Performance of Nested vs. Non-nested SVM Cross-validation Methods in Visual BCI: Validation Study

Mohammed J. Abdulaal*, Alexander J. Casson† and Patrick Gaydecki‡
School of Electrical and Electronic Engineering,
The University of Manchester, Manchester, UK
*mohammed.abdulaal@postgrad.manchester.ac.uk
†alex.casson@manchester.ac.uk
‡patrick.gaydecki@manchester.ac.uk

***Abstract*—Brain-Computer Interface (BCI) is a technology that utilizes brainwaves to link the brain with external machines for either medical analysis, or to improve quality of life such as control and communication for people affected with paralysis. The performance of BCI systems depends on classification accuracy, which influences the Information Transfer Rate. This motivates researchers to improve their classification accuracy as best possible. A bias problem in reporting accuracies by using non-nested cross-validation methods was thought to increase accuracy. The aim of this paper was to validate and quantify such a concept by using a low-cost commercial EEG recorder to classify visually evoking face vs scrambled pictures, and report high accuracy using non-nested cross validation. The algorithm employed Independent Component Analysis followed by feature extraction with sample covariance matrices. The data were then classified using Support Vector Machines. The accuracy was tested with nested and non-nested cross-validation methods; accuracies obtained were 63% and 76%, respectively.**

I. Introduction

Brain Computer Interface (BCI) is the technology that utilizes brainwaves to link the brain with machines for various applications; including medical analysis, control of the environment, communication for those who are affected with partial or total paralysis, or any other directed purpose. BCIs use electroencephalography (EEG) as a means of measuring electrical activity in the brain via non-invasive electrodes that require no surgery or long preparation [1]. The most common BCI modalities are: Motor Imagery, P300-oddball, and Steady State Visually Evoked Potentials (SSVEP). For these systems to work efficiently, classification accuracies, often reported in Area Under Curve (AUC) percentages, must be high enough to maintain an acceptable level of Information Transfer Rate (ITR). This led to great motivation to make every effort to investigate accuracy improvements. The accuracy of visually evoked BCIs hugely depends on several factors such as the quality of the EEG recorder, experimental setup, nature of stimuli, and algorithm development.

The use of research-grade EEG recorders enhances the accuracy significantly, compared to commercially inexpensive EEG recorders due to differences in signal-to-noise ratios (SNRs). One study tested the classification accuracy using Emotiv EPOC (commercial EEG) with 14 electrodes, and a Biosemi headset (research-grade) with 32 electrodes employing the oddball paradigm [2]. They have found the accuracy of the 32-channel Biosemi headset to be 88.5% and the Emotiv to be 61.7%. Many other studies have examined performance of commercial EEG recorders employing other visually evoked BCIs like SSVEP and obtained similar results [3], [4].

A second important aspect is the type of visual stimuli and area affected by different classes stimuli. For example, many accuracies above 95% have been reported for the P300-oddball and SSVEP paradigms [5], [3], [4]. Meanwhile, face recognition based classification accuracies are rather inferior to this, just as discussed in the following.

Another important aspect is the classification algorithm, which is a multistage problem. Taking a Magnetoencephalography (MEG) classification competition of face vs scrambled images as a benchmark, which is a similar technology to EEG, the three best classification accuracies reported for subject-independent classification were 75%, 73%, and 71%, respectively. The winner also reported a subject-dependent classification accuracy of 86%. They utilized Event-Related Potential (ERPs) sample covariance matrices as features, and vectorization using tangent space along with a logistic regression (LR) classifier. The second place work involved down-sampled filtered raw-data as features with LR and random-forest (RF) classifiers combined. Whereas the third place used Support Vector Machines (SVMs). The leader-board is available via [6]. A more recent study [7] has reported the use of non-linear SVMs and an RF classifier with XxDAWN spatial filtering and reported a 71% accuracy based on EEG and 82% using MEG.

In this paper, we would like to shed light on another problem that affects reported accuracies, which is the use of nested vs non-nested cross-validation methods. This topic has generally been explained in [8]. The non-nested cross-validation method divides the data into training and testing parts, while nested cross-validation divides the data into training, validation and testing parts, forming two cross validation steps. The training data allow the classifier to learn the parameters and tune them

ISBN 978-90-827970-1-5 © EURASIP 2018

for testing the validation data, without access to test data. Non-nested cross-validation makes use of testing data for validation stage and report best accuracies, thereby increasing overall accuracy.

