Article

Recognition of mental workload of pilots in the cockpit using EEG signals

Aura Hernández-Sabaté [1,2]*, José Yauri [1,2], Pau Folch [3], Miquel Àngel Piera [4] and Debora Gil [1,2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona
[2] Computer Science Department, Universitat Autònoma de Barcelona
[4] Telecommunications and Systems Engineering Department, Universitat Autònoma de Barcelona
[3] Aslogic, Parc de Recerca UAB
* Correspondence: aura@cvc.uab.cat

† This paper is an extended version of our paper published in 23rd International Conference of the Catalan Association for Artificial Intelligence.

Abstract: The commercial flightdeck is a naturally multi-tasking work environment, one in which interruptions are frequent and of various forms, contributing in many cases to aviation incident reports. Automatic characterization of pilots workload becomes essential to prevent these kind of incidents. As well, minimising the physiological sensor network as much as possible remains both a challenge and a requirement. Electroencephalogram (EEG) signals have shown a high correlation to specific cognitive and mental states like workload. However, there is not enough evidence in the literature to validate how well models generalize in case of new subjects performing tasks of a workload similar to the ones included during model's training. In this paper we propose a convolutional neural network to classify EEG features across different mental workloads in a continuous performance task test that measures a portion of working memory and working memory capacity. Our model is valid at a general population level and it is able to transfer task learning to a pilot mental workload recognition in a simulated operational environment.

Keywords: Cognitive states; Mental workload; EEG analysis; Neural Networks, Multimodal data fusion

Citation: Hernández-Sabaté, A.; Yauri, J.; Folch, P.; Piera, MA.; Gil, D. Recognition of mental workload of pilots in the cockpit. *Appl. Sci.* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

A fundamental aspect of multiple task management is to attend to new stimuli and integrate associated task requirements into an ongoing task set; that is, to engage in interruption management [1]. Interruptions often negatively affect human performance. Specifically, most laboratory and applied experiments demonstrate that interruptions increase post-interruption performance times [2] and error rates [3], increase perceived workload [4] , and motivate compensatory behavior [5].

The commercial flightdeck is a naturally multi-tasking work environment, one in which interruptions are frequent and of various forms. Further, interruptions have been cited as a contributing factor in many aviation incident reports. External and aircraft events, as well as interactions with other operators, compete for pilots' attention and require pilots to integrate performance requirements associated with these unexpected prompts with ongoing flightdeck tasks.

For that, the study of workload becomes essential to prevent accidents, since it could compromise human task performance [6]. Since workload involves cognitive, neuro-physiologic, and perceptual processes to resolve a task, it is affected by individual capabilities, motivation, as well as, its physical and emotional state [7]. Although this multifaceted nature of workload prevents to study workload directly, it is feasible to be inferred from different quantifiable variables [8]. There exist many proposals for recognizing workload based on physiological features, such as hearth rate, eye movement and dilation, electroencephalogram (EEG) and electrocardiogram (ECG) [9, 10].

Besides, the recent emerging of low cost EEG headsets has driven to new researches (like interaction with home devices, teaching-learning educative methods or mentally control robotic arms) further than medical screening of neurological disorders. In the particular case of cognitive state assessment, EEG alone is becoming the preferred sensor for addressing its characterization [11–13]. However, there is not enough evidence in the literature to validate how well models generalize in case of new subjects performing tasks of a workload similar to the ones included during model's training.

The goal of this paper is to characterize the mental workload of flying pilots in the cockpit from the analysis of EEG signals.

The remainder of this paper is organized as follows: Section 2 presents relevant related works, Section **??** explains the process followed to collect the data. Section **??** presents our proposal for the analysis of EEG generalization capabilities, while Section 6 is devoted to present the results of our experiments. Finally, Section **??** outlines the conclusion and future work.

2. Related work

The most generalized mechanisms to measure workload can be split in two main categories [9,14,15]: subjective measures based on the subject perception and objective scores based on physiological responses.

On the one hand, subjective measures are still the most used to assess mental workload, being the NASA Task Load Index (TLX) [16] the most prominent test to gain insight about the perceived workload levels while a subject works with various human-machine interface systems [17,18]. This questionnaire measures the mental workload based on a weighted average of six sub-variables: mental demand, physical demand, temporal demand, performance, effort and frustration and it is widely used in aviation to assess mental workload of pilots while interacting with plane controls [19,20].

On the other hand, physiological measures provide a more reliable data of workload by measuring physiological dynamic changes which cannot be controlled consciously, so they are becoming more popular among researchers in recent years [21–23]. The most common sensors to record physiological data are: electrocardiogram (ECG) to register heart's electrical activity, electromyograph to read skeletal muscles electrical activity, electroencephalogram (EEG) to detect electrical activity in the brain, photoplethysmography to register volumetric changes in the blood flow, respiration rate sensors, electro-dermal activity (EDA) to read skin surface temperature, oxygen density in the blood in the brain, and eye movement trackers, among others [24]. TLX surveys allow to assess the perceived workload [16], but it is highly subjective. However, physiological data occurs spontaneously, and, together with TLXs, provide a more reliable information [9,17,21].

The combination of several physiological sensors to classify workload states gives better results than using a single one. The approach proposed in [25] combines EEG, ECG, and electrooculography (EOG) and results show a highest predictive power for their combination (80%) rather than the analysis of each one independently (70%). Besides, the study in [10] reports an accuracy average of 85.2 (± 4.3%) combining EEG, ECG, respiration rate, and EDA to classify 4 mental states. The work in [26] still shows better results combining EEG, ECG and EDA than using only EEG signal from classifying four mental states, although results from the single sensor are promising (86.66%).

At that point, Deep Learning (DL) approaches are gaining ground over more classical machine learning techniques due to their ability to automatically extract the features [24,27]. For instance, the study in [6] proposes a concatenated structure of deep recurrent and 3D convolutional neural networks to combine both raw and spectral EEG data and assess two degree of mental workloads reporting an average accuracy of 88.9% in a cross-task assessment.

