2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Mental Effort Estimation by Passive BCI: A Cross-Subject Analysis

Nicolina Sciaraffa, Daniele Germano, Andrea Giorgi, Vincenzo Ronca*, *Member, IEEE*, Alessia Vozzi, *Member, IEEE*, Gianluca Borghini, Gianluca Di Flumeri, Fabio Babiloni, *Member, IEEE* and Pietro Aricò

***Abstract*—Despite the technological advancements, the employment of passive brain computer interface (BCI) out of the laboratory context is still challenging. This is largely due to methodological reasons. On the one hand, machine learning methods have shown their potential in maximizing performance for user mental states classification. On the other hand, the issues related to the necessary and frequent calibration of algorithms and to the temporal resolution of the measurement (i.e. how long it takes to have a reliable state measure) are still unsolved. This work explores the performances of a passive BCI system for mental effort monitoring consisting of three frontal electroencephalographic (EEG) channels. In particular, three calibration approaches have been tested: an intra-subject approach, a cross-subject approach, and a free-calibration procedure based on the simple average of theta activity over the three employed channels. A Random Forest model has been employed in the first two cases. The results obtained during multi-tasking have shown that the cross-subject approach allows the classification of low and high mental effort with an AUC higher than 0.9, with a related time resolution of 45 seconds. Moreover, these performances are not significantly different from the intra-subject approach although they are significantly higher than the calibration-free approach. In conclusion, these results suggest that a light (three EEG channels) passive BCI system based on a Random Forest algorithm and cross-subject calibration could be a simple and reliable tool for out-of-the-lab employment.**

I. Introduction

Mental effort, mental workload and mental strain are the most widely used terms to define the relationship between the cognitive resources of a subject performing a task and the difficulty of the task itself [1]. The interest in such a cognitive state was born in the field of human factors, where the monitoring of an operator's cognitive state is crucial to avoid onerous consequences. Thanks to the convergence of human factors requirements and neuroscience techniques, neurophysiological measures have been proposed as a valid tool to provide an objective and continuous, as well as online measurement of an operator's mental effort, leading to the concept of passive brain computer interface (BCI). A typical passive BCI translates the brain activity unconsciously modulated by the subject's cognitive state into an output aiming to trigger the surrounding environment [2]. On the one hand, the acquisition of brain activity through electroencephalography (EEG) can be easily performed thanks to the recent technological improvement of minimally invasive systems, with few and gel-free electrodes [3]. On the other hand, the measurement of mental effort during real tasks by means of EEG signals is still challenging. The employment of Machine Learning (ML) techniques has allowed a step forward in this direction, by paving the way to decode and characterize task-relevant brain activity modulation and to distinguish it from noninformative features. However, there are still practical issues preventing the employment of passive BCI out of the lab.

One crucial aspect is related to calibration. In a supervised approach, labelled observations must be available in order to calibrate the passive BCI, before using it as a predictor for new data [4]. In this regard, a reliable system needs a calibration phase, which is long and difficult to be performed during a realistic use, considering also that to work properly ML needs to be re-calibrated frequently [5]. In recent years, many studies have been focused on finding a methodology to avoid or at least reduce the calibration phase, proposing unsupervised techniques that at the moment suffer from a lower performance compared to supervised ones [6].

A second important aspect is the temporal resolution requested from the passive BCI, that is how long it takes to have a reliable state measure. In safety-critical applications, like for example pilots monitoring, the passive BCI is required to provide answers within few seconds. Conversely, if used during the operator's training phases it should react also in longer times, for example, every 30 seconds or even minutes. A low temporal resolution usually corresponds to a high level of algorithm accuracy. Therefore, it is necessary to find a good compromise between a proper temporal resolution, able to guarantee an acceptable classification accuracy, depending on the specific context requirements.

In this work, we compared the performance of a passive BCI system monitoring mental effort using different temporal resolutions and three different approaches for calibration. Firstly, we performed an *intra-subject* calibration, which is the standard approach during which the system is trained using the calibration dataset available from the subject him/herself. Secondly, to mitigate the calibration issue, we tested a c*ross-*

N. Sciaraffa is with Dept. Molecular Medicine, Sapienza University of Rome, Italy (e-mail: nicolina.sciaraffa@uniroma1.it).

