Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks

Pietro Aricò *Member, IEEE*, Gianluca Borghini *Student Member, IEEE*, Gianluca Di Flumeri, Alfredo Colosimo, Ilenia Graziani, Jean-Paul Imbert, Géraud Granger, Railene Benhacene, Michela Terenzi, Simone Pozzi, and Fabio Babiloni *Member, IEEE*

***Abstract*—Machine-learning approaches for mental workload (MW) estimation by using the user brain activity went through a rapid expansion in the last decades. In fact, these techniques allow now to measure the MW with a high time resolution (e.g. few seconds). Despite such advancements, one of the outstanding problems of these techniques regards their ability to maintain a high reliability over time (e.g. high accuracy of classification even across consecutive days) without performing any recalibration procedure. Such characteristic will be highly desirable in real world applications, in which human operators could use such approach without undergo a daily training of the device. In this work, we reported that if a simple classifier is calibrated by using a low number of brain spectral features, between those ones strictly related to the MW (i.e. Frontal and Occipital Theta and Parietal Alpha rhythms), those features will make the classifier performance stable over time. In other words, the discrimination accuracy achieved by the classifier will not degrade significantly across different days (i.e. until one week). The methodology has been tested on twelve Air Traffic Controls (ATCOs) trainees while performing different Air Traffic Management (ATM) scenarios under three different difficulty levels.**

I. Introduction

In the recent decades, research is focusing on the evaluation of user's mental states based on his/her neurophysiological activity in operating environments. In this regard, the mental workload (MW) monitoring is of particular interest especially in safety-critical applications where human performance is often the least controllable factor. In fact, as the MW increases, it became harder to maintain the user's task performance within an acceptable range, resulting then into an increasing of errors' occurrence [1].

P. Aricò is with the Dept. Physiology and Pharmacology, University "Sapienza" of Rome, Italy and with the IRCCS Fondazione Santa Lucia, Rome, Italy (corresponding author to provide phone: +39 3292973269; e-mail: p.arico@hsantalucia.it).

G. Borghini is with the Dept. Physiology and Pharmacology, University "Sapienza" of Rome, Italy and with the IRCCS Fondazione Santa Lucia, Rome, Italy

I. Graziani is with the Dept. Physiology and Pharmacology, University "Sapienza" of Rome, Italy.

G. Di Flumeri and A. Colosimo are with the Dept. Anatomical, Histological, Forensic & Orthopedic Sciences, University "Sapienza" of Rome, Italy.

J.P. Imbert, G. Granger, R. Benhacene are with Ecole Nationale de l'Aviation Civile Toulouse, France

M. Terenzi, S. Pozzi are with the Deep Blue Research and Consulting Rome, Italy.

F. Babiloni, is with the Dept. Molecular Medicine, University "Sapienza" of Rome, Italy and BrainSigns srl. Rome, Italy.

The theta (θ: over the frontal and the occipital sites) and alpha (α: over the parietal sites) rhythms of the electroencephalographic (EEG) signal have been taken into account in several studies because of their strong correlation with the MW variations [2], [3].

In this regard, the application of machine-learning techniques (MLTs) to the MW evaluation based on the measurement of brain activity is growing continuously. In general, the use of these techniques allows to assess subjects' MW in a short time (few seconds), reaching high binary discrimination accuracy (DA~90%) [2], [5]. These algorithms are able to extract from a big amount of physiological data within a training dataset the most significant features closely related to the user's mental state (i.e. MW). Then, based on those features it should be possible to assess the user's MW level and to keep the DA stable across different days. Actually, one of the big concerns of the MLTs is related to the capability of such algorithms to extract from the training dataset only those neurophysiological features by which the reliability of the measure could remain stable over time, providing a high DA across different days. Nowadays, the effects of day-to-day fluctuations in the operator's brain signals have not been thoroughly assessed while operators are engaged in complex tasks. Different studies showed that the performance of classifiers in evaluating the different MW levels of the user dramatically decrease over days [6],[7]. Despite the EEG is not a stationary process, let us assume that few neurophysiological features of the user related to the MW could remain enough stable over time. In this way, a classifier trained with those features should not degrade his DA over time. On the contrary, it will be indicative that it becomes too specific to the training dataset (*overfitting)*. The reduction of features used by the classifier during its calibration phase could mitigate the overfitting and improve the reliability and stability of the DA over time [8].