2.1. Contributions

AI methods characterizing WL from EEG signals must face several challenges. First, in order to properly be trained and tested, it is mandatory to have data with unambiguous annotations (known as ground truth, GT). The collection of this annotated data is complex because the concept of WL itself is subjective and difficult to determine in an objective systematic manner. Second, for an optimal performance of the system these should properly combine the signals recorded from the different EEG electrodes. Finally, a main issue that a ML system involving humans should consider is the generalization power of ML systems, which includes reproducibility of results and capability for transfer learning. That is, to what extend a general model trained over a set of individuals can successfully predict a new unseen individual performing a different task than the ones used for training the system [28].

This work contributes to the 3 challenges as follows:

1. **Unambiguous Annotated Dataset.** In order to generate data with unambiguous annotation we have serious games and flight scenarios in a A320 simulator. The serious game was a modified N-back-test [29] of increasing memory demand. The level of difficulty of the test is our GT for training models. Such level of difficulty was cross-checked with the difficulty perceived by the player assessed using NASA TLX questionnaire. Models were trained using N-back-test data recorded from a population that did not include pilots. Task and population transfer of systems were validated in cockpit simulation exercises designed to have different levels of complexity, as well as, unexpected unsaved situations known to substantially drop pilots' performance.
2. **Models able to recognize two levels of workload with high generalization capability.** Two different architectures are proposed for the fusion of EEG sensor signals (channels) at two different levels [30]: input data (labelled input projector model) and convolutional feature (labelled feature projector model) . Both architectures consist of an input block managing fusion at input level, a convolutional block and an output block for fusion of convolutional features. For each architecture model several classification problems (including an increasing number of WL classes) are trained on N-back-test data using a one-subject-out scheme and tested in binary problem for detection of WL on flight simulations.

Results show that regardless of the projection level, best performers are models trained to classify the largest number of WL classes. Between the two architecture models, projecting convolutional feature channels achieves higher performance, with 81.81% of sensitivity and 85% specificity in WL detection in N-back-test leave-one-out subject evaluation and good task transfer with the detected WL increasing with the number of interruptions.

3. Data annotation and Ground Truth Generation

In this paper, we provide two different automatically annotated datasets that serve to train, validate and verify learning and population transfer of models. The first dataset was recorded from a group of non-pilot subjects playing a memory demanding serious game with increasing demand of WL. The second dataset was recorded from pilots flying scenarios of different complexity on a A320 flight simulator.

3.1. Dual n-back test

N-Back-tests are memory demanding games requiring the resolution of tasks according to a stimulus presented N trials before. We used three variants of the N-Back-tests to induce low, medium, and high mental workload:

1. *Position 1-back for low workload.* A square appears every few seconds in one of eight different positions on a regular grid over the screen. Players must press a keyboard

key in case the position of the square on the current screen is the same as the square of the previous grid.

2. *Arithmetic 1-back for medium workload*. An integer number between 0 and 9 appears every few seconds on the screen while an audio message says an arithmetic operation (plus, minus, times and divide). Players have to solve this operation using the current number and number that appeared in the previous screen.
3. *Dual arithmetic 2-back for high workload*. This test combines the two previous ones. An integer number between 0 and 9 appears every few seconds in one of eight different positions on a regular grid. At the same time, for each number that appears on screen, an operator is presented with an audio message. As before, players have to solve this operation using the current number and number that appeared in two screens before. In addition, players have to press a key in case the position of the current number is the same as the position of the number shown two screens before.

The neurophysiological response of a subject against mental demanding tasks depends on its baseline state, which is prone to vary across time. In order to account for differences in the baseline state of subjects, previous to the N-back-tests participants watched a relaxing video for 10 minutes. For each experiment (1-low, 2-medium and 3-high workload), we call the video watching stage, baseline phase, and the N-back-test, workload phase. Thus, we call BL1, BL2, and BL3 the baseline phases of each experiment, while WL1, WL2, and WL3 are devoted for the workload phases of each experiment.

After the game, participants answered a TLX questionnaire to collect their subjective perception of game difficulty and workload. Results presented in [31] showed that the level of difficulty of the games was correlated to the performance of players and also to the subjective perception of WL computed using NASA-TLX questionnaire.

A total of 20 subjects participated in the experiment. Subjects were adults between 20 and 60 years, all of them were healthy without any condition that might have cause an imbalance in the data recorded. The sequence of tasks were randomly assigned to subjects, and recording of each session was in different days and hours.

3.2. Flight simulations

The experiments have been designed considering the importance to collect experimental data that could be useful to quantify the impact of a task load increment to both pilots through operational interruptions of Air Traffic Controller (ATC), cabin crew (TCP) and Electronic Centralized Aircraft Monitor (ECAM) warnings, in order to assess to what extend the system presented to discriminate between low and high workload can be transferred to a more complex environment.

Four scenarios with different levels of complexity have been designed, all of them assuming Pilot Monitoring (PM) incapacitation in order to check how interruptions can overload PF.