D. Germano and A. Giorgi are with BrainSigns srl, Rome, Italy

A. Vozzi and V.Ronca are with BrainSigns srl, Rome, Italy and Dept. Anatomical, Histological, Forensic & Orthopedic Sciences, Sapienza University of Rome, Italy (corresponding author: phone: +39 06 51501163; e-mail: vincenzo.ronca@uniroma1.it).

G. Borghini, P. Aricò, and G. Di Flumeri are with Dept. Molecular Medicine, Sapienza University of Rome, Italy, BrainSigns srl, Rome, Italy (e-mail: [gianluca.borghini; pietro.arico; gianluca.diflumeri]@uniroma1.it.

F. Babiloni are with Dept. Molecular Medicine, Sapienza University of Rome, Italy, BrainSigns srl, Rome, Italy, and the College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China (e-mail: fabio.babiloni@uniroma1.it).

subject approach, training the model with the calibration data of other subjects performing the same task. In fact, after recording brain activity from a specific pool of subjects, it could be possible to avoid the calibration from a new coming subject. Finally, we computed the mental effort neurometric (the average of frontal activity in the theta band) as a calibration-free method [7]. To mimic a real situation, in which few EEG channels should be employed, we used just three frontal EEG channels to realize the previously mentioned approaches.

II. Material and Methods

A. Dataset

In the current work, we used the EEG data described in [8]. The experiment was conducted following the principles outlined in the Declaration of Helsinki of 1975, as revised in 2000. It received the favourable opinion from the Ethical Committee of the National University of Singapore (NUS), Centre for Life Sciences (NUS-IRB Ref. No: 13-132, NUS-IRB Approval No: NUS 1864). In particular, eight subjects performed the NASA - Multi Attribute Task Battery (MATB, Fig. 1), a computer-based multi-task designed by the NASA to evaluate the cognitive operational capability during simulated conditions requiring different levels of mental effort. For this analysis we aimed to discriminate between the Easy and Hard task levels, classifying respectively the Low and the High mental effort. In particular, four repetitions of 2.5 minutes long Easy and Hard conditions were available for each subject.

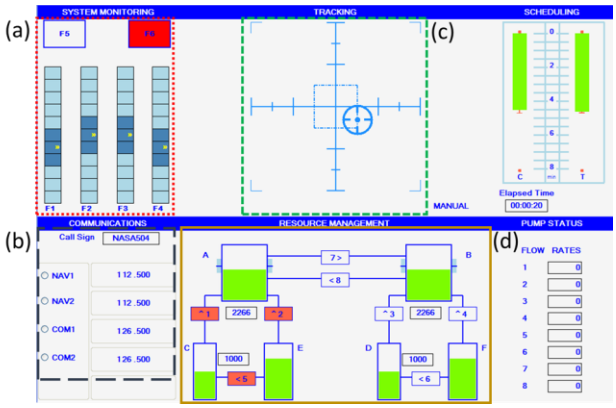

Figure 1. Multi Attribute Task Battery (MATB) interface. There is (a) the emergency lights task; (b) the task of cursor tracking; (c) the radio communication task and, finally, (d) the fuel levels managing

B. Features Computation

The EEG signal was band-pass filtered (1-30 Hz) and then segmented into epochs of 2 seconds, shifted by 0.125 seconds. The FPz channel has been used to remove eye-blinks contribution from each epoch of the EEG signal, by using the Reblinca algorithm [9]. Therefore, a threshold criterion has been applied: if the EEG signal overcomes the threshold of 100 (μV) it has been marked as artifact and removed from the analysis. After that, for each epoch, the power spectral density (PSD) was calculated using a periodogram with Hanning window (2 seconds). As mentioned before, the PSD values in the theta band and for three channels (F3, F4, Fz) has been used as spectral features.

C. Algorithm and Tuning

For this analysis, the Random Forest (RF) has been employed. The RF is a nonlinear classifier [10] belonging to the ensemble methods. This family of classifiers allows generalizing well to new data constructing sets of individual and independent tree classifiers. For this analysis, we used the *RandomForestClassifier* function contained in the Scikit-learn package [11]. This function allows us to define the input parameters described in Table 1.

