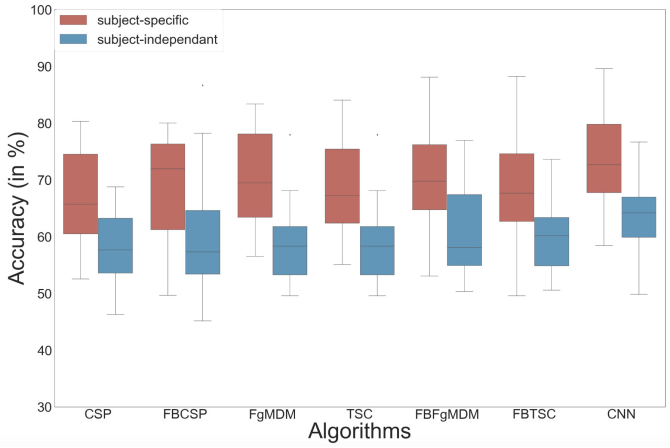

Fig. 2. Classification accuracy of each algorithm on the workload data set.

Post-hocs analyses - Student t-test for paired samples - with Bonferroni adjustments showed no significant differences between algorithms in the subject-specific or subject-independent studies. However, performances obtained suggested better (but non-significantly so) results with the CNN compared to other algorithms, in both subject-specific and subject-independent studies. Riemannian geometry classifiers (RGC), in particular the newly proposed ones (FBfgMDM and FBTSC) provided the second best performances, just after the CNN. On the other hand, the baseline CSP+LDA obtained the worst results.

B. Valence

The balanced classification accuracies obtained are reported on Figure 3. We ran a 2-ways ANOVA for repeated measures to evaluate the impact of *Algorithm* on the emotion-valence data set, regarding the *Calibration Type*. The results showed significant differences in *Algorithm* [GG(1,32)=6.918, p=0.002], *Calibration Type* [F(1,32)=21.732, p¡0.0001] and *Calibration Type*Algorithm* [GG(1,32)=5.374, p=0.003].

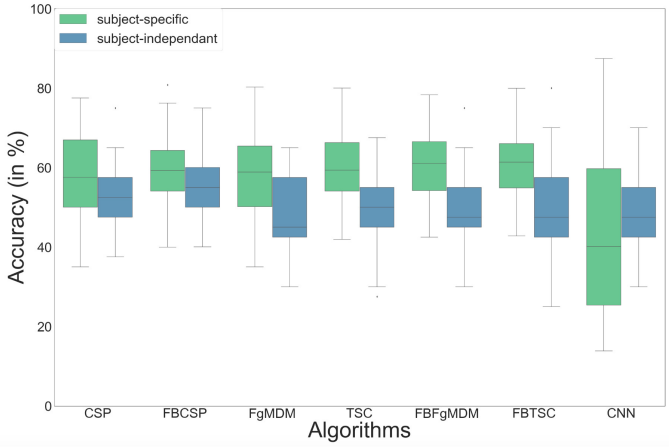

Fig. 3. Balanced classification accuracy on the emotion-valence data set.

Post-hoc analyses - Student t-test for paired samples - with Bonferroni corrections showed a significant difference between FBTSC and CNN for subject-specific calibration [$perf_{FBTSC}$ = 61.09%, $perf_{CNN}$ = 46.32%; $p \leq 0.05$]. No algorithm showed better results than others with the subject-independent calibration. Overall, FBFgMDM and FBTSC obtained the best accuracy (both about 61%) for subject-specific calibration, while FBCSP obtained the best performances for the subject-independent one (55.2%).

C. Arousal

The balanced classification accuracies for the emotion-arousal data set are reported on Fig. 4. We then performed a 2-way ANOVA with repeated measures, with factor *Algorithms* and *Calibration Type*. Results revealed significant effects for *Algorithms* [GG(1,32)=9.177, $p \leq 0.0001$], *Calibration Type* [F(1,32)=4.262, p=0.048] and *Algorithms*Calibration Type* [GG(1,32)=3.894, p=0.008].