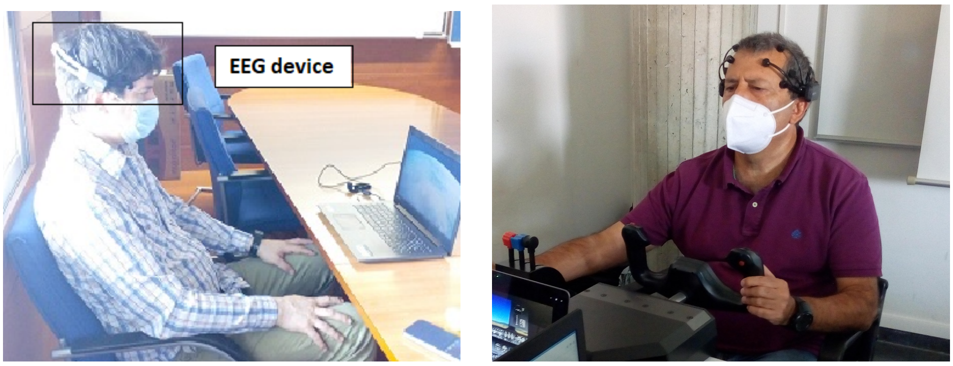
- *Flight 1*. It is based in a nominal standard flight. This experiment is used to take reference parameters. Thus, nominal flight without considering any interrupting event from abnormal procedures due to systems failure neither ATC vectoring instruction such as a direct to TEBLA. In this scenario ATC provides a minimum number of instructions to which the pilots are used to. This scenario the lowest complexity and is considered as the BL class.
- *Flight 2*. It also relies on the approach phase and it is modified from the nominal scenario, by three different interruptions which increase PF workload. This scenario has an overall high WL demand.
- *Flight 3*. This scenario is based on previous experiment with similar interruptions, but they are slightly advanced or delayed at time windows in which the PF workload is low and can attend the interruption without a negative performance impact. Given that interruptions were issued at the most appropriate times, this scenario has a lower level of WL demand than Flight 2.

- *Flight 4*. This last scenario is based on previous experiment with the same interruptions, but unfortunately, they are fired at a time in which PF is attending other concurrent actions, increasing considerably the workload with an impact on the PF performance. This scenario has a similar or greater WL than Flight 2.

Functional Resonance Analysis Method (FRAM) [32] is an agent based modeling framework to identify those factors that affects the performance of the pilot flying cockpit functionalities considering different socio-technical operational conditions. According to this agent, the impact of an interruption on the Pilot Flying (PF) workflow depends largely on the time at which the interruption occurs. Consequently, FRAM provides a reliable measure of the workload that will be faced by the pilot and, thus, it was used to design simulation scenarios with interruptions triggered at times when the pilot had a low and high WL peaks and, thus, have realistic flying situations of controlled difficulty. As well, FRAM output (both, number of tasks and its complexity) was used to assess the capability of ML models to detect WL peaks associated to highly demanding tasks.

In this case a single pilot flew the 4 scenarios.

Figure 1 illustrates a volunteer during a session for the dual n-back test task (a) and a pilot during a simulated flight session (b).



(a) (b)

Figure 1. Data collection with Emotiv Epoc+ headset. (a) A volunteer during a N-Backtest and (b) a pilot during simulated flight session.

4. Workload Recognition

In this section we present our model, able to recognize between baseline and workload. Falta donar-li una volta i explicar què volem classificar

The recognition pipeline follows a usual pipeline of machine learning recognition module: first, raw input data is preprocessed to obtain the proper input data Later, these signals are fed into the network model to automatically extract the features that will be furtherly combined in a classifier step to discriminate among the number of classes previously determined (baseline vs workload in our case).

4.1. Extracting input data from EEG signals

For EEG recording, an EMOTIV EPOC+ headset [33] has been used, which has 14 electrodes placed according to the 10/20 system. This sensor provides both raw data and power spectra for the main brain frequencies (θ, α, β_{low}, β_{high}, and γ). Given that proposed N-back tasks are memory demanding stressing games and baseline phases consist in watching a relaxing video, the theta wave [34] is the best candidate for discriminating the different mental loads of our experimental phases. In this work, we use the power spectrum of theta wave (4–8 Hz) sampled at 8 Hz.

Eye blinking and sudden head movements introduce abrupt sharp peaks of large amplitude in the power spectra wave that should be filtered before using them as predictors of a mental state [21]. In particular, we use an Inter Quartile Range (IQR) [35] filtering strategy to detect outlier values associated to muscular movement wave peaks. Our IQR filtering is based on setting the value of the 99% percentile of the distribution to all points above it.

To ensure a high quality of signals, we further filter data according to the quality of the EEG during recordings provided by the headset itself. For each sensor and recorded sample, Emotiv reports the quality of the recording in a discrete scale with values in the range 0-4 indicating how good the contact between sensor and head is: 4 for optimum - 0 for none. For the sake of data with the highest possible quality while keeping a reasonable sample size signals with a 25% of bad recordings are discarded (< 3). Further, since there is no evidence about what are the most discriminating sensors that best correlate to the detection of mental workload, the whole phase is discarded if the signal of two or more of the sensors has a low quality. Finally, a subject is discarded if either all its base line or its workload phases are discarded, since, in this case, there is not enough data to define the binary classification. After this quality filtering, only 16 of the 20 subjects were selected for models training and testing.

In order to feed data to models, θ signals are cut in temporal windows. Notice that the size and overlap of the temporal windows might be a critical issue in order to properly include workload peaks [36]. For that we have used several window widths with different overlaps, obtaining the best results with 40 seconds windows overlapped 30 seconds. Thus, the input data of the networks are the concatenation of the 40 second windows for the 14 EEG sensors (14 * 40 = 560-dimensional feature space). In order to account for the difference in units and magnitudes, input data has been standardized using the mean and standard deviation of the training set.

4.2. Network architectures

The spatio-temporal representation of EEG signals is an issue that any classification ML system has to face. The simplest question is when to combine the signals, before or after extracting features? We propose two architectures that differ in the moment when EEG sensor signals (channels) are projected: one projects input EEG sensors (input projector model) and the other one projects the convolutional features extracted from each EEG sensor (feature projector model). Each model has one input unit projecting EEG channels, a convolutional unit equal for both models and an output unit projecting the convolutional features extracted from each EEG sensor. This output unit has a fully connected layer with sigmoid activation and output the number of classes. To account for different window lengths, we apply an average pooling before the classification layer. All convolutional layers use kernels of size 3 and stride 1 and have Relu activation.

The convolutional unit has 3 blocks consisting of one convolutional layer with max pooling and having, respectively, 16, 32 and 64 neurons for each convolutional layer. The classification layer has 256 neurons. For the input projector model, the projection unit has one convolutional layer with 16 neurons. For the feature projector model, the output unit has 2 blocks consisting of one convolutional layer before the classification layer. The first one has 64 neurons, the second one projects convolutional features also using 64 neurons.