TABLE I. Random Forest Parameters

Parameter	Values
criterion	gini or entropy
n estimators	From 1 up to 100
min samples split	From 5 up to 100
min sample leaf	From 5 up to 100
max leaf nodes	From 2 to 20
min impurity decrease	From 0.00005 to 0.01
max depth	From 2 up to 100

An iterative process was built in order to optimize a limited number of parameters at each step. Once the parameters were divided into blocks, *GridSearchCV* functions of the Scikit-learn [11] library has been recalled setting the cross-validation parameter to 5. At each iteration the optimization process is divided into two steps: after the optimization of a block of parameters, they have been optimized again in a restricted range of values.

D. Intra-Subject Approach

In this case, labelled observations of each subject have been used to calibrate the model. The first 30 seconds of the training dataset (one of the four repetitions of Easy and Hard conditions) was used for calibration. We set 30 seconds, to simulate a realistic condition in which users can dedicate a low amount of time for the calibration phase. To assess the performance of the model a cross-validation approach has been used. In particular, the training dataset has been divided into 4 sessions and one session has been iteratively selected for training to resemble a realistic approach when there are a low number of data available for training.

E. Cross-Subject Approach

The cross-subject approach differs from the intra-subject because the training phase employed a set of data coming from subjects different from the subject of the testing phase, i.e. in this case for each subject a calibration dataset coming from the remaining subjects was generated (i.e. leave-one-subject-out calibration). The calibration dataset was built taking 7 subjects and the remaining subject was used to test the performance of the classifier.

F. Mental effort Neurometric

When calibration data are not available, mental effort can be assessed through a neurometric, an index based on the physiological knowledge underpinning mental effort. In particular, most of the studies showed that the brain electrical activity mainly involved in the mental effort analysis is the theta brain rhythms gathered from the Pre-Frontal Cortex (PFC) region [12], [13]. Therefore, the mental effort neurometric has been computed averaging the PSD value on the three frontal channels and in the theta band.

G. Performance evaluation

The performances among the different conditions have been assessed by means of the Area under Curve (AUC,[14]). In particular, for each data point of testing data, it has been calculated the probability of belonging to the Hard class in output from the RF model, both for the intra and cross-subject approaches. Instead, for the neurometric evaluation, the PSD averaged over the 3 frontal electrodes on the same testing dataset has been used. It was then applied a moving average on these values, to simulate different time resolutions, starting from 0.125 seconds, up to 64 seconds (maximum value to be able to evaluate a reliable AUC value).

The Wilcoxon signed-rank test has been used to compare the performance obtained through the cross-subject approach respectively with the intra-subject and the neurometric.

III. RESULTS

Figure 2 shows the statistical comparison between the cross-subject and the intra-subject performance for different time resolutions (from 0.125 to 64 seconds). As hypothesized, as the time resolution decreases the accuracy increases, reaching a plateau around 30 seconds. The intra-subject calibration reaches 0.98, whereas the cross-subject slightly exceeds 0.9. For any time resolutions, the AUC values are not significantly different between the two approaches.

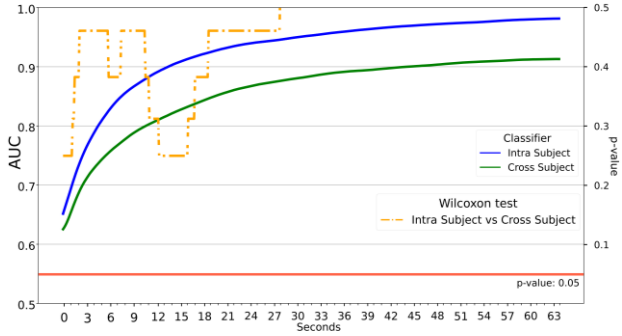

Figure 2. Averaged AUC values of intra-subjects (blue) and cross-subjects (green) results for different time resolutions (0.125-64 seconds). The dashed yellow line represents the results in terms of p-values of the Wilcoxon test. The threshold of significance (p-value = 0.05) has been represented in red and for readability p-values>0.5 have not been shown.