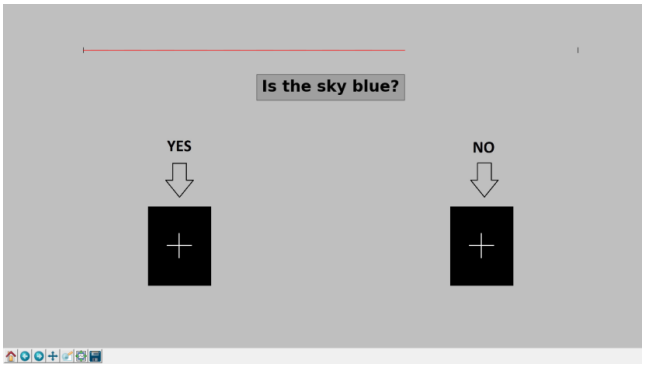
The problem lies in the fact that it is difficult to know which method was used unless clearly indicated creating an unjustified gap between high accuracies (above 95%) and medium accuracies (70-80%). In this study we will test the possibility of obtaining robust accuracies using a low-cost EEG recorder and visual perception. This will be compared using nested and non-nested cross-validation methods, which we hope will indicate that not all high accuracies reported using low-cost recorders are actually accurate and feasible for a real-time application. This will be achieved by using the best algorithm methods from the literature and assessing them with a dataset that was made public, then it will be tested on our data using the aforementioned cross-validation methods in the hopes it will enlighten the processes used by other researchers and motivate them to clearly indicate their reporting methodologies. We will also show pseudo-codes to indicate how both methods are utilized based on the Python platform.

This is not a new problem. It could be considered a fact that in machine learning, reporting training accuracy (non-nested cross-validation) results in higher but less generalizable accuracies than nested cross-validation. However, in BCI application we believe this is still used. Otherwise, the existence of large gaps in reporting accuracies cannot be justified. The aim here is to quantify this problem in practice using SVMs in a visual BCI application and motivate researchers to progress with it.

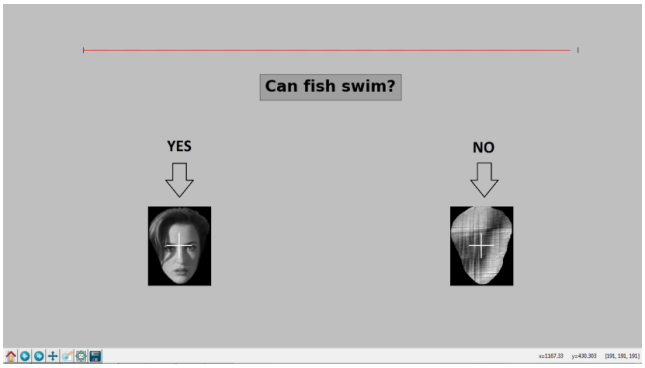
The overall methodology used in this paper was to design an ERP recognition system in Python and collect data using a synchronized Emotiv EPOC+. That will be followed by explaining the classification algorithm consisting of preprocessing, feature extraction and cross-validated classification using SVMs. The results will include analysis of a dataset of faces vs. non-faces, using different set-ups for comparison purposes. It will then include assessing the data collected employing the commercial EEG.

II. Methodology

Participants were requested to answer a list of questions by staring at pictures. Their visual perception determined their EEG behavior and this was utilized to enable communication. The software posted each question by presenting a message window that asked simple yes/no questions. The answer screen would appear with two options Yes (left) and No (right). The participant answered the question by staring at the cross sign beneath the words Yes and No accordingly, see Figure 1a. A timer represented by a red growing bar at the top of the screen indicated when the pictures would appear. At the end of the timer, two random images appeared where the cross signs were located. One showed a face and the other was a scrambled picture. Both were randomly chosen by the software; further, the association of a face with a yes/no was randomized so the participant could not predict the picture presented. The pictures were present for 500 ms and then were replaced by pictures of circles, see Figure 1b. The database used for the pictures was obtained from the dataset published by [9]. There were 100 questions in total for each subject.