Figures 2 and 3 show both architectures.

Although our main problem is a binary one, to ensure generalization capabilities of the classifier (including task transfer) we increase the diversity of the classifier by increasing the number of classes used to train the network. That is, our architecture has been trained as classifiers to discriminate between a BL and WL classes using 4 different grouping of the data recorded from the 3 n-back tests:

1. **Binary problem** (noted BLs-WL2) given by BL=(BL1,BL2,BL3) and WL2, that is the BL class is defined by aggregating the baselines for the 3 games and WL class defined by the workload phase of the second experiment.

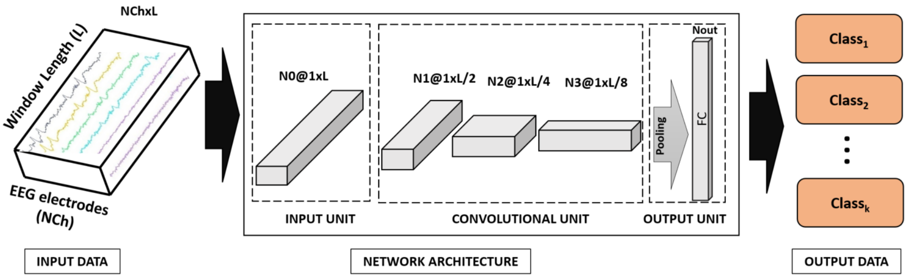


Figure 2. Architecture of the Input Projector Model

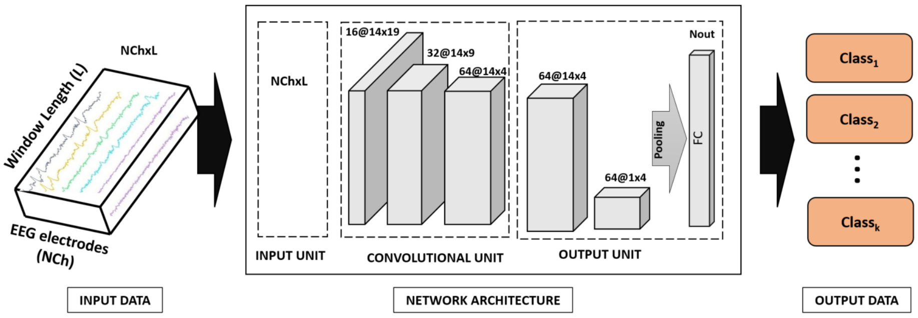


Figure 3. Architecture of the Feature Projector Model

2. **Three class problem 1** (noted BLs-WL2-WL3) given by BL=(BL1,BL2,BL3), WL2 and WL3, that is a BL class defined as before and two WL classes given by the workload phase of the second and third experiments.
3. **Three class problem 2** (noted WL1-WL2-WL3) given by WL1, WL2 and WL3, that is, a BL class defined by the workload phase of the first experiment and two WL classes given by the phase 2 of the second and third experiments.
4. **Four class problem** (noted BLs-WL1-WL2-WL3) given by BL=(BL1,BL2,BL3), WL1, WL2 and WL3, that is a BL class defined as in the first configuration and also defined by the workload phase of the first experiment and two WL classes given by the workload phase of the second and third experiments.

Unlike binary problems, in multiclass settings, the classifier does not predict the probability of belonging to each class. It rather gives a score of belongingness. It follows that the class predicted is not the one having a score above 0.5 (as it is the case of binary problems), but the one having the largest value of the score predicted by the classifier. In our case, since the final class prediction is binary, we compute the binary class labels in the multiclass settings by binarizing first the output probabilities and then taking the maximum between the two as the final class label. The transformation between classifier output and BL-WL classes scores is as follows:

1. **BLs-WL2-WL3**: The probability of BL is directly the probability of the train BL class, while the probability of the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.
2. **WL1-WL2-WL3**: The probability of the class BL is given the probability of the class WL1, while for the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.
3. **BLs-WL1-WL2-WL3**: The probability of the class BL is the maximum probability of the BL and WL1 classes, while for the class WL it is the maximum of the probabilities of the WL2 and WL3 classes.

5. Experimental Design

In order two validate the proposed models, two experiments have been conducted:

5.1. Training and Validation using N-back-test data.

To assess to what extend a model trained over a set of individuals can successfully predict a new unseen individual, we have considered a Generalist Population Model, where a single model using all subjects is trained to assess whether inter subject variability can be properly modelled. The validation of the capability for modelling a population has been tested using a leave-one-out scheme to allow statistical analysis. Models were trained using a batch size of 750, a weighted cross-entropy loss to compensate unbalances between baseline and workload phases, Adam [37] as optimization method, 100 epochs and a learning rate of 0.0001.

The performance of the different approaches for detection of mental workload has been assessed using the accuracy (or sensitivity) for each class:

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP = number of True Positives and FN = number of False Negatives. Sensitivity measures the capability of the system to detect BL and WL classes. Since we have a binary classification problem with WL the positive class, then the sensitivity for BL corresponds to the specificity of the model.

5.2. Task Transfer Verification using Flight Simulator data.

To assess the capability of our model for transfer learning experiments are devoted to show that the model trained to detect WL in a memory demanding task (n-back test) can detect an increase of WL associated to multitask procedures with interruptions decreasing performance.

The EEG signals of the flight dataset explained in section 3 are intended to assess:

1. Correlation of WL recognition to the number of tasks carried on by the pilot. Since we expect that the proportion of samples classified by our model as medium-high WL is higher in the intervals where the PF performs more tasks we show the percentage of predictions for BLs and each WL in correspondence to the number of tasks demanded.
2. Correlation of WL recognition to flight complexity. Flights 2 and 4 are designed to have more workload than flight 3 (flight 1 is considered as baseline) so that the hypothesis is that the proportion of samples classified by the model as medium-high WL is higher than in flight 3.