In this work, we hypothesized that if a simple classifier is calibrated by using a low number of brain spectral features between those ones strictly related to the MW (i.e. Frontal and Occipital θ, Parietal α rhythms), those features will make the classifier DA high across the days. In particular, the DA of a linear classifier across an entire week has been investigated by using EEG data related to MW collected in a professional school of Air Traffic Management (ATM) while Air Traffic Controllers (ATCos) trainees were performing several ATM scenarios characterized by increasing level of difficulty.

II. Materials and Methods

A. Experimental protocol

Twelve ATCo trainees (age 25±3) from ENAC (Ecole Nationale de l'Aviation Civile, Toulouse, France) have been involved in this experimentation. In particular the experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board. The experimental task used in this study was the Labyrinth (LABY, Figure 1), a functional simulated ATM environment, developed by ENAC [9]. The difficulty of the task can be altered according to how many aircraft the participant have to control, the number and type of clearances required over the time and the number/trajectory of other interfering flights. Subjects were asked to execute the LABY task under three difficulty conditions (Easy [E], Medium [M] and Hard [H]) chosen by ATM Experts at ENAC. Controllers have been trained to use LABY before starting with the experimentation. The experimental protocol has been composed of three recording sessions performed on three different days. The first two sessions have been performed in two consecutive days named hereafter as *Day 1* and *Day 2*. The last session has been performed after one week from the last one (*Day 9*). Each session consisted in twelve runs, in which subjects performed the three LABY conditions (E, M, H) four times each in a randomized sequence. Each condition lasted 3 minutes. Also, in order to avoid habituation and expectation effects, some task parameters have been randomly changed across the experimental sessions. In summary, the whole dataset has been composed of twelve triplets of conditions (4 triplets of E, M, H conditions for each of the three experimental days).

Figure 1. The LABY task, developed by ENAC

B. EEG-based workload index

Neurophysiological signals have been recorded by the digital monitoring BEmicro system (*EBNeuro, Italy*). Thirteen EEG (FPz, F3, Fz, F4, AF3, AF4, P3, Pz, P4, POz, O1, Oz, O2) and one vertical EOG channels have been collected simultaneously with a sampling frequency of 256 (Hz). All the EEG electrodes have been referenced to both the earlobes, and the impedances of the electrodes were kept below 10 (kΩ). The acquired EEG signals have been digitally band-pass filtered by a 4^{th} order Butterworth filter [1÷30] (Hz) and the EOG signal has been used to remove eyes-blink artifacts from the EEG data. The EEG signal has been then segmented into epochs of 2 seconds, shifted of 0.125 seconds and the Power Spectral Density (PSD) has been calculated for each EEG epoch by using only the frequency bands directly correlated to the MW (frontal and occipital θ and parietal α bands, [3-12] (Hz)).

C. Classifier features selection

In this study we chose to use a StepWise Linear Discriminant Analysis (SWLDA) regression [2] that is one of the best outperforming classifiers, in fact with respect to other methods it has the advantage of having automatic features extraction, so that insignificant terms are statistically removed from the model. In particular, a three-classes SWLDA has been used to select the most relevant EEG spectral features, within the training dataset, to discriminate the MW level among the three task conditions (E [Label = 0], M [Label = 0.5] and H [Label = 1]), and then the linear discriminant function has been evaluated to test the reliability of the feature selection criteria. For each PSD epoch, only the frequencies strictly related to the MW have been considered to train and test the classifier.