The cross-subject approach has been compared with the neurometric. Figure 3 shows the statistical comparison between the cross-subject and the neurometric performances for different time resolutions. For every time-resolution higher than 2 seconds the performances obtained with the cross-subject approach are significantly higher than those obtained without calibration, namely with the neurometric. In particular, the neurometric starting from 18 seconds settles AUC to 0.73.

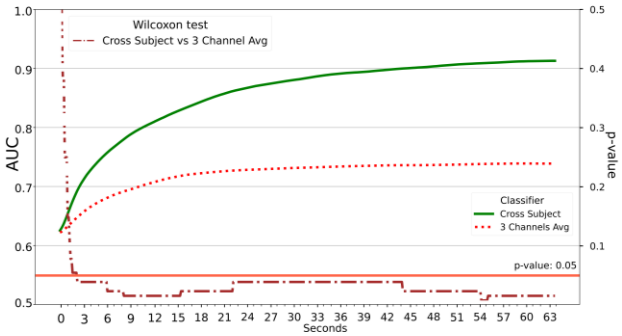

Figure 3. Averaged AUC values of cross-subjects (green) and neurometric results (dotted red) for different time resolutions (0.125-64 seconds). The dashed burgundy line represents the results in terms of p-values of the Wilcoxon test. The threshold of significance (p-value = 0.05) has been represented in red.

IV. DISCUSSION

The aim of this work was to test the performance of a light passive BCI system in classifying mental effort associated with two levels of multitasking. In particular, we tested different calibration approaches and time resolutions, comparing the cross-subject performances with those obtained through intra-subject calibration and a calibration-free approach based on physiological knowledge, namely the mental effort neurometric [7].

We observed that intra-subject calibration reaches the highest values of performance, higher than 0.95, overcoming 0.9 even at moderate temporal resolutions (15 seconds). The cross-subject approach did not provide significantly different performance compared to intra-subject, however we observed that to obtain an AUC of 0.9 it is necessary to decrease the time resolution to 45 seconds. Finally, the neurometric showed the lowest performance: even for the lowest time resolution (64 seconds) the AUC did not exceed 0.73 and is significantly lower than the cross-subject approach.

Even if several steps forward have been made since it has been accepted the possibility that neurophysiological measures would be employed to predict operator's cognitive state, nowadays is still not possible to use a passive BCI to assess the mental effort out-of-the-lab. A recent review dealing with workload recognition using EEG and machine learning raised three open issues that require further investigation: real-time design, generalizability, and interpretability of the models [15]. The current work fulfils these needs.

Firstly, the number of channels usually used to classify workload varies from 2-128 [15]. In the case of real-time experimental design, it has been highly recommended to prefer the low- number of channels approach to reduce both set-up and algorithm related cost of time. In this work, we considered only three frontal electrodes. In fact, most of the recent EEG systems off-the-shelf consist mainly of a low number of frontal electrodes, therefore methodological approaches should consider this technological aspect to increase the usability of the passive BCI out-of- the-lab.

Secondly, the cross-subject approach typically has lower performances compared to the intra-subject approach, due to

the higher complexity and individual variabilities [15]. Therefore, the generalizability of models needs to be investigated further also developing more robust machine learning models. In this work, we compared intra-subject and cross-subject approaches to allow for a direct comparison of the obtained performance. Moreover, testing also the neurometric performances we highlighted the advantages of employing a machine learning method based on cross-subject calibration compared to calibration-free approach. In fact, it has been observed that by balancing the time resolution it is possible to obtain classification performances higher than 0.9. Only 7 subjects have been used to train the model, however, all the available training sessions have been employed (about 10 minutes easy and 10 minutes hard per subject). This could explain the fact that the obtained performances are higher compared to a previous work, where accuracy of 80% has been obtained employing a hierarchical Bayes model with cross-subject calibration during a similar MATB task on 8 subjects [16]. When there is the possibility to perform a calibration on the single subject, the obtained results proved that the proposed intra-subject approach allows the classification of low and high mental effort using just 30 seconds of calibration, which is in line with the results already obtained in the same conditions (30 seconds of calibration to discriminate 2 classes) [17].

Thirdly, regarding the importance of the models interpretability[15], the employment of the Random Forest model allows for higher interpretability compared to other less "transparent" models.

