On the use of machine learning for EEG-based Workload assessment: algorithms comparison in a realistic task

Nicolina Sciaraffa[1,2,3], Pietro Aricò[2,3,4], Gianluca Borghini[2,3,4], Gianluca Di Flumeri[2,3,4], Antonio Di Florio[2] and Fabio Babiloni[2,4].

[1] Dept. Anatomical, Histological, Forensic & Orthopedic Sciences, "Sapienza" University of Rome, Italy [2] BrainSigns srl, Rome, Italy [3] IRCCS "Fondazione Santa Lucia", Rome, Italy [4]Dept. of Molecular Medicine, "Sapienza" University of Rome, Italy

Abstract. The measurement of the mental workload during real tasks by means of neurophysiological signals is still challenging. The employment of Machine Learning techniques has allowed a step forward in this direction, however, most of the work has dealt with binary classification. This study proposed to examine the surveys already performed in the context of EEG-based workload classification and to test different machine learning algorithms on real multitasking activity like the Air Traffic Management. The results obtained on 35 professional Air Traffic Controllers showed that a KNN algorithm allows discriminating up to three workload levels (low, medium and high) with more than 84 % of accuracy on average. Moreover, in such realistic employment it emerges how important is to opportunely choose the set of features to ward off that task-related confounds could affect the workload assessment.

Keywords: Workload, EEG, Machine Learning, Real settings, ATM.

1 Introduction

Reading the Special Issue for the golden anniversary of the "Multiple Resources and Mental Workload" theory of Christopher D. Wickens [1] allows to retrace the history of the concept of workload from the difficulty of its definition up to the need to measure it within Human Factors, passing through the definition of a workload model. In fact, since the '70s, when the term workload began to appear in scientific publications, several terms and definitions have overlapped and followed one another. The mental workload, mental effort and mental strain were the most widely used terms to define the relationship between the cognitive resources of a person who is required to perform a task and the difficulty of the task itself. One of the first definitions of workload was "the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort" [2]; another typical description define the workload as "the difference between the capacities of the information processing system that are required for the task performance to satisfy performance expectations and the capacity available at any given time" [3]. In each of the definitions given to the workload to date there is the term "capacity" which implies a finite amount of resources [1], in this case cognitive resources. The pool of

cognitive resources referred to is not unique, but it is the set of different pools that allows to explain the link between performance and difficulty in the case of multitasking. In fact, many of the actions we perform are multitasking, such as observing a picture to describe it, talking while driving and typically multitasking activities such as those carried out in an aircraft cockpit.

1.1 Multifaceted aspects of workload

It is precisely in these safety-critical environments that the need to evaluate an operator's workload was firstly felt. In 1981, again Wickens pointed out that the development of increasingly complex technologies had radically changed the role and load to which an operator was subjected, leading to the dual need to exploit the model of multiple resources to optimize the processing of human operator information in the definition of tasks ("Should one use keyboard or voice? Spoken words, tones, or text? Graphs or digits? Can one ask people to control while engaged in visual search or memory rehearsal?"[1]) and measure the operator's workload. From that moment on, the measure of the workload has spread from the aeronautical [4][5], the educational [6][7] and to the clinical [8] fields. Even the aims of workload measurement have evolved: the ultimate goal in all environments is mainly the Workload Adaptation, the process of workload management to aid learning, healing or limiting human errors. Moreover, workload measurement affects both the design and management of interfaces. On the one hand, by testing the workload of subjects during the use of web interfaces [9], for example, it is possible to direct the design. On the other hand, in the field of adaptive automation, it is the continuous monitoring of the workload level of the subject that allows the system to vary the feedback in response to the mental state of the operator [10], [11].

1.2 Workload Measurements

The workload of an operator can be measured in three ways: by administering questionnaires, analyzing the performance of a subject or through psychophysiological measures. Since the workload has different aspects (e.g. mental workload is different from physical workload) the questionnaires preferably used are multidimensional ones, such as NASA TLX [12] and SWAT [13]. However, these are subjective measures and require subjects to be trained in interpreting and judging their condition. Moreover, they can only be assessed after a task, not online. Similarly, performance measures do not represent a direct indicator of the workload status of the subject as they do not allow to know the amount of resources used and therefore the residual resources to reach that performance value [14]. Moreover, measuring performance on a task does not allow to obtain this measure of a differential nature (the remaining resources) so it is always necessary to use a secondary task. However, even the use of secondary tasks very often remains too closely linked to typical laboratory tasks and makes what really happens in multitasking implausible [15]. The main objective of neurophysiological measurements is to provide an objective and continuous, as well as an online measurement of an operator's workload. Thanks to

the possibility of making neurophysiological measures less and less intrusive, so far have been correlated with workload values almost all neurophysiological known measures like the Electrocardiogram (ECG), the Eye movements, the Pupil diameter, the Respiration, the Galvanic Skin Response (GSR) and the brain activity. Summarizing the evidence, the ECG, the GSR and the ocular activity measurements highlighted a correlation, not only with workload, but also with different mental states like stress, mental fatigue, drowsiness. Therefore, they were demonstrated to be useful and robust only in combination with other neuroimaging techniques directly linked to the Central Nervous System (CNS), that is the brain [16]. Consequently, the electroencephalogram (EEG) and the functional Near-Infrared Spectroscopy (fNIRS) as measures of the brain activity, are the most likely candidates that can be straightforwardly employed to monitor the workload in real environments [17]. Between the two, the EEG is usually preferred for the workload assessment for its high temporal resolution. Moreover, it has been proved that EEG features provide higher accuracy respect to ECG and GSR ones [18], [19].