In a SWLDA regression, the input features are usually weighted by using ordinary least-squares regression to predict the target class label (i.e. 0, 0.5, 1). At each step, a new term can be added to or deleted from the model (if p-value < α_{ENTER} or if p-value > α_{REMOVE}). This process goes on until the predefined number of significant features is reached ($Iteration_{MAX}$), unless there are no more features satisfying the entry (α_{ENTER}) and the removal (α_{REMOVE}) conditions [10]. Normally, it is possible to optimize a SWLDA regression by tuning all or few of the three parameters available in the algorithm (α_{ENTER}, α_{REMOVE} and $Iteraction_{MAX}$). For reducing the degrees of freedom of the problem, we chose to not impose constrains over the first two parameters α_{ENTER}, α_{REMOVE}, but to tune only the $Iteraction_{MAX}$ value. In other words, a "forward SWLDA" has been implemented. In this way, it was possible to impose the SWLDA to select a features' number equal to the number of iterations ($Iteration_{MAX}$). Despite no constrains on the α_{ENTER} and the α_{REMOVE} parameters, features are included in the model in order of significance (i.e. the first feature added into the model will be the most significant one, and so on). In this way, we can be sure that the first features included in the model are also the most significant ones. In particular, we performed simulations by using three different values of the $Iteration_{MAX}$ parameter, that is 5%, 50% and 100% of the available features (Frontal and Occipital θ, Parietal α). We used the 5% to test our hypothesis, so that to have a low number of features used for the calibration of the classifier. As we stated before, we expected that by using a low number of features (i.e. 5%) the DA of the classifier would be reliable over time. On the contrary, in the other cases (50% and 100%) the DA should decrease over time.

D. Mental workload index based on EEG activity

The linear discriminant function ($y_{test}(t)$) for each window of 2 seconds has been computed by using the coefficients (weights: w_{itrain} and bias: b_{train}) returned by the SWLDA function (equation 1, where $f_{itest}(t)$ represents the PSD matrix of the testing dataset at the time sample t, and of the i^{th} feature). Finally, we applied a moving average of 8 seconds (*8MA*) to the $y_{test}(t)$ function in order to smooth it out by

reducing the variance of the measure, and we defined it as *EEG-based workload index* (W_{EEG}). Here below the SWLDA discriminant function (1) and the W_{EEG} index (2) equations are reported.

$$y_{test}(t) = \sum_i w_{i\,train} \cdot f_{i\,test}(t) + b_{train} \quad (1)$$

$$W_{EEG} = 8MA\,(y_{test}(t)) \quad (2)$$

E. Cross-validations between days

For each subject, different cross-validations have been performed by training the classifier with one triplet of E, M, H conditions and by testing it over the other triplets. In particular, to investigate the stability of the measure across different days, we considered three types of cross-validations. The *Intra* cross-validation type, where the training and testing triplets belonged to the same day; the *Short term* cross-validation type, where the training triplets belonged to Day 1 (Day 2) and the testing triplets to Day 2 (Day 1); and finally, the *Medium term* cross-validation type, where the training triplets belonged to Day 1 or Day 2 (Day 9) and the testing triplets to Day 9 (Day 1 or Day 2).

F. Performed analyses

Discriminant Accuracy (DA) analysis: We performed analysis by using different $Iteration_{MAX}$ values, the 5%, 50% and 100% of the available features. For each testing triplet, we calculated the W_{EEG} indexes. At this point, *Area Under Curve* (AUC) values of the *Receiver Operating Characteristic* (ROC, [11]) have been calculated by considering couples of W_{EEG} distributions (E vs H, M vs H and E vs M) in order to test the DA of the classifier.

Workload distribution analysis (W_{EEG}): For each subject we evaluated the W_{EEG} distributions over the testing dataset, by considering the best $Iteration_{MAX}$ value resulting from the DA analysis. Two two-ways repeated measures ANOVA (CI = .95) analyses have been performed, one on the AUC values and the other one on the W_{EEG} distributions. In the first one, we averaged the AUC values related to the three difficulty levels (E vs H, M vs H and E vs M) considering as *within* factors the "$Iteration_{MAX}$ values" (5%, 50%, 100%) and the "three cross-validation types" (Intra, Short term, Medium term). In the second one, we fixed the $Iteration_{MAX}$ value (based on the first ANOVA results) and we considered the W_{EEG} by using as *within* factors the three "conditions" (E, M, H) and the three "cross-validation types" (Intra, Short term, Medium term). Post-hoc tests (Bonferroni correction for multiple comparisons) have been performed to assess significant differences among all pairs of levels of the considered factors. Before every statistical analysis, we used the *z-score* [12] correction formula to normalize the different behaviors of the subjects. In particular, we calculated this score by using the mean and the standard deviations of the related values (i.e. AUC, scores) over the conditions (i.e. classifiers, cross-validations, difficulty levels).