6. Results

6.1. Training and Validation using N-back-test data

Tables 1 and 2 summarize the recalls of baseline (BL) and workload (WL2) for the binarized models trained on different class problems for, respectively, the feature and input projector models. Tables show ranges for WL and BL detection computed for the 24 subjects and removing 3 outlying cases that all approaches fail to correctly predict.

For all cases, performance increase with the number of classes used to train models. Regarding projection approaches, models projecting features achieve higher performance. In particular the 4-class feature projector model achieved an average detection of, both, BL and WL of 85%.

6.2. Task Transfer Verification using Flight Simulator data

Barplots in figures 4 and 5 show the percentage of WL detection as function of the number of interruptions. The expected pattern would be an increasing percentage of WL detection with the number of interruptions. For both projection models, the 3-class problem WL1_WL2_WL3 is the only model that does not follow the expected increasing pattern. For the remaining problems, both architectures seem to behave equally.

Table 1: Input projector model binarized

		All population	**80% of population**
BL-WL2	BL	66.57 ± 13.11	66.10 ± 13.87
	WL	65.42 ± 25.59	73.95 ± 18.62
BLs-WL2-WL3	BL	78.16 ± 10.83	75.5 ± 10.29
	WL	78.63 ± 16.59	84.35 ± 10.88
WL1-WL2-WL3	BL	72.94 ± 18.08	70.58 ± 19.29
	WL	77.34 ± 16.72	82.85 ± 11.48
BLs-WL1-WL2-WL3	BL	80.75 ± 9.87	79.42 ± 10.07
	WL	76.44 ± 16.81	80.96 ± 13.16

Table 2: Feature projector model binarized

		All population	**80% of population**
BL-WL2	BL	66.57 ± 13.11	66.10 ± 13.87
	WL	65.42 ± 25.59	73.95 ± 18.62
BLs-WL2-WL3	BL	80.00 ± 10.19	77.92 ± 10.22
	WL	83.34 ± 15.61	88.54 ± 9.80
WL1-WL2-WL3	BL	65.69 ± 37.19	61.07 ± 39.78
	WL	72.69 ± 36.15	87.15 ± 19.79
BLs-WL1-WL2-WL3	BL	85.34 ± 7.27	84.62 ± 7.77
	WL	81.91 ± 14.21	85.42 ± 11.69

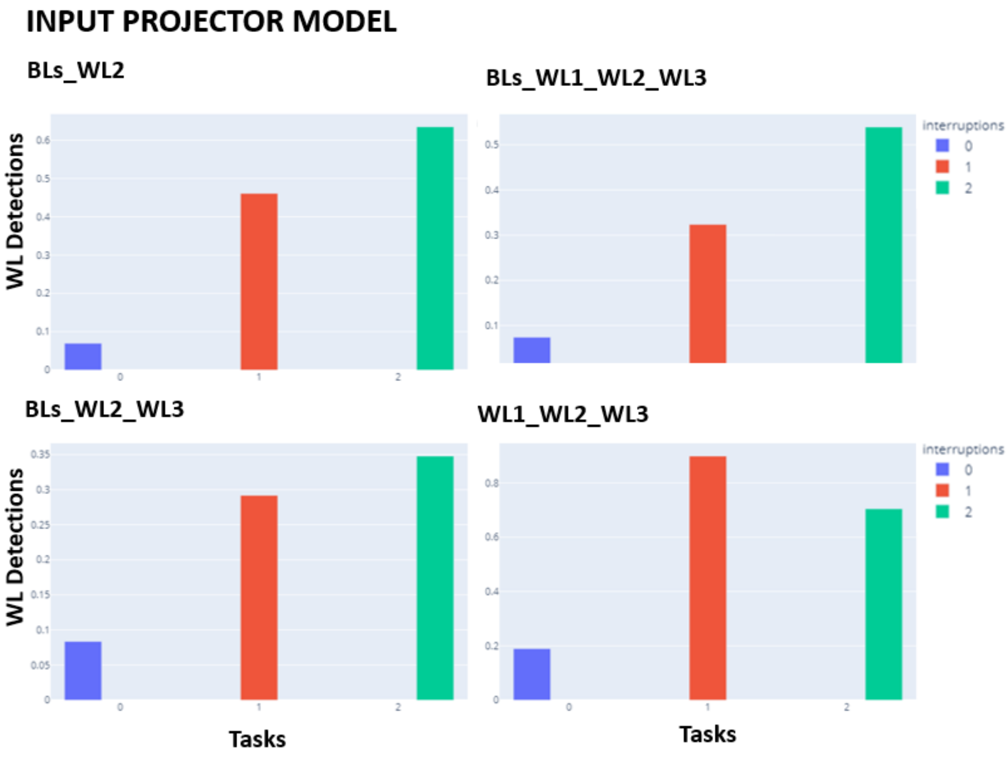


Figure 4. FRAM tasks barplots of WL predictions for the Input Projector model

350 Figures 6 and 7 show the barplots for the number of BL and WL predictions for the 4 flights. The expected pattern would be to have the lowest number of detection for Flight 1, Flight 2 and Flight 4 with similar amount of detected WL and Flight 3 presenting an increase in detected WL with respect these 2 flights, as expected. The most significant differences between flights are evident in the 3-class problem BLs_WL2_WL3, followed