The electroencephalogram is the measure of brain electrical activity that in a non-invasive way can be performed by means of electrodes placed on the scalp. To date, there has been a strong improvement in technology oriented towards minimally invasive systems, with few electrodes, and possibly, dry electrodes [20]. The analysis of EEG signals is usually aimed at studying the variance of the spectral power in the conditions of interest. In the case of Workload, it has emerged that a higher task demand corresponds to an increase in frontal theta band activity and a decrease in parietal alpha band activity [16].

1.3 Machine learning to get back out-of-the-lab

Therefore, the concept of workload was born for practical needs, has been modeled in the laboratory essentially using dual-task procedures, but then the need to measure it in realistic contexts as in operational, educational and clinical returns overwhelmingly. The practical implications of applying a workload measurement in a realistic environment define the characteristics that an automatic workload measurement system must have. Firstly, especially in the applications in real environment, it is difficult to create a direct link between the mental state of the subject and his brain activity, or more generally his physiological state since there is no control condition typical of the laboratory. The employment of secondary cognitive task (e.g. the n-back) during real activities does not fit the realistic conditions and may increase the actual workload level [21]. Moreover, because of the high individual variability of physiological responses, traditional statistical tests are not able to discover the relationship between cause and effect, so it is necessary to employ techniques that allow to take into account the individual characteristics to correctly define the level of workload, such the machine learning techniques [22]. Such methods allow to extract the features mostly influenced by the mental state variation, and then use this information to classify the specific workload level. Secondly, since by definition the workload is linked to the performance by the inverted “U” model[23], the ideal would be to be able to distinguish at least 3 levels of workload, one suboptimal that concerns the workload too low, one optimal, and the

threshold that defines the overload condition. However, most of the work in literature is limited to classify two levels of workload, the low and high. In these cases, the levels of accuracy reached are generally very high, greater than 80% [18], [22], [24]–[29]. Much less are the examples of multiclass-classification[17], whose highest number of workload levels classified has been 7 [30] and, almost all, have been obtained by means of n-back and arithmetic tasks in a laboratory context[31], [32]. In this context, the majority of methods used to define the level of mental workload of a subject are supervised machine learning techniques. The process that leads from the recording of EEG signals to an indication of the workload level passes through the use of signal analysis methods that allow to extract the informative features of the phenomenon to be investigated. Regarding the measurement of the workload have been used in several studies both spectral, temporal and spatial features [33]. The use of spectral features remains the most suitable for the temporal continuity required by the workload monitoring, since the brain activity induces variations in its spectral power which, unlike ERPs used for time domain analysis [34], does not need to be triggered with a certain timing [35]. Taking into account the nature of the features, there are countless different examples of configurations, in terms of number of channels and frequencies used in the literature. The number of electrodes can vary from 64 [24], [36] to 6 [37], and even the bands considered vary from 2 (Theta and Alpha, [9]), to 7 (0–4 Hz, 4–7 Hz, 7–12 Hz, 12–30 Hz, 30–42 Hz, 42–84 Hz, 84–128 Hz [38]), up to considering all the single frequency bins that define the spectrum [39]. Several studies have shown that it does not necessarily take more than 5-10 electrodes to classify the workload [24]. Especially when dealing with a high number of channels and high spectral resolution, the number of used features increases exponentially and leads to the so-called "curse of dimensionality"[40]. Many researchers have therefore highlighted the need to make a selection of features both to decrease the computational cost of machine learning algorithms and to use this as additional information in experimental setups. In fact, if the analysis shows that some electrodes are not useful for the classification of the workload, it is possible to remove them and then make the instrumentation lighter and less invasive. In this case the most used methods for the selection of features are those recursive, such as recursive feature elimination [18], [41], sequential forward feature selection [24] or methods that take into account the dependence between features such as the Minimum Redundancy Maximum Relevance selection [37], [42], [43], or unsupervised method (Locally linear embedding, [44]). Once defined a meaningful set of features, it is necessary to choose a model to define the workload level. In the literature innumerable algorithms, essentially of a supervised nature, have been used to define the workload level of a subject starting from his brain signals, belonging to both the so-called Shallow learning and deep Learning domains [45], [46]. In all cases the efficiency of such algorithms is usually presented in terms of accuracy. However, the accuracy obtained in different studies are not directly comparable, since the calculation of accuracy includes several factors like the task employed to elicit the workload, the number of subjects and the number of EEG signals recorded (and also the kind of instrumentation employed), the features extracted, the methods used to eventually select them and, only at the end, the algorithm used to classify the workload. Only taking into account all this information, it could be possible to

compare the results obtained so far in different works employing machine learning techniques to classify the level of workload.