(a) The software screenshot prior to stimulus



(b) The software screenshot during stimulus

Fig. 1: The software developed in Python to enable communication using visual perception

The device used in this experiment was an Emotiv EPOC+ with a sampling frequency of 128 Hz. Ten volunteers participated. Experiments last from 30 to 45 minutes. Electrode locations (using the 10-20 electrode location system) were AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4 with references at P3 and P4 (CMS and DRL respectively). Saline was employed to wet all electrodes. For reference electrodes, smell- and color-free water-soluble based gel was also used to ensure conductivity if the saline dried out. The setup time took less than five minutes for each participant. A synchronization circuit was needed to allow precise timings to the Emotiv apparatus with channels T7 and T8 thats sends triggers when the photos appeared on the screen. A battery-based system was developed to provide a direct link to the electrodes from the stimulus software, developed with Python. Similar work has been reported by [10]. The software sent a command using serial communication to a USB-connected microcontroller (via a photodiode attached to the monitor) to transmit a radio-frequency (RF) signal to the receiver attached to the EEG recording machine. The receiving device which

ISBN 978-90-827970-1-5 © EURASIP 2018

was enabled with RF communication triggered one of the electrodes of the EEG machine (T8) with a number of pulses of 330 μV for 8 ms, equivalent to 1 sample with a sample frequency of 128 Hz. The ground of the circuit was connected to another electrode (T7). Both electrodes were biased to the DRL electrode using 500 $k\Omega$ resistors. In the case of a failure, the resistors would limit any current flow to 9 μA. This limit was less than 10 μA for the CF Applied Part according to the IEC 60601-1 requirements. The overall system is presented in Figure 2.

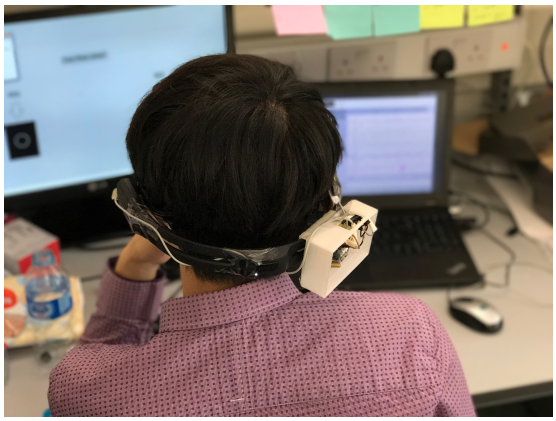


Fig. 2: The setup for the system. The EEG recording machine along with the synchronization circuit inside a three-dimensional (3D) printed enclosure. The thickness of the walls is less than 1mm making the total weight low and not affecting the balance of the headset

III. Algorithm

The classification algorithm consisted of three main components; filtering, feature extraction and classification which includes training and testing.

Raw data was fed into a 3rd order Butterworth band pass filter of cutoff frequencies between 2 and 20 Hz. This is required to remove low frequencies such as offsets and undesired high frequencies, like mains at 50 Hz. At this stage plotting the averages of the trials data in most cases did not reveal any ERP components, such as P300 or N170. More advanced filtering was needed. Independent Components Analysis (ICA) was then applied compute 12 independent components corresponding to the 12 recording channels. Independent components associated with visible ERPs were kept, either at the front with positive potential (given the reference point was at P3) and negative potentials at the occipital part the brain. The remaining components were removed. The process was carried out manually for all subjects. The noise was successfully removed and ERPs could now be seen (see Figure 3).

The ERP covariances of the data were then calculated. This resulted in a feature matrix of size $E \times E$ for each trial, where $E = 12$ is the number of electrodes. The covariances,