V. Conclusion

Our results confirmed that a passive BCI system consisting of just three frontal EEG electrodes and a random forest model, calibrated using a cross-subject approach, could be a valuable and simple tool that can prevent performing a subject-dependent calibration. This could pave the way to the definition of an online index for mental effort assessment. Notwithstanding these encouraging results, the cross-subject approach is only one of the possible solutions to avoid passive BCI subject-dependent calibration. Other approaches such cross-task calibration, or other methods like unsupervised classification and transfer learning should be tested further to have a complete overview of calibration-free possibilities. In addition, it has to be underlined that the MATB is something close to a laboratory-based task. In this regard, it would be relevant to test the same approach on a task closer to a real-world situation (e.g. user driving a car, or a pilot flying an airplane).

Acknowledgment

This work is co-financed by the European Commission by Horizon2020 projects "MINDTOOTH" (GA n. 950998). H2020-SESAR-2019-2 projects: ''ARTIMATION,'' (GA n. 894238); "WORKINGAGE" (GA n. 826232); "SI-MUSAFE" (GA n. 723386); "SAFEMODE" (GA n. 814961), "BRAIN-SAFEDRIVE" (Italy-Sweden collaboration) with a grant of Ministero dell'Istruzione dell'Università e della Ricerca della Repubblica Italiana.

References

[1] N. Sciaraffa, P. Aricò, G. Borghini, G. Di Flumeri, A. Di Florio, and F. Babiloni, "On the Use of Machine Learning for EEG-Based Workload Assessment: Algorithms Comparison in a Realistic Task," in *International Symposium on Human Mental Workload: Models and Applications*, 2019, pp. 170–185.

[2] P. Arico, G. Borghini, G. Di Flumeri, N. Sciaraffa, A. Colosimo, and F. Babiloni, "Passive BCI in Operational Environments: Insights, Recent Advances, and Future Trends," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1431–1436, 2017.

[3] G. Di Flumeri, P. Aricò, G. Borghini, N. Sciaraffa, A. Di Florio, and F. Babiloni, "The Dry Revolution: Evaluation of Three Different EEG Dry Electrode Types in Terms of Signal Spectral Features, Mental States Classification and Usability," *Sensors*, vol. 19, no. 6, p. 1365, 2019.

[4] F. Schettini, F. Aloise, P. Aricò, S. Salinari, D. Mattia, and F. Cincotti, "Self-calibration algorithm in an asynchronous P300-based brain–computer interface," *J. Neural Eng.*, vol. 11, no. 3, p. 35004, 2014.

[5] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, no. 2, pp. 387–399, 2011.

[6] B. Blankertz *et al.*, "The Berlin brain-computer interface: progress beyond communication and control," *Front. Neurosci.*, vol. 10, p. 530, 2016.

[7] P. Aricò *et al.*, "Human-Machine Interaction Assessment by Neurophysiological Measures: A Study on Professional Air Traffic Controllers," in *EBMC 2018, 40th International Engineering in Medicine and Biology Conference*, 2018.

[8] G. Borghini *et al.*, "A new perspective for the training assessment: Machine learning-based neurometric for augmented user's evaluation," *Front. Neurosci.*, vol. 11, p. 325, 2017.

[9] G. Di Flumeri, P. Aricó, G. Borghini, A. Colosimo, and F. Babiloni, "A new regression-based method for the eye blinks artifacts correction in the EEG signal, without using any EOG channel," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, 2016, pp. 3187–3190.

[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[12] A. Gevins, M. E. Smith, L. McEvoy, and D. Yu, "High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice.," *Cereb. cortex (New York, NY 1991)*, vol. 7, no. 4, pp. 374–385, 1997.

[13] G. Borghini *et al.*, "Frontal EEG theta changes assess the training improvements of novices in flight simulation tasks," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 6619–6622.

[14] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Psychol.*, vol. 12, no. 4, pp. 387–415, 1975.

[15] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, and D. Zhang, "Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review," *IEEE Trans. Cogn. Dev. Syst.*, 2021.

[16] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, and W. D. Gray, "Cross-subject workload classification with a hierarchical Bayes model," *Neuroimage*, vol. 59, no. 1, pp. 64–69, 2012.

[17] 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," *Conf. Hum. Factors Comput. Syst. - Proc.*, no. December 2014, pp. 835–844, 2008.