Leveraging on the theoretical comparison of the works done so far with regard to the classification of the workload through electroencephalographic signals, it is necessary to highlight an issue. In any case, starting from a very large set of features or making a blind selection of them, very high accuracy of classification could be obtained. However, it could not be directly associated with a change in the workload level. To be sure that the system has actually classified workload, it is therefore necessary to perform two fundamental actions. Firstly, it is necessary to perform an excellent calibration of the machine learning algorithm avoiding task-related confound, like for example movements or the influences of other mental states. However, during a real task that is typically highly multitasking, it may not be possible to perform a rigorous calibration of the system, as the ideal conditions provided by the typical control conditions of the laboratory task are lacking. Therefore, in these cases the calibration could be dirtied by task-related confounds. To solve this problem cross-task calibration has been proposed, but the results produced so far have not shown that it is possible to use a laboratory task to effectively classify a real task and the performance is much lower than that obtained in a within-subject condition [41]. Secondly, it should be preferred a careful selection of features related to the phenomenon, and possibly not to the other mental state variations, respect to a blind one. In the case of workload, for example, many works have identified in the activity of the frontal brain areas in the theta band and parietal areas in the alpha band [47] the most informative features. Even more accurately, [30] have carried out a selection of the channels through source localization analysis, and it has been possible for them to classify 7 levels of workload.

Therefore, the practical aim of this work is to provide a comparison of five different machine learning algorithms and two different sets of EEG features to discriminate three different levels of workload during a real task of Air Traffic Management. The Air Traffic Management (ATM) is a highly multitasking activity. In fact, air traffic controllers are continuously engaged in visual activities (airplane control on the radar) and auditory (as they communicate with the pilots of different aircrafts). The ATM represents one of the fields where the evaluation of workload has a fundamental role both in training aspects and in operational conditions[48], [49]. In fact, it has been established that it is a high demanding work during which the slightest error could have very ominous effects [50]. Changes in the traffic manipulation, complexity or volume, produced changes in mental resources required and therefore in workload. In ATM domain, different tasks have been used to investigate workload changes and to create a workload index based on biosignals, even though most of the results are related to laboratory environment. However, the present study takes advantages of realistic task in a highly realistic context and of 35 professional Air Traffic Controllers.

2 Methods

2.1 Experimental Protocol

An experimental protocol with high realistic ATM scenarios has been settled up. The controller position is similar to the ones used in the real operational center. The controller working position has two screens, one 30" to display radar image and a 21" to interact with the radar image (zoom, move, clearances and information) and the voice communication between controller and pseudo-pilot uses the same hardware and software like the one used in training (headphones, microphone and push-to-talk command), very similar to what normally used into operations. The radar picture shows the sector (light grey), routes, waypoints and flights displayed according to their coordination state (white ones are assumed). Information on neighbor flights is displayed in the list. The experimental task consists in an ATM scenario in which different air-traffic conditions take place. For instance, the task could start with an increasing traffic complexity until a hard condition, and then decreasing until a condition similar to the initial one by passing through a medium complexity condition. The variation of the task complexity is necessary to evaluate if the system is able to differentiate the different workload levels. Each scenario lasts globally 45 minutes, while each session of low (L), medium (M) and high (H) workload level lasts about 15 minutes. The different levels of traffic, defined by subject matter experts (SME), vary according to the number of aircrafts, traffic geometry, the number of conflicts and subjective assessment of controller's skill. Three different scenarios have been proposed, with compatible events in order to induce the three mentioned difficulty levels (Easy, Medium, Hard), but not exactly the same, to not induce habituation or expectation effects on the experimental subjects. The experimental protocol involves 35 professional ATCOs. ATCOs have been selected in order to have a homogeneous sample in terms of sex, age and expertise. Sixteen EEG channels (FPz, F3, Fz, F4, AF3, AF4, C3, Cz, C4, P3, Pz, P4, POz, O1, Oz, O2) have been recorded by the digital monitoring BEmicro system (EBNeuro system) with a sampling frequency of 256 Hz. All the EEG electrodes have been referenced to both earlobes, and the impedances of the electrodes have been kept below 10 kΩ.

2.2 Signal Processing

The EEG signals have been digitally band-pass filtered by a 4th order Butterworth filter (low-pass filter cutoff frequency: 30 Hz, high-pass filter cutoff frequency: 1 Hz) and the Fpz signal has been used to correct eyes-blink artifacts from the EEG data by means of the Reblinca algorithm [51]. It should be underlined that normally in a realistic environment, different sources of artifacts could affect neurophysiological recorded signals, more than in the laboratory environments. For instance, ATCOs normally communicate verbally and perform several movements during their operational activity. Then each trial having an amplitude exceeds 100 μV (threshold

criteria), or the slope trend higher than 3, or a sample to sample difference higher than 25 μV have been marked as "artifact" and then rejected, with the aim to have clean EEG signals from which estimate the brain parameters for the different analyses. The aforementioned parameters and related techniques have been set following the methodology available on the EEGLAB toolbox [52].