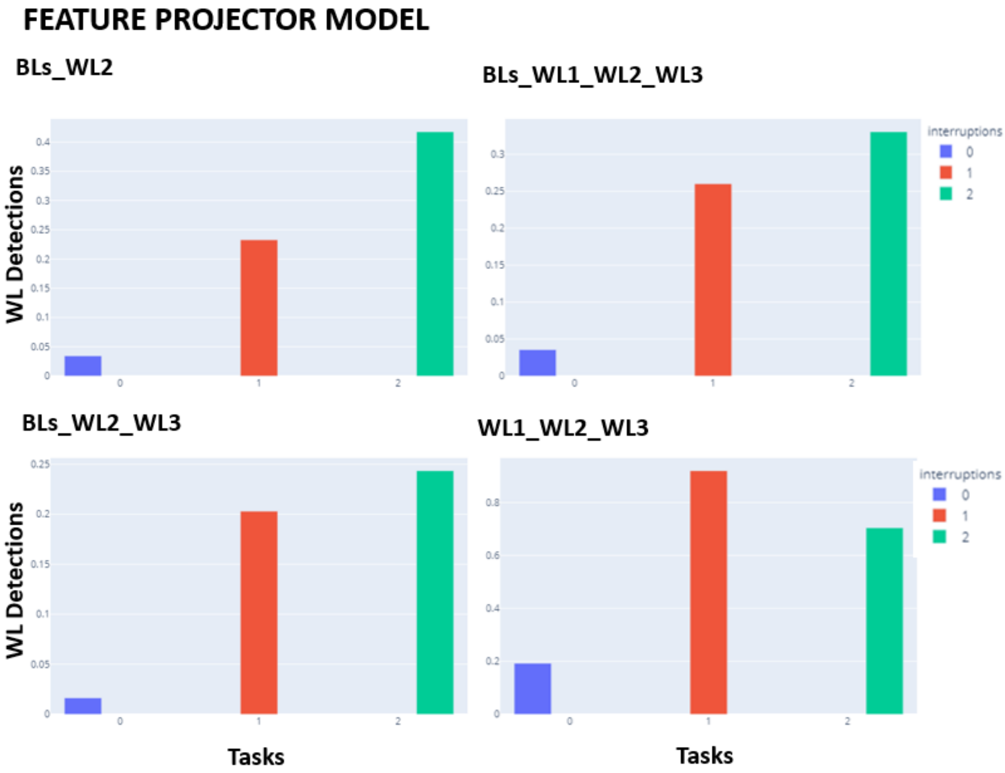


Figure 5. FRAM tasks barplots of WL predictions for the Feature Projector model

by the 4-class problem. The 3-class problem WL1_WL2_WL3 does not apparently detect any difference among them.

7. Conclusions

In this paper, we have presented two different approaches to the fusion of EEG sensor signals. Models have been trained and validated on own-designed games (one serious game and one flight simulator with specific scenarios) to ensure unambiguous annotations. Models were trained and validated on the serious game using one-subject-out scheme, while simulator data gathered from a subject not included in the training data was used to evaluate transfer capability.

Results show that regardless of the projection level, best performers are models trained to classify the largest number of WL classes. Between the two architecture models, projecting convolutional feature channels achieves higher performance, with 81.81% of sensitivity and 85% specificity in WL detection in N-back-test leave-one-out subject evaluation and good task transfer with the detected WL increasing with the number of interruptions. Although these results provide evidence of the ability of the EEG sensor to discern between more and less demanding tasks, as well, increasing the evidence of pointing to the robustness of the EEG and its ability to transfer tasks, the fact that the 3-class problem WL1_WL2_WL3 does not correlate to flight complexity suggests the following improvements.

A delicate issue that has an impact in the performance of methods is the filtering of signals required to remove muscular motion peaks and other artefacts. EEG preprocessing approaches have not been standardized, and even small changes in artefact removal strategy may result in differences with large effects on particular portions of the signal. In this study, we have adopted a filtering approach based on signal probabilistic distribution for outlier removal in the temporal space. We consider that muscular motion could be filtered calibrating muscular signals before test recording to set either the values or the frequency ranges associated to muscular motion.