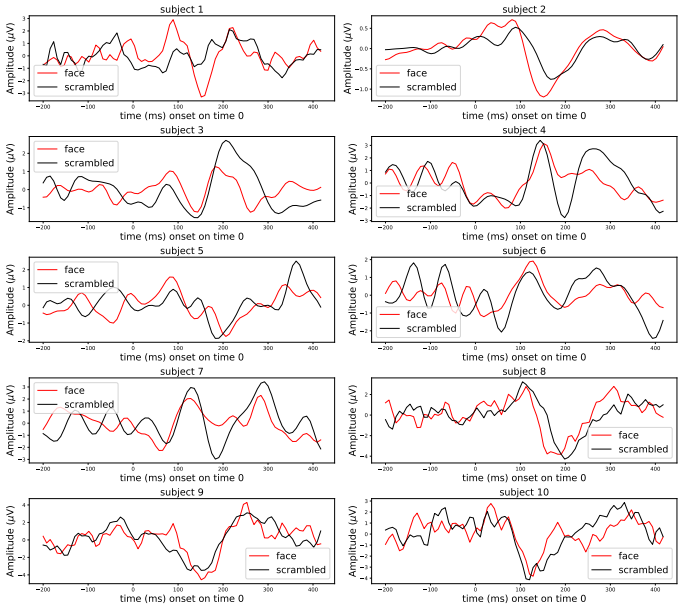


Fig. 3: Shows the average trials for both face and non-face trials using the average of four electrodes (P7, O1, O2, P8).

as discussed in [11], were calculated by firstly concatenating each trial $\mathbf{z}_i$ with the averages of each class $\mathbf{p}(1)$ and $\mathbf{p}(2)$:

$$\tilde{\mathbf{z}}_i = \begin{bmatrix} \mathbf{p}(1) \\ \mathbf{p}(2) \\ \mathbf{z}_i \end{bmatrix}$$

The spatial covariance matrix $\sigma_i \in \mathbb{R}^{12\times12}$ was therefore defined as:

$$\sigma_i = \frac{1}{N}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T \tag{1}$$

However, these features are in matrix form and need to be converted into vector form. To do this we needed to use Riemannian Geometry, which is explained in [12]. The Riemannian distance for two covariance matrices σ_1 and σ_2, representing classes 1 and 2, is defined by [13] as:

$$\delta(\sigma_1,\sigma_2) = \| \log({\sigma_1}^{-1/2}\sigma_2{\sigma_1}^{-1/2}) \|_F = [\sum_{e=1}^{E} \log^2 \lambda_e]^{1/2} \tag{2}$$

where λ_e , $e = 1...E$ are the real eigenvalues of ${\sigma_1}^{-1/2}\sigma_2{\sigma_1}^{-1/2}$ and $E = 12$ is the number of electrodes. Thus the Riemannian mean of the I covariance matrices is the matrix minimizing the sum of the squared Riemannian distances defined in [14] as:

$$\arg\min_{\sigma} \sum_{i=1}^{I} \delta_R^2(\sigma,\sigma_i) \tag{3}$$

to feed the features to the classifier it is necessary to project matrices in a vector Euclidean space, referred to as tangent space, leading a covariance matrix of size $E \times E$ to be represented by vectors of dimension $E(E+1)/2$. In this case E was 12.

ISBN 978-90-827970-1-5 © EURASIP 2018

an SVM classifier was used. SVMs were explained thoroughly in [15]. Using a Radial Basis Function (RBF) kernel, parameters conventionally known as C and γ needed tuning. Parameter C represents the cost i.e. the classification surface smoothness to compromise misclassification of training trials to gain a simpler decision surface, where γ represents the impact of individual samples on choosing support vectors. Tuning could be carried out in two ways that result in major differences in reporting accuracies; nested and non-nested cross-validation methods.

In nested cross-validation, the parameters are tuned using the training data without access to the testing data. Unlike non-nested cross-validation where the test data is used to optimize the parameters, and report scores based on best accuracies. We tested both methods and reported accuracies for comparison purposes. The values of parameters of C ranged from 1 to 1000, and γ ranged from 0.0001 to 0.1, with 10 folds each. The reported accuracy metric was in the Area Under the Curve (AUC), which is a well-known technique and more information on it could be found easily.

The pseudo code in Algorithm 1 describes one way for reporting cross-validation methods. Lines 2-5 indicate preparing the data for classification. Line 6 implements a grid-search method (based on Python) using 10-fold shuffle split of SVM. Line 7 signifies the training of the machines using the grid-search method. The difference lies in Line 8 where the non-nested best accuracy of the grid search is reported as system accuracy, whereas the nested methodology has an extra layer of cross-validation method using 10 folds and reports average accuracy.