2.3 Features extraction

The EEG signals have been segmented in periods of 2 seconds, 0.125 seconds shifted. After that, for each period, the power spectral density (PSD) by using the Fast Fourier Transform has been computed in the Theta and Alpha frequency bands because it has been stated that they are the most related to workload effects[9], [47]. The EEG frequency bands were defined accordingly with the Individual Alpha Frequency (IAF) value estimated for each subject. Since the alpha peak is mainly prominent during rest conditions, the participants were asked to keep the eyes closed for a minute before starting with the experiment. In particular, the theta and alpha bands have been respectively defined as (IAF-6 ÷ IAF-2) and (IAF-2 ÷ IAF+2). Two different sets of features have been considered. In the first case, the PSD values in the theta and alpha bands have been computed for 14 EEG electrodes, to imitate what is usually done in several studies in literature [9], [41], [43], [53]. In this work, 28 PSD Features (14 Channels x 2 Bands) have been computed. The second set consisted of 9 features, 5 describing the frontal theta activity and 4 the alpha parietal activity. These features have been chosen according to the literature, because it is proven these are the features most correlated to workload [6], [47]. In both cases the features have been normalized because the differences in ranges affect the calibration and the functioning of some algorithms [54] (e.g. K nearest neighbor).

2.4 Machine Learning Algorithms

Five different machine learning techniques have been employed to discriminate between three different levels of workload (i.e. Low, Medium and High level). The starting dataset shows balanced classes (6000 instances per class) and has been used in a within-subject manner, as this approach is allowed in case of long-lasting recordings of a single subject. The total amount of occurrences available for each subject has been divided in an optimization set, which occurrences have been used for the optimization of model parameters by using grid search and 3-fold cross validation, and an evaluation set, where the performance of the algorithms have been evaluated computing the accuracy through 5-fold cross-validation. Optimization and evaluation of machine-learning techniques have been computed by Phyton Scikit-learn library[55]. In particular five different techniques have been trained to cover a wide range of algorithms types: a regression-based method (the multinomial logistic regression), a linear method without any optimization procedure (LDA), a linear classifier with a cost parameter (SVM), an instance-based method (the k nearest neighbor) and an ensemble method (the Random Forest).

Logistic Regression (LR) is a model used for the prediction of the probability of occurrence of an event. In this case it has been used in its multiclass configuration, the multinomial logistic regression and the value of l_2 penalization has been chosen in the log space between -3 and 3

Linear Discriminant Analysis (LDA) is a linear algorithm that allows to create hyperplanes in n-dimensional space according to the number of features, to discriminate 2 or more classes. Its advantage is that it has not any parameter to optimize, however, it could finally try to describe only linear problems.

Support Vector Machine (SVM) is a supervised algorithm that allows to create hyperplanes in n-dimensional space according to the number of features, to discriminate 2 or more classes. It could be a linear or nonlinear classifier (or regressor) according to the employed kernel. In this case it has been used a linear kernel in a multiclass configuration using the Crammer and Singer method () and the optimal cost parameter has been chosen in a log space between -3 and 3

K- Nearest Neighbor (KNN) is a nonlinear instance-based algorithm. Its main idea is to predict the class on the basis of the distance between the observation and the first k neighbors and does not assume a priori the distribution of the dataset. The advantage of this algorithm is that it is optimized locally, and it is not affected by the complexity of the entire phenomena. The weakness is that the computational cost could be as high as the amount of features increase. Only the number k of neighbors has been chosen in a range from 1 to 20.

Random Forest Classifier (RF) is a nonlinear classifier [56] belonging to the ensemble methods. This family of classifiers allows to generalize well to new data [57] and are more robust to overfitting than individual trees because each node does not see all the features at the same time[56]. Several parameters could be optimized, however, in this case only the number of trees ([10,100,200]) have been chosen.

In the latter case, the algorithm allows also to obtain the information related to the feature importance that could be used to explain how the model is affected by each feature. Therefore the topographies showing the feature importance in the case of 28-feature and 9-feature sets have been compared.

3 Results

The classification results for both sets of features are shown in Fig 1 and 2. Each box plot represents the value of the accuracy of the population, while the single mean value of the accuracy obtained for each subject is shown in the black dot. The Friedman test has been performed to statistically assess if there is any significant difference between the algorithms, because the sphericity requirement was not met for both conditions (Mauchly's test $p<10^{-4}$). For the 28-feature set the Friedman test provided χ^2 ($N=35$, $df=4$)$=123.0171$ ($p<10^{-4}$) and for the 9-feature set χ^2 ($N=35$, $df=4$)$=126.2857$ ($p<10^{-4}$). The multiple comparison Bonferroni corrected has been performed and the results are shown in Table 1. In both cases the mean accuracy provided by the KNN is significantly higher than all the other algorithms. The Random Forest provided significantly higher accuracy than the LR, LDA and SVM. Such 3 methods did not show significant differences in their performances when the

28-feature set was used, while the SVM performed significantly worse than the other algorithms when 9 features were used. In Fig 3 are shown the topographies of the feature importance computed with the Random Forest algorithm for theta and alpha band. The values have been normalized. The scalp maps show that the most important features are the central and occipital PSD values in alpha band in the case of 28-feature set and the parietal activity in alpha band in case of the 9-feature set.