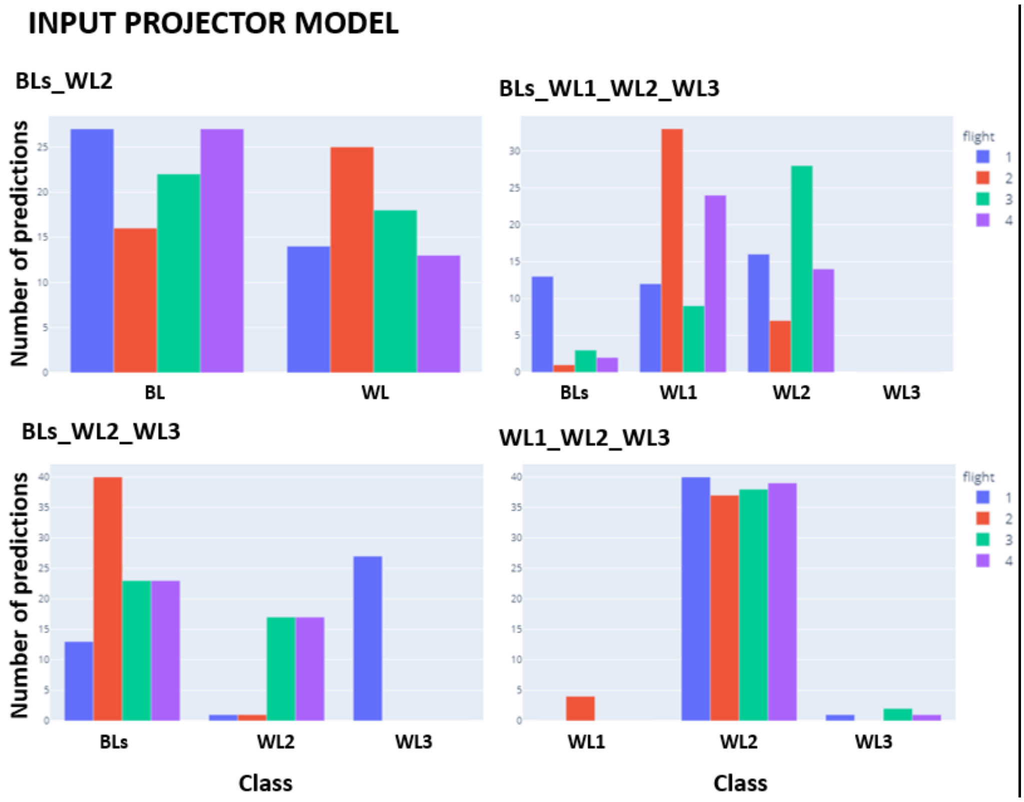


Figure 6. Flight test barplots of WL predictions for the Input Projector model

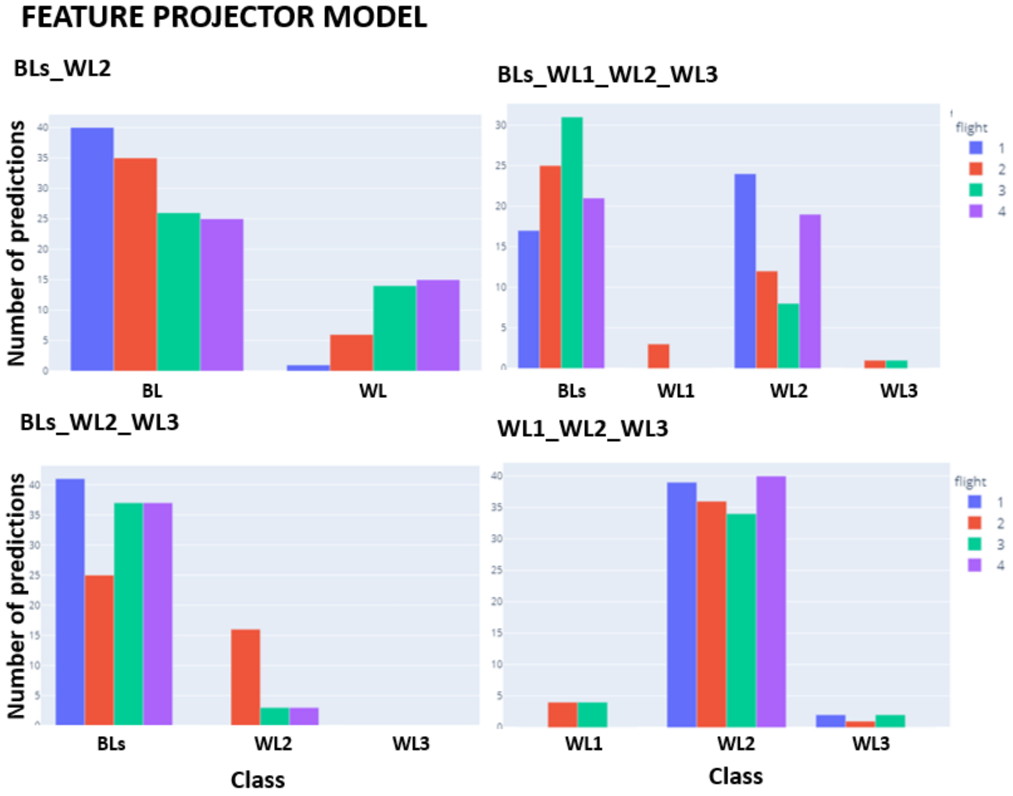


Figure 7. Flight test barplots of WL predictions for the Feature Projector model

Given that the way EEG sensors are fused has a direct impact in performance of models, alternative architectures should be further investigated. In this context, a direct improvement would be to consider ensemble models processing each sensor separately with own-learned weights. Also more recent architectures like convolutional/LSTM and Lambda Nets including attention modelling should be also studied.

Author Contributions: Conceptualization, D.G., and MA.P.; methodology, P.F., A.H-S. and D.G.; software, P.F, J.Y.; validation, J.Y, P.F, A.H-S and D.G.; formal analysis, A.H-S and D.G.; investigation, P.F, A.H-S and D.G.; resources, MA.P.; data curation, J.Y and P.F.; writing—original draft preparation, J.Y and A.H-S.; writing—review and editing, A.H-S. and D.G.; visualization, J.Y and A.H-S.; supervision, D.G., A.H-S. and MA.P.; project administration, MA.P. and D.G.; funding acquisition, MA.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Cleansky grant number 831993, Ministerio de Economía, Industria y Competitividad, Gobierno de España grant number RTI2018-095209-B-C21, Agència de Gestió d'Ajuts Universitaris i de Recerca grant numbers 2017-SGR-1597 and 2017-SGR-1624 and CERCA Programme / Generalitat de Catalunya.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://iam.cvc.uab.es/portfolio/e-pilots-dataset/

Acknowledgments: DGil is a Serra Hunter Fellow.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Latorella, K.A. *Investigating interruptions: Implications for flightdeck performance*; Vol. 99, NASA, 1999.
2. Foroughi, C.K.; Werner, N.E.; McKendrick, R.; Cades, D.M.; Boehm-Davis, D.A. Individual differences in working-memory capacity and task resumption following interruptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **2016**, *42*, 1480.
3. Oulasvirta, A.; Saariluoma, P. Long-term working memory and interrupting messages in human–computer interaction. *Behaviour & Information Technology* **2004**, *23*, 53–64.
4. Kirmeyer, S.L. Coping with competing demands: interruption and the type A pattern. *Journal of Applied Psychology* **1988**, *73*, 621.
5. Cellier, J.M.; Eyrolle, H. Interference between switched tasks. *Ergonomics* **1992**, *35*, 25–36.
6. Zhang, P.; Wang, X.; Zhang, W.; Chen, J. Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2019**, *27*, 31–42. doi: 10.1109/TNSRE.2018.2884641.
7. Li, D.; Wang, X.; Menassa, C.C.; Kamat, V.R. Understanding the impact of building thermal environments on occupants' comfort and mental workload demand through human physiological sensing. In *Start-Up Creation*; Elsevier, 2020; pp. 291–341.
8. Hendy, K.C.; Liao, J.; Milgram, P. Combining time and intensity effects in assessing operator information-processing load. *Human Factors* **1997**, *39*, 30–47.
9. Heine, T.; Lenis, G.; Reichensperger, P.; Beran, T.; Doessel, O.; Deml, B. Electrocardiographic features for the measurement of drivers' mental workload. *Applied ergonomics* **2017**, *61*, 31–43.
10. Han, S.Y.; Kwak, N.S.; Oh, T.; Lee, S.W. Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering* **2020**, *40*, 324–336.
11. Zhang, P.; Wang, X.; Chen, J.; You, W.; Zhang, W. Spectral and Temporal Feature Learning with Two-Stream Neural Networks for Mental Workload Assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2019**, *27*, 1149–1159. doi: 10.1109/TNSRE.2019.2913400.
12. Lee, D.H.; Jeong, J.H.; Kim, K.; Yu, B.W.; Lee, S.W. Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network. *IEEE Access* **2020**, *8*, 121929–121941.
13. Wu, E.Q.; Peng, X.; Zhang, C.Z.; Lin, J.; Sheng, R.S. Pilots' fatigue status recognition using deep contractive autoencoder network. *IEEE Transactions on Instrumentation and Measurement* **2019**, *68*, 3907–3919.