Algorithm 1 Nested vs non-nested cross-validation methods

```
1: procedure GETACCURACY
2:     cv ← ShuffleSplit(n_splits=10)
3:     X ← data (trials,channels,samples)
4:     y ← labels
5:     X_features ← calculate features of X
6:     model ← GridSearchCV(estimator=SVM, cv)
7:     model.fit(X,y)
8:     non_nested score ← model.best_score
9:     nested score ← cross_val_score(model, X,y,cv)
```

IV. Results

A. Research-grade dataset validation

To test the algorithm and better analyze the original data, nested cross-validation SVMs were applied to an external dataset. This dataset was obtained from the OpenfMRI database. Its accession number is ds000117. The dataset included 16 subjects and used a total of 74 EEG electrodes and 306 MEG electrodes [9]. To properly confirm the algorithm, the same 12 electrodes were also tested in the analysis using nested cross-validation.

Table I shows the different accuracies obtained using different setup electrodes. It shows that the best accuracies of 86% and 85% were obtained using the 74 EEG electrodes and 306 MEG electrodes, respectively. Another column was added to compare the same 12 electrodes used in the Emotiv EPOC+. The test was also conducted using different number of training trials to test the effect of having different experiment lengths; 100 (similar to the EPOC+ experiment), 200, and 300 to find the point of accuracy convergence. The results demonstrated that using 300 training trials the accuracy was 77%. Further, using 200 training trials, the accuracy was 75% and using 100 training trials the accuracy was 70%. These results were used as a benchmark for our system accuracy.

TABLE I: Average accuracies for different setups: EEG or MEG (number of electrodes), and number of training trials .

Training Trials	MEG (306)	EEG (74)	EEG (12)
300	85.72	86.02	77.39
200	81.48	83.28	75.27
100	69.26	77.37	70.44

B. Commercial-grade collected data

Table II lists the accuracies obtained by running the algorithm for both the nested and non-nested cross-validation methods. Non-nested cross validation suggests a large superiority to nested cross-validation. The average accuracy for non-nested cross-validation was 76% and the average for nested cross-validation was 63%. The greatest difference was for subjects 6 and 9 at an increase of 19%. The lowest increase was for subject 8 at 6%.

TABLE II: Accuracies for different subjects.

Subject	Non-nested SVM	Nested SVM	Difference
1	0.70	0.56	0.14
2	0.75	0.61	0.13
3	0.86	0.77	0.09
4	0.76	0.68	0.08
5	0.76	0.60	0.16
6	0.85	0.67	0.19
7	0.64	0.55	0.09
8	0.74	0.68	0.06
9	0.78	0.59	0.19
10	0.74	0.59	0.15
Average	0.76	0.63	0.13

V. Discussion

The algorithm accuracies for the dataset were similar to those obtained by the winner at the competition for the same dataset. Combining SVM with covariances resulted in similar accuracies garnered by an LR classifier. Over-fitting of the RBF kernel, owing to limitations in the number of trials and less generalizable conditions, might become an advantage if reporting non-nested cross-validation accuracies. However, analysis of the non-nested classification of the research-grade dataset was not included in this paper based on space limitations. The number of electrodes affected the accuracy;

ISBN 978-90-827970-1-5 © EURASIP 2018

306 MEG and 74 electrodes resulted in similar accuracies. Reducing the number of electrodes to 12 diminished the accuracy from 86% to 77%. This shows the benefit of machine learning where more data, with localization and spatial filtering, improves accuracy. In addition, the effect of the number of training trials is significant. However, high cost is associated with having large number of trials, including subjects and time.

The accuracy of the system is not as high as 95% as reported by studies using P300 and SSVEP paradigms. This could be caused by a number of factors such as the ERP pattern differentiation between the target and non-target classes. The face and non-face ERP patters are very similar to each other when looking at N170 and P300 components. However, for the oddball paradigm, the ERP patterns of the target class have significant P300 components while there are no ERP components in the non-target class. The same applies to SSVEP where the frequencies are different for various stimuli classes.