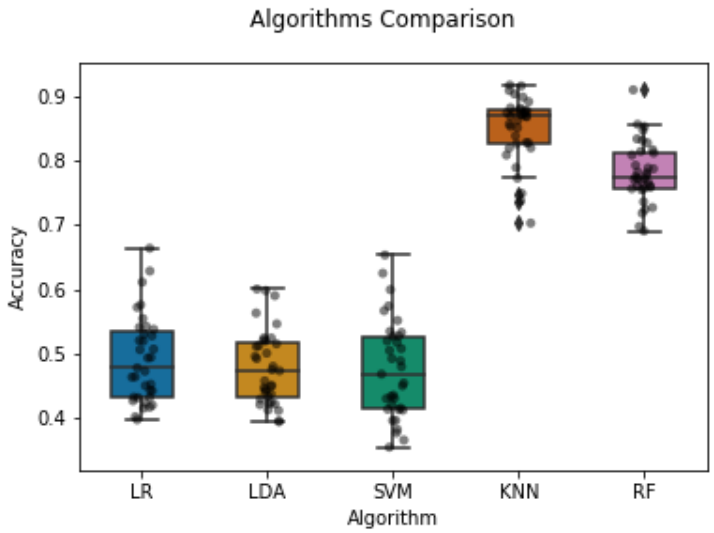

Fig. 1. Accuracy distribution for each algorithm in case of 28 features. The black dots represent the accuracy value for each subject.

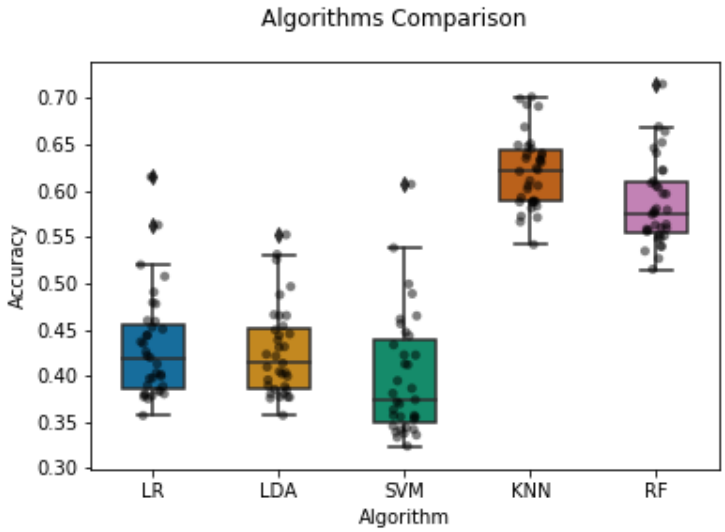

Fig. 2. Accuracy distribution for each algorithm for 9-feature set. The black dots represent the accuracy value for each subject.

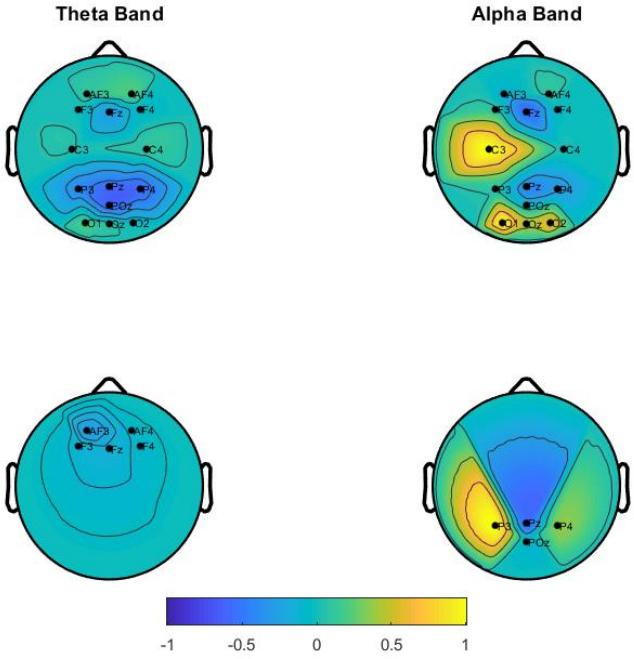

Fig. 3. Topographies of Feature importance according to Random Forest algorithm in theta and alpha band. In the first row are represented the results for 28 features and in the second row for 9 features.

Table 2. Mean accuracies and standard deviation for each algorithm and multiple comparison p-value (Bonferroni corrected). When the results for the two sets of features are different both p-value have been reported.

			Accuracy (%)		p-value 28-Feat/9 Feat		
Algorithm	28-Feat Mean (SD)	9-Feat Mean (SD)	LR	LDA	SVM	KNN	RF
LR	49.176 (6.74)	43.096 (5.72)		1.000	0.874/ **10^{-7}**	**0.000**	**0.000**
LDA	47.889 (5.74)	42.645 (4.88)	1.000		1.000/ **$4.9*10^{-5}$**	**0.000**	**0.000**
SVM	47.645 (7.57)	39.914 (6.55)	0.874/ **10^{-7}**	1.000/ **$4.9*10^{-5}$**			**0.000**
KNN	84.947 (5.05)	62.112 (3.98)	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
RF	78.214 (4.68)	58.617 (4.57)	**0.000**	**0.000**	**0.000**	**0.000**	