Post-hoc analyses -Student t-test for paired samples- with Bonferroni corrections showed significant differences with the subject-specific calibration between CNN and all other classifiers (see results in the supplementary material). No algorithm showed better results than others with the subject-independent calibration. Overall the best results were all obtained by RGCs, FBFgMDM and FBTSC for the subject-specific calibration, and FgMDM for the subject-independent one.

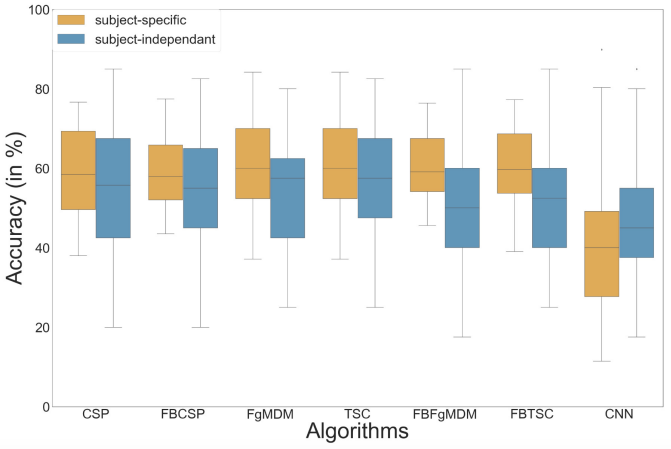

Fig. 4. Balanced classification accuracies on the emotion-arousal data set.

IV. Discussion, Conclusion and Future Work

In this paper, we explored promising classification algorithms, both existing and new ones, to classify mental workload and emotions (valence and arousal) from EEG signals, with both subject-specific and subject-independent calibration. Altogether we studied CSP+LDA, FBCSP+LDA, four RGCs (FgMDM, TSC and two new variants proposed here: FBFgMDM and FBTSC), and CNN.

The first results to highlight are the CNN classification performances we obtained across the different conditions and data sets. Indeed, this algorithm has a higher mean accuracy (although non-significantly so) than the original authors' results, the baseline CSP+LDA, and more importantly than both FBCSP and Riemannian methods, with both subject-specific and subject-independent calibrations on the workload data set. Moreover, obtaining reasonable performances in a subject-independent calibration from only two seconds of EEG data and only 21 users for calibration, makes the CNN particularly interesting to design calibration-free neuroadaptive technologies in the future. By contrast, this algorithm significantly under-performed with both subject-specific and subject-independent calibrations on both the valence and arousal data sets. All algorithms indeed outperformed this CNN in all conditions on the emotion data sets.

Multiple factors could explain the observed algorithm performances. First, the number of trials that are used for training models is important. In [33], authors tested the Shallow ConvNet on multiple motor-imagery data sets (from 288 to 1168 trials), and often obtained significantly better performances with the CNN than with FBCSP. In our study, the workload data set contained 720 training trials whereas both valence and arousal data sets contained 39 training trials only (with cross validation calibration). This might suggest that the CNN could be useful for mental state classification, but only when large amount of training trials are available (around 700 in our study), which is not always possible. However, other factors also differ between both data sets studied and could also explain differences in CNN performances, including the EEG epochs length (2s epochs for workload and 60s epochs for emotions), and the nature of the mental states studied (workload vs emotions). Indeed, emotions are thought to originate from deep brain areas [5] and are thus known as being difficult to estimate from EEG. In the future, deeper analyses would thus be needed to fully disentangle these factors, by systematically varying the mental states studied, the EEG epoch length and the number of training trials.

Another relevant result is the promising classification performances of the proposed RGCs. Indeed, FBTSC and FBFgMDM outperformed the results from the data sets' authors in most conditions/data sets. Moreover, FBFgMDM with subject-specific calibration, and FBFgMDM and FBTSC with subject-independent calibration, reached higher mean accuracies than all other algorithms, except the CNN on the workload data set. More interestingly, the low number of trials in the emotion data sets did not seem to affect their performances since they also reached the highest mean accuracies on both the emotion-valence and emotion-arousal data set, both with subject-specific calibration. These promising results compared to standard RGCs (TSC and FgMDM), are probably due to the extra spectral information extracted with the filter bank, and our study enabled us to quantify this gain.