III. Results

Discriminant Accuracy (DA): Figure 2 represents the error bars (CI = .95) related to the mean AUC values of the classifier over the E vs H, M vs H and E vs M conditions by using the three $Iteration_{MAX}$ values (5%, 50%, 100%) and the three cross-validation types (Intra, Short term and Medium term). ANOVA results highlighted a significant main effect between the two factors ($F(4,44)=3.81$, $p=.01$). The post-hoc test highlighted a significant decrement of the AUC values within all the cross-validation types (Intra, Short term and Medium term) by considering both the 50% and the 100% of the available EEG features (all $p<10^{-3}$), but no significant differences across the different cross-validation types have been found by considering the 5% of features (all $p=1$). Also, despite there was no significant difference among the AUC values achieved by using the 5%, 50% and 100% of the features within the Intra cross-validations (all $p>.3$), the post-hoc test highlighted a significant decrement of the AUC values between the 5% of features and the 50% and 100% of features within the Short term and the Medium term cross-validation types (all $p<10^{-3}$).

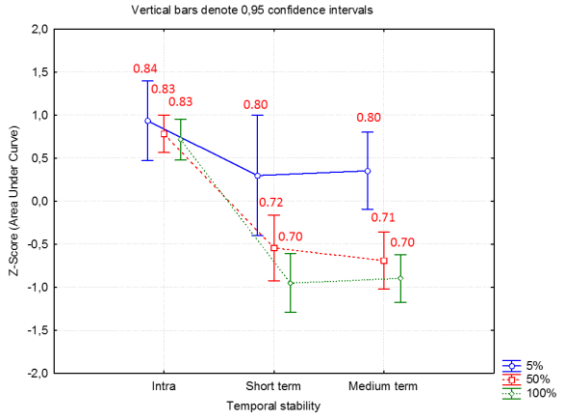

Figure 2. Error bars (CI = .95) related to the normalized AUC values of the classifier over the E vs H, M vs H and E vs M conditions by using the three $Iteration_{MAX}$ values (5%, 50%, 100%) and the three cross-validation types (Intra, Short term, Medium term). The absolute AUC values have been also reported (red writings).

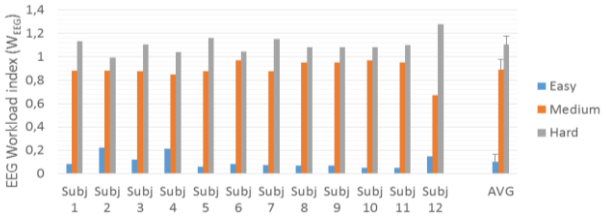

Figure 3. W_{EEG} values distribution across the three difficulty levels across subjects.

Neurophysiological workload distribution (W_{EEG}): ANOVA results showed a significant interaction between the two factors ($F(4,44)=19.14$, $p=10^{-6}$). The post-hoc test highlighted that the W_{EEG} distributions related to the three difficulty levels (E, M, H) were significantly different for each cross-validation type (all $p<10^{-6}$). Furthermore, no significant differences have been shown between each W_{EEG} distribution related to the three MATB difficulty levels among the cross-validation types (all $p>0.6$). Figure 3 represents the W_{EEG} values measured for each difficulty level (E, M, H) and for each subject.

IV. Discussion

The ATM environment imposes multiple concurrent demands on the operator, including air traffic monitoring, anticipating loss of separation between aircrafts, and intervening to resolve conflicts. The assessment of the MW associated with operating in such complex systems has long been recognized as an important issue [13]. MLTs have been widely used to assess the MW of the operators by using their brain activity [2], [5]. One of the big problems in using these approaches is that the MW measure does not remain stable over time, unless a frequent recalibration procedure. The reliability of such methodologies is of great importance for their effective use in real work contexts. The necessity to re-calibrate the system every day makes such kind of approach unsuitable in operative environments.

In addition, there are evidences in literature in which it has been demonstrated that, as long as a subject is trained to do a specific task, the cognitive processes required for performing such task would be always present over time [3]. In other words, it would be possible to take into account certain cerebral features strictly related to the MW of the user (Frontal and Occipital θ, Parietal α), that remain enough stable over time. In this work, we observed that if a classifier is calibrated by using a low number of cerebral features (among the ones strictly related to the MW of the operator), the reliability of the system will not degrade across days. In this context, we could speculate that the classifier is able to identify those brain features always involved in the execution of the proposed task, since the DA remained stable over time. The described algorithm has been tested on twelve ATCos trainees while performing an ATM task under three different difficulty levels (E, M, H) resembling different possible ATM scenarios. Results demonstrated that if a SWLDA regression is calibrated by using a low number of features (5% out of the total available features) a stable MW measure across the days is obtained. The consequence of this finding is that within a week, it will not be necessary to recalibrate the algorithm and the classification system with new user's EEG data. In addition, the proposed algorithm has been able to differentiate significantly the MW over the three experimental conditions (E, M, H).