4 Discussion

This work aims to classify three different levels of workload during real multitasking activities like the ATM. According to the theory, a correct modulation of the workload in a laboratory environment should be based on a dual task to allow an evaluation of the subject's residual cognitive resources and consequently, by definition, of his workload. However, in a real environment, it is very difficult to integrate a control condition, as well as to take into account the variability underlying cognitive phenomena both intra and inter subject. Therefore, the application of Machine Learning has been considered the solution for classifying the workload and overcoming these issues typical of real applications. The preliminary analysis of the works carried out so far in this context has shown that it is possible to discriminate with acceptable accuracy only two levels of workload [6], [18], [22], [24], [25], [27]–[29], [33], [36]–[39], [42], [45], [46], [58], even though, above all in view of a practical application of the workload measurement, it is necessary to establish at least the value of two thresholds to define the underload and the overload state. The most frequently employed features are the spectral ones, because they can be calculated

with a high temporal resolution (up to one second) and allow to monitor brain activity in a quantitative manner without temporal triggers. Therefore, in this work the values of the PSD calculated in time windows of 2 seconds have been used, averaging the values of PSD in each band to limit the number of features and keep at the same time under control the collinearity[59]. Due to this condition, in fact, the information provided by very close frequency bins could be superimposable, which would lead to introduce a bias in the creation of the model. Since the number of observations available is in the order of thousands, it was decided not to use any feature selection algorithm, but to provide a posteriori information on the feature importance. One of the chosen algorithms, the Random Forest, allowed to have this information. Taking advantage of this potentiality, it has been highlighted that the most discriminating features of the concerned model are in the alpha band. In particular, when the higher number of features has been used, the most important features cover the central and occipital brain areas. This aspect can be explained considering that the alpha band intervenes twice in the considered task. In fact, in the alpha band it is possible to find both the motor alpha pattern, due to the activity of the sensorimotor area and generally strongly lateralized [60], and the pattern associated to the visual area. This set of features is not directly referring to the typical workload topographies, whose purpose is to measure the net of cognitive resources used by the subject, but rather these features provide the information derived from the movements of the subjects to define the level of workload imposed by the task itself, which does not necessarily correspond to that perceived by the subject. On the other hand, when only frontal theta and parietal alpha features have been used, the most important features are related to parietal activity, that is usually associated with the attentional alpha pattern that reflects the increase of brain activity in areas afferent to the posterior attentional networks [60]. Therefore, it has been demonstrated that, especially when the task is real and the algorithm does not take advantage of a rigorous calibration to avoid the task related confounds, the role of the features chosen a priori becomes essential and recalls the concept of "no free lunch" [61] in machine learning: the necessity to use prior knowledge to optimize the algorithm functioning, but at the cost of generality. However, it is necessary to highlight that reducing the number of features there is a decreasing in the performance of each tested algorithms. In fact, to classify three levels of workload it was decided to test different algorithms of machine learning, which in the first place can be divided into two categories: regression and classification. Among the classifiers, we can further distinguish linear (LDA and SVM) and non-linear (KNN and RF) methods. Although regression has been proposed as a method that avoiding the strict equality between classes allowing to have higher performances, especially in case of cross-task classification [41], it provided an accuracy value equal to 50%, but still higher than the chance (33.33%). On the other hand, linear methods both in the case of optimization (SVM) and in the case of non-optimized method (LDA) provided significantly lower accuracy than nonlinear methods. The linear methods are appropriate when limited data and limited knowledge about the data itself is available[62]. However, a linear classifier does not work well in the presence of strong noise or outliers, if the dimensionality of the features space is too high, if regularization was not done well or the problem is intrinsically non-linear [17]. If there are large amounts of data, non-linear methods are suitable to find potentially complex structure in the data. In this work the KNN not

only provides the highest accuracy (84%), but it also has different advantages: it is a method that does not require the calculation of the covariance matrix as in the case of the LDA and is therefore mathematically very simple [57]. On the other hand, it does not need time for training (because it just memorizes the training set) and then could be used for an online application when there are a few features. In fact, in the case of a large number of features this method does not allow to easily manage those irrelevant and at the same time becomes computationally expensive to calculate the distance. In addition to the high accuracy provided, even the choice of Random Forest as classifier in realistic multitasking could be advantageous essentially for two reasons. First, because it is an ensemble method, it tends to generalize well and is not subjected to overfitting, Second, it allows to have the information regarding the feature importance, that increase the possibility to know what the system is actually classifying. However, the final choice of a classifier should be made after a systematic evaluation of other different performance parameters, such sensitivity, specificity, recall and precision.

5 Conclusion

With this work it has been proved that it is possible to reach very high accuracy to distinguish between three levels of workload during a real task only by using the EEG signals. However, according to the literature the high accuracy is only one of the optimal characteristics required for an out-of-the-lab classifier besides the none or at most few data samples for training the classifier and higher temporal reliability. Therefore, several other questions need to be pointed out in realistic contexts.

7 Acknowledgment

This work is co-financed by the European Commission by Horizon2020 projects "WORKINGAGE: Smart Working environments for all Ages" (GA n. 826232); "SIMUSAFE": Simulator Of Behavioural Aspects For Safer Transport (GA n. 723386); "SAFEMODE" (GA n. 814961); "BRAINSAFEDRIVE: A Technology to detect Mental States during Drive for improving the Safety of the road" (Italy-Sweden collaboration) with a grant of Ministero dell'Istruzione dell'Università e della Ricerca della Repubblica Italiana.