Finally, FBCSP+LDA obtained a higher mean accuracy than CSP+LDA, although not significantly so, in all conditions/data sets, and the higher overall mean accuracy for valence classification with subject-independent calibration. However, it did not obtain higher mean accuracies than others in any other condition. It should be noted that such results reflect the performances obtained in offline evaluation. As such they are likely to be similar to performances obtained in offline or open-loop mental state monitoring, e.g., for noeuroergonomics (ex: mental workload monitoring) or neuromarketing (ex: emotion monitoring). The performances are likely to change in closed-loop applications, with neuroadaptive technologies, and will thus need to be evaluated in this context as well.

Such results enable us to suggest guidelines about which algorithm to use for mental states classification from EEG. First, the CNN is recommended for mental workload classification with both subject-specific and subject-independent calibration, but seems to need a large amount of training trials (at least several hundreds). It should thus probably be avoided for data sets with little training data (i.e., a few dozens). Second, Filter Bank RGCs (FBTSC and FBFgMFM) should also be recommended to obtain good classification performances notably with subject-specific calibration, for both workload and emotion classification, whatever the amount of training data. However, such methods do not seem suitable for subject-independent classification with little training data and/or for emotion classification. They seem suitable for subject-independent classification of workload with large amount of training data though. Our results also confirmed that passive BCIs with subject-independent calibration is possible but very challenging and with much lower accuracies. Similarly, affective state classification in EEG is possible but much more challenging than workload estimation. However, those suggestions imply computational costs that will differ from an algorithm to another. Indeed, using the FB RGCs or the CNN will require a long calibration time, when the testing phase might also be time consuming and has to be considered before

to go towards online uses. See the supplementary material for more information.

For the emotion data set, we labelled trials as in the original paper to allow comparisons, i.e., with a global subject-independent partition between low/high valence/arousal trials, based on the SAM ratings. Note that better methods for partitioning low/high trials in a per-subject basis can also be used [42] in the future, to limit the class imbalance. Still in the future, other deep learning architectures, notably Recurrent Neural Networks (RNN) [34] may prove promising for EEG classification and passive BCIs as well. It would also be interesting to study whether CNN and RGCs can be used to estimate robustly other cognitive states such as fatigue, curiosity or engagement, and how well the proposed RGCs perform on motor imagery data for active BCIs. Altogether, our results suggested that CNN and the proposed filter bank RGCs are valuable machine learning tools for scientists aiming at decoding cognitive and affective states from EEG signals.

Acknowledgements: This work was supported by the European Research Council (grant ERC-2016-STG-714567) and the Japanese Society for the Promotion of Science.

REFERENCES

[1] M. Clerc, L. Bougrain, and F. Lotte. *Brain-Computer Interfaces 1*. ISTE-Wiley, 2016.

[2] J. R. Millán, R. Rupp, G. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, F. Cincotti, A. Kübler, C. Neuper, K.-R. Müller, and D. Mattia. Combining Brain-Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges. *Front in Neuro*, 2010.

[3] T.O. Zander and C. Kothe. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *J Neur Eng*, 8, 2011.

[4] C. Mühl, C. Jeunet, and F. Lotte. EEG-based workload estimation across affective contexts. *Frontiers in Neuroscience*, 8:1–15, 2014.

[5] C. Muhl, B. Allison, A. Nijholt, and G. Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 2015.

[6] J. Frey, M. Daniel, M. Hachet, J. Castet, and F. Lotte. Framework for electroencephalography-based evaluation of user experience. *CHI*, 2015.

[7] E. Peck, B.F. Yuksel, A. Ottley, R. Jacob, and R. Chang. Using fNIRS brain sensing to evaluate information visualization interfaces. *CHI*, 2013.