V. Conclusion

In this work, we have provided evidences that if a SWLDA classifier is used to select a low number of EEG spectral features related to the MW of the user, those features will make the classifier DA stable across different days. In other words, if the classifier selects certain cerebral features on Monday, those features can be used within one week, without degradation of the system's reliability. However, the study has several limitations. The first one is that the reliability of the measure has been tested only in a short period of one week. A second limitation is that different variables have been kept under control for the purpose of this experimentation. For example, each subject took part to the experiment at the same time over the different days; in addition the simulation scenario was not too much ecological in terms of time duration (the work shift for an ATCo is of ~40 minutes) and performed tasks (the ATCos use to talk with their colleagues and with the pilots). In addition, an open question that has to be investigated following the results of the actual study is if it exists the possibility to identify a subset of EEG features that are maintained stable even between different subjects. Taking into account all these limitations, there is the need to perform further experiments to test the proposed approach in a more real ATM work scenario, by considering a longer period of stability ("Long-term") than one week.

Acknowledgment

This work is co-financed by EUROCONTROL on behalf of the SESAR Joint Undertaking in the context of SESAR Work Package E - NINA research project. The grant provided by the Italian Minister of University and Education under the PRIN 2012 scheme to F.B. is also gratefully acknowledged.

References

[1] D. A. Norman and D. G. Bobrow, "On the analysis of performance operating characteristics," Psychol. Rev., vol. 83, no. 6, pp. 508–510, 1976.

[2] P. Aricò, G. Borghini, I. Graziani, F. Taya, Y. Sun, A. Bezerianos, N. V. Thakor, F. Cincotti, and F. Babiloni, "Towards a multimodal bioelectrical framework for the online mental workload evaluation," Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 2014, pp. 3001–3004, 2014.

[3] G. Borghini, P. Aricò, I. Graziani, S. Salinari, Y. Sun, F. Taya, A. Bezerianos, N. V. Thakor, and F. Babiloni, "Quantitative Assessment of the Training Improvement in a Motor-Cognitive Task by Using EEG, ECG and EOG Signals," Brain Topogr., Jan. 2015.

[4] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with EEG pattern recognition methods," Hum. Factors, vol. 40, no. 1, pp. 79–91, Mar. 1998.

[5] J. Kohlmorgen, G. Dornhege, M. Braun, B. Blankertz, K.-R. Müller, G. Curio, K. Hagemann, A. Bruns, M. Schrauf, and W. Kincses, "Improving human performance in a real operating environment through real-time mental workload detection," 2007. .

[6] J. C. Christensen, J. R. Estepp, G. F. Wilson, and C. A. Russell, "The effects of day-to-day variability of physiological data on operator functional state classification," NeuroImage, vol. 59, no. 1, pp. 57–63, Jan. 2012.

[7] L. K. McEvoy, M. E. Smith, and A. Gevins, "Test-retest reliability of cognitive EEG," Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol., vol. 111, no. 3, pp. 457–463, Mar. 2000.

[8] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer New York, 2000.

[9] J.-P. Imbert, H. M. Hodgetts, R. Parise, F. Vachon, and S. Tremblay, "The LABY Microworld A Platform for Research, Training and System Engineering," Proc. Hum. Factors Ergon. Soc. Annu. Meet., vol. 58, no. 1, pp. 1038–1042, Sep. 2014.

[10] N. R. Draper and H. Smith, Applied Regression Analysis, Third. Wiley-Interscience, 1998.

[11] 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, Nov. 1975.

[12] J.-H. Zhang, T. D. Y. Chung, and K. R. Oldenburg, "A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays," J. Biomol. Screen., vol. 4, no. 2, pp. 67–73, Apr. 1999.

[13] D. Gopher and E. Donchin, "Workload: An examination of the concept," in Handbook of perception and human performance, Vol. 2: Cognitive processes and performance, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. Oxford, England: John Wiley & Sons, 1986, pp. 1–49.