14. Averty, P.; Collet, C.; Dittmar, A.; Athènes, S.; Vernet-Maury, E. Mental workload in air traffic control: an index constructed from field tests. *Aviation, space, and environmental medicine* **2004**, *75*, 333–341.
15. da Silva, F.P. Mental Workload, Task Demand and Driving Performance: What Relation? *Procedia - Social and Behavioral Sciences* **2014**, *162*, 310–319. doi:10.1016/j.sbspro.2014.12.212.
16. Hart, S.G. NASA-task load index (NASA-TLX); 20 years later. Proceedings of the human factors and ergonomics society annual meeting. Sage publications Sage CA: Los Angeles, CA, 2006, Vol. 50, pp. 904–908.
17. Borghini, G.; Astolfi, L.; Vecchiato, G.; Mattia, D.; Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews* **2014**, *44*, 58–75.
18. Index, L. Results of empirical and theoretical research. *Advances in* **1990**.
19. Wickens, C.D. Situation awareness and workload in aviation. *Current directions in psychological science* **2002**, *11*, 128–133.
20. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making* **2008**, *2*, 140–160.
21. Wang, Z.; Yang, L.; Ding, J. Application of heart rate variability in evaluation of mental workload. *Chinese journal of industrial hygiene and occupational diseases* **2005**, *23*, 182–184.
22. Stanton, N.; Salmon, P.M.; Rafferty, L.A. *Human factors methods: a practical guide for engineering and design*; Ashgate Publishing, Ltd., 2013.
23. Jang, E.H.; Park, B.J.; Kim, S.H.; Chung, M.A.; Park, M.S.; Sohn, J.H. Classification of human emotions from physiological signals using machine learning algorithms. Proc. Sixth Int'l Conf. Advances Computer-Human Interactions (ACHI 2013), Nice, France. Citeseer, 2013, pp. 395–400.
24. Rim, B.; Sung, N.J.; Min, S.; Hong, M. Deep learning in physiological signal data: A survey. *Sensors* **2020**, *20*, 969.
25. Ziegler, M.D.; Russell, B.A.; Kraft, A.E.; Krein, M.; Russo, J.; Casebeer, W.D. Computational Models for Near-real-time Performance Predictions Based on Physiological Measures of Workload. In *Neuroergonomics*; Elsevier, 2019; pp. 117–120.
26. Secerbegovic, A.; Ibric, S.; Nisic, J.; Suljanovic, N.; Mujcic, A. Mental workload vs. stress differentiation using single-channel EEG. In *CMBEBIH 2017*; Springer, 2017; pp. 511–515.
27. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep Learning for Time Series Classification: A Review. *Data Mining and Knowledge Discovery* **2019**, *33*, 917–963, [1809.04356]. doi:10.1007/s10618-019-00619-1.
28. Ziegler, M.D.; Kraft, A.; Krein, M.; Lo, L.C.; Hatfield, B.; Casebeer, W.; Russell, B. Sensing and assessing cognitive workload across multiple tasks. International Conference on Augmented Cognition. Springer, 2016, pp. 440–450.
29. Jaeggi, S.M.; Buschkuehl, M.; Jonides, J.; Perrig, W.J. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 6829–6833. doi:10.1073/pnas.0801268105.
30. Bokade, R.; Navato, A.; Ouyang, R.; Jin, X.; Chou, C.A.; Ostadabbas, S.; Mueller, A.V. A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. *Expert Systems with Applications* **2021**, *165*, 113885.
31. Yauri, J.; Hernández-Sabaté, A.; Folch, P.; Gil, D. Mental Workload Detection Based on EEG Analysis. Artificial Intelligence Research and Development. IOS Press, 2021, pp. 268–277.
32. Piera, M.A.; Ramos, J.J.; Muñoz, J.L. A socio-technical holistic agent based model to assess cockpit supporting tools performance variability. *IFAC-PapersOnLine* **2019**, *52*, 122–127.
33. Emotiv. EMOTIV EPOC+ 14-Channel Wireless EEG Headset.
34. Addante, R.J.; Watrous, A.J.; Yonelinas, A.P.; Ekstrom, A.D.; Ranganath, C. Prestimulus Theta Activity Predicts Correct Source Memory Retrieval. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 10702–10707. doi:10.1073/pnas.1014528108.
35. Wasserman, L. *All of statistics : a concise course in statistical inference*; Springer, 2010.
36. Gupta, S.S.; Taori, T.J.; Ladekar, M.Y.; Manthalkar, R.R.; Gajre, S.S.; Joshi, Y.V. Classification of cross task cognitive workload using deep recurrent network with modelling of temporal dynamics. *Biomedical Signal Processing and Control* **2021**, *70*, 103070.
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* **2014**, [1412.6980].