[8] T. Gateau, G. Durantin, F. Lancelot, , and F. Dehais. Real-time state estimation in a flight simulator using {fNIRS}. *PloS one*, 10(3), 2015.

[9] S. Fairclough. BCI and physiological computing for computer games: Differences, similarities & intuitive control. In *Proc ACM CHI*, 2008.

[10] B.F. Yuksel, K.B. Oleson, L. Harrison, E.M. Peck, D. Afergan, R. Chang, and R.J.K. Jacob. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. *CHI*, 2015.

[11] H. Ayaz and F. Dehais. *Neuroergonomics: The Brain at Work and in Everyday Life*. Elsevier, 01 2019.

[12] J. Frey, M. Hachet, and F. Lotte. EEG-based neuroergonomics for 3D user interfaces: opportunities and challenges. *Le Travail Humain*, 2017.

[13] K.K. Ang, Z.Y. Chin, C.C. Wang, C.T. Guan, and H.H. Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, 6:1–9, 2012.

[14] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1), 2008.

[15] R Mahmoud, T Shanableh, I P Bodala, and N Thakor. Novel Classification System for Classifying Cognitive Workload Levels under Vague Visual Stimulation. 1748(c):1–12, Sensors 2017.

[16] Anne-Marie Brouwer, Maarten A Hogervorst, Jan B F van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. Estimating workload using {EEG} spectral power and {ERPs} in the n-back task. *JNE*, 9(4):45008, 2012.

[17] P. Zarjam, J. Epps, F. Chen, and N. H. Lovell. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in Biology and Medicine*, 43(12):2186–2195, 2015.

[18] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. N. Rao. Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. *Proc ACM CHI*, pages 835–844, 2008.

[19] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty. *IEEE Trans Syst Man Cyb, Part A*, 41(6):1052–1063, 2011.

[20] B. Reuderink, C. Muhl, and Poel M. Valence, arousal and dominance in the EEG during game play. *IJAACS*, 6(1), 2013.

[21] P. Rani, N. Sarkar, and C. Liu. Maintaining optimal challenge in computer games through real-time physiological feedback. *ICHCI*, 2005.

[22] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[23] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[24] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, March 2019.

[25] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[26] H. Candra, M. Yuwono, R. Chai, H. T. Nguyen, and S. Su. EEG emotion recognition using reduced channel wavelet entropy and average wavelet coefficient features with normal Mutual Information method. *Proc. IEEE EMBC*, pages 463–466, 2017.

[27] A. Al-Nafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil. Recognition of affective states via electroencephalogram analysis and classification. *Advances in Intelligent Systems and Computing*, 722:242–248, 2018.

[28] S. Koelstra, C. Mühl, M. Soleymani, Jong S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis Using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[29] C. A. Kothe, S. Makeig, and J. A. Onton. Emotion recognition from EEG during self-paced emotional imagery. *Proc. ACII*, 2013.

[30] F Lotte, C Jeunet, R Chavarriaga, L Bougrain, D.E. Thompson, R Scherer, Md R. Mowla, A Kübler, M Grosse-Wentrup, K. Dijkstra, and N. Dayan. Turning negative into positives! Exploiting "negative" results in Brain-Machine Interface (BMI) research. *Brain-Computer Interfaces*, 2020.

[31] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, and P. D. Bamidis. Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Trans. on Info Technology in Biomedicine*, 14(3):589–597, 2010.

[32] F. Yger, M. Berar, and F. Lotte. Riemannian approaches in Brain-Computer Interfaces: a review. *IEEE TNSRE*, 25(10), 2016.

[33] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 2017.

[34] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[35] A. Appriou, A. Cichocki, and F. Lotte. Towards robust neuroadaptive HCI: exploring modern machine learning methods to estimate mental workload from EEG signals. In *CHI*, 2018.

[36] H. Peng, F. Long, and C. Ding. Feature Selection Based On Mutual Information: Criteria of Max-Dependency,Max-Relevance, and Min-Redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27, 2005.