6 Bibliography

[1] C. D. Wickens, "Multiple resources and mental workload," *Hum. Factors*, vol. 50, no. 3, pp. 449–455, 2008.

[2] R. Curry, H. Jex, W. Levison, and H. Stassen, "Final report of control engineering group," in *Mental workload*, Springer, 1979, pp. 235–252.

[3] D. Gopher, "In defence of resources: On structures, energies, pools and the allocation of attention," in *Energetics and human information processing*, Springer, 1986, pp. 353–371.

[4] B. H. Kantowitz and P. A. Casper, "Human workload in aviation," in *Human*

Error in Aviation, Routledge, 2017, pp. 123–153.

[5] I. Bargiotas, A. Nicolaï, P.-P. Vidal, C. Labourdette, N. Vayatis, and S. Buffat, "The Complementary Role of Activity Context in the Mental Workload Evaluation of Helicopter Pilots: A Multi-tasking Learning Approach: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Pape," 2019, pp. 222–238.

[6] P. Gerjets, C. Walter, W. Rosenstiel, M. Bogdan, and T. O. Zander, "Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach," *Front. Neurosci.*, vol. 8, no. DEC, pp. 1–21, 2014.

[7] A. Byrne, "The Effect of Education and Training on Mental Workload in Medical Education: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers," 2019, pp. 258–266.

[8] A. Byrne, "Measurement of mental workload in clinical medicine: a review study," *Anesthesiol. pain Med.*, vol. 1, no. 2, p. 90, 2011.

[9] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during web browsing," *Sensors (Switzerland)*, vol. 18, no. 2, pp. 1–26, 2018.

[10] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Pozzi, and F. Babiloni, "A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks," in *Progress in brain research*, vol. 228, Elsevier, 2016, pp. 295–328.

[11] P. Aricò *et al.*, "Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment," *Front. Hum. Neurosci.*, p. 539, 2016.

[12] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, Elsevier, 1988, pp. 139–183.

[13] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," in *Advances in psychology*, vol. 52, Elsevier, 1988, pp. 185–218.

[14] G. F. Wilson, "Operator functional state assessment for adaptive automation implementation," in *Biomonitoring for Physiological and Cognitive Performance during Military Operations*, 2005, vol. 5797, pp. 100–105.

[15] H. A. Colle and G. B. Reid, "Double trade-off curves with different cognitive processing combinations: Testing the cancellation axiom of mental workload measurement theory," *Hum. Factors*, vol. 41, no. 1, pp. 35–50, 1999.

[16] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neurosci. Biobehav. Rev.*, vol. 44, pp. 58–75, 2014.

[17] P. Aricò, 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.

[18] H. Zhang, Y. Zhu, J. Maniyeri, and C. Guan, "Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine," *2014 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC 2014*, pp. 2985–2988, 2014.

[19] P. Aricò *et al.*, "Towards a multimodal bioelectrical framework for the online mental workload evaluation," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2014, pp. 3001–3004.

[20] 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.

[21] J. Heard, C. E. Harriott, and J. A. Adams, "A Survey of Workload Assessment Algorithms," *IEEE Trans. Human-Machine Syst.*, vol. 48, no. 5, pp. 434–451, 2018.

[22] C. Dijksterhuis, D. De Waard, K. A. Brookhuis, B. L. J. M. Mulder, R. De Jong, and S. E. Kerick, "Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns," vol. 7, no. August, pp. 1–11, 2013.

[23] A. Bruggen, "An empirical investigation of the relationship between workload and performance," *Manag. Decis.*, vol. 53, no. 10, pp. 2377–2389, 2015.

[24] S. Mathan, A. Smart, T. Ververs, and M. Feuerstein, "Towards an index of cognitive efficacy: EEG-based estimation of cognitive load among individuals experiencing cancerrelated cognitive decline," *2010 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC'10*, pp. 6595–6598, 2010.

[25] C. L. Baldwin and B. N. Penaranda, "Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification," *Neuroimage*, vol. 59, no. 1, pp. 48–56, 2012.

[26] Y.-T. Wang *et al.*, "Developing an EEG-based on-line closed-loop lapse detection and mitigation system ," *Frontiers in Neuroscience* , vol. 8. p. 321, 2014.

[27] D. De Massari *et al.*, "Fast mental states decoding in mixed reality," *Front. Behav. Neurosci.*, vol. 8, no. November, pp. 1–9, 2014.

[28] M. Schultze-Kraft, S. Dähne, M. Gugler, G. Curio, and B. Blankertz, "Unsupervised classification of operator workload from brain signals," *J. Neural Eng.*, vol. 13, no. 3, p. 36008, 2016.

[29] G. N. Dimitrakopoulos *et al.*, "Task-Independent Mental Workload Classification Based Upon Common Multiband EEG Cortical Connectivity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1940–1949, 2017.

[30] P. Zarjam, J. Epps, F. Chen, and N. H. Lovell, "Estimating cognitive workload using wavelet entropy-based features during an arithmetic task," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2186–2195, 2013.