Accuracy obtained by research-grade EEG recorders is expected to be superior to commercial EEG recorders. The algorithm accuracy when tested on the dataset with the research-grade EEG recorder resulted in an accuracy of 86% using a nested cross-validation method. This, however, was obtained by using 300 trials for training and 74 EEG electrodes. With the same 12 electrodes, as in the Emotiv EPOC+ and a similar number for training trials at 100 led to an accuracy of 70%, which is 7% more than the data obtained by the Emotiv EEG recorder of 63%. This difference is justified by the difference of SNR between the recorders. However, low-cost easy-to-use EEG is also favorable when time funding resources are limited.

On the other hand, the non-nested classification accuracy 76% is higher than expected when compared with the same number of training trials and electrodes using the research-grade EEG. This indicates that it is possible to report high accuracies with low-cost hardware and setup. The big difference in accuracies of the utilization of nested vs non-nested is an reflection of the importance of assessment tools and their implication in real-life application, where access to validation data is limited to validation and cannot be used for reporting accuracies.

VI. Conclusion

This paper analyzed different factors that affect accuracy performances in visually evoked BCI systems. It focused on the aspect of reporting accuracies by using nested vs. non-nested cross-validation methods. To analyze the matter in more detail, we collected face recognition EEG data associated with looking at pictures of faces and scrambled images with a synchronized commercial EEG recorder. It showed the possibility of reporting relatively high accuracies with such low-cost equipment by simply changing the reporting methodology. The algorithm for classifying faces and non-faces was designed based on the state-of-the-art in face recognition. The algorithm was tested on a dataset and deemed acceptable. The accuracy of the system using the collected data had accuracies of 63% and 76% using nested and non-nested cross-validation methods, respectively. The nested cross-validation accuracy of 63% was compared to obtaining 12 EEG electrodes from a research-grade EEG that resulted in 70% accuracy. The non-nested accuracy of 76% was thought to be higher that the usual, which was accomplished by reporting bias. This raised the question of whether all reported accuracies in the literature obtained by either low-cost or research-grade EEG that were of very high accuracy, were accurate representations of reality. This paper hoped to encourage researchers to clearly indicate their cross-validation methodology to reduce confusion caused by reading different accuracies in the literature.

Acknowledgment

This work was part of a PhD research sponsored by King Abdulaziz University, Jeddah, Saudi Arabia.

References

[1] J. Wolpaw and E. W. Wolpaw, *Brain-computer interfaces: principles and practice*. Oxford University Press, 2012.

[2] F. Nijboer, B. Van De Laar, S. Gerritsen, A. Nijholt, and M. Poel, "Usability of three electroencephalogram headsets for brain–computer interfaces: a within subject comparison," *Interacting with computers*, vol. 27, no. 5, pp. 500–511, 2015.

[3] N. Masood and H. Farooq, "Emotiv-based low-cost brain computer interfaces: A survey," in *Advances in Neuroergonomics and Cognitive Engineering*. Springer, 2017, pp. 133–142.

[4] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the emotiv epoc headset for p300-based applications," *Biomedical engineering online*, vol. 12, no. 1, p. 56, 2013.

[5] Y. Liu, X. Jiang, T. Cao, F. Wan, P. U. Mak, P.-I. Mak, and M. I. Vai, "Implementation of ssvep based bci with emotiv epoc," in *Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS), 2012 IEEE International Conference on*. IEEE, 2012, pp. 34–37.

[6] "Decmeg2014 - decoding the human brain." [Online]. Available: https://www.kaggle.com/c/decoding-the-human-brain/leaderboard

[7] S. Fatima and A. M. Kamboh, "Decoding brain cognitive activity across subjects using multimodal m/eeg neuroimaging," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 3224–3227.

[8] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[9] D. G. Wakeman and R. N. Henson, "A multi-subject, multi-modal human neuroimaging dataset," *Scientific data*, vol. 2, 2015.

[10] P. de Lissa, S. Sörensen, N. Badcock, J. Thie, and G. McArthur, "Measuring the face-sensitive N170 with a gaming EEG system: A validation study," *Journal of neuroscience methods*, vol. 253, pp. 47–54, 2015.

[11] L. Korczowski, M. Congedo, and C. Jutten, "Single-trial classification of multi-user p300-based brain-computer interface using riemannian geometry," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 1769–1772.

[12] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2012.

[13] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 735–747, 2005.

[14] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

ISBN 978-90-827970-1-5 © EURASIP 2018