[37] M. Congedo, A. Barachant, and R. Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

[38] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Riemannian geometry applied to BCI classification . *LVA/ICA*, pages 629–636, 2010.

[39] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, and V. Dubourg. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12, 2011.

[41] E. Combrisson and K. Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 250:126–136, 2015.

[42] A. Clerico, A. Tiwari, R. Gupta, S. Jayaraman, and T. H. Falk. Electroencephalography amplitude modulation analysis for automated affective tagging of music video clips. *Front. Comp. Neuro.*, 11, 2018.

Aurelien Appriou Aurelien Appriou received the M.Sc. (with honors), in cognitive science, from the University of Bordeaux, France, in 2015. As a PhD candidate he is part of Inria Bordeaux-Sud-Ouest, and member of the European Research Council (ERC) project "BrainConquest", aiming at improving Brain-Computer Interfaces (BCIs) users training. During his PhD, he is focusing on two main topics, first measuring learning-related mental states through both electroencephalographic and physiological signals, and second investigating machine learning algorithms for decoding such signals. His research interests include brain-computer interfaces, human-computer interaction, machine learning and cognitive science.

Andrzej Cichocki Andrzej Cichocki received the M.Sc. (with honors), Ph.D. and Dr.Sc. (Habilitation) degrees, all in electrical engineering, from Warsaw University of Technology in Poland. He worked several years at University Erlangen-Nuerenberg in Germany as an Alexander-von-Humboldt Research Fellow and Guest Professor. In 1995-2018, he was a team leader and the head of the laboratory for Advanced Brain Signal Processing, at RIKEN Brain Science Institute in Japan. Under the guidance of Professor Cichocki, the new Laboratory Tensor Networks and Deep Learning for Applications in Biomedical Data Mining is established at SKOLTECH. The mission of the Laboratory is to perform cutting-edge innovative research in the design and analysis of deep neural networks, tensor networks and multiway component analysis for biomedical applications. He is author of more than 500 papers and six books in English. He is among the most cited Polish computer scientists and he is or has been associate editor of the international journals in signal processing, computational neuroscience and neural networks. His publications currently report over 42,000 citations, with an h-index of 93 according to Google Scholar. He is Fellow of the IEEE since 2013.

Fabien Lotte Fabien Lotte obtained a M.Sc., a M.Eng. and a PhD degree in computer sciences, all from the National Institute of Applied Sciences (INSA) Rennes, France, in 2005 (M.Sc., M.Eng.) and 2008 (PhD). His PhD Thesis received both the PhD Thesis award 2009 from AFRIF (French Association for Pattern Recognition) and the PhD Thesis award 2009 accessit (2nd prize) from ASTI (French Association for Information Sciences and Technologies). In 2009 and 2010, he was a research fellow at the Institute for Infocomm Research (I2R) in Singapore, working in the Brain-Computer Interface Laboratory. From January 2011 to September 2019, he was a Research Scientist (with tenure) at Inria Bordeaux Sud-Ouest, France, in team Potioc (http://team.inria.fr/potioc/). He obtained an Habilitation (HDR) in Computer Science from the University of Bordeaux in September 2016. Between October 2016 and January 2018, he spent 1-year as a visiting scientist at RIKEN Brain Science Institute, Japan, in Cichocki's laboratory for advanced brain signal processing. Since October 2019, he is a Research Director (DR2) at Inria Bordeaux Sud-Ouest, France, still in team Potioc. His research interests include Brain-Computer Interfaces (BCI), human-computer interaction, pattern recognition and brain signal processing. He is part of the editorial boards of the journals Brain-Computer Interfaces (since 2016) and Journal of Neural Engineering (since 2016). He co-edited the books "Brain-Computer Interfaces 1: foundations and methods" and "Brain-Computer Interfaces 2: technology and applications", published both in French and in English in 2016, as well as the book "Brain-Computer Interfaces Handbook: Technological and Theoretical Advance" published in 2018 . In 2016, he was the recipient of an ERC Starting Grant (project BrainConquest) to develop his research on BCI.