[31] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring Mental Workload with

EEG+fNIRS," *Front. Hum. Neurosci.*, vol. 11, no. July, pp. 1–20, 2017.
[32] B. Rebsamen, K. Kwok, and T. B. Penney, "Evaluation of cognitive workload from EEG during a mental arithmetic task," *Proc. Hum. Factors Ergon. Soc.*, vol. 5, pp. 1342–1345, 2011.
[33] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 1, pp. 31–42, 2019.
[34] P. Aricò, F. Aloise, F. Schettini, S. Salinari, D. Mattia, and F. Cincotti, "Influence of P300 latency jitter on event related potential-based brain--computer interface performance," *J. Neural Eng.*, vol. 11, no. 3, p. 35008, 2014.
[35] T. Radüntz, "Dual frequency head maps: A new method for indexing mental workload continuously during execution of cognitive tasks," *Front. Physiol.*, vol. 8, no. DEC, pp. 1–15, 2017.
[36] P. K. Jao, R. Chavarriaga, and J. D. R. Millan, "Using Robust Principal Component Analysis to Reduce EEG Intra-Trial Variability," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2018-July, no. 1, pp. 1956–1959, 2018.
[37] F. Dehais *et al.*, "Monitoring pilot's mental workload using erps and spectral power with a six-dry-electrode EEG system in real flight conditions," *Sensors (Switzerland)*, vol. 19, no. 6, 2019.
[38] A. J. Casson, "Artificial Neural Network classification of operator workload with an assessment of time variation and noise-enhancement to increase performance," vol. 8, no. December, pp. 1–10, 2014.
[39] J. Fan *et al.*, "A Step towards EEG-based brain computer interface for autism intervention," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 3767–3770.
[40] R. Bellman and R. Kalaba, "Dynamic programming and statistical communication theory," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 43, no. 8, p. 749, 1957.
[41] Y. Ke *et al.*, "An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," *Front. Hum. Neurosci.*, vol. 8, no. September, pp. 1–10, 2014.
[42] C. Mühl, C. Jeunet, F. Lotte, and M. A. Hogervorst, "EEG-based workload estimation across affective contexts," vol. 8, no. June, pp. 1–15, 2014.
[43] M. Arvaneh, A. Umilta, and I. H. Robertson, "Filter bank common spatial patterns in mental workload estimation," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2015-Novem, pp. 4749–4752, 2015.
[44] Z. Yin and J. Zhang, "Identification of temporal variations in mental workload using locally-linear-embedding-based EEG feature reduction and support-vector-machine-based clustering and classification techniques," *Comput. Methods Programs Biomed.*, vol. 115, no. 3, pp. 119–134, 2014.
[45] 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, 2012.
[46] S. Yang, Z. Yin, Y. Wang, W. Zhang, Y. Wang, and J. Zhang, "Assessing

cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders," *Comput. Biol. Med.*, vol. 109, no. April, pp. 159–170, 2019.

[47] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience and Biobehavioral Reviews*, vol. 44. Elsevier Ltd, pp. 58–75, 2014.

[48] T. Radüntz, N. Fürstenau, A. Tews, L. Rabe, and B. Meffert, "The Effect of an Exceptional Event on the Subjectively Experienced Workload of Air Traffic Controllers: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers," 2019, pp. 239–257.

[49] T. Edwards, L. Martin, N. Bienert, and J. Mercer, "The Relationship Between Workload and Performance in Air Traffic Control: Exploring the Influence of Levels of Automation and Variation in Task Demand," 2017, pp. 120–139.

[50] P. Arico *et al.*, "Human Factors and Neurophysiological Metrics in Air Traffic Control: a Critical Review," *IEEE Rev. Biomed. Eng.*, Apr. 2017.

[51] G. Di Flumeri, P. Aricò, 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*, vol. 2016. 2016.

[52] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[53] W. L. Lim, O. Sourina, and L. P. Wang, "STEW: Simultaneous task EEG workload data set," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2106–2114, 2018.

[54] N. J. Nilsson and N. J. Nilsson, *Artificial intelligence: a new synthesis*. Morgan Kaufmann, 1998.

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

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

[57] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interact. Comput.*, vol. 24, no. 3, pp. 154–172, 2012.

[58] L. G. Hernández, O. M. Mozos, J. M. Ferrández, and J. M. Antelis, "EEG-Based Detection of Braking Intention Under Different Car Driving Conditions," *Front. Neuroinform.*, vol. 12, no. May 2018, pp. 1–14, 2018.

[59] T. Næs and B. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *J. Chemom. A J. Chemom. Soc.*, vol. 15, no. 4, pp. 413–426, 2001.

[60] M.-P. Deiber, E. Sallard, C. Ludwig, C. Ghezzi, J. Barral, and V. Ibañez, "EEG alpha activity reflects motor preparation rather than the mode of action selection," *Front. Integr. Neurosci.*, vol. 6, p. 59, 2012.

[61] D. H. Wolpert and W. G. Macready, "No free lunch theorems for search," Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[62] J. R. Wolpaw *et al.*, "Brain-computer interface technology: a review of the

first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